Industrial Organization

a Contract Based Approach

Nicolas Boccard

2010/12/17

Outline

A Introduction 11
1 About the Book 12
2 Microeconomic Foundations 27
B Market Power 67
3 Monopoly 68
4 Differential Pricing 87
C Strategic Interaction 123
5 Imperfect Competition 124
6 Strategic Moves 154
7 Economic Rivalry 184
D Antitrust Issues 216
8 Legal Framework 217
9 Anti-Competitive Practices 240
10 Barriers to Entry 257
E Differentiation and Innovation 282
11 Differentiation and Competition 283
12 Research and Development 318
F Integration 349
13 Firms vs. Markets 350
14 Vertical Integration 380
15 Horizontal Integration 411
G Public Oversight 445
16 The State 446
17 Regulation 473
18 Natural Resources 505
H Incentives and Information 528
19 Risk and Uncertainty 531
20 Moral Hazard 549
21 Adverse Selection 572
22 Auctions 599
23 Entrepreneurship 618
I Network Industries 644
24 Standards and Components 645
25 Network Congestion 681
26 Appendix 709
Bibliography 771
Index 801

ISBN 84-609-9337-X

Contents

Acronyms and Symbols . . . 7

A Introduction *11*

1 About the Book 12
1.1 Foreword . . . 12
1.2 Purpose . . . 13
1.3 Summary . . . 17
1.4 Methodology . . . 20
1.4.1 Scientific Methodology of Economics . . . 21
1.4.2 Individual Rationality . . . 22
1.4.3 Prerequisites . . . 25

2 Microeconomic Foundations 27
2.1 Supply . . . 27
2.1.1 Equi-marginal Principle . . . 27
2.1.2 Production . . . 28
2.1.3 Cost . . . 30
2.1.4 Competitive Supply . . . 33
2.1.5 Miscellanies . . . 36
2.2 Demand . . . 38
2.2.1 Utility . . . 39
2.2.2 Willingness to pay . . . 42
2.2.3 Market Revenue . . . 44
2.2.4 Consumer Surplus . . . 45
2.3 Equilibrium . . . 47
2.3.1 Producers Surplus . . . 47
2.3.2 Equilibrium and Welfare . . . 48
2.3.3 Welfare Economics . . . 49
2.4 Game Theory . . . 50
2.4.1 Simultaneous Games . . . 50
2.4.2 Sequential Games . . . 57
2.4.3 Negotiation: Sharing & Bargaining . . . 58
2.4.4 Free Riding & Social Dilemma . . . 64

B Market Power *67*

3 Monopoly 68
3.1 Optimal Behavior . . . 68
3.1.1 Typology . . . 68
3.1.2 Dominant Position . . . 69
3.1.3 Market Power . . . 70
3.1.4 Output Distortion . . . 71
3.2 Inefficiency . . . 74
3.2.1 Welfare Loss . . . 74
3.2.2 X-efficiency . . . 75
3.2.3 Extensions . . . 77
3.3 Quality . . . 78
3.3.1 Monopoly and Quality . . . 79
3.3.2 Transportation † . . . 82
3.3.3 Monopolistic Taxation . . . 83

4 Differential Pricing 87
4.1 Premices . . . 88
4.1.1 Definition . . . 89
4.1.2 Typology . . . 91
4.1.3 Consumer Surplus Extraction . . . 92
4.2 Direct Differential Pricing . . . 96
4.2.1 Perfect Discrimination . . . 96
4.2.2 Imperfect Discrimination . . . 97
4.2.3 Endogenous Segmentation . . . 101
4.2.4 Proviso . . . 103
4.3 Indirect Differential Pricing . . . 106
4.3.1 Quantity Rebates . . . 107
4.3.2 Optimal Single Tariff . . . 108
4.3.3 Optimal Quantity Discrimination . . . 112
4.3.4 Discrimination by Differentiation . . . 115
4.3.5 Inter-temporal Discrimination . . . 116
4.3.6 Behavior-Based Discrimination . . . 120

C Strategic Interaction *123*

5 Imperfect Competition 124
5.1 Quantity Competition . . . 125
5.1.1 Symmetric Cost . . . 125
5.1.2 Asymmetric cost . . . 128
5.1.3 Oligopoly . . . 129
5.1.4 Welfare . . . 131
5.2 Price Competition . . . 132
5.2.1 Pure Competition: Homogeneous Goods . . . 133
5.2.2 The Hotelling model of Differentiation . . . 136
5.2.3 Duality of Cournot and Bertrand . . . 138
5.3 Contract Competition . . . 141
5.3.1 Supply Function † . . . 142
5.3.2 Clubs . . . 145
5.3.3 Platform Economics . . . 148
5.3.4 Competition and Discrimination . . . 151

6 Strategic Moves 154
6.1 Dynamic of Entry and Exit . . . 154
6.1.1 Perfect Competition . . . 155
6.1.2 Bertrand Competition . . . 157
6.1.3 Cournot Competition . . . 157
6.1.4 Hotelling Competition . . . 159
6.1.5 Generalization † . . . 160
6.1.6 Market Integration . . . 161
6.1.7 Contestability . . . 163
6.2 Stackelberg Leadership . . . 167
6.2.1 Intuition . . . 167
6.2.2 Formal Analysis . . . 168
6.2.3 The value of Commitment . . . 169
6.2.4 Business Strategies † . . . 171
6.3 Forward Market . . . 173
6.3.1 Introduction . . . 173
6.3.2 Analysis . . . 173
6.3.3 Comparisons . . . 175

6.4 Vertical Relationship . . . 176
6.4.1 Managerial Compensation . . . 177
6.4.2 Vertical Integration . . . 179
6.4.3 Strategic Debt Commitment . . . 180

7 Economic Rivalry 184
7.1 Introduction . . . 184
7.1.1 Typology . . . 184
7.1.2 Origins of Economic Rivalry . . . 186
7.1.3 On Rivalry and Efficiency . . . 187
7.2 Wasteful Conflict . . . 189
7.2.1 Wealth Dissipation . . . 189
7.2.2 Asymmetry . . . 191
7.2.3 Canonical Rivalry . . . 194
7.2.4 Enforcing Agreements . . . 196
7.2.5 Productive Conflict . . . 198
7.3 Political Economy . . . 201
7.3.1 Lobbying . . . 201
7.3.2 Capture and Pressure . . . 203
7.3.3 Collective Action . . . 205
7.3.4 Optimal size of the Firm . . . 206
7.3.5 Coase Theorem . . . 207
7.4 Patent Races and Attrition . . . 209
7.4.1 Patent Race . . . 210
7.4.2 Attrition . . . 211
7.4.3 Performance based Compensation . . . 212

D Antitrust Issues *216*

8 Legal Framework 217
8.1 Rule of Law . . . 217
8.1.1 Property Rights . . . 218
8.1.2 Regulation vs. Litigation . . . 221
8.1.3 Coase Theorem . . . 223
8.2 European Union . . . 225
8.2.1 Historical Development . . . 225
8.2.2 Institutions of the EU . . . 229
8.2.3 Economic Principles . . . 231
8.2.4 Current Competition Law . . . 232
8.3 US Antitrust Laws . . . 235
8.3.1 Historical Development . . . 235
8.3.2 Modern Application . . . 237
8.3.3 Comparison of the EU & US . . . 238

9 Anti-Competitive Practices 240
9.1 Cartel and Collusion . . . 240
9.1.1 Price Fixing . . . 241
9.1.2 History of Cartels . . . 243
9.1.3 Collusion . . . 247
9.2 Vertical Agreements . . . 249
9.2.1 Best Price Clauses . . . 249
9.2.2 Vertical Restraints . . . 251
9.2.3 Price Discrimination . . . 253
9.2.4 Antitrust Activity . . . 253

10 Barriers to Entry 257
10.1 Standard Theory . . . 257
10.2 Preemption . . . 260
10.2.1 Exclusive Contracts . . . 260
10.2.2 Contracts as Barriers to Entry . . . 262
10.2.3 Preemption with Capacity Building . . . 267
10.3 Foreclosure . . . 270
10.3.1 Theory . . . 270
10.3.2 Foreclosure strategies . . . 272
10.3.3 Cases and Remedies . . . 273
10.4 Predation . . . 275
10.4.1 Introduction . . . 275
10.4.2 Limited Entry . . . 276
10.4.3 Limit Pricing . . . 277
10.5 Industrial Policy . . . 279

E Differentiation and Innovation *282*

11 Differentiation and Competition 283
11.1 Horizontal Differentiation . . . 283
11.1.1 Differentiation Principles . . . 284
11.1.2 Competition for Location . . . 285
11.1.3 Oligopoly & the Circular City . . . 288
11.1.4 Urban Economics . . . 289
11.2 Location and Variety . . . 292
11.2.1 Variety and Opportunity Cost . . . 292
11.2.2 Multi-dimentional Differentiation . . . 294
11.2.3 Monopolistic Competition . . . 296
11.2.4 International Trade . . . 297
11.3 Vertical Differentiation: Quality . . . 300
11.3.1 Quality and Market Power . . . 300
11.3.2 Price Competition . . . 301
11.3.3 Quality competition . . . 303
11.4 Drivers of Differentiation † . . . 304
11.4.1 Differential Pricing . . . 304
11.4.2 Cost Edge . . . 308
11.5 Advertising . . . 308
11.5.1 Opposite Views . . . 309
11.5.2 Theories . . . 313

12 Research and Development 318
12.1 Social and Legal Matters . . . 319
12.1.1 Social Value . . . 319
12.1.2 Legal Background . . . 320
12.1.3 The Patent System . . . 321
12.2 The Pace of Innovation . . . 323
12.2.1 Innovation in a static market . . . 323
12.2.2 Patent Licensing . . . 326
12.2.3 Innovation Race . . . 331
12.2.4 Market Rivalry . . . 333
12.2.5 Public Subsidies † . . . 336
12.3 Intellectual Property Rights . . . 338
12.3.1 Patent . . . 339

12.3.2 Copyright . . . 340
12.3.3 Trademarks . . . 341
12.3.4 IPR Cases . . . 342

F Integration *349*

13 Firms vs. Markets 350
13.1 Inside the Firm . . . 351
13.1.1 What is a Firm ? . . . 351
13.1.2 Explicit Incentives: Compensation . . . 352
13.1.3 Implicit Incentives: Motivation . . . 355
13.1.4 Internal Organization of the Firm . . . 359
13.2 Contract Theory . . . 362
13.2.1 Incomplete Contracts . . . 362
13.2.2 Simple Contracts . . . 364
13.2.3 Flexible Contracts . . . 366
13.3 Theories of the Firm . . . 367
13.3.1 Organizations vs. Markets . . . 367
13.3.2 The Neoclassical Firm . . . 369
13.3.3 Transaction Costs Economics . . . 370
13.3.4 Property Rights Theory . . . 376

14 Vertical Integration 380
14.1 Bilateral Monopoly . . . 380
14.1.1 Complementary Monopolies . . . 381
14.1.2 Vertical Monopolies . . . 382
14.1.3 Double Marginalization . . . 384
14.1.4 Resale Price Maintenance . . . 386
14.2 Specific Investment . . . 388
14.2.1 Ex-post Opportunism . . . 388
14.2.2 Buy-Out . . . 391
14.3 On Property Rights . . . 395
14.3.1 The Control of Incentives . . . 395
14.3.2 Relation Specific Investments . . . 396
14.4 On Transaction Cost . . . 400
14.4.1 Contractual Design . . . 401
14.4.2 Limited liability . . . 405
14.4.3 Employment vs. Performance . . . 408

15 Horizontal Integration 411
15.1 Merger Activity . . . 411
15.1.1 Large Firms . . . 411
15.1.2 Reasons to merge . . . 415
15.1.3 Merger Scrutiny . . . 418
15.2 Merger Paradox . . . 420
15.2.1 Cournot Competition . . . 420
15.2.2 Bertrand Competition . . . 422
15.2.3 Efficiencies and Welfare . . . 424
15.3 Measures of Concentration . . . 428
15.3.1 What is a Market? . . . 429
15.3.2 Concentration Indices . . . 433
15.3.3 Industry Cases . . . 435
15.3.4 Size Distribution . . . 441

G Public Oversight *445*

16 The State 446
16.1 Missions and Means . . . 446
16.1.1 Missions . . . 446
16.1.2 Means . . . 449
16.2 Rationalizing the State . . . 450
16.2.1 Public Interest . . . 451
16.2.2 Public Choice . . . 452
16.2.3 Regulatory Capture . . . 453
16.2.4 Rent Planning . . . 454
16.3 Rent-seeking . . . 458
16.3.1 Tipology . . . 458
16.3.2 Corruption and Lobbying . . . 460
16.3.3 Inefficiency . . . 462
16.3.4 Remedies . . . 463
16.4 Liberalization . . . 464
16.4.1 Grants and Concessions . . . 465
16.4.2 Nationalization and Municipalization . . . 467
16.4.3 Privatization . . . 468
16.4.4 Deregulation . . . 470

17 Regulation 473
17.1 Why Regulate . . . 473
17.1.1 Context . . . 473
17.1.2 Pricing Quandary . . . 477
17.2 Ideal Regulation . . . 480
17.2.1 Homogeneous Demand . . . 481
17.2.2 Heterogeneous Demand . . . 482
17.2.3 Price vs. Quantities . . . 485
17.2.4 Public Provision . . . 487
17.3 Practical Regulation . . . 488
17.3.1 Cost Based methods . . . 489
17.3.2 Incentive Regulation . . . 493
17.3.3 Regulatory Constraints . . . 498
17.3.4 Public Service Obligation . . . 501

18 Natural Resources 505
18.1 Exhaustible Resources . . . 505
18.1.1 Identities . . . 506
18.1.2 Optimal Resource Extraction . . . 509
18.1.3 Examples . . . 511
18.1.4 Commons . . . 513
18.2 Renewables Resources . . . 516
18.2.1 Introduction . . . 516
18.2.2 Stock, Flow, Extraction . . . 517
18.2.3 Static Equilibrium . . . 520
18.2.4 Dynamic Equilibrium . . . 523

H Incentives and Information *528*

19 Risk and Uncertainty 531
19.1 A Framework for Uncertainty . . . 532
19.1.1 Introduction . . . 532
19.1.2 Time and Money . . . 532

19.1.3 Probability Theory . . . 534
19.2 Choice under Uncertainty . . . 537
19.2.1 Expected Utility and Risk . . . 537
19.2.2 Subjective Risk Measure . . . 539
19.2.3 Objective Risk Measure . . . 541
19.3 Firm Behavior under Risk . . . 544
19.4 Advanced Topics† . . . 546
20 Moral Hazard 549
20.1 The Agency Relationship . . . 550
20.1.1 Framework . . . 550
20.1.2 Moral Hazard . . . 552
20.2 Managerial Incentives . . . 554
20.2.1 Individual Compensation . . . 554
20.2.2 Misaligned Incentives . . . 556
20.2.3 Rank-Order Tournaments † . . . 557
20.2.4 Multi-Tasking . . . 558
20.2.5 Career Concerns . . . 560
20.3 State of Nature Approach . . . 561
20.3.1 The Second Best program . . . 561
20.3.2 Resolution . . . 563
20.3.3 The Mirrlees Approach † . . . 564
20.3.4 Automobile Insurance . . . 565
20.4 Renegotiation and Auditing . . . 567
21 Adverse Selection 572
21.1 Information Unraveling . . . 572
21.1.1 Market for Used Cars . . . 573
21.1.2 Corporation vs Partnership . . . 575
21.1.3 Signaling . . . 577
21.2 Screening & Self-Selection . . . 580
21.2.1 Procurement . . . 580
21.2.2 Non Linear Pricing . . . 583
21.2.3 Public Firm Regulation . . . 586
21.2.4 Procurement and Moral Hazard . . . 589
21.2.5 Insurance Cream Skimming † . . . 594
22 Auctions 599
22.1 Purpose of Auctions . . . 599
22.1.1 Origins . . . 599
22.1.2 The case for auctioning . . . 600
22.2 Comparing Auctions . . . 602
22.2.1 Typology . . . 602
22.2.2 Standard Auctions . . . 604
22.3 Optimal Auctions . . . 606
22.3.1 Revenue Equivalence . . . 606
22.3.2 Optimal Selling Mechanism . . . 608
22.3.3 Bilateral Trade under Uncertainty † . . . 614
23 Entrepreneurship 618
23.1 Agency Cost of Equity Finance . . . 619
23.1.1 Background . . . 619
23.1.2 Incentives to Under-invest . . . 621
23.1.3 Equity Underpricing . . . 623
23.1.4 Signaling Quality . . . 625
23.2 Agency Cost of Debt Finance . . . 627
23.2.1 Asset Substitution and Free Cash Flow . . . 627
23.2.2 Debt Overhang . . . 630
23.2.3 Debt as the Optimal Security . . . 632
23.2.4 Credit rationing . . . 635
23.3 Managerial Incentives . . . 637
23.3.1 Debt as a Signal of Profitability . . . 638
23.3.2 Debt as a Signal of Obedience . . . 639
23.3.3 Efficiency Wage . . . 640
I Network Industries *644*
24 Standards and Components 645
24.1 Components: Tie and Bind . . . 646
24.1.1 Tying . . . 646
24.1.2 Bundling . . . 647
24.2 Creation of a Standard . . . 649
24.2.1 Substitutes . . . 650
24.2.2 Complements . . . 650
24.2.3 Compatibility . . . 651
24.2.4 Adoption of a Standard . . . 652
24.2.5 On Deregulation . . . 655
24.2.6 Network and Competition . . . 656
24.3 Critical Mass . . . 656
24.3.1 Pros and Cons of joining a Standard . . . 657
24.3.2 Consumption Externality . . . 658
24.3.3 Big Push . . . 661
24.3.4 Rent-Seeking . . . 664
24.3.5 Myths . . . 665
24.4 Two sided platforms . . . 669
24.4.1 Game consoles . . . 669
24.4.2 Portable Devices . . . 670
24.5 Social interaction . . . 670
24.5.1 Status seeking . . . 671
24.5.2 Rationing . . . 673
24.5.3 Group behavior . . . 675
24.5.4 Social Cohesion . . . 678
25 Network Congestion 681
25.1 Roots . . . 681
25.1.1 Seasonality . . . 682
25.1.2 Congestion . . . 683
25.1.3 Arbitrage . . . 683
25.2 Capacity Expansion . . . 685
25.2.1 Hidden Cost . . . 685
25.2.2 Application to Communications . . . 686
25.2.3 Application to Air Transport . . . 688
25.3 Peak Load Pricing . . . 690
25.3.1 Variability and Seasonality . . . 690
25.3.2 Optimal Policies . . . 692
25.3.3 Demand Side Management . . . 695
25.3.4 Nodal Pricing in Electricity † . . . 696

25.4 Road Congestion . . . 699
25.4.1 Origin . . . 699
25.4.2 Solutions . . . 700
25.4.3 Expansion . . . 701
25.4.4 Cost Benefit Analysis † . . . 704
26 Appendix 709
26.1 Miscellanies . . . 709
26.1.1 Returns to Scale . . . 709
26.1.2 Constant Elasticity of Substitution . . . 710
26.1.3 Rubinstein's bargaining model . . . 711
26.1.4 Coase Conjecture . . . 711
26.1.5 Oligopoly . . . 713
26.2 Risk and Uncertainty . . . 716
26.2.1 Time Discounting . . . 716
26.2.2 Euler Equation . . . 716
26.2.3 Risk Aversion . . . 717
26.2.4 Choice under Uncertainty . . . 719
26.3 Auctions and Finance . . . 721
Notes . . . 726
Bibliography *771*
Index *801*

Acronyms and Symbols

Convention

In all bilateral relationships, parties are referred to “she” and “he” in order to abridge and clarify exposition.

Acronyms

- § chapter or section
- † harder topic
- ■ end of a proof
- aka also known as
- c. circa
- cf. compare with or see also
- e.g. for instance
- i.e. that is to say
- p. page
- resp. respectively
- w.l.o.g. without loss of generality
- w.r.t. with respect to
- FOC first order condition
- DRS decreasing returns to scale
- SCP structure conduct performance
- WTP willingness to pay
- EU European Union
- EC European Commission
- US United States (of America)
- UK United Kingdom

Logic and Maths

- $A \Rightarrow B$ A implies B
- $A \because B$ A because B
- $A \Leftrightarrow B$ A is equivalent to B
- $A = B$ equality of A and B
- $A \equiv B$ A defined by B
- du differential of u
- $\frac{\partial u}{\partial z}$ derivative of u w.r.t. z
- $\dot{u}$ time derivative of $u(x,t)$
- u' spatial derivative of $u(x,t)$
- $\tilde{x}, \tilde{y}$ random variables
- G, H distribution functions
- g, h density functions
- $\mathbb{E}$ expectation operator
- $\mathbb{V}$ variance operator

Economic Symbols

- L, l labour quantity
- K, k capital quantity
- w wage
- r, ρ interest rates
- $\Phi(K, L)$ production function
- p, p_1 prices
- q, q_1 individual quantities
- Q aggregate quantity
- C cost function
- C_m marginal cost
- c constant marginal cost
- AC average cost function
- J, F fixed cost
- i, j, l labels for economic agents
- m, n # of firms or consumers

- π, Π profit
- BR best reply function
- u, U utility
- D, D^1 demand functions
- a, b, d demand parameters
- α, β, γ generic parameters
- $P(q)$ willingness to pay
- ϵ demand elasticity
- R revenue
- R_m marginal revenue
- W_d, W_D consumer surplus (demand side)
- W_S producer surplus (supply side)
- W welfare (market surplus)
- v individual value of an item
- V aggregate value of an item
- z^* efficient superscript
- z^M monopoly superscript
- z^B Bertrand superscript
- z^C Cournot superscript
- $\bar{z}, \hat{z}$ special values for z

Dedication

This monograph grew out from my teaching notes into a full-fledged book as I felt compelled to offer my students a coherent and exhaustive vision of the field. I wish to acknowledge the positive influence of Sylvain Sorin, Patrick Rey and Bruno Jullien during my PhD years at ENSAE–CREST–INSEE (Paris, France) and of Jean Gabszewicz, Claude d'Aspremont and Jean François Mertens during my Postdoc years at CORE (Louvain, Belgium).

This manuscript was prepared during a decade on four Apple laptops using many freewares related to the LaTeX typesetting project (TeXShop, teTeX, Ghostscript, Smultron, BibDesk, Excalibur), spell-checked in US-English and typed with the free fonts *Utopia* from Adobe Systems Incorporated and *Fourier* from the GUT*enberg* association.

To my father René (1^{st} edition)

To my brother Gilles (2^{nd} edition)

Part A

Introduction

Chapter 1

About the Book

In this first chapter, we summarize the content of the book to give the reader a broad perspective of what Industrial Organization (aka Industrial Economics) is about. Next, we discuss our methodology from several points of view and the prerequisites needed to follow the exposition.

1.1 Foreword

The first thing to do before reading a book about the organization of industries is to learn a little about the main sectors of economic activity in the advanced economies we aim to study. Figure 1.1 illustrates how services have gain the upper hand over manufacturing during last century in the US.

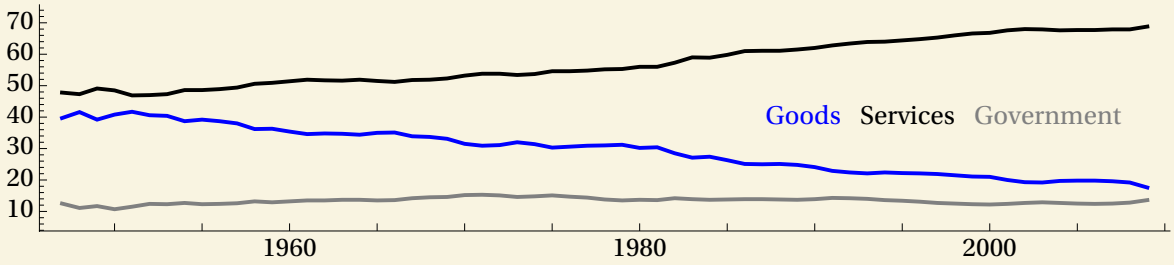


Figure 1.1: Goods vs. Services in the US (in % of GDP)

Table 1.1 using data from EuroStat and the US BEA goes into finer detail to compare these major economies.[1@] Given the annual GDP of the EU, each ‰ represents 12bn€ of value creation or roughly 24bn€ of sales. Likewise for the US, each ‰ represents 14bn$ of value creation or roughly 28bn$ of sales.[2@]

Area	**EU**			**US**		
Sector	*1995*	*2000*	*2007*	*1998*	*2003*	*2008*
Agriculture	2.7	2.2	1.7	1.1	1.0	1.1
Mining	0.9	0.9	0.8	0.9	1.2	2.1
Food, Tobacco	2.4	2.2	1.9	1.6	1.6	1.3
Apparel	1.2	1.0	0.6	0.6	0.3	0.2
Wood, Paper	1.2	1.1	0.9	0.9	0.7	0.6
Publishing	1.2	1.2	0.9	1.5	1.4	1.3
Fuel	0.3	0.3	0.3	0.6	0.7	1.0
Chemical	2.1	2.0	1.8	1.7	1.5	1.5
Plastic	0.9	0.9	0.7	0.7	0.6	0.5
Mineral	1.0	0.9	0.7	0.5	0.4	0.3
Metal	2.7	2.6	2.5	1.8	1.2	1.3
Machinery	2.2	2.1	2.0	1.3	0.8	0.9
Electronica	2.3	2.5	2.0	0.4	0.4	0.3
Vehicle	2.0	2.0	1.8	1.2	1.1	0.6
Furniture	0.8	0.8	0.6	0.9	0.9	0.7
Energy, Water	2.7	1.8	2.1	1.9	1.7	1.8
Construction	6.0	5.5	6.3	4.4	4.6	4.4
Wholesale trade	7.0	7.0	6.9	6.3	5.7	5.7
Retail trade	4.6	4.5	4.1	7.1	6.9	6.0

Area	**EU**			**US**		
Sector	*1995*	*2000*	*2007*	*1998*	*2003*	*2008*
Hotel	2.6	2.9	3.0	2.7	2.9	2.8
Land transport	2.7	2.7	2.5	1.4	1.3	1.3
Air, water transport	0.5	0.6	0.5	0.7	0.6	0.5
Support for transport	1.0	1.2	1.4	1.1	1.0	1.1
Telecommunications	2.5	2.4	2.3	2.5	2.4	2.4
Banks	3.7	3.5	3.6	3.2	4.1	3.4
Insurance	0.9	0.8	1.1	2.6	2.5	3.2
Financial services	0.5	0.6	0.9	1.5	1.6	1.8
Real estate	10.2	10.7	11.7	12.1	12.7	12.8
Renting services	0.9	1.0	1.0	1.4	1.5	1.5
Computer services	1.2	1.7	1.8	1.4	1.6	1.7
R&D	0.6	0.7	0.7	3.9	4.2	4.8
Other business	6.4	7.2	8.0	3.8	4.0	4.1
Administration	6.8	6.3	6.1	12.5	12.9	12.8
Education	5.1	5.1	5.0	0.8	1.0	1.0
Health	6.4	6.6	7.2	6.0	6.7	6.9
Refuse	0.5	0.6	0.6	2.7	2.9	2.9
Leisure	2.2	2.4	2.4	0.9	0.9	1.0
Personal services	1.4	1.4	1.4	2.8	2.6	2.5

Table 1.1: Industry Shares of GDP (in %)

1.2 Purpose

Our aim in this book is best understood by looking at the historical development of the industrial organization field. According to the standard definition, the economic science studies the *allocation of scarce resources among alternative uses to achieve desired objectives*. Industrial Organization focuses on the firms that are involved in this process.

On Competition

The synonyms for *competition* found in a dictionary are rivalry, contest, conflict, strife, struggle and quarrel. All convey negative connotations of waste, violence, compulsion, insanity and above all irrationality ; as such they can hardly be associated with the advancement of humanity. Yet, the *competitive market* is the Holy Grail of economists and serves as the benchmark for normative analysis (here and elsewhere). This paradox takes its root in the oblivion of the conditions surrounding the idea of economic competition.

It is only in a society where violence and coercion are checked by a benevolent but powerful State that the pursuit of selfish interest is canalized into hard-work, innovation and risk-taking instead of theft, fraud, deception or influence. Most of this book is set in this ideal world whose practical incarnation is the free market democracy (e.g., US, EU). However, we are lead to deal at the margin with non productive behaviors such as

collusion, predation, exclusion or rent-seeking. It is therefore necessary to have a good understanding of how the free market democracy ever appeared and what is the current legal framework in these societies. The first part of this book undertakes this task by succinctly studying institutions, to wit the state apparatus (and all the actors revolving around it) and the legal framework bearing on firms.

Tradition

The economic theory tradition focuses on private property, free markets and the price system as means to produce and distribute goods and services. This institutional arrangement is backed formally by the general equilibrium theory and the welfare theorems; however reality often fails to conform to their preconditions and predictions. Industrial Organization is the field of economics that tries to address some of these discrepancies. To take just one obvious issue, we observe oligopoly, that is to say a few active firms in the vast majority of markets, instead of the large number of competing firms that perfect competition calls for. As a consequence, the rational minded entrepreneurs who run these few firms do not act as price-takers but strategically; they recognize their interdependence. The neat conclusions of the perfect competition model therefore cease to hold.

Starting with Cournot (1838) and Marshall (1890), economists have developed a body of theory to address these inconsistencies; the paradigm that has been in vigor until the 1970s is called *structure conduct performance* (SCP).

Structure studies how far is the actual industrial activity from being a competitive one. One looks at the number and size of sellers and buyers in the various markets that form the production chain, at the differentiation of products and finally at the entry conditions.

Conduct studies how firms compete when they cease to be price-takers (as a consequence of the non competitive structure). We are interested by all strategic behaviors like collusion, differential pricing, advertising or R&D.

Performance deals with the departure from the productive efficiency achieved in a competitive market. From a static point of view, we are interested by the effects of market power (profitability and cost efficiency), while the dynamic point of view raises the difficult question of the optimal pace of innovation. For instance, how long shall a patent protect the innovator to encourage innovation.

A major implication of the SCP paradigm is to justify governmental interference in the economic life. Antitrust laws and regulations are passed with the aim of restoring the efficiency lost when firms take advantage of their market power. In the US, throughout

the XXth century, legal activity has been intensive and enfored by the federal commissions and the supreme court. In Europe similar rules started to be enforced actively by the Commission in the 1970s.

Critique

The dominant SCP paradigm was challenged by the Chicago school of thought in the 1960s on two grounds. Firstly, governmental intervention is based on loosely demonstrated market failures like the abuse of market power; the remedies proposed need not improve on the problem if for instance, pressure groups succeed to capture their regulator, thereby reinforcing their inefficient dominant position. Secondly, the SCP paradigm is entirely driven by price theory and anonymous market transactions, thus it totally ignores the complex *contractual* bilateral relations that exist inside firms, between firms or among firms and their providers, their clients or their regulator. For instance, a firm can voluntarily restrict the number of its suppliers to limit competition among them, it can set-up a complex pricing formula rather than a fixed per-unit price to screen customers, or an employer can offer its employees a wage schedule contingent on performance in order to give them stronger work incentives. The persistence of these practices must be explained and their degree of efficiency must be explored.

The explanatory power of "price theory" for this endeavor is limited because it studies how actors interact inside a market, being forced to obey its rule i.e., they are allowed to choose only one variable like a price, a wage or a quantity. However, it is well known that firms juggle with many decisions at the same time, with issues that belong to different dimensions. Contract theory encompasses this complexity by extending the available choices of the actors and by explicitly considering other transaction mechanisms beyond the market one. At a higher level, institutional economics search for the factors that lead firms to adopt a different trading mechanism. For instance, why do firms integrate or why do government privatize formerly public industries. Transaction cost or property rights are among the proposed answers. These new insights, originally based on cognitive logic, have been integrated into the mainstream of microeconomics, a fact we wish to reflect here.

The Chicago paradigm also takes a more dynamic view of economic affairs and argues that the inefficiency generated by market power is of a temporary nature and is bound to be eliminated by the entry of more innovative firms. It therefore calls for a sequential analysis that considers the competition **for** the market on top of the competition **in** the market, the idea of contestable markets. To summarize, the Chicago contention is that competition might be the worst economic system except all those other forms that have been tried previously in history (misquoting Winston Churchill on democracy).

Synthesis

Received Industrial Organization implicitly focused on *market competition* as the sole productive behavior conducive of Pareto improvements for society. Modern research has shown that direct *contractual relationships* (e.g., bilateral monopoly, integration, agency theory) offer an alternative path for wealth creation. They have made their way into IO volumes but mostly as side dishes. We have tried to rethink the structure of the text in order to integrate this facet in a more coherent way.[3@]

Another recent avenue of microeconomics research we contemplate is unproductive *rivalry* over existing wealth. Organized contests (e.g., auctions, regulation) or spontaneous conflicts (e.g., rent-seeking or patent races) are interactions mediated by either formal or informal rules. The large literature on this topic has yet to be incorporated into an IO textbook in a systematic fashion; ours is a first intent.

For the reasons exposed above, it is our belief that modern Industrial Organization theory ought to acknowledge the role of contracts, explicit as well as implicit. We thus take a *Contract Based* approach to *Industrial Organization* (hence the acronym IOCB). The SCP framework of analysis remains our guiding scheme but the study of conduct is extended from pricing behavior to contracting behavior and other forms of rivalry.

We mostly adopt a positive view as we describe how competing firms operate in order to identify the extent of their market power. We then inquire the implications for all actors, including consumers or government. When possible, we use the consumer surplus and market welfare concept to make normative judgments. Thus beyond, trying to explain how firms acquire or maintain market power, we also seek to shed light on what policy makers can do about it. The notable absence in this book are empirical studies of IO; a good starting point is Einav and Levin (2010).

Regarding the style adopted for the crafting of this text, two features are worth explaining. Contrariwise to the ahistorical perspective of most economics monographs, we have tried to associate each theory or concept with its original author in order to link the appearance of novel ideas to their epoch. We believe that political and technological developments in the advanced economies where these scientists lived had a definite influence over what to study and how to approach issues. In a sense, economic theory is a lively science very much related to the real economy, rather than an ordeal escaped from an ivory tower.

The other novel feature of this book is technical as we extensively cross-reference sections but also apparently unrelated topics who nevertheless share the same basic model or derive from the same fundamental economic problem. Thus, only the original model is fully developed while later applications are more succinctly presented. Our desire here is to illuminate how a few bare bones support a wealth of theories. This also

allows the reader to quickly find all the sections where a concept is being used.

1.3 Summary

As the adage goes, a picture is worth a thousand words. We thus try with Figure 1.2 to convey a graphical overview of the topics we approach in this book. The 8 gray boxes indicate Parts while the other boxes more or less refer to the chapters. Locations on the chart do not respect truthfully the content, rather they loosely reflect connexions with each other.

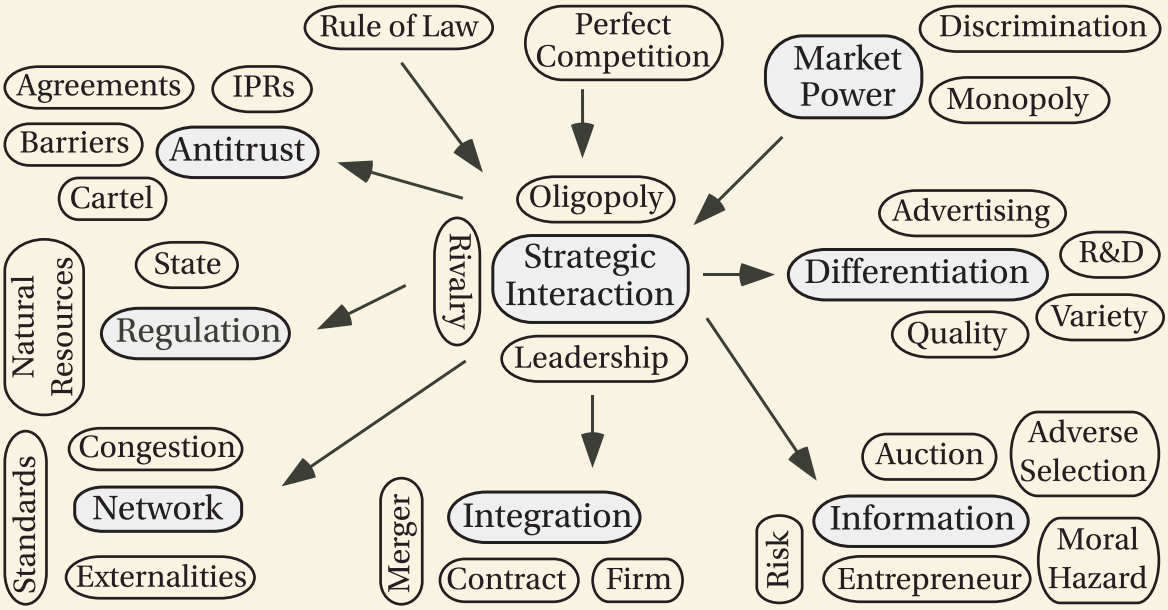


Figure 1.2: Book Scheme

After this introductory chapter, we recall the basics of the perfect competition paradigm and present the normative concept of welfare used throughout the book to judge and compare economic outcomes (ch. 2). We also review quickly some elements of game theory.

Market power is the ability to raise the market price of its product without losing all customers to competitors. We delve into this issue by focusing on the least competitive market structure where stands a unique seller, the *monopoly*. We characterize the optimal conduct and the resulting allocative inefficiency of the monopoly (ch. 3). We then study *differential pricing*, the wide range of discrimination strategies that a firm can adopt to take advantage of her market power (ch. 4).

An *oligopoly* is a market structure where a limited number of firms with market power compete among themselves, realizing their interdependence when taking strategic decisions on price, output and quality. We introduce the concept of strategic behavior

and the game theoretical concept of Nash equilibrium to study the determinants of this competition in a variety of situations (ch. 5). On top of being empirically relevant, the oligopoly is also the theoretical long term outcome of a perfectly competitive market where entry is only limited by the existence of fixed costs. We thus examine this entry dynamic as well as other strategic decisions that firms can take to improve their lot (ch. 6). Lastly, we complement market driven rivalry with pure economic rivalry (ch. 7).[4@] We offer a variety of models dealing with conflicts, contests, patent races, rent-seeking, political pressure.

Armed with the previous theories and their predictions, we take on *anti-competitive* behavior. At the root is the fact that competition harms profits so that a firm may rationally decide to adopt practices harmful for society in order to obtain or defend a profitable incumbent position. Yet, fair and undistorted competition is a cornerstone of any market economy. We thus present the playing field upon which our theories are constructed, that is to say the legal framework of the European Union and of the United States of America (ch. 8). Their *antitrust* policies pursue the goal of defending and developing effective competition. The next two chapters are devoted to modeling potentially anti-competitive behaviors. We first deal with the elimination of concerted practices such as cartel, collusion or restrictive agreements (ch. 9). We explain why they reduce welfare and how they are combated. Then we look at abuse of their dominant position when firms try to block entry or force exit of challengers with limit pricing, preemption, predation, foreclosure or attrition (ch. 10).

The next part focuses on non-price strategies whereby firms seek alternative ways to relax competition. *Differentiation* and *innovation* aim to make a product nearer to consumers wants so as to make it relatively more attractive than competing proposals and thereby obtain a lawfully exploitable market power. We first study how competition is affected when products are differentiated (ch. 11). The *vertical* or quality differentiation amounts to design a product to make it more attractive to all potential buyers while *horizontal* or characteristics differentiation amounts to expand a basic product into a whole range to address the heterogeneity of customers. We also treat *advertising* which is to make one's product known to all, and *branding* which is to associate the product with a particular group of consumers. Lastly, chapter 12 deals with innovation understood as a broad concept, of either a technical or subjective nature. Firms invest in research and development (*R&D*) to develop new products, new features or new ways of doing things (faster or cheaper). We look at the framework where innovation takes place, then study the incentives for firms to invest in that activity; we end with a very important byproduct of the innovation process is its legal protection which is treated in §12.3 on intellectual property rights (IPRs).

The part dedicated to *integration* gathers all the theories dealing with the expansion (and contraction) of firms, how they grow from handicraft to conglomerates via merger and acquisition or why they suddenly decide to concentrate on core activities and home markets by divesting unwanted units. Stated in other words, we try to identify the boundaries of firms. It is customary in economics to distinguish *vertical* and *horizontal* integration i.e., whether the two firms about to merge make *complementary* or *substitute* products. In the first case, the decision to acquire a provider or a distributor is related to the age-old "make or buy" quandary: what are the transactions more efficiently conducted inside a firm, on a market or face-to-face with another firm? Horizontal integration regards mergers involving similar firms who look to reach a "critical mass" and take advantage of scale economies and complementarity w.r.t. geography, brands, or segments.

Both dimensions of integration are therefore treated together in this part. We define a firm and how it works, before presenting the various economic theories of the firm as well as a sketch of contract theory (ch. 13). Then, we tackle issue of vertical integration such as the bilateral monopoly, specific investment and the hold-up problem (ch. 14). We also compare formally the so-called transaction cost and property rights theories of the firm. Lastly, we treat horizontal integration by presenting the numerical methods used by antitrust authorities to decide whether competition is endangered or not by a merger (ch. 15).

One of the novelty in this book is to dedicate a part to public oversight. The state (ch. 16) is the superstructure gathering all the organizations in charge of mediating the economic activity; it includes the legislative, judicial and executive bodies and within the latter, the government, its bureaucracy and regulatory agencies. We seek to explain its dual role as judge and party and how its origin impinges on today's legal framework. We present succinctly the State and look at the various rationalizations for its encroachment upon economic life. We then look at the related activities of rent-seeking and the liberalization trend of the last decades. The next chapter is devoted to *regulation*, the direct oversight of the State upon markets thought to suffer from failures such as the exercise of market power (ch. 17). We end this part with a synthetic treatment of natural resources as these markets display market power and externalities, and thus stand as perfect candidates for public oversight (ch. 18).

This next part introduces the reader to *information economics*; it builds on incentives and asymmetric information to fruitfully complement the traditional price theory approach of Industrial Organization. This part is the largest in the book due to the wealth of material to be exposed and the necessity to introduce some background first. As in the previous part, we relinquish the market playing field to focus on the firm itself.

Our first task is to update microeconomics concepts regarding *risk* and *uncertainty* where imperfect information is added to the standard microeconomic theory (ch. 19). We then focus on incentives and hidden action known as *moral hazard* (ch. 20) before moving more specifically to asymmetric information with the concept of hidden information known as *adverse selection* (ch. 21). We then give an account of auctions (ch. 22), a competitive trading mechanism used to extract or reveal information; it extends adverse selection. Lastly, we encroach upon the field of corporate finance to examine how these asymmetric information phenomena impinge upon the financing of an *entrepreneur*-owned firm (ch. 23).

We wind up the book with *network industries*, characterized by network externalities, whether on the demand side or on the supply side. The study of externalities does not belong per-se to the field of Industrial Organization but their presence in many oligopolistic or regulated industries shape the strategies of firms and regulators. The technological features giving rise to externalities are quite different according to whether we study goods or services so that different chapters are called for. Chapter 24 on standards and components deals with oligopolistic markets for goods where a positive externality is generated by the demand side. The main issue for the competing firms is to decide on their degree of differentiation; they can either make products compatible with their's challengers or not. Chapter 25 deals with service industries relying on a physical networks e.g., transportation, energy or communications. Both positive and negative externalities are present here. The positive one comes from the supply side as the technology displays increasing returns to scale (natural monopoly). The negative network externality is the congestion created by excessive use. This singularity requires a specific market design since a competitive market would never emerge.

1.4 Methodology

We start with the methodology of economic science before narrowing our scope to the topic of this book and conclude with the prerequisites for a fruitful reading. Readers accustomed with economic methodology may want to skip this section; we nevertheless invite them to (re)read the following timeless quotations.

Knight (1933) There is no more important prerequisite to clear thinking in regard to economics itself than is recognition of its limited place among human interests at large.

Pareto (1897) Economic science does not attempt to establish any particular method of economic organization, and it is not the business of science to do so. Science

does, however, attempts to solve problems of the following kind: (1) What are the effects of a regime of free competition? (2) What are those of a regime of monopoly? (3) Those of a collectivist regime? All these questions must, of course, be treated, not from a polemical point of view, but solely for the purpose of ascertaining what results would follow upon their installation. It is especially necessary for us to discover what relation these results bear to the aggregate well being of humanity.

Poincaré (1905) Science is facts; just as houses are made of stones, so is science made of facts; but a pile of stones is not a house and a collection of facts is not necessarily science.

Solow (1956) All theory depends on assumptions which are not quite true. That is what makes it theory. The art of successful theorizing is to make the inevitable simplifying assumptions in such a way that the final results are not very sensitive. A "crucial" assumption is one on which the conclusions do depend sensitively, and it is important that crucial assumptions be reasonably realistic. When the results of a theory seem to flow specifically from a special crucial assumption, then if the assumption is dubious, the results are suspect.

Dasgupta (2002) Critics of modern Economics take refuge in such aphorisms as that "it is better to be vaguely right than precisely wrong". What this misses, however, is that you won't even know if you are vaguely right if you operate within a framework in which you cannot be precisely wrong; there is no way to controvert a vague statement.

1.4.1 Scientific Methodology of Economics

Social sciences are often called the "soft" sciences in opposition to the "hard" sciences formed by physic, chemistry, biology and all their derivatives. Economic theorists like to see their field as the hardest of the soft sciences because it has adopted the modus operandi of pure science, namely the construction of theories and models that are worked out using the hypothetico-deductive method (cf. Poincaré (1905)). It is worthwhile noticing that most of the authors who contributed the material presented in this book were trained as engineers or mathematicians before turning to economics; their achievements based on the use of advanced mathematics have cleared questions that intuition alone could not dominate.

An economic model is a reflection upon past observations and experiences aiming at understanding the existing relationships among economic agents or between different situations. It tries to single out principles or laws that could apply to a class of problems;[5@] this way we can anticipate the outcome of new circumstances as they appear or

propose policy changes that are likely to generate a movement towards a more (socially) desirable situation. Practically, an economic model should be based on assumptions that we derive from the observed reality, taking care to select the most relevant ones leaving aside details. There should be enough components (bare bones) to enable the derivation of non trivial results but not too much to avoid inextricable complications.[6@]

The analysis of the model is the step where economics enters the domain of logic. The implications we derive may confirm or invalidate our original views; in the latter case the analysis helps us to comprehend what and where something got wrong, whether a false assumption or an invalid idea. The interaction between ideas and models is thus bidirectional; it is a process of trial and error that permits to devise models that come always closer to explaining the real situation. Lastly, whether a model turns out to be useful depend on how and why we use it. Computing the trajectory of a tennis ball is quite different from that of a satellite; it should therefore not be used as a proxy of the satellite's trajectory because one would almost surely lose the satellite.

Our approach in this book is to bring forth in the simplest way theories and results that, after being challenged and tested by academics and practitioners, have passed the test of time. We mostly draw on the recent literature who authors aim at uncovering the simplest assumptions needed to generate interesting conclusions. We expose them without excessive mathematical rigor in order to devote more space to motivation, historical cases, relevance and above all economic intuition.

1.4.2 Individual Rationality

Thesis

As we already explained in the book's summary, our fundamental tool is game theory;[7@] its use is warranted by our adherence to the principle of *individual rationality* which is almost universal in modern economics and derives itself from the wider philosophical doctrine of Methodological Individualism. The basic tenet of the latter is that all collective phenomena are the outcome of the actions and interactions of individual decision makers as well as of the traditions created and preserved by them.

Individual rationality states that people make rational choices not foolish ones. More precisely, a decision-maker chooses the best action according to her preferences, among all the actions available to her. No qualitative restriction is placed on the decision-maker's preferences; her "rationality" lies in the consistency of her decisions when faced with different sets of available actions, not in the nature of her likes and dislikes. An economic decision maker is thus seen as a different breed of human called Homo Economicus whose main, if not unique, motivation is greed.

Antithesis

Everyone will agree with Leijonhufvud (1993) that *economic theory* describes the behavior of *incredibly smart people in unbelievably simple situations* whereas reality is more one of *believably simple people coping with incredibly complex situations*. Over the last decades, the study of the human brain has shown that people's mental limitations rule out the perfect rationality embodied in "Homo Economicus". This is because many of our decisions and behavioral responses are in fact reflexive and emotional (we have no time to think), thus bound to be ridden with mistakes. Yet, the learning process we experience through life is precisely aimed at integrating useful and valuable reflexes to avoid future losses. Thus, our mistake-ridden decision processes are, on balance, efficient and welfare enhancing. This aspect is further enhanced when individual behavior is aggregated among numerous people; it does not matter if a few people make mistake when the vast majority follows a rational path, the overall pattern of behavior will be almost rational. In a nutshell, the interaction of "ordinary" people in markets tends to produce "incredibly smart" results.

An important factor, as far as this book is concerned, is that the decision maker is not a lonely human being but a collectivity known as the firm. Thus, although perfect rationality poorly matches human behavior, it does a better job when it comes to firms and organizations. In the same vein, many empirical irrationalities are macro phenomena with a rational micro foundation i.e., people act rationally given the informative and processing constraints they face, but market or government failures aggregate their individual behavior in an inappropriate manner. Most of these instances share an analytical similarity with the prisoner dilemma whose solution occupies a prominent place in this work.

The obsession of "Homo Economicus" with money is another source of criticism. A wide range of experiments based on behavioral game theory have shown that honesty, integrity, intrinsic job satisfaction, and peer recognition are powerful motivators, and lead to better results for contracting parties than reliance on financial incentives alone. Hence, we rationally use a wide variety of medium beyond money to interact in society. It must be understood though that other motivations are not denied or deemed inconsequential; rather, they are ignored for the time being to allow analytical progress. In a sense, this happens because economists have themselves limited intellectual faculties and are thus forced to simplify their theoretical models. In this respect, economics is no different than other sciences and over the last decades, it has started integrating the full range of human motivations within the basic staple of rational choice.[8@] We echo this effort at some points in the book.

Next "Homo Economicus" is too polite because for him "a handshake is a handshake".

More generally, the standard framework of neoclassical economics is a dream world disconnected from reality. In Walrasisan models, agents are anonymous because they transact at no cost on markets, not with actual people. There is no hierarchy in the firm i.e., capital may hire labor or the reverse. Firms have no power as they only produce to serve the wants of consumers using the price signals transmitted by markets. Similarly, people have only purchasing power. Life as we know it is far bleaker and full of power play. As recalled by Williamson (1985) (p. 51), modern economics must go beyond market exchange where one only seeks to buy cheap and sell dear to include "the full set of ex ante and ex post efforts to lie, cheat, steal, mislead, disguise, obfuscate, feign, distort and confuse". This novel aspect is fully dealt with in Chapter 7 and more generally, the previously mentioned weaknesses are tackled along the book.

Synthesis

In view of the stream of criticism, one may wonder why has the “Homo Economicus” paradigm been so fruitful for economic analysis and so lonely. One possible answer is to observe with Myerson (1999) that no one has yet developed an accurate and tractable theory of the inconsistency and foolishness in human behavior; so our best analytical models are based on the rationality assumption for lack of any better foundation. But a more compelling answer starts by recognizing that social science is not about predicting human behavior, but about analyzing social institutions and evaluate proposals to reform them.

When looking for potential flaws in a social institution like the senate of a democratic country, it is helpful to assume that the agents interacting in the institution, the senators, are not themselves flawed (or insane). Otherwise, when citizens complain about the institution, we cannot say whether the institution should be reformed or whether senators should get better education! To do any kind of analytical social theory, we must describe institutions and predict the individual’s likely behavior inside these institutions. To handle normative questions (what should be done), there must also be some concept of human welfare in our model. If we assume that some individuals are not motivated to maximize their own welfare (as measured in our model) or that some individuals do not understand their environment (as predicted in our analysis), then any loss of welfare that we find in our analysis can be blamed on such dysfunctional or misinformed individual behavior, rather than on the structure of social institutions. This means that a defense to reform social institutions (rather than reeducating individuals) is most persuasive when based on a model which assumes that individuals intelligently understand their environment and rationally act to maximize their own welfare. Only in that case can they complain about the poor performance of their institutions in a manner

that is meaningful.

1.4.3 Prerequisites

This textbook intends to provide a complete panorama of Industrial Organization building on many separate modules forming, we hope, a coherent whole (cf. Figure 1.2). The prerequisites for a successful reading are low but not nil; some knowledge of microeconomics[9@] will help the reader to grasp IO ideas faster. For completeness, chap. 2 provides a quick review of all the useful concepts. Next, we assume that actors (firms, managers, consumers, governments) are rational so that their interactions can be analyzed by the models of *game theory*. To avoid duplication with excellent material already published, we only offer a minimal introduction in §2.4.[10@]

The conciseness achieved in this book has been made possible by recurring to a certain amount of (mathematical) formalism, in an amount sensibly greater than for similar economics manuals. We therefore assume some acquaintance with mathematics. The reader should know a minimum of functional analysis and linear algebra e.g., the concepts of function, derivative, maximization and linear systems of equations.[11@] It is the author's belief that mathematical formulae ought to be kept to a minimum in economic textbooks; as a consequence most space is devoted to intuitions, informal theories and examples. The other side of the coin is a certain dryness in the formalization but it ought not be a difficulty, neither for scientifically trained readers nor for economic students who have already seen the basic tools of optimization in economics.

Mathematical Optimization

The one fundamental tool the reader should keep in mind is exposed on Figure 1.3. In a typical economic problem, an agent cares for some dimension y such as profit or utility that is determined by some variable x such as quantity, price or quality through the relationship $y = f(x)$; it is shown identically over the four panels. The unconstrained maximum of f is achieved at x_0, the solution to the first order condition $f'(x) = 0$ (FOC) i.e., when the curve becomes flat. This mathematical optimum is NOT the economic optimum, it is only a candidate because in most situations, the decision maker faces restrictions such as $x \geq x_1$ and $x \leq x_2$ that arise from his interaction with other economic agents or with the market.

On panel ❶, these conditions define two stripped areas of allowed values; their intersection, the shaded area, is the economic domain from which the decision maker can pick. Since x_0 belongs to this domain, the economic optimum is $x^* = x_0$. On panels ❷ and ❸, the economic domain does not contain the mathematical optimum so that the economic

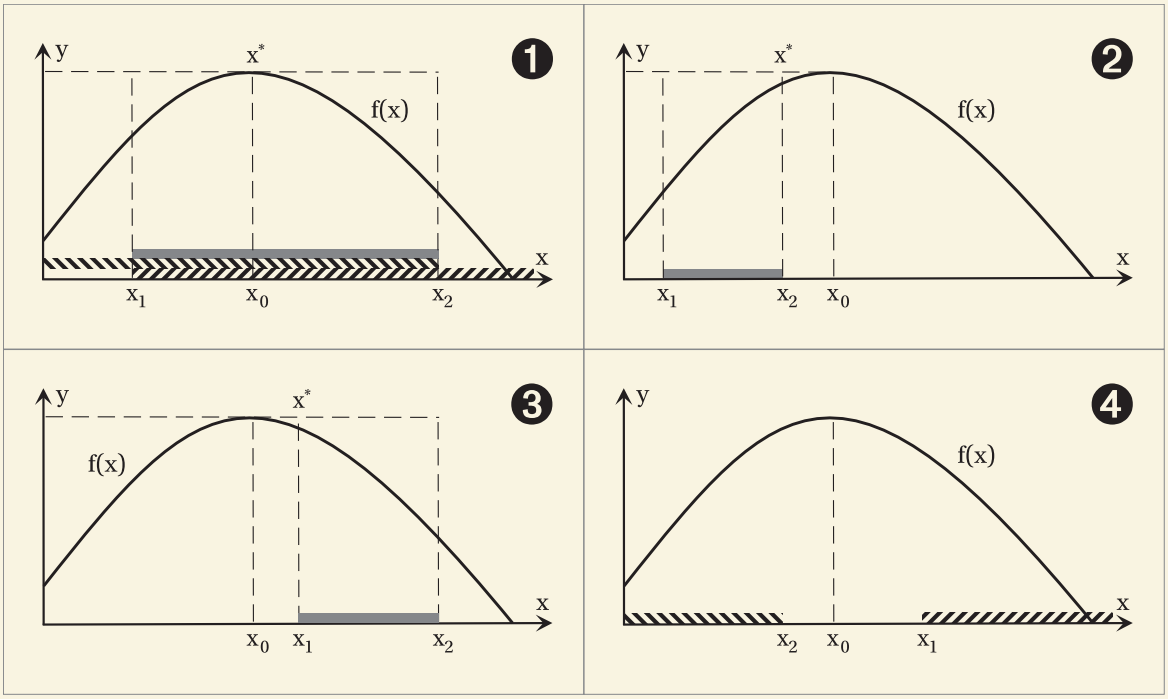


Figure 1.3: Mathematical Optimization

optimum is a corner solution, respectively the maximum and minimum allowed levels for the endogenous variable ($x^* = x_2$ resp. x_1). Lastly, panel ❹ displays the case where the restrictions are incompatible among themselves and leave no way for the agent to make a pick. His optimization problem, in that case, has no solution. In retrospect, we may say that the economic problem was badly formulated.

Chapter 2

Microeconomic Foundations

This chapter recalls some key results of microeconomic theory and game theory. The concepts of marginal cost and willingness to pay are basic tools for the analysis of firms behavior while the efficiency of the competitive equilibrium will stand as a benchmark for our (limited) normative analysis (cf. §2.3.3 for a discussion).

2.1 Supply

The neo-classical theory of supply encompasses the notions of production, cost, competitive supply and some miscellaneous properties relating to the multiplicity of plants and products.

2.1.1 Equi-marginal Principle

The equi-marginal principle (aka Gossen (1854)'s second law) is a staple of neoclassical economics worth recalling here. In many static optimization problems, a rational decision maker must allocate scarce resources over a variety of employments. The "manna" trick is a procedure or algorithm that repeatedly applies arbitrage to derive an equi-marginal principle as a property of an optimal allocation. It works as follows:

- Compute the value of a small gift when allocated to a specific employment
- Compare these values for all possible employments and allocate the gift to that with the highest value.
- Observe that the marginal value of the winner employment has decreased
- Prove that the wedge between winner and forerunner is reduced
- Divide a large gift into a series of smaller ones
- Reason that incremental allocation will bring all marginal values in line
- Conclusion: at the optimum, marginal values in all employments are equal

2.1.2 Production

In this section, we look at technology per-se and productive efficiency.

Technology

The neoclassical theory reduces a firm to a single, rational minded, economic agent, the entrepreneur. He/she owns or knows a production *technology* completely and uniquely characterized by the maximal quantity $\Phi(K,L)$ that can be produced by using K units of capital measured in €'s and L units of labor measured in hours of unqualified work.[1@]

The productivity of the input factor K (resp. L) is defined as $\Phi_K \equiv \frac{\partial \Phi}{\partial K}$ (resp. Φ_L), it is positive and often assumed decreasing with the input (the quantity of the other input being fixed). An isoquant is the locus of capital/labor pairs (K,L) such that production $\Phi(K,L)$ remains constant. On the isoquant, total differentiation[2@] yields the marginal rate of technical substitution (MRTS):

$$-\frac{dL}{dK} = \frac{\Phi_K}{\Phi_L} \tag{2.1}$$

which values one unit of K in units of L; it is the subjective[3@] price of K (in units of L).

The technology exhibits constant returns to scale (CRS), decreasing returns to scale (DRS) or increasing returns to scale (IRS) if an increase of $\lambda\%$ of the factors generate an increase of production equal, lesser or superior to $\lambda\%$. In the presence of DRS, several small production units are call for in order to achieve productive efficiency. On the contrary, all inputs should be grouped into a single production unit under IRS (cf. check when each applies).

A frequently used technology is Cobb and Douglas (1928)'s $\Phi(K,L) = AK^\alpha L^\beta$ with $\alpha, \beta > 0$. Notice that the A parameter can always be normalized to unity by a convenient choice of physical unit of production (but then the price cannot be treated likewise). There are DRS if the sum $\alpha+\beta$ is lesser than unity, IRS if greater and CRS if equal. Notice that when the technology parameters satisfy $0 < \alpha, \beta < 1$, we have Φ_{LL} and $\Phi_{KK} < 0$ which corresponds to the usual assumption of "decreasing marginal productivity". However, the sum $\alpha+\beta$ can be anywhere between zero and 2, thus any kind of returns to scale are possible; this is so because the complementarity between the inputs mutually reinforce their productivity although each by itself displays DRS.

Productive Efficiency

Productivity Analysis is one of the most important econometric instrument at the disposition of public authorities to improve the effectiveness of an industry both in terms of

costs and quality of service. As we shall later comment in §17, regulators use productivity analysis to implement yardstick competition among regulated firms belonging to the same sector or industry. We present below the intuition underlying the theory originally developed by Farrell (1957) to judge in a practical way of the productive efficiency of (comparable) firms.[4@]

Consider a good or service and the "state of the art" technology to produce it. The minimal combinations of factors (inputs) K and L necessary to produce an output Q form an efficient isoquant displayed on the left panel of Figure 2.1. Any firm in the industry will have to use a larger combination like A to reach the same goal; thus we can speak of the ratio $\frac{OB}{OA} \in [0;1]$ as the *technical efficiency* of A. If factor prices are such that the cost minimizing combination is C (tangency between cost line and isoquant) then the ratio $\frac{OD}{OB} \in [0;1]$ can be interpreted as the *price efficiency* of the factors proportion used in A (and B). Now, the product ratio $\frac{OD}{OA} \in [0;1]$ can be meaningfully regarded as an index of *productive efficiency* obtained by multiplication of the previously defined indexes. If the technology under scrutiny has constant returns to scale then the ratios of two different firms can be compared although the size of firms (their output) are different.

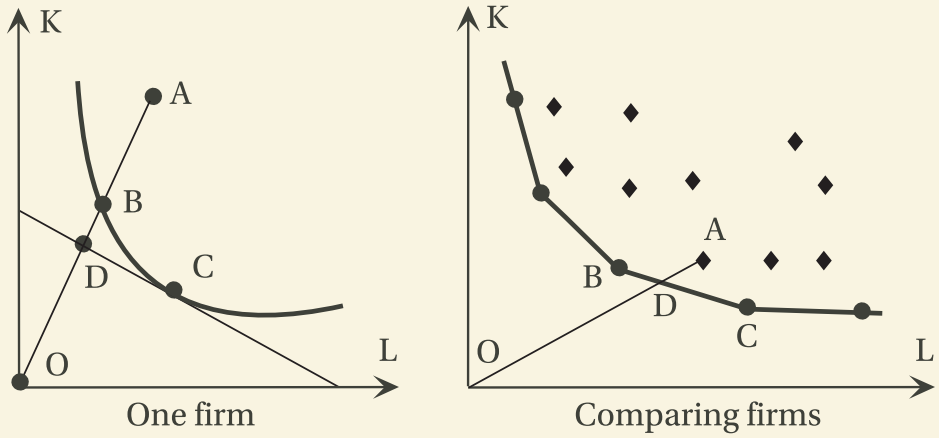


Figure 2.1: Productive Efficiency

The previous construction is very elegant but hinges on some "state of the art" technology that even a careful engineering study would fail to adequately identify; an easy way out of this awkwardness is to build an empirical production function based on the best results observed in practice. Since firms sales are different, the quantities of factors they use have to be scaled by total output value (measured by sales) so that the axes on the right panel of Figure 2.1 are "inputs per unit of output". Plotting together the combinations of factors used by different firms is meaningful if we keep assuming constant returns to scale. The lower contour of the scatter of points defines the empirical efficient isoquant (for one unit of output); firms on that curve (dots) are deemed efficient while

the others (diamonds) are deemed inefficient. For one of them such as A, an algorithm identifies the nearest efficient points B and C and use the ray to the origin to construct D, a blend of B and C, which uses the factors in the same proportion as A; the index of efficiency is as before $\frac{OD}{OA} \in [0;1]$.

2.1.3 Cost

In this section, we review the notions of fixed cost, marginal cost, impact of long-run and economies of scale.

Fixed Cost

To start its productive activity, a firm has to incur a start-up costs F including expenditures on Research and Development (R&D), a government license, organizing a marketing department and most importantly building or acquiring facilities for production, transportation, communications and distribution.

Accounting and economics differ in their treatment of the cost of using an asset for production in two fundamental aspect. Firstly, the *opportunity* cost of the asset is the net revenue from the best alternative use (i.e., pursuing another activity with this asset). Secondly, part of the fixed costs are *sunk cost* in the sense that they have been committed and cannot be recovered. The reason is that some activities require specialized assets that cannot readily be diverted to other uses. This applies to the capital assets listed above. Examples are investments in machines which can only produce a specific item, the development of products for specific customers, advertising expenditures and R&D expenditures. For economic reasoning, those sunk cost are not taken into account to take future decisions i.e., we adhere to the proverb “let bygones be bygones”. Accounting, on the other hand, keeps them in the books.

Another important characteristic of a fixed cost is its indivisibility. Most industrial projects require an investment of a minimum size in which case the fixed cost is said to be *lumpy*. Outsourcing (using subcontractors) is a way to develop an activity without bearing its associated fixed costs but at an obviously higher marginal cost.[5@]

Marginal Cost

The marginal cost is the cost of producing an additional unit while the average cost is the “typical” cost of any of the units that have been produced so far. To define precisely these concepts we first go trough the derivation of the cost function.

The neoclassical approach to costs uses the production function and the existence of perfectly competitive markets for inputs, mainly capital and labour. Given the prices r and w of the inputs K and L, the cost of producing a quantity q is

$$C(q) \equiv \min_{K,L} F + rK + wL \quad \text{s.t.} \quad \Phi(K,L) \geq q$$

Notice first that because Φ is increasing with both inputs, the constraint must be an equality at the solution (otherwise it would be possible to reduce input expenditure a little and still meet the production requirement). Our second observation is that the optimal capital/labor mix equalizes the *subjective* relative price $\frac{\Phi_K}{\Phi_L}$ of K in units of L to the *objective* relative price $\frac{r}{w}$ observed on the market. If it was not so, one would be able to substitute a little of the relatively expensive input by the cheaper one while maintaining the production constant; total cost would then be reduced.

If we apply these considerations to the Cobb and Douglas (1928) technology $\Phi(K,L) = K^\alpha L^\beta$ then $\frac{\Phi_K}{\Phi_L} = \frac{\alpha L}{\beta K}$ and we deduce $L = K\frac{r\beta}{w\alpha}$ which can be substituted into the production constraint $\Phi(K,L) = q$ to yield the conditional capital demand $K^* = q^{\frac{1}{\alpha+\beta}}(\frac{w\alpha}{r\beta})^{\frac{\beta}{\alpha+\beta}}$. Reporting this value into the FOC, we obtain the conditional labour demand L^* and finally

$$C(q) = rK^* + wL^* = \left(1 + \frac{\alpha}{\beta}\right) rK^* = \left(\Omega r^\alpha w^\beta q\right)^{\frac{1}{\alpha+\beta}} \tag{2.2}$$

where $\Omega \equiv \alpha^\beta \beta^{-\beta} + \alpha^{-\alpha}\beta^\alpha$ (the manual derivation of these formula is a lengthy but rewarding exercise).

The key concept of *marginal cost* is defined as the derivative of the cost function:

$$C_m(q) \equiv \frac{\partial C(q)}{\partial q}$$

and is related to the optimal factor mix condition by

$$\frac{r}{\Phi_K} = \frac{w}{\Phi_L} = C_m \tag{2.3}$$

We have thus obtained an equi-marginal principle (cf. §2.1.1) stating that whenever two factors like machines, building, land, skilled or unskilled workers have a degree of substitutability, their market compensated marginal productivities have to be equalized (to $1/C_m$). Reworking (2.3) as $\frac{w}{r} = \frac{\Phi_L}{\Phi_K}$, we conclude that every firm, whatever the good she produces, will equate her MRTS of any two factors to the ratio of their relative prices. Observe lastly that total cost of producing some quantity q is the sum of the fixed cost F

and the (marginal) cost of all units, from first to last; in formulas:

$$C(q) = F + \sum_{x=1}^{q} C_m(x) = F + \int_0^q C_m(x)\,\mathrm{d}x \tag{2.4}$$

Short vs. Long Term Cost

In the short-term, the current level of capital K_o cannot be adjusted so that only labour L remains flexible to meet the production objective $q = F(K, L)$. The short-term cost function is therefore greater than the long-term one since a degree of freedom in the search for cost minimizing inputs is lost. However, when the long-term demand for capital K^* is exactly equal to the short-term availability K_o, the short-term cost is equal to the long-run one; this occurs for only one production level q_o because the demand for capital K^* is increasing in q. At this point, short-term and long-term marginal cost are also equal.[6@]

In our Cobb-Douglas example, the long-term cost function is given by (2.2); the short-term demand for labour is $L = q^{\frac{1}{\beta}} K_o^{\frac{-\alpha}{\beta}}$ (we isolate L from the production equation $q = K_o^{\alpha} L^{\beta}$) so that the short-term cost function is $C^{st}(q) = F + rK_o + wK_o^{\frac{-\alpha}{\beta}} q^{\frac{1}{\beta}}$. Figure 2.2 plots both cost and marginal cost functions for the case where $\alpha = \beta = \frac{1}{2}$ and $F = 0$; we have $C^{st}(q) = rK_o + \frac{wq^2}{K_o}$, $C_m^{st} = \frac{2wq}{K_o}$, $C^{lt}(q) = 2\sqrt{rw}q$, $C_m^{lt} = 2\sqrt{rw}$ and $q_o = 2\sqrt{r/w}K_o$.

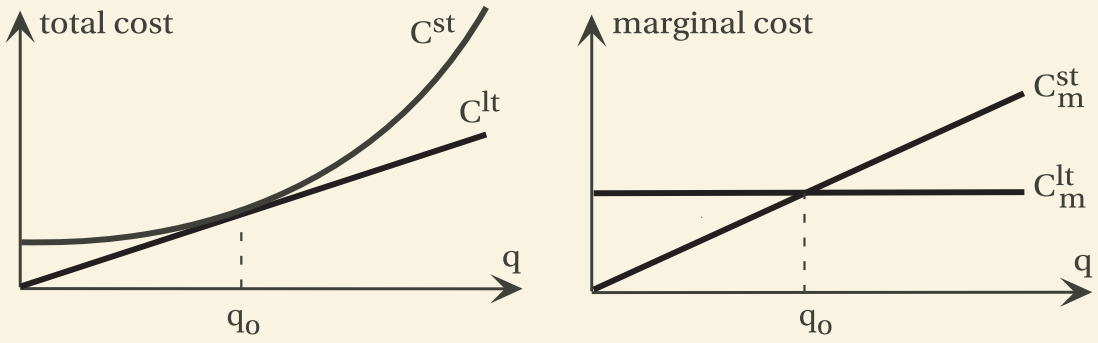


Figure 2.2: Short vs. Long Term Costs

Scale Economies

Scale economies capture the idea that duplicating production does not duplicate cost, hence production ought to take place in a single plant instead of several in order to reduce total cost. The formal condition is $C(\lambda q) < \lambda C(q)$ for any $\lambda > 1$ (the verbal definition uses $\lambda = 2$). It is simple algebraic exercise to check that economies of scale are exactly equivalent to the property $C_m < AC$ which is itself equivalent to a decreasing average cost (cf. Fig. 2.3).[7@] Diseconomies of scale correspond to the reverse situation ($C_m >$

$AC \Leftrightarrow C(\lambda q) > \lambda C(q)$ for any $\lambda > 1$) and tell us that production ought to be divided among many small identical plants.

Examples of scale economies include most activities involving a physical transportation network. Meanwhile the network capacity is not exhausted, the marginal cost is more or less constant c; the average cost is then decreasing as the fixed cost of the network is passed upon more and more units.

For most activities though, the initial range of output displays scale economies ($C_M \searrow$), then constant returns to scale (C_m cte) over a range of normal business and finally diseconomies of scale ($C_M \nearrow$) when output overshoots the plant's capacity. This gives rise to Edgeworth (1913)'s celebrated U-shaped average cost curve which crosses the marginal cost curve at its minimum, the *minimum efficient scale* (cf. §2.1.4).

2.1.4 Competitive Supply

In this section, we first recall the definition of a competitive market before looking at the optimal production of a competitive firm and its competitive supply.

Competitive Market

Edgeworth (1881) defines a *perfectly competitive* market by four conditions:

- $H1$: Homogeneous goods are for sale
- $H2$: Large number of buyers and sellers
- $H3$: Perfect price information for buyer and sellers
- $H4$:No barrier to entry for potential buyers or sellers (cf. §10.2)

An equilibrium of this market is a situation where everyone performs the trade she wishes given the prices she observes and the decisions taken by other agents. Conditions $H1$ and $H3$ imply a unique price in equilibrium because every buyer can see the lowest price between several identical goods and therefore buys the cheapest. Condition $H2$ implies that no one has *market power* i.e., can credibly threaten anyone to buy at a lower price or to sell at a larger price, hence the price is taken by everybody as given. With this *price-taking* behavior the equilibrium price equates demand and supply. As long as profits are positive on this market, $H4$ enables entry of economic agents seeking a better remuneration than elsewhere, thus in the long run sellers profits are nil.[8@]

Optimal Production

A firm participating in a competitive market completely lacks market power and rationally adopts the *price-taking* behavior. The price p becomes a parameter imposed by the

market while the supplied quantity q is the only variable to be chosen by the firm. Using the cost decomposition (2.4), the profit reads

$$\pi(p,q) = qp - C(q) = -F + \sum_{x=1}^{q} p - C_m(x) \tag{2.5}$$

and we define the summation term as the *producer surplus* (equivalently, it is profits nets of fixed cost i.e., $\pi(p,q)+F$). The difference $p - C_m(x)$ is the profit derived from unit $\#x$, it is called the *marginal profit*. The optimal production q, maximizing profit π (or producer surplus), solves

$$p = C_m(q) \tag{2.6}$$

and is intuitively derived from (2.5) by noticing that profit increases with additional production meanwhile the marginal profit $p - C_m(x)$ remains positive. As can be grasped on Figure 2.3, the competitive supply curve is just the marginal cost curve but since we are looking for a quantity as a function of price, the exact formula is the inverse function i.e., $s(p) = C_m^{-1}(p)$.

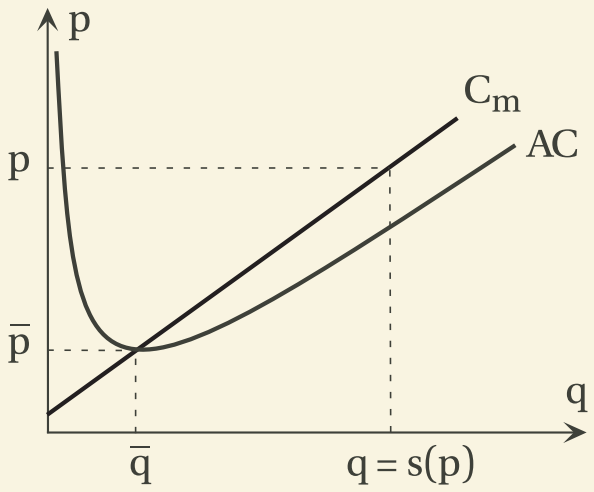


Figure 2.3: Individual Supply

The Cobb-Douglas technology $q = K^\alpha L^\beta$ provides an easy application in the present case of decreasing returns to scale ($\alpha+\beta<1$). We previously derived the marginal cost function $C_m(q) = \frac{1}{\lambda} q^{\left(\frac{1}{\alpha+\beta}-1\right)}$ for some $\lambda > 0$ (cf. eq. 2.2); the competitive supply is thus its inverse $s(p) = \left(\lambda p\right)^{\frac{\alpha+\beta}{1-\alpha-\beta}}$.

At the optimum $q = s(p)$, firm profit becomes a sole function of the market price $\Pi(p) \equiv \pi(p, s(p))$.[9@] The producer surplus is then equal to the triangle area between the price and the marginal cost curve but beware that this is true only when the quantity is the competitive supply.

Optimal Behavior

The optimal quantity $s(p)$ has been derived as the solution of the "$p = C_m$" equation which implicitly assumed that the firm was going to be active i.e., produce at least one unit. More formally, we have maximized the producer surplus which differs from profit by the fixed cost. Whenever the latter is very large, the truly optimal strategy is to shut down temporarily because the maximum producer surplus is lesser than the fixed cost which can be avoided by not starting production (recall the convention we adopted to remove any sunk cost from the fixed cost). This situation is characterized by $\Pi(p) < 0$.

The *average cost* $AC(q) \equiv \frac{C(q)}{q}$ is useful to account for fixed cost and check whether producing something is the optimal policy. Indeed, profit can be factorized by production into $\pi(p,q) = q\big(p - AC(q)\big)$ so that positive profits are obviously equivalent to average cost below the market price. It is a simple exercise to check that the minimizer $\bar{q}$ of AC satisfies $C_m(\bar{q}) = AC(\bar{q}) \equiv \bar{p}$;[10@] this quantity is called the *minimum efficient scale* while the corresponding marginal or average cost $\bar{p}$ is the long term limit of the equilibrium price in a perfectly competitive market with free entry of firms using the technology Φ.[11@] These concepts are illustrated on Figure 2.3 for some positive fixed cost.[12@]

Hotelling's lemma

Differentiating Π and using the envelope theorem, we obtain *Hotelling's lemma*: $s(p) = \Pi'(p)$ stating that the firm's behavior reveals her preferences; indeed, if we are able to follow the evolution of profits with the evolution of the market price, then we can recover the competitive supply function which is to say, the marginal cost function that tells us almost everything about the firm's technology.

It is sometimes convenient to skip the construction of the cost function and derive the demand for inputs and the supply of output directly from the profit function $\pi = p\Phi(K,L) - rK - wL - F$. The prices p, r and w are exogenous parameters while the inputs K and L are the endogenous choices to be made; the FOCs $\Phi_L = w/p$ and $\Phi_K = r/p$ yield a unique pair (K^*, L^*) function of the relative prices w/p and r/p. The competitive supply is then simply $\Phi(K^*, L^*)$.

Applied to the Cobb-Douglas technology $q = K^\alpha L^\beta$, we first use the marginal principle $\frac{w}{r} = RMST_{K/L} = \frac{\beta K}{\alpha L}$ to obtain $\frac{w}{p} = \Phi_L(K,L) = \beta K^\alpha L^{\beta-1}$ and $\beta\left(\frac{\alpha w}{\beta r}\right)^\alpha L^\alpha L^{\beta-1} = \frac{w}{p} \Leftrightarrow \alpha^\alpha \beta^{1-\alpha} w^{\alpha-1} r^{-\alpha} p = L^{1-\alpha-\beta}$

$$L^* = \left(\alpha^\alpha \beta^{1-\alpha} w^{\alpha-1} r^{-\alpha} p\right)^{\frac{1}{1-\alpha-\beta}}$$

$$K^* = \frac{\alpha w}{\beta r} L^* = \left(\alpha^{1-\beta} \beta^\beta w^{-\beta} r^{\beta-1} p\right)^{\frac{1}{1-\alpha-\beta}}$$

$$S(p,w,r) = \Phi(K^*, L^*) = \left(\alpha^\alpha \beta^\beta w^{-\beta} r^{-\alpha} p^{\alpha+\beta}\right)^{\frac{1}{1-\alpha-\beta}}$$

2.1.5 Miscellanies

Factor Demand

Firms sell output in competitive markets but also buy inputs, the factors of production. At the optimal level of production where price equates marginal cost, the marginal productivity of each factor is equal to the ratio of this factor's price to the output price; for instance $F_L = \frac{w}{p}$ and $F_K = \frac{r}{p}$. Now, since we generally assume marginal productivity to be decreasing, it must be the case that factor demand decreases with the factor price; this is indeed the case for the Cobb-Douglas technology seen above. To conclude, the firm, as a buyer of factors, behaves like a regular consumer i.e., bid a decreasing quantity-price schedule in competitive factor markets like those for labour, capital or raw materials.

Multi-plant Firm

The previous issue of grouping or not production into a single plant leads us to consider the related matter of the *multi-plant* firm which is faced with the problem of allocating total production (of the same good) among its various production centers to minimize overall cost i.e., we are looking for

$$C(q) \equiv \min_{q_1,\dots,q_n} \sum_{i \le n} C^i(q_i) \quad \text{s.t.} \quad q = \sum_{i \le n} q_i \tag{2.7}$$

If at least one plant has economies of scale then all production should take place in the cheapest one. Otherwise all plants present diseconomies of scale and the solution is quite simple: given the actual level q_i in each plant $i = 1$ to n, rename plants to obtain $C_{m,1}(q_1) \le C_{m,2}(q_2) \le \dots \le C_{m,n}(q_n)$. The optimal way to produce more is to use the first plant, hence bit by bit its marginal cost will come to equate that of the second plant. If we want to further increase production we will allocate units among plants 1 and 2 to maintain $C_{m,1} = C_{m,2}$ but both marginal cost will increase to the point where they equate $C_{m,3}$. Generalizing this equi-marginal principle (cf. §2.1.1), we find that the optimal allocation is characterized by equality of al marginal costs.

$$C_m^1(q_1) = C_m^2(q_2) = \dots = C_m^n(q_n) \tag{2.8}$$

If all plants are identical with $C^i(.) = \phi(.)$ then, by symmetry, production is identical in all plants so that $C(q) = n\phi(q/n) < \phi(q)$ (since the common technology presents diseconomies of scale).

Multi-products Firm

Consider now a firm using factors such as capital and labour in several independent plants to produce a variety of commodities labeled $j = 1, .., k$. The total profit is $\pi = \sum_{j \leq k} \pi_j$ where the branch j profit is $\pi_j = \Phi^j(K_j, L_j) - wK_j - rL_j$.

As we already saw in equation (2.3), maximizing the branch profit π_j leads to $p_j = \frac{w}{\Phi^j_L} = \frac{r}{\Phi^j_K} \Rightarrow p_j \Phi^j_L = w$. Since the cost of labour, the wage w, is the same in all branches, we obtain $p_j \Phi^j_L = p_i \Phi^i_L$ for any two branches i and j.

Let us now introduce the *marginal rate of transformation* from good j to i (MRT) as $\frac{\Phi^j_L}{\Phi^i_L}$; this ratio expresses how much production of good j must be given up to produce one additional unit of good i without changing the overall quantity of labour.

We may now say that the overall profit is maximal only if the chosen quantities of factors are such that

$$\frac{p_i}{p_j} = \frac{\Phi^j_L}{\Phi^i_L}$$

for if the MRT is not equal to the price ratio, then there exists an arbitrage opportunity that amounts to transfer an employee from one branch to the other. The same marginal principle holds for other factors such as capital (cf. §2.1.1). Summarizing, we have

$$\frac{\Phi^j_L}{\Phi^i_L} = \frac{p_i}{p_j} = \frac{\Phi^j_K}{\Phi^i_K} \tag{2.9}$$

Notice that if the branches were independent firms, (2.9) would nevertheless hold as a consequence of (2.3) because firms adjust their factor purchases through the factor prices: $p_i \Phi^i_L = w = p_j \Phi^j_L$ and $p_i \Phi^i_K = r = p_j \Phi^j_K$.

Economies of Scope

Increasing the variety of goods and services offered for sale need not be extremely costly for a firm when it is possible to share components or use the same facilities and personnel to produce several products. An assembly chain in a car factory, an oven in a bakery or a train track can all be used to produce different products or services in response to evolving demand. There is thus a potentially large saving on capital costs, the downside being the need for more experienced workers and a more flexible organization. Analytically, the economies of scope occur if $C(q_1, q_2) < C(q_1, 0) + C(0, q_2)$ where q_1 and q_2 are quantities of variants of the same product (produced within the same plant or office).

Economies of scope are also present in all service industries where demand is seasonal for a trivial reason: you cannot consume peak and off-peak services at the same

time, thus the capital goods are never in rivalry to serve demand. It is more economical to have a single large hotel on the beach open all year and use only the first floors during winter rather than having the same hotel opened only during summer and a small boardinghouse aside opened only during winter. Seasonality is more thoroughly studied in §25.1.

Constant Elasticity of Substitution

With the Cobb and Douglas (1928) technology, the relation between productivity of labour $\frac{Q}{L}$ to relative wage w/p is bound to be linear when labour and product markets are competitive; indeed, profit maximization yield $w/p = \Phi_L = \alpha\frac{Q}{L}$. Now, econometric studies covering production data from many industries and countries have shown an increasing but concave relationship. If we write the equation in logarithm (normalizing p to unity) as

$$\ln\frac{Q}{L} = \ln a + b\ln w + \varepsilon \tag{2.10}$$

then the parameter b is empirically lesser than unity. If the relation is exact (i.e., $\varepsilon = 0$), then b is the elasticity of labour productivity $y \equiv \frac{Q}{L}$ to wage w (this is a simple consequence of differentiating (2.10)). Arrow et al. (1961) then impose constant returns to scale and look for all the production functions such that the elasticity of substitution is constant: they obtain the Constant Elasticity of Substitution technology (cf. proof).

2.2 Demand

We shall always assume a large number of potential buyers for the good or service under consideration. These purchasers can be households at the retail level or firms at the wholesale level; each individual demand is the result of the maximization of the relevant objective, either utility[13@] for end-users or profits for firms. The sum of individual demands form an aggregate demand that *negatively* relates market price and total sales.

In this section, we first study succinctly the theory of preferences, utility and individual demand. Then, we study the properties of the aggregate demand function and define two key concepts for the analysis of markets: *revenue* for firms and *surplus* for consumers.

2.2.1 Utility

Preferences and Utility

How many chocolate bars would you exchange for an apple? How many apples would you give up for a chocolate bar? If you go to an apple auction, how many chocolate bars would you bid for an apple? These examples involve the idea of substitution between goods. If your answer to the last question is 3, you should answer 1/3 and 3 to the previous ones in order to be consistent. The dimensionless number 3 is your rate of substitution from chocolate bars to apples while 1/3 is your rate of substitution from apples to chocolate bars. Considering smaller quantities (as if apples were divisible into tiny quarters), we obtain the central concept of consumer theory, the *marginal rate of substitution* (MRS). Unlike this example seems to suggest, the MRS need not be constant. In most cases, it changes as the quantities of both goods hold by the individual change. For instance, if you now own more apples than previously, then your MRS from apples to chocolate bars should fall from 3 to, say 2.

Modern economic theory assumes that goods and services can be clearly identified by labels j running from 1 to k and measured by quantities x_j in well specified units so that we can meaningfully speak of a basket $\mathbf{x} \equiv (x_j)_{j \leq k}$ of goods (the supermarket analogy is useful for illustrations). Next, it is assumed that consumers are empowered with the ability to compare all baskets; more specifically, for each individual there exists a binary relation $\succ$ where $\mathbf{x} \succ \mathbf{y}$ means "$\mathbf{x}$ is preferred to $\mathbf{y}$ by the individual". These preferences satisfy:

Completeness either $\mathbf{x} \succ \mathbf{y}$ or $\mathbf{y} \succ \mathbf{x}$ or $\mathbf{x} \sim \mathbf{y}$

Transitivity if $\mathbf{x} \succ \mathbf{y}$ and $\mathbf{y} \succ \mathbf{z}$ then $\mathbf{x} \succ \mathbf{z}$

Monotony if $\mathbf{y}$ adds 1 unit of one good to $\mathbf{x}$ then $\mathbf{y} \succ \mathbf{x}$

Convexity if $\mathbf{z} = \frac{\mathbf{x}+\mathbf{y}}{2}$ then $\mathbf{z} \succ \mathbf{x}$ and $\mathbf{z} \succ \mathbf{y}$

Then, one can demonstrate the existence of a utility function u representing the consumer's preference in the sense that $\mathbf{x} \succ \mathbf{y} \Leftrightarrow u(\mathbf{x}) > u(\mathbf{y})$ which eases the task of working with the preferences of consumers. By monotony, the *marginal utility* of good j is positive i.e., $u_j \equiv \frac{\partial u}{\partial x_j} > 0$.[14@]

The indifference curve through $\mathbf{x}$ is defined as the set of all baskets $\mathbf{y}$ indifferent to $\mathbf{x}$ for the consumer i.e., the solutions in $\mathbf{y}$ of the equation $u(\mathbf{y}) = u(\mathbf{x})$. If we pick $\mathbf{y} = \mathbf{x} + d\mathbf{x}$ (a small variation), then $u(\mathbf{y}) = u(\mathbf{x}) + u_i dx_i + u_j dx_j$ so that to maintain utility constant, we need $0 = u_i dx_i + u_j dx_j \Rightarrow \frac{-dx_j}{dx_i} = \frac{u_i}{u_j} > 0$. Observe now that if we set $dx_i = +1$, the ratio of marginal utilities is exactly how much good j the consumer is ready to sacrifice to

get one more unit of the good i, hence the marginal rate of substitution from j to i is $MRS_{j/i} \equiv \frac{u_i}{u_j}$.

It is obvious that the utility concept is a non tangible measure of a consumer felicity since any increasing transformation of u gives rise to another acceptable utility function (e.g., take $v(\mathbf{x}) = u(\mathbf{x})^2$); it thus make no sense at all to say "utility increased twofold" as opposed to the same statement relative to income. Because of this multiplicity, one could worry about the coherence of our definition of the MRS. Hopefully the latter is independent of the choice of the utility function because whether u or v represent the preferences of the consumer, we have $\frac{u_i}{u_j} = \frac{v_i}{v_j}$.[15@]

Market Choice

Introducing markets for all goods and services, prices $\mathbf{p} \equiv (p_j)_{j \leq k}$ and income w for the consumer, we are able to tackle the selection of the optimal basket which will procure the greatest satisfaction. Letting $\mathbf{p}.\mathbf{x} = \sum_{j \leq k} p_j x_j$ denote the cost of basket $\mathbf{x}$ at prices $\mathbf{p}$, the maximization of utility under the budget restriction amounts to solve

$$V(w) \equiv \begin{cases} \max_{\mathbf{x}} & u(\mathbf{x}) \\ \text{s.t.} & \mathbf{p}.\mathbf{x} \leq w \end{cases} \tag{2.11}$$

To solve (2.11), we first observe that the constraint trust be binding at the optimum for otherwise the money leftover could be used to buy more of any good and further increase utility (by the monotony property of preferences). Next, we apply the "manna" trick and derive an equi-marginal principle (cf. §2.1.1). We divide w into many small gifts of one monetary unit. Each gift allows to buy $\frac{1}{p_j}$ units of good j and get $\frac{u_j}{p_j}$ additional utility. In order to successfully apply the manna trick, we must assume decreasing marginal utility i.e., $u_{jj} < 0$ because marginal utility u_j must decrease when money is used to buy good j in order to reduce the wedge with other employments. At the optimum, when all w has been spent, we have[16@]

$$\frac{u_1}{p_1} = \frac{u_2}{p_2} = ... = \frac{u_k}{p_k} \tag{2.12}$$

analytically identical to (2.3). Furthermore, this constant is the marginal utility of income $V'(w)$ because this is exactly how much final utility increases when income rises by one unit.[17@] From (2.12), we see that the consumer equates his RMS to the price ratio, hence the equi-marginal principle holds economy wide across all consumers.

At the solution of (2.11), each quantity of good depends on all prices and the income but not on the particular utility function used since only the RMS matter and it is inde-

pendent of the particular utility function representing the preference of the consumer. The *money-metric* utility or minimum income function, $\tilde{u}(\mathbf{x}) \equiv \min_{\mathbf{z} \succeq \mathbf{x}} \mathbf{p}.\mathbf{z}$, avoids introducing arbitrary utility functions, thus enables comparisons among consumers at the cost of being based on the current set of prices.

For the practical resolution of (2.11) (e.g., exercise, numerical estimates), Lagrange (1797) adds a new variable λ, the multiplier and builds the (now called) Lagrangian $\mathscr{L}(\mathbf{x}, \lambda) = u(\mathbf{x}) + \lambda(w - \mathbf{p}.\mathbf{x})$. It appears that the FOCs for a saddle point of $\mathscr{L}$ are $u_j = \lambda p_j$ for $j \leq k$ and $\mathbf{p}.\mathbf{x} = w$ i.e., exactly those characterizing the original solution.[18@] In economic applications of the Lagrange method, the multiplier is called the shadow price (of the constraint). An interpretation sees an auctioneer coming out of the shadow and offer to pay an interest λ for the slack money $z = w - \mathbf{p}.\mathbf{x}$; the consumer then freely maximizes $\mathscr{L}$ over $\mathbf{x}$. Since this possibility is just a mind experiment, the two approaches coincide only if the consumer neither wish to lend nor borrow at his optimum. We thus obtain an additional condition on the shadow price, $z^* = 0$. Under this interpretation, the shadow price at the optimum appears to be the marginal utility of money.

Substitution and Complementarity

We present here an example which also serves as the foundation of Singh and Vives (1984)'s oligopoly model studied in §5.2.3.

Varieties 1 and 2 of a good are sold alongside an aggregate of all other goods called money m. The representative consumer utility is $U(q_1, q_2, m) = m + q_1 + q_2 - \frac{1}{2}\mu_1 q_1^2 - \frac{1}{2}\mu_2 q_2^2 - \gamma q_1 q_2$. Given his income M, the consumer maximizes $U(q_1, q_2, m)$ under his budget constraint $M \leq R - p_1 q_1 - p_2 q_2$. His demand (q_1, q_2) satisfies the FOCs over quantities q_1 and q_2

$$\begin{cases} p_1 = 1 - \mu_1 q_1 - \gamma q_2 \\ p_2 = 1 - \mu_2 q_2 - \gamma q_1 \end{cases} \quad \Rightarrow \quad q_i = \frac{\mu_j - \gamma - \mu_i p_i + \gamma p_j}{\mu_1 \mu_2 - \gamma^2} \text{ for } i = 1, 2$$

We assume $\mu_1 \mu_2 > \gamma^2$ to guarantee that demand be decreasing in own-price and $|\mu_i| > |\gamma|$ to guarantee that demand be positive for zero prices. Summing over all consumers, the demand addressed to firm $i = 1, 2$ is $D_i(p_i, p_j) = a_i - b_i p_i + d_i p_j$ for some coefficients $a_i > 0, b_i > 0$, $|b_i| > |d_i|$ and d_i the sign of which is given by γ if it is the same for all consumers.

Market Demand

Demand theory, by interrelating the various goods and services is useful to understand the process of substitution that a consumer might embark on in response to a change of price or of income. That knowledge is crucial for a firm wishing to anticipate the effect

that raising her price will have on the demand she receives from consumers or for a government wanting to know the effect of increasing taxation (for whatever purpose) on consumers behavior. The general rule is that a change in price of one good, say cheaper gasoline, generates a wealth effect (the purchasing power increases) and a substitution effect (gasoline consumption increases because it is *relatively* cheaper than substitute energies or transportation modes).

We shall often isolate the relationship between the quantity x_j of a good and its price p_j, thereby obtaining our workhorse, the *individual demand curve* which is decreasing.[19@] Generally, we treat changes in other prices or income as shifts or movements of the entire demand curve. A good is complementary (resp. substitute) to another if increasing the price of one causes the demand for the other to fall (resp. increase). A good is deemed inferior if the curve goes down when income increase; among normal goods (i.e., not inferior), luxury goods are those for which total spending increases faster that the buyer's income.[20@]

Although the buying power of 1€ is independent of wealth, its "satisfying" power is not. Once we get to satisfy our basic needs, we look forward to consume more onerous goods and services. Thus, the richer we are, the less valuable is a 1€ coin. This observation which dates back to Aristotle is a major assumption of economic theory: economic agents display *decreasing marginal utility of wealth*. The relation with the previous theory of rational consumption is that holding prices constant, the indirect utility derived from the optimal basket is a function $v(w)$ of his income which satisfies $v'' < 0 < v'$.

As further reading, the website of the *Association for Consumer Research* provides entertaining studies such as:

- how we build our preferences
- why smoking has the status of a forbidden fruit for teens
- whether time is money for everyone
- why do we spend so much on weddings

2.2.2 Willingness to pay

Dupuit (1844)'s concept of *willingness to pay* (WTP) plays a central role in this book and thus warrants some space. We introduce the idea using the original example of this author and then relate it to the demand.

Imagine that continuous improvement in technology enable to reduce the price at which water is supplied to a city; inhabitants will start to consume more and more water for less and less pressing uses as shown in Table 2.1. From this observed behavior, we deduce the resident's willingness to pay for additional cubic meters of water. We can

therefore attribute a personal and monetary value to each cubic meter consumed by the resident noticing that each additional unit is devoted to satisfy an evermore futile need.

Price	Qty	Marginal use	WTP
5€/m^3	1 m^3	drinking, washing	$\geq$ 5€/m^3
3€/m^3	4 m^3	weekly car wash	3 to 5€/m^3
2€/m^3	10 m^3	daily garden watering	2 to 3€/m^3
1€/m^3	20 m^3	running water fountain	1 to 2€/m^3

Table 2.1: Water Use

These observations confirm Aristotle's claim (expressed in modern words) that *marginal utility diminishes with consumption*. As a direct consequence, the demand for water of the resident increases as the unit price diminishes. Notice however that the demand for a free good will never be infinite because there is an opportunity cost of consuming it. Indeed, there always exists alternative and competing ways to use our time, so that instead of spending the whole day watering his lawn, the resident will use some time to mow. These simple deductions hold for any good or service.

To understand the aggregation of individual demands, we recur to a new example. The left panel of Figure 2.4 represents the monthly demand schedule for movie rentals of a consumer, say Julie. If the price last month p was such that she rented 5 movies, it must have been the case that her willingness to pay for a sixth one, p_6, was lesser than p for otherwise she would have rented that sixth unit; her pattern of behavior also tells us that her willingness to pay for the fifth one p_5 was greater than p.[21@] If the price this month rises a bit above p_5, she reduces his consumption down to 4 rentals because the fifth one is not worthwhile anymore to her. By moving up and down the price, we see that her demand curve is also a decreasing *willingness to pay* curve.

Summing the individual demand schedules of all consumers for that market, we obtain a smooth decreasing relationship between quantity and price $D(p)$ called the market demand as shown on the right panel of Figure 2.4; the only change from a graphical point of view is that the scale of the horizontal axis is now in thousands of monthly rentals.[22@] The inverse of the demand function is the decreasing price function $P(q)$ which exactly measures the willingness to pay for an additional unit of the individual who bought the last unit ($\#q$). For that reason, the inverse demand is rightfully called a (market) *willingness to pay* function. All our examples in this book use a linear specification for demand with

$$D(p) = a - bp \Leftrightarrow P(q) = \frac{a-q}{b} \tag{2.13}$$

Parameter a is a proxy for the market size since it measures exactly how many units would be consumed if the good was free. Parameter b then measures the reactiveness of

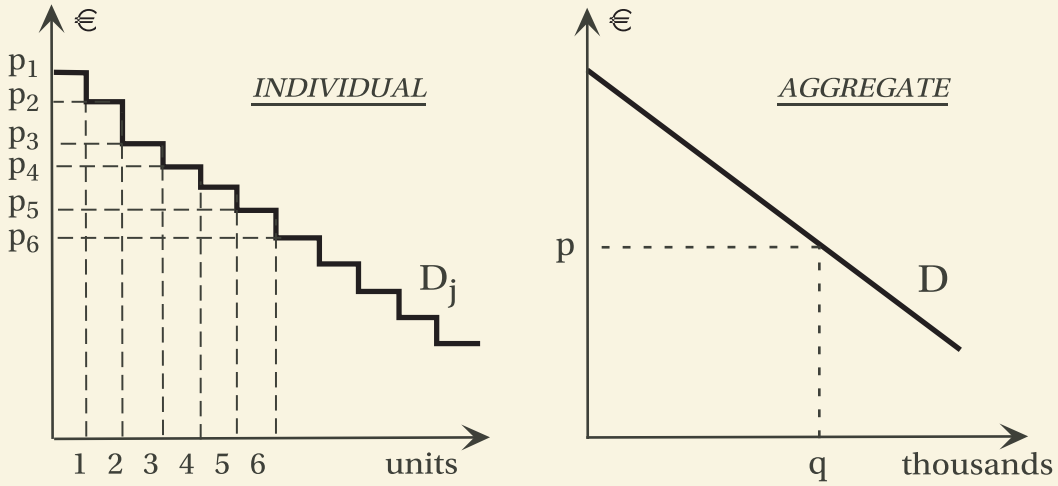


Figure 2.4: Demand Schedule for Movie Rentals

demand to changes in prices.[23@]

An important concept for practical purposes is the *elasticity* of demand with respect to price which is the percentile variation of demand in response to a percentile variation of price; formally,

$$\epsilon \equiv \left| \frac{\Delta D}{D} \Big/ \frac{\Delta p}{p} \right| = \left| D'(p) \right| \frac{p}{D(p)} = \frac{P(q)}{q \left| P'(q) \right|} \tag{2.14}$$

The consumer demand is said to be elastic if $\epsilon > 1$, isoelastic if $\epsilon = 1$ and inelastic otherwise. Be very careful that the elasticity varies along the demand curve. In our linear example, we have $\epsilon = \frac{bp}{a-bp} = \frac{a-q}{q}$, so that the demand is elastic meanwhile $q < \frac{a}{2}$ and inelastic otherwise.

2.2.3 Market Revenue

The total revenue of sellers is by definition the total expenditure of buyers and is called the market revenue. Since sales depends on price, so does revenue; the useful idea of Cournot (1838) is to relate market revenue with sales by defining $R(q) \equiv qP(q)$. The marginal revenue is the additional revenue accruing to sellers when an additional unit is sold. Formally, it is the derivative of revenue with respect to sales; it satisfies

$$R_m(q) = P(q) + qP'(q) = \left(1 - \tfrac{1}{\epsilon}\right) P(q) < P(q) \tag{2.15}$$

This concept, complementary to that of marginal cost, will prove crucial to study the performance of a market independently of the number of firms; it makes clear that the revenue of an industry is not a free variable, it depends on the consumers and their preferences for this good versus the alternatives or substitutes they can have for their

money (in case they refrain from buying that one).

2.2.4 Consumer Surplus

Unit Surplus

When looking at a product in a store, we can always ask ourselves how much we would be ready to pay to get it. If this subjective price, say 10€, is greater than the objective (market) price, say 8€, then we would be right to actually buy the item; in doing so we would derive a surplus of 2€[24@] . The force of this simple comparison is to measure in monetary terms something that is difficult to gauge: the excess of satisfaction we derive from consuming the item instead of using the 10€ to buy a worse alternative.

Dupuit (1844) offers with the concept of *consumer surplus* to decompose an individual consumption into a sum of marginal choices.[25@] Consider Julie's demand for movie rentals displayed on the left panel of Figure 2.4. Applying the previous idea, we can say that her surplus of the first rental is $p_1 - p$, that of the second rental is $p_2 - p$ and so on until the last rented unit for which the surplus is almost zero (because p_5 and p are close).

Individual Surplus

Julie's consumer surplus is then simply defined as the sum of unitary surpluses for all the units she consumes:

$$(p_1 - p) + (p_2 - p) + (p_3 - p) + (p_4 - p) + (p_5 - p) = \sum_{x=1}^{5} (P_j(x) - p)$$

where P_j is Julie's WTP function. When the good is infinitely divisible, the (individual) *consumer surplus* of consumer j is

$$cs_j(p) \equiv \int_0^{D_j(p)} (P_j(x) - p)\,\mathrm{d}x = \int_p^{+\infty} D_j(y)\,\mathrm{d}y \tag{2.16}$$

The reader can check the validity of (2.16) on Figure 2.4 using vertical bars and horizontal ones for integration. The raw surplus summing the WTP of all units consumed is called utility or TWTP (total WTP);[26@] it is, by construction, independent of the price. The integral form is

$$u_j(q) \equiv \int_0^{q} P_j(x)\,\mathrm{d}x \tag{2.17}$$

Consumer Surplus

The (aggregate) *consumer surplus* at price p is simply the sum of all individual surpluses and since market demand sums individual demands ($D = \sum_{j \le k} D_j$), we obtain[27@]

$$W_D(p) \equiv \sum_{j \le k} cs_j(p) = \int_p^{+\infty} D(y)\,\mathrm{d}y \tag{2.18}$$

with $W_D'(p) = D(p)$ i.e, margin surplus is demand.

This seemingly trivial step hides an important normative choice: interpersonal utility comparison. By adopting a *utilitarian* concept, we are implicitly assuming that 1€ gives the same satisfaction to everyone, which is clearly not the case when comparing a bourgeois and a pauper. Marshall (1890) argues however that this neglect of income and substitution effects is acceptable for goods and services that represent a small share of an individual's expenses (cf. Willig (1976) for an exact derivation and Binmore (2007) for a discussion).[28@] Notice that whenever the item under consideration is an intermediate good like "crude oil" or a service like "office cleaning", its buyers are firms and their consumer's surplus is in fact a monetary profit so that the previous critic has absolutely no bite.

One could also worry that the calculus of W_D (whose notation shall become clear in §2.3.2) depends very much on the shape of demand at very low quantities, an information almost impossible to obtain. This is irrelevant because the practical use of the concept involves comparison of two nearby situations, hence depends only on the slope of demand around the current point of consumption.

Finally, we use the duality of demand and willingness to pay, $q = D(p) \Leftrightarrow p = P(q)$ to express the consumer surplus as a function of the total quantity absorbed by the market,

$$W_D(q) \equiv \int_{P(q)}^{+\infty} D(y)\,\mathrm{d}y = \int_0^q (P(x) - P(q))\,\mathrm{d}x \tag{2.19}$$

Observing[29@] that $W_D'(q) = -qP'(q)$, we can rewrite equation (2.15) in a way that will help us understand the issue of market power in the next chapter:

$$R_m(q) = P(q) - W_D'(q) \tag{2.20}$$

Equation (2.20) also tells us how the price of an additional unit (P) is divided among sellers (R_m) and buyers (W_D'); it makes clear the impossibility for sellers to appropriate themselves the entire willingness to pay of buyers if the transaction takes places in a market (i.e., all units are sold at the same price).

In the linear demand example, we obtain $W_D(p) = \frac{(a-bp)^2}{2b}$ or $W_D(q) = \frac{q^2}{2b}$ (check this

formula using the right panel of Figure 2.4).

2.3 Equilibrium

This section, inspired by Marshall (1890),[30@] considers the market for a single good or service. We shall derive the aggregate supply and match it with the aggregate demand to deduce the equilibrium and its properties. Then we relate these findings to the theories of general equilibrium and welfare in order to infer the efficiency tool that will be central in this book: the *market welfare*.

2.3.1 Producers Surplus

We saw previously that the individual competitive supply of a firm with *decreasing returns to scale*[31@] is an increasing function of the product's price. When firms $i = 1,..,n$ are active in this market, their aggregate competitive supply is $S(p) \equiv \sum_{i \leq n} s_i(p)$ where s_i denotes the supply function of firm i. If, for example, all firms share the same technology with decreasing returns to scale characterized by the cost function $C^i(q) = \frac{cq^2}{2}$, then the common marginal cost is $C^i_m(q) = cq$ so that the common supply function is $s_i(p) = \frac{p}{c}$; lastly, we deduce the aggregate supply $S(p) = \frac{np}{c}$.

More generally, the aggregate supply curve S is increasing and its inverse, denoted C_m, can be rightly called the *marginal cost curve of the industry*. Indeed, if all firms were subsidiaries of a single (monopolistic) firm, her marginal cost would be exactly C_m.[32@] The *producers surplus* which sums the surpluses of all producers can now be precisely defined as the sum of the marginal profits made on each sold unit

$$W_S(p,q) \equiv \int_0^q \big(p - C_m(x)\big)\,\mathrm{d}x \tag{2.21}$$

When the quantity is the aggregate competitive supply $q = S(p)$, the producers surplus is given by the area between the horizontal price line and the industry supply curve from zero to the quantity supplied at that price. It should be noticed that the consumer surplus depends on either p or q but not both since the point (p,q) belongs to the demand curve; this is so because consumers have no market power. The producer surplus, on the other hand, depends on two free variables because producers may exercise market power which means that the point (p,q) need not pertain to the competitive supply curve. Compare for instance the pairs $(p_1,\hat{q})$ and $(p_2,\hat{q})$ on the left panel of Figure 2.5 below.

2.3.2 Equilibrium and Welfare

The equilibrium price p^* is reached when supply equals demand as can be seen on the right panel of Figure 2.5 where the competitive demand and supply schedules are plotted. It is readily observed that the equilibrium occurs for the quantity q^* that equates the willingness to pay $P(q)$ of consumers with the industry marginal cost $C_m(q) = S^{-1}(q)$. In the example where $D(p) = a - bp$ and $S(p) = \frac{np}{c}$, it is a simple exercise to find $p^* = \frac{ac}{bc+n}$; the total quantity produced is $q^* = \frac{na}{bc+n}$ and the individual one is $q_i^* = \frac{a}{bc+n}$ for $k \leq n$.

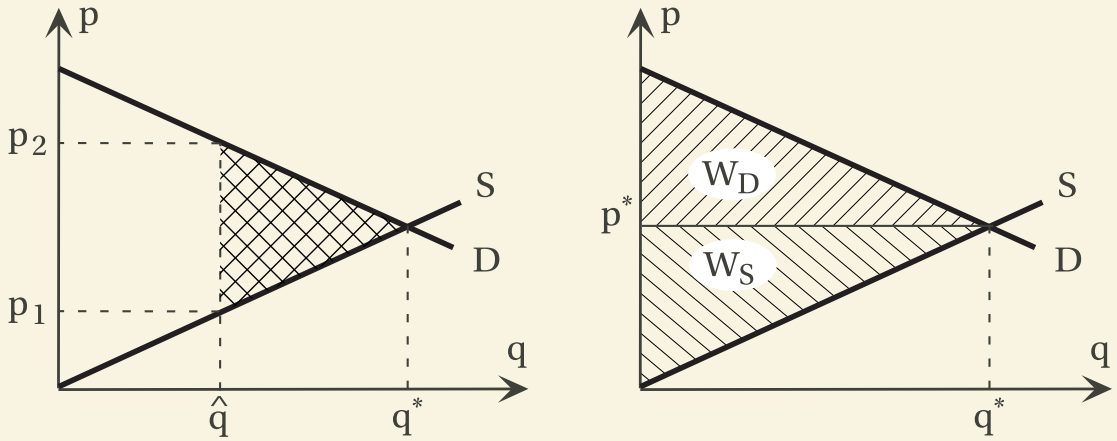


Figure 2.5: Competitive Equilibrium

As we explain in the next subsection, the key concept for the *normative* analysis in this book is that of total surplus or *welfare* which adds the surpluses of consumers (2.19) and producers (2.21):

$$\begin{aligned} W(p,q) &\equiv W_D(q) + W_S(p,q) \\ &= \int_0^q (P(x) - C_m(x))\,\mathrm{d}x \quad \text{if} \quad q = D(p) \end{aligned} \tag{2.22}$$

At the competitive equilibrium (p^*, q^*), welfare reaches its maximum denoted $W^* \equiv W(p^*, q^*)$. Indeed, whenever the price is p_1 or p_2 different from p^*, the traded quantity $\hat{q}$ is the minimum of demand and supply which generates a *welfare loss* shown by the hatched triangle on the left panel of Figure 2.5. In equilibrium, the distribution of welfare between consumers and producers is given by the equilibrium price, the large upper triangle is W_D while the large lower triangle is W_S (cf. right panel of Figure 2.5). In our linear example, the welfare at the equilibrium can be computed and simplifies into

$$W^* = W_D(q^*) + n\left(q_i^* p^* - C_i(q_i^*)\right) = \frac{1}{2b}\frac{na^2}{n+bc} \tag{2.23}$$

If the industry marginal cost is constant (flat supply curve), a price rise Δp above

the efficient level p^*, due for instance to a tax, will reduce the quantity exchanged by $\Delta q = -b\Delta p$ so that a welfare loss of $\frac{1}{2}\Delta p \Delta q = \frac{-b}{2}(\Delta p)^2$ is generated.[33@]

The competitive price p^* implements an allocative efficiency i.e., the "right" number of units is produced, sold and consumed. We can also take the longer view and wonder how many firms ought to participate in this market i.e., seek dynamicefficiency. The answer is not "as many as possible" because some scarce resources are needed to set up a firm, there is a fixed cost. The efficient number of firms is then that which maximizes the long term welfare net of entry cost $W^* - nF$ i.e., one has to balance the fixed cost of introducing an additional copy of the technology with the welfare gain derived from its use by lowering the industry supply curve. This issue is studied in §6.1 on the dynamic of entry and exit.

2.3.3 Welfare Economics

As we already noted, our normative approach is rooted in the *utilitarian* concept of consumer surplus. Our main instrument to assess allocative efficiency and judge among alternative outcomes is *market welfare* W, the contribution of the market under study to national income (GDP). This criteria for guiding social choice is not innocuous since it justifies taking away everything from someone if he is blocking the progress of society, whether he is rich or poor. This is quite the opposite of the *rawlsian* approach aiming at benefitting first the person with the lowest level of satisfaction in society (cf. Dasgupta (2007) for a discussion).

A more technical issue is the fact that W is a partial equilibrium concept. We thus need to relate it to the findings of general equilibrium theory in order to assert its aptness. Indeed, it could be the case that restoring efficiency in one market through regulation, taxes or breaking-up a monopoly creates worse distortions in related markets. The *second best* theory shows formally such a possibility; however, this fear has never materialized itself beyond contrived general equilibrium models with no bearing upon reality.[34@] In practice, when two markets are strongly interdependent (e.g., car maker and car electronics), the partial equilibrium analysis can be augmented to account for their relationship; total (market) welfare then sums those of the two markets. We may therefore synthesize our normative approach as follows:[35@]

> The received paradigm of Industrial Organization states that accurately defining the relevant market and solving its inefficiencies increase the welfare possibilities of the entire economy, given the current redistribution mechanisms at work.

2.4 Game Theory

Firms often collaborate to create value but then compete to divide it up. Game theory, using mathematical formalism, tackles these interactions known as *cooperative* and *non cooperative* (Aumann (1985) offers a panorama on the aims and method). We briefly present the equilibrium concepts that enable the analysis of simultaneous interactions; simultaneity here means that an economic agent has no information relative to the action of his challengers, only guesses. We then move on to concepts related to sequential interactions, the main novelty being that each decision maker is able to observe some or all of the actions taken by those who played before himself. Holding such an information can be advantageous for me, as a player, because I am able to make "better informed" choices. Yet, as we shall see, it can also be a limitation because everybody knows what I know myself, thus everybody can guess what my course of action will be, as if they all possessed more information. The last section deals with cooperative settings, negotiations, bargaining and problems of wealth division.

2.4.1 Simultaneous Games

Reduction

A game can involve any number of players[36@] but since interactions in IO are typically bilateral, we shall content ourselves with two players. Furthermore, we reduce the interactions to the bare minimum of a binary choice. The graphical representation displayed on the left panel of Figure 2.6 leads to speak of the *row* (R) and *column* (C) players whose strategies are respectively *north* (N) or *south* (S) and *east* (E) or *west* (W). The geographic labels have obviously no meaning and serve to underline the power of abstraction of game theory. Each player holds preferences over the four possible outcomes $\{NE, NW, SE, SW\}$ synthesized by four utility levels or *payoffs*. A 2×2 matrix game is thus characterized by 8 figures. If the gain of one party is exactly the loss of the other, the game is zero-sum. When parties share a prize (it may have negative value as in war), the game is constant-sum.

$R \setminus C$	E	W
N	NE	NW
S	SE	SW

$R \setminus C$	E	W
N	a, b	$0, 0$
S	$0, 0$	c, d

Figure 2.6: Game in Normal Form

We can now apply Stinchcombe (2007)'s powerful simplification to classify 2×2 matrix games. Whatever the equilibrium concept, it appears that the only thing that matters is

the utility differential contingent on the opponent's action. Thus, adding a constant contingent on the opponent's action, is neutral for the equilibrium analysis. We may then zero the out-of-diagonal payoffs for both players. The resulting payoff matrix, synthesized by just 4 figures $\{a, b, c, d\}$, is shown on the right panel of Figure 2.6: the first entry is the row player's payoff while the second entry is the column player's payoff. When the interaction of the players is worthy of analysis, each is sensitive to his opponent's action, hence none of the 4 figures is zero. This means that figures are either positive or negative, so that we have $2^4 = 16$ different games.

Strategic Classification

The analysis of strategic interaction displayed on Figure 2.7 uses arrows indicating the desire of players to change their action, conditional on the opponent's action. A ☹ occurs as soon as one player would prefer to leave the entry while a ☺ indicates a stable outcome known as a Nash equilibrium. When the two arrows of a player go in the same direction, they indicate a dominant action i.e., one action pays more than the other independently of what the opponent does (the column or row is crosshatched).[37@] This situation occurs each time the payoffs for one player have opposite signs. Checking out the possible combinations, we have[38@]

❶ Both players have a dominant strategy; we may erase the dominated row or column for each. One entry remains. It is a dominant strategy equilibrium.

❷ One player has a dominant strategy; we may erase the dominated row or column leaving one player choosing his preferred action. There is a unique stable outcome.

❸ Multiple equilibria either in the diagonal or off it. The corresponding issues are, respectively, coordination and leadership.

❹ No stable outcome as optimal choices cycle clockwise or counter-clockwise.

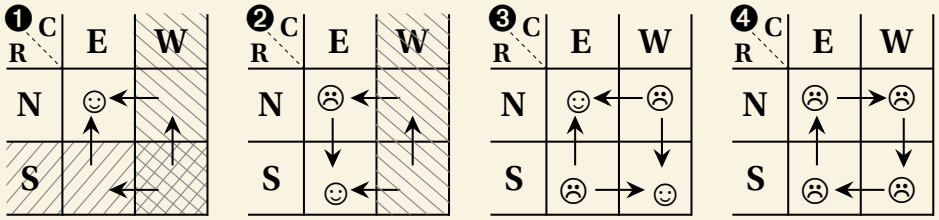


Figure 2.7: Games in Normal Form

Economic Classification

We now analyze the previous cases from an economic point of view using convenient payoffs to convey better intuitions.

① **Prisoner Dilemma** The following situation is typical of life in societies (biology, sociology, psychology and economics). Two parties would like to agree on some course of mutually beneficial action but each sees that particular action as a dominated one, thus plays the other one, triggering a bad outcome. The agreement cannot be carried because there is no way to enforce it.

$R \setminus C$	Sit	Stand
Sit	$2,2$	$0,4$
Stand	$4,0$	$1,1$

Stand-up trap

$R \setminus C$	Deny	Rat
Deny	$-2,-2$	$-8,0$
Rat	$0,-8$	$-5,-5$

Prisoner Dilemma

$R \setminus C$	No	Ads
No	$3,3$	$1,4$
Ads	$4,1$	$2,2$

Advertising

- If we all sit or stand, we watch the show equally well, the only difference being exhaustion ($2 > 1$). If one rises when others sit, he sees much better (4) at the expense of who sits behind (0). Standing is the dominant action for both.
- The light Hawk-Dove game of biology uses the same matrix: a resource of value 4 can be shared peacefully (Dove attitude = Sit) or contested (Hawkish attitude = Stand). If both contest, half of the value is destroyed. Hawkish attitude is dominant. Notice that if too much is destroyed, then we have a heavy version which is a game of leadership (cf. ④).
- Prisoner Dilemma (Original version): payoffs are negative since players evaluate jail terms. Denial by both suspects entails a smaller jail term (2 years) than mutual confession (5 years). Yet, rating on the other (lone confession) is rewarded by freedom for the informant and by charging the betrayed accomplice with the 10 years of jail. Betrayal (rating) is the dominant action.
- Prisoner Dilemma (Business version): pure advertising. The first to advertise gains market share at the expense of his challenger. Then, the loser starts advertising too in order to regain his clients. Once all advertise, market shares return at their initial level but advertising expenses have been incurred, so all lose.
- Trust in commerce: when parties are not face to face, it is a dominant strategy for each to renege on its promise which is to send the good for the seller and to send the payment money for the buyer. If so there is no trade which is why trust building is so important.

② **Second Mover Advantage** This is the least known case, probably because it is the least frequent one. In some situations involving weak W and strong S players, the highest earner can be the weakest as shown by Markides and Geroski (2004). The reference situation has payoffs normalized to zero. Investment into a copiable innovation costs 2 and generates a market value of 8. When both firms invest, S grabs 6 out of 8 so that payoffs are $6-2=4$ and $2-2=0$ for W. When a single firm invests, she is left alone polishing the new product and supporting the cost of its final development while her competitor is then able to copy for free the final design. This means that the distribution of the market value 8 is tilted against the investor by one unit. If only the weaker firm invests, she gets 1 and earns a net $1-2=-1$; S earns 7. If only S invests, she gets 5 and earn $5-2=3$ while S earns 3. Waiting thus appears to be a dominant strategy for the weaker player. In practical terms, W can successfully convince S that he will never invest, thus forcing him to go alone. The key to the result is the -1 entry i.e., that investing only to be overtaken by the strong follower is not worthwhile for the weak player.

$W \setminus S$	Act	Wait
Act	0,4	−1,7
Wait	3,3	0,0

Ambush

③ **Coordination** In games of coordination, there are several equilibria.[39@] differing wrt. how players rank these outcomes.

$R \setminus C$	Stag	Hare
Stag	2,2	0,1
Hare	1,0	1,1

Stag Hunt

$R \setminus C$	Movie	Football
Movie	1,2	0,0
Football	0,0	2,1

Battle of the Sexes

- Rousseau (1755)'s Stag Hunt illustratescoordination issues in society: a lonely hunter can catch a hare but it takes two hunters to catch a valuable stag (deer). If one is going to hunt hare, it is pointless for the other one to try to catch a stag. There are two equilibria and both players rank them likewise; yet if there is a history of hare hunt, stag hunt is unlikely to emerge.
- Bank run (macroeconomic version): I keep my deposit in the bank if everybody does so. Yet, I rush to take it out once I become convinced that everyone will do the same. There are two equilibria, chaos and stability, that savers rank alike.
- Insurance or Mutualization: a risk, unlikely to hit all members of a community at the same time, can be eliminated if all contribute their actuarially fair share of

the expected cost. Yet, there is another equilibrium where none contributes i.e., a coordination failure.

- Battle of the Sexes: Spouses prefer to spend their evening together. Joint activities are the only (pure strategy) equilibria but since each partner has an ideal activity, they prefer different equilibria.
- Joint venture: in this business equivalent, firm join forces to develop a product around a core but each would like to impose his original technology as the core.

④ **Leadership** Although strategically identical to games of coordinations, games of leadership display asymmetric behavior in equilibrium. There are two stable outcomes, inversely ranked by the participants so that again, history is likely to decide who gets to pick the equilibrium.

$R \setminus C$	Defy	Bend
Defy	0,0	3,1
Bend	1,3	2,2

Chicken

$W \setminus G$	Act	Wait
Act	2,2	1,3
Wait	3,1	0,0

Snowdrift Removal

$R \setminus C$	Ham	Cheese
Ham	0,0	1,2
Cheese	2,1	0,0

Local Monopoly

- The Chicken game illustrates political science and economics; outcomes are ranked as follow: the best is to be strong against a weak opponent (3), then both weak (2) because it is a sage attitude where no one looses face, followed by being weak in front of a strong one (1) as one is ridiculed (called a "chicken"). Finally, mutual fighting has dire consequences (0). As noted above, the Hawk-Dove game with destruction outstripping initial value is a game of chicken.
- The Snowdrift game illustrates public good finance: both parties should remove the snow but it is always tempting to free ride on a volunteer. However, the action is still beneficial for its author (as opposed to ①), so that two asymmetric equilibria exist.
- In the Local Monopoly game, firms can either produce ham or cheese flavored crackers. If they produce the same kind, they fiercely compete "à la Bertrand" (cf. §5.2.1) and earn zero profit whereas if they differentiate their crackers, each enjoys market power and profits (cf. §5.2.2). Still, consumers favor cheese-flavored crackers so that the producer of this particular type earns more (2 vs. 1).

⑤ **Cycling** These games of strategic interaction have no pure stable outcome since there is always a player willing to change his action given what the other is about to do.

$R \setminus C$	Head	Tail
Head	$1,-1$	$-1,1$
Tail	$-1,1$	$1,-1$

Matching Pennies

$R \setminus C$	Audit	Pass
Lie	$0,1$	$2,0$
Truth	$1,0$	$1,0$

Auditing

$R \setminus C$	Music	Book
Music	$0,1$	$1,0$
Book	$1,0$	$0,1$

Differentiation/Imitation

- In the Matching Pennies, players show the Head or Tail of a coin; if the faces agrees, R pays 1€ to C, otherwise C pays R. Rock, Paper, Scissors is an extension to more strategies.
- Auditing between the IRS and a tax payer involves a cycle: if I declare truthfully my income, the IRS has no point in auditing me, thus I'd better start evading taxes but then the IRS should audit me systematically at which point i'd better abide to avoid punitive damages.
- Alternative settings: in the workplace between a controller (potential auditor) and an employee (potential shirker), safety regulation between firms and the inspector. If the cost of behaving is too high wrt. the expected penalty from being caught then shirking becomes a dominant strategy and we are back to ③ (the outcome depends on the auditor incentives in front of a known infraction).
- Imitator vs. Innovator: if firms choose the same product (C imitates R), they enter a Bertrand competition that C wins thanks, say, to its cost advantage in production. If on the other hand, firms choose different products then R's innovative approach to branding enables him to win the entire market. The idea here is that each firm has a specific know-how and would like to turn it into a market advantage but this outcome depends on the behavior of the competitor.

The Nash equilibrium (cf. below) of cycling games sees each player mixing between his two actions in the proportion that makes the other player indifferent between his own two actions. Using the reduced form with payoff parameters a, b, c, d, the conditions are

$$\left.\begin{array}{rcl} \alpha_S a + (1-\alpha_S)0 & = & \alpha_S 0 + (1-\alpha_S)c \\ \alpha_R b + (1-\alpha_R)0 & = & \alpha_R 0 + (1-\alpha_R)d \end{array}\right\} \Rightarrow \left\{\begin{array}{l} \alpha_S = \frac{c}{a+c} \\ \alpha_R = \frac{d}{b+d} \end{array}\right.$$

which are well defined here given that a and c have the same sign (likewise for b and d).

General Theory

A game in "strategic form" Γ consists of a set of players $i = 1, ..., n$, a set of strategies S_i for each player i, a function g that maps a full profile of strategies $s = (s_i)_{i \leq n}$ into an outcome $d = g(s)$ and finally a utility function u_i for each player i that is defined over the set of

possible outcomes D. Notice that we may consider the payoff function $\pi_i(s) = u_i\big(g(s)\big)$ in order to analyze the game. We denote $s_{-i} = (s_j)_{j\neq i}$ the vector s without the i^{th} coordinate s_i, hence for all $i \leq n$, we always have $s = (s_i, s_{-i})$.

In a game Γ, each player is rational, thus looks forward to choose the strategy that gives him the largest payoff. An optimal strategy for player i also known as a *best response* depends on the vector s_{-i} of strategies that other players are using; a best response is a particular strategy $\hat{s}_i \in S_i$ that maximizes his payoff $\pi_i(s_i, s_{-i})$ when varying s_i overall possibilities from S_i. A best response need not be unique. The short notation[40@] is $BR_i(s_{-i}) \equiv \text{argmax}_{s_i}\pi_i(s_i, s_{-i})$. Thus, to pick his strategy, a player must account for what others are doing themselves i.e., he has to form a belief about the others' actions. We assume that each player knows his opponents will behave rationally, like him.

We can now define a *Nash Equilibrium* as a profile of strategies s^* such that no player i can do better by choosing an action s_i different from s_i^*, given that every other player j adheres to s_j^*. In formulas:

$$\forall i \leq n, \forall s_i \in S_i, \quad \pi_i(s_i, s_{-i}^*) \leq \pi_i(s^*)$$

Alternatively, a Nash equilibrium is a fixed point of the best response correspondences i.e., for all i, s_i^* belongs to $BR_i(s_{-i}^*)$. The Nash equilibrium epitomizes a stable "social norm": if everyone else adheres to it, nobody wishes to deviate from it *unilaterally*.

A strategy for each player is a complete *plan* that specifies a move at each stage where he is active, as a function of his information at that stage. A rational player can choose his strategy before the game begins, with no loss of generality. Indeed, a strategy lets him specify a different move for every situation in which he might find himself during the game. In other words, he need not wait to observe the consequences of other players' decisions to decide on his own in response, he can anticipate every single possibility and write down in his plan, what he would do in each conceivable case. Hence, each player designs his strategy without being informed of the other players' strategy choices. The concept of Nash Equilibrium therefore presumes that all players remit an envelope containing their strategy (the master plan) to an referee who will open them and apply the instructions they contain to perform a play of the game.

Regarding existence, there are often more than one Nash equilibrium which is problematic for economic interpretation as we have no clue on which one to pick. In some case, as the matching pennies, there is no Nash equilibrium and we must introduce mixed strategies (i.e., enlarge the game) to find an equilibrium. The idea is that players use a statistical distribution over the set of strategies S. It has been shown by Harsanyi (1973) that this is akin to pick a strategy with certainty in a game where one lacks

information regarding the payoff of others.

2.4.2 Sequential Games

There are many instances of economic interaction where agents act in a sequential manner. For instance, one firm enters a market and later on, another can do likewise, or pass. The method we shall use to study these interactions and find the Nash equilibrium is called the *backward induction*. We explain the idea using the example of the two farmers.

Each farmer requires his neighbor's help to harvest his field when the time comes, or else half will rot in the field. Since the south-looking field ripens earlier, cooperation is technically feasible: N should first help S and later on S should help N. The problem lies in the incentive to do that. Plato already recognized that S has nothing to gain from helping N when it's his turn to help; he would be better of sparing himself the hard labor of a second harvest. Hobbes (1651) went further to suggest that S should not help N because it is irrational to honor an agreement made with someone else who has already fulfilled his part of the agreement. Of course, if this analysis is correct, then would not N anticipate this "double crossing" and act accordingly? This is precisely Hume (1740)'s analysis: since N cannot expect S to return his aid later on, he will not help S in the first place when S's corn ripens first, and of course S will not help N when N's corn ripens later.

This story can be modeled as game which is represented in *extensive form* on the left panel of Figure 2.8. The payoffs entries are (π_N, π_S). The value of a fully harvested field is 80; it obtains only if the other farmer helps while if the other cheats, half of the crop is lost. To these profits, one must possibly retract 3 which is the value of time and sweat lost in harvesting someone else's crop. The selfish solution we previously identified is easily understood with the help of the tree representation. We start by looking at the last decision to be taken before the end of the game when all decisions are translated into consequences and final payoffs. After having benefited from N's help (left side), S prefers to cheat to earn 80 rather than 77. Even if he has not benefited form N's help (right side), he still prefers to cheat to earn 40 instead of 37. We now move backward at the moment where N must decide whether to help S or not. We are assured that N knows what S will do because everybody is rational and everybody knows that everybody is rational. Hence, N knows for sure that his decision will either lead to the second or fourth outcome, hence cheating is also the optimal strategy for him.

The backward induction enables to identify strategies for both players which form a Nash equilibrium of the game which is called *subgame perfect* (after Selten (1975))

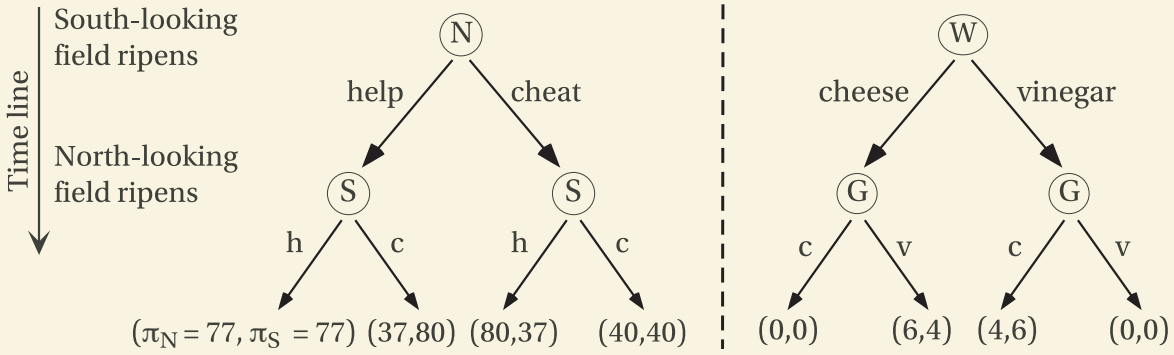


Figure 2.8: Games in Sequential Form

because it rules out non credible strategies at any node of the tree.

To clarify this latter idea, consider the competition for cocktail crackers among two firms, Wallace (W) and Gromit (G) displayed on the right panel of Figure 2.8. W can commit to produce either vinegar (v) or cheese (c) flavored crackers. Then, G acting in full knowledge of W's decision, picks a flavor to produce. If they produce the same flavor they fiercely compete and earn a nil benefit while if they choose different flavors, each one enjoys market power and derives handsome profits. Still it is well known that cheese-flavored crackers are the favorite of consumers so that the producer of this particular type earns more (6 vs. 4). The backward induction method remains simple to apply: being rational, Gromit will choose v after c and c after v. Knowing that Gromit is rational, Wallace can anticipate this course of action and therefore reduce the game tree to the second and fourth outcomes; it is then evident that c yields a higher final profit.

There exists another Nash equilibrium where Gromit commits to "always c" while Wallace plays v. Yet this equilibrium is imperfect because Gromit is making a *non-credible threat* when pretending that after observing the choice c by Wallace he would still stick to c which we all know would be foolish. The backward induction method therefore selects the most natural equilibrium among all possible Nash equilibria.

2.4.3 Negotiation: Sharing & Bargaining

One of Smith (1776)'s claim, put in modern language, was that "no big bills are ever left on the sidewalk" i.e., opportunities for value creation tend to be identified and realized. Yet, this outcome, conducive of allocative and even dynamic efficiency, requires a negotiation of the stakes among the parties involved.

This section deals with the sharing of a valuable item and follows the original insights of Edgeworth (1881)'s theory of exchange. We begin with a string of examples involving two parties which corresponds to the situation most frequently encountered in this book.

We study the division issue first from a normative point of view i.e., *how should it be done*? Later on, bargaining will approach the problem from a descriptive point of view i.e., *how shall it be done*? In the last section, we show how the presence of asymmetric information reduces the efficiency of negotiation i.e., parties sometime fail to reach an agreement although it would be socially desirable.

Bilateral Monopoly

In a perfectly competitive market, welfare is maximized and distributed among participants through the price mechanism. Industrial organization on the other hand studies mostly imperfectly competitive situations where actors are free to negotiate over many issues such as price, quantity or quality; this happens whenever they identify an opportunity for cooperation i.e., an opportunity to take coordinated actions that increase their joint profit. Examples are:

- price to be paid by a city to a builder for delivering a bridge.
- negotiation between a union and the owner of a factory over wage schemes, schedules or working conditions.
- negotiation between the EU and a candidate country for accession.
- Barter of wheat against milk between a grower and a farmer.
- bargaining between this book's author and his publisher to set the retail price and the royalties.
- dispute between a polluter and his victim regarding abatement.
- negotiation between the maker of a component and the prospective buyer of this specific item.

One speaks of a *bilateral monopoly* because each party has the monopoly or exclusive ownership of one good or service that is of interest to the other party. Obviously, the value created by engaging into such a form of cooperation has to be *shared* and this is the main point of contention.

To introduce the bilateral monopoly in a more concrete manner, consider farmers whose plots of land are face-to-face in a valley where runs a river, one plot looking south S, the other looking north N. If there is no bridge to cross the river, both farmers are forced to live in autarky. When the harvest time comes, each lonely farmer hasten to reap as much as he can but he nevertheless loses one half of the crop. Assuming that the southward oriented land is more productive, farmer S's income e.g., $\pi_S = 60$, will be greater than farmer N's e.g., $\pi_N = 40$. Industry surplus is $\pi = \pi_S + \pi_N = 100$.

If the gentle lord of the valley build a bridge over the river, new opportunities arise. Farmers can take advantage of the fact that the southward oriented crop ripens first to join forces and make a full harvest in each field in turn, thereby avoiding any loss. This way, the total value of the crops is doubled to $\pi^* = 200$. It remains for the farmers to sign an agreement or follow an ancient custom to divide the additional income $\delta = \pi^* - \pi = 100$, generated by the introduction of the bridge (technical progress). The intuitive solution of equal sharing is also the most likely whether we take a cooperative or strategic approach to the issue.

Cooperative Approach

Let us look for properties that a cooperative or normative solution should satisfy with the help of Figure 2.9. Firstly, according to *individual rationality* (IR), a party will never agree to an allocation giving him less than what he gets in case of disagreement. The latter is an amount of money that can be interpreted as an *opportunity cost* of participating to an agreement, in other words, what he would be able to earn if he were to perform his most rewarding alternative activity. In the farming example, opportunity cost are $\pi_S = 60$ and $\pi_N = 40$. This particular outcome is called the *disagreement* or threat point. Graphically, the postulate says that the final outcome will be inside the triangle with origin at the disagreement point. We illustrate this with two examples, a bold dot and a black square corresponding to initial surpluses π and π'.

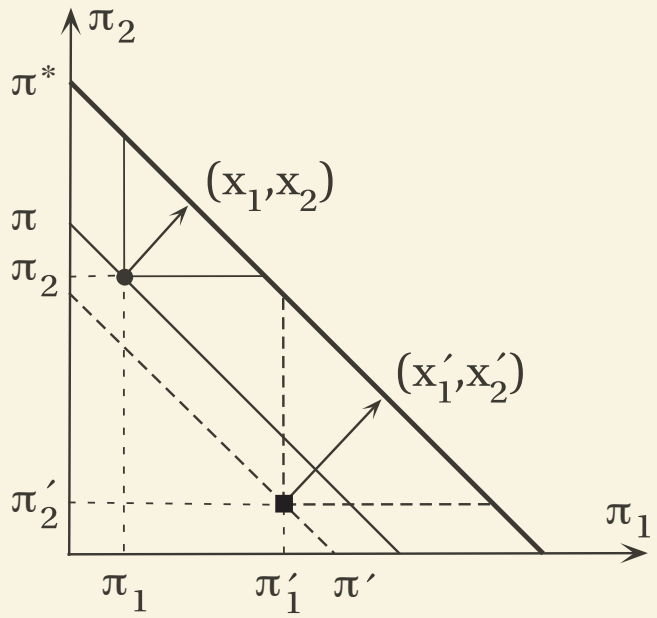


Figure 2.9: Pareto improving Bargaining

A second natural postulate is *Pareto efficiency* or joint rationality: parties are able to

identify the outcome maximizing joint profit and choose to carry it on; hence the solution is a Pareto optimum.[41@] This axiom amounts to assume that renegotiation of any agreement is possible and costless; hence if the current agreement does not maximize joint profit, the parties enter a renegotiation round to replace it by a better one. This process lasts up to the moment where there is no alternative jointly preferred to the agreement in place. On Figure 2.9, efficiency means the final outcome will be on the bold line cutting the two axes at $\pi^* = 200$.

The sharing problem is thus summarized by the opportunity costs π_S and π_N (disagreement point) and the prize for cooperation $\delta \equiv \pi^* - \pi$ (graphically, the distance between the two anti-diagonals).[42@] aka quasi-rent[43@] Any division of this prize among the two parties will satisfy the two rationality axioms; graphically the final outcome will be anywhere on the outer segment of the triangle with origin at the disagreement point. Of these many outcomes, only one treats both parties democratically or symmetrically; it is the Nash solution of equal prize sharing where farmer i receives $x_i = \pi_i + \frac{1}{2}\delta$ for $i = S, N$. This apportionment is said to preserve the differences since $x_S - \pi_S = x_N - \pi_N$ which means that the increase of farmer S's payoff due to the help of farmer N is equal to the increase of farmer N's payoff due to the help of farmer S.

Non-cooperative Approach: alternating offers

We now turn to the descriptive analysis. What can we expect to happen between two rational decision markers? One farmer, say S, may ask N to pay him 39 in oder to exchange a few days of work at harvesting time. Farmer N would be likely to accept since he would harvest his whole field for a gain of 80, then pay 39 and earn finally 41 which is more than 40, the profit in autarky. Farmer S would then earn $120 + 39 = 159$. The sum of profits is maximum at 200, thus efficiency has been achieved; notice however that the proposer, farmer S, grabs most of the prize for cooperation δ thanks to his clever offer. The weakness of this reasoning is that there is no reason for farmer N to accept this offer since he could work the numbers himself and make an entirely symmetrical counter-offer where he would grab the prize δ.

Rubinstein (1982) studies such a sequence of alternatives offers for the division of a monetary prize, normalized to unity, between S and N as follows: S asks a share q_S for himself which N can either accept or reject. In the latter case, after one period elapses, the roles are reversed and N asks q_N for himself. The alternating offers go on until one player agrees to the offer made to him. Let δ_S and δ_N be the discount factors (cf. §19.1.2). Notice that a small value indicates impatience.[44@]

It can be shown that in the unique subgame perfect equilibrium, the first player to speak, say S, offers $q_S^* = \frac{1-\delta_N}{1-\delta_S\delta_N}$ for himself which is accepted by N who gets the comple-

mentary part of the unit prize (cf. proof). One observes that greater patience, $\delta_S > \delta_N$, implies a greater payoff $q_S^* > q_N^*$ (comparing S and N as first speakers). If players are equally patient, the first speaker earns $q^* = \frac{1}{1+\delta}$ whereas the other player gets δq^*. If both parties become infinitely patient, their share tend to 50% which is the cooperative Nash solution.

Many agency situations studied in Part H display a principal relatively more impatient than the agent. This is the case for a government, a public body or a CEO on issues where swift decisions are expected by stakeholders (e.g., consumers, voters, clients) whereas the agent is a contractor who is not immediately threatened to go bankrupt by a delayed payment. The negotiation regarding the final payment associated with the delivery of the good or service turns often at the favor of the agent who holds-up the principal.

Asymmetric information

According to a popular proverb, "virtue is in the eye of the beholder". This irrational feature of human nature (cf. §1.4.2), also called "self-serving bias", may lead negotiating parties to exaggerate the importance of their contribution to the joint benefit up to the point where the sum of "apparent" opportunity costs overshoots it, thereby making impossible an agreement. This problem may be re-interpreted as one of asymmetric information because parties are unsure of their opportunity cost or of the value of cooperation.

More generally, when information relative to the bargaining is asymmetrically distributed, the negotiation may fail so that the exchange is not performed when it would be efficient to do so. This frequent situation generates *ex-post inefficiency* i.e., waiting to resolve issues as they unfold does not always result in a Pareto efficient outcome as claimed by the Coase theorem studied in the next paragraph. Nevertheless, in most of the book, we overlook this problem and assume on the contrary that parties to a bargaining are able to trike an efficient deal.

We present below a simple case of bargaining with one sided private information and show that the result is not always efficient. Dealing with asymmetric information on both sides is more involved and thus deferred to §22.3.3 in the auctions chapter.

Consider an agent A who operates an asset of unitary value on behalf of a principal B (boss).[45@] Due to the complexity of the environment, it is not possible to specify ex-ante the precise service that A must render to B by operating the asset, nor the price to be paid for that. Ex-post, once all relevant information is known, B specifies the exact service she expects from A and makes a payment offer. The agent can either accept the offer or reject it and sell instead his services to an alternative employer. However, the

agent has previously invested into the asset to improve its value within the relationship with B, thus the payment A gets in his best alternative is only $\theta \leq 1$. Equivalently, one can adopt a rent-seeking view of the negotiation and assume that when A contests B's offer, he is able to divert the asset (e.g., steal) and recover a share θ of its unitary value for himself; in the process $1-\theta$ of value is destroyed. The information asymmetry lies in the fact that A has private knowledge of θ whereas from B's point of view, this value is uniformly distributed over $[0;1]$.

Since A can guarantee himself θ, he will accept B's payment proposal p only if $p \geq \theta$. The probability of this event being p, B's profit is $p(1-p)$ which is maximum for $p = \frac{1}{2}$. This choice means that A refuses the offer half of the time. In that case, he contests the offer to realize his larger default payoff θ.The agent's expected default payoff is thus $\mathbb{E}\left[\max\{\theta, \frac{1}{2}\}\right] = \frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{3}{4} = \frac{5}{8}$. The expected surplus is $\frac{1}{2}\mathbb{E}\left[1|\theta \leq \frac{1}{2}\right] + \frac{1}{2}\mathbb{E}\left[\theta|\theta > \frac{1}{2}\right] = \frac{1}{2} + \frac{3}{8} = \frac{7}{8} < 1$ i.e., there is an inefficiency due to asymmetric information.

Under symmetric information regarding θ, the boss pays exactly θ so that negotiation never fails and there is efficiency. The agent's payoff is then $\mathbb{E}[\theta] = \frac{1}{2}$, thus the information advantage translates into a pecuniary advantage at the cost of a lessening of efficiency because the principal's profit drops from $\frac{1}{2}$ to $\frac{1}{4}$. The result remains qualitatively correct when if the (informed) agent has some bargaining power and is able to make a "take-it-or-leave-it" offer to B with probability α (as long as $\alpha < 1$).

Coase Theorem

The theoretical and political issues surrounding the Coase Theorem are developed in §8.1.3. We contend here with its relation to bargaining. The basic situation considered by Coase (1960) is when firm N, while producing generates an externality, positive or negative, affecting firm S's profits without generating a countervailing transaction. In the absence of communication, firm N will maximize profits and ignore the externality. This outcome defines opportunity costs π_S and π_N. If now a technological or institutional improvement enables parties to communicate and negotiate at no cost, they will be able to internalize the externality and restore efficiency; their joint profit will rise. [46@] In this ideal world, no State intervention is warranted.[47@]

Making firm N liable for the externality caused to firm S only changes the opportunity costs, not the final efficient outcome. Observe indeed that under this new rule, firm S can guarantee itself a greater default profit $\pi_S^{'} > \pi_S$ by requiring that firm N sets the externality at the level optimal for him. However, this particular solution ignores that this restriction cuts back firm N's profits down to $\pi_N^{'} < \pi_N$. As illustrated on Figure 2.9, the sum of profits may increase but it still falls short of the maximum because the externality level has not been chosen to equate its marginal value to marginal cost. To

conclude, the externality can be internalized only if parties are able to communicate either directly or through some institutional arrangement.

2.4.4 Free Riding & Social Dilemma

A social dilemma is a situation where following *individual rationality* (IR) leads to a worse outcome than following *social rationality* (SR). Such a situation may appear in a public or a private setting. The former case is the most celebrated and also the more intuitive one; it is presented first.

Public setting

The archetypal case is a public service or public good (cf. §16.1.1) such as Hume (1740)'s draining of a meadow or Rousseau (1755)'s stag hunt (cf. §2.4.1). It is well agreed that each who stands to enjoy the public work should contribute to it but it is also well understood that no-one will miss the opportunity to get away with the shore; this is the essence of free riding.

At the family or village level (small group), *peer pressure* forces everyone to behave i.e., to follow SR instead of IR, because each and everyone becomes a controller of the proper contributive behavior of others.[48@] At the city or country level (large groups), collective action is harder to implement as it is comparatively easier to shirk and implicit control become diluted if not ineffective altogether. Authorities have, from time immemorial, solved this conundrum by forcing everyone to contribute time or money to the construction and maintenance of public works; later on, in kind coerced contribution is replace by taxation which allows to fund the modern State' activities. Noted author de Condorcet (1788) explicitly justifies coercive taxation to avoid free riding.

Private setting

Social dilemma are also pervasive in private markets but are referred to as prisoner dilemma. They are a great concern for Industrial Organization since they relate to the efficiency of market equilibria and the resulting justification for State regulation. As we show formally in §9.1, the best for an industry selling a standard private good or service is to coordinate marketing strategies so as to implement the monopoly price and earn the associated maximum profit (which is then apportioned among members appropriately). Yet, every single member firm deviates from the cartel agreement if she follows solely her short term interest. Depending on how we model the strategic interaction between firms, the ensuing equilibrium lies between the Cournot and the perfectly competitive

outcomes. What really matters though is that all firms end up earning less (when compared to the successful cartel). Unless the industry, as a collectivity of economic agents, finds a way to discipline its members, it will suffer a *prisoner dilemma* and this constitutes the essence of the social dilemma in IO.

Misconception

Tuck (2008) shows that until the advent of game theory c. 1930, economists believe that rationality is socially based i.e., the bonding of the individual with his peer group is so complete that he takes the group's objective as his own. People or firms acting differently, typically in a selfish manner, are deemed irrational; they are either fools, ignorants (of the workings of the market), uninformed (about other firms), unable to coordinate (because of the large number of participants) or overemphasize the short term.[49@] This misconception means that these authors see social dilemma as episodes of bounded rationality. Now, if these bad outcomes result from the imperfections of man, there is the possibility to improve ourselves with education; the obnoxious IR behavior will then recede forever. It is our opinion that this lasting misunderstanding of the true forces underlying social dilemma have lead economists to neglect this aspect of economic interaction. Thanks to Olson (1965)'s book on the logic of collective action, social dilemma are now at the center stage in economics.

We conclude this section with a recollection of the opinions held by the important IO thinkers with respect to competition and how the modern conception used in this book emerged. Smith (1776) believes that oligopolists will take every opportunity to collude. He sees price wars, an instance of IR, as irrational behavior. His hopes to uphold competition therefore rest with freedom of entry.[50@] Mill (1848) exposes neatly the free rider problem faced by a labor union but refuses to see the individual worker behavior as rational; rather he calls for the law to enforce the union agreement on the ground that SR is the real objective of the workers.

Cournot (1838)'s model of duopolisitic competition based on IR behavior contains the first formalization of the prisoner dilemma (cf. §2.4.1) but since the author ignores altogether the potential for cartelization, he does not address the underlying social dilemma.[51@] This change of paradigm from SR to IR is not well accepted. Bertrand (1883), for instance, is of the opinion that in such a market, there is either collusion leading to monopoly or a price war leading to perfect competition. The latter outcome is however deemed "unwise" given that the short-term benefit of free riding on the cartel (with a price cut) is always lesser than the long-term benefits of collusion.

Edgeworth (1881) conclude his bargaining theory with the belief that collusion, not competition is rational for firms. Pareto (1896-97) makes a progress by assuming that

economic agents either follow SR or IR, the discriminating criteria being the size of their group (industry, community). Marshall (1890) stresses the benefits of combination, particularly through their ability to capture scale economies. Knight (1921) also believes that cartelization is the inescapable outcome when the conditions of perfect competition hold. Italian and Scandinavian public economists at the turn of the XX^{th} century also view free riding as a pathological behavior and, in any case, believe that a unanimous consent can be forged for financing public endeavors (cf. Musgrave (1939)'s account). Like de Borda (1781)'s voting method made to be used by gentlemen, these authors still believe in a social form of rationality. In a nutshell, they hold as inconsistent a person who values an outcome but do not to wish to contribute towards it.

Eventually, Chamberlin (1929) (better know as Chamberlin (1933)) introduces the concept of *oligopoly* for competition between a limited number of independent firms that fail to internalize the negative externalities their market behavior impose onto others. He thus argues that monopoly can be sustained, not because firms follow SR but because when following IR, they have the means to retaliate to a price cut or any behavior threatening monopolization. He also concludes that perfect competition will arise once there are so many firms that each can safely sell as much as it wishes because the impact unto others will be too small to trigger retaliation.

Part B

Market Power

Chapter 3

Monopoly

In his book Politics, the Greek philosopher Aristotle juxtaposes "sole" and "seller" to create *monopoly*, a concept he proceeds to illustrate with a tale: the clever mathematician Thales, to demonstrate his business wit, studied astronomy to foresee a soon to come abundant crop of olives, he then contracted out the services of all the olives presses in the city and made a huge profit by subletting the presses at a high price to heedless farmers when harvest came. Aristotle concludes that the acquisition of monopoly is the natural aim of any rational businessman, a claim we shall confirm.[1@]

In the first section, we define precisely monopoly and market power in economic language; then, we characterize the optimal behavior of a pure monopolist before assessing the welfare consequences of the exercise of market power. Lastly, we contemplate a number of extension of this basic theory.

3.1 Optimal Behavior

3.1.1 Typology

A monopoly can be either *de facto* or *de jure*, that is to say either conquered from contenders or bestowed by public authorities. Among the latter we distinguish among *private*, *regulated* and *public* monopolies. Let us explain the distinctions through a series of examples.

- de facto: The Boeing 747 held a monopoly over the long-haul high capacity aircraft market for 35 years before the appearance of the Airbus 380 contender because no other aircraft maker was able to produce a comparable plane.[2@]

- de jure private: The patented drug Nurofen held an almost worldwide monopoly over the painkiller market for 20 years because patents are recognized and enforced by almost all countries; patents are studied in §12 on R&D.

- de jure regulated: In most cities, taxicab services is not a perfectly competitive market because the city council delivers a limited number of licenses and sets the tariffs. Whether there are many individual drivers or one firm operating a fleet, we face a regulated monopoly as far as economic efficiency is concerned.

- de jure public: When the water network of a city is not operated by a private regulated firm but by a municipal agency, we face a public monopoly. Postal services is another example of public monopoly at the national scale.

The first two categories can be brought together under the heading of *private monopoly* which permits to drop the "de jure" prefix from the last two categories and derive a new classification based on the objective and constraints faced by the firm which fits better our economic study:

- The private monopoly is free to seek profit.
- The regulated monopoly is constrained in his pursuit of profit.
- The public monopoly cares for welfare.

3.1.2 Dominant Position

There are few cases of pure *de facto* monopolies but in many markets there is a dominant firm; an overt example is Microsoft over the market for operating systems (OS) of personal computers (PC). The European Commission clusters these two cases into the concept of *dominant position* whereby a firm is able to behave independently of its competitors, customers, suppliers and of final consumers. Notice further that a monopoly exists only insofar there is no close substitute to its product; for instance, the Nurofen painkiller has imperfect subtitutes like Aspirin or Paracetamol. As we shall see, focusing on the pure monopoly case involves no loss of generality to characterize the optimal behavior of firms holding market power (or dominant positions).[3@] Likewise, the *monopsony* (market with a sole buyer) being an entirely symmetric situation, we do not ascribe it much space (cf. §3.2.3).

It will be shown that with respect to the benchmark case of perfect competition,the monopoly earns an additional profit called the *rent*.[4@] This prize may motivate a *de facto* monopoly to abuse its dominant position to maintain it i.e., block entry or eradicate competition by anti-competitive means. This issue is taken on in Part D on *antitrust* laws. In this respect, recall that in the vernacular language, "monopolize" is synonymous of unfair exploitation or conspiracy to raise prices unduly and restrict output; it used to

be judged as sinful as usury. Modern laws are thus an economic update of the Christian doctrine according to which only two price setting methods are acceptable, perfect competition and the fair price set by public authorities.

For a regulated monopoly, the rent is set by the regulator so that the firm may be tempted to sway (e.g., lobby) her to increase the size of the rent, an issue we study in §16.3 on *rent-seeking*.

3.1.3 Market Power

In the remnant of this chapter, we deal with a private monopoly; regulated and public monopolies are the object of §17.

Awareness

Monopolies and competitive firms are similar in many respects. Their profit functions both read $pq - C(q)$ where p is the market price and q are the individual sales. Furthermore, both understand that the market demand D is a decreasing function of the market price p. At a given price p, no more than $Q = D(p)$ units can be sold. Inversely, the demand will absorb Q units only if the price is less than $p = P(Q)$, the (market) willingness to pay.

The only but fundamental difference lies in the relation between the price and the firm's sales. By definition, the supply q of a perfectly competitive firm is small in front of the total market sales Q and has therefore almost no effect on price, thus this firm rationally chooses to act as a price taker. For her, the price $p = P(Q)$ is unrelated to her own sales q.

The monopoly, on the other hand, is aware of his *market power*, the simple fact that his sales constitute the entirety of market sales i.e., $Q = q$. The equilibrium price is thus $p = P(q)$ and is a direct consequence to his sales decision q; this permit to rewrite the monopolist's profit as

$$\Pi^M(q) \equiv P(q)q - C(q) \tag{3.1}$$

Since price and quantity are linked by the demand equation, we could as well see the monopoly profit as a function of the price he wants to charge, which is more or less the way real firms behave. It is however easier to work out the theory using quantity as the primary decision variable. Unless stated otherwise, the fixed cost of production is zero throughout this chapter, so that profit equals producer surplus.

Intuition

Before searching for the optimal quantity and price, let us try, using the left panel of Figure 3.1, to understand the motivation of a monopoly. Consider a raise from the competitive price p^*; the rectangle grey area is the additional profit of selling the good at a higher price, being collected on all infra-marginal units it is called a *volume effect*. The small triangle, on the order hand, called the *price effect*, is the loss of profit generated by the loss of consumers; it is a negligible effect at p^* because the area of the triangle is very small compared to that of the rectangle.[5@] Hence,

A monopoly finds it profitable to raise his price above the competitive level.

The limit to the price increase by the monopoly is attained when the volume and price effects are equalized as shown by the grey areas on the right panel of Figure 3.1: the difference with the first case is that now, the consumers lost due to the price increase (those around q^M) were very profitable ones because the unit margin $p - C_m$ was large. The price effect now counterbalances the volume effect. We now proceed to characterize the optimal sales.

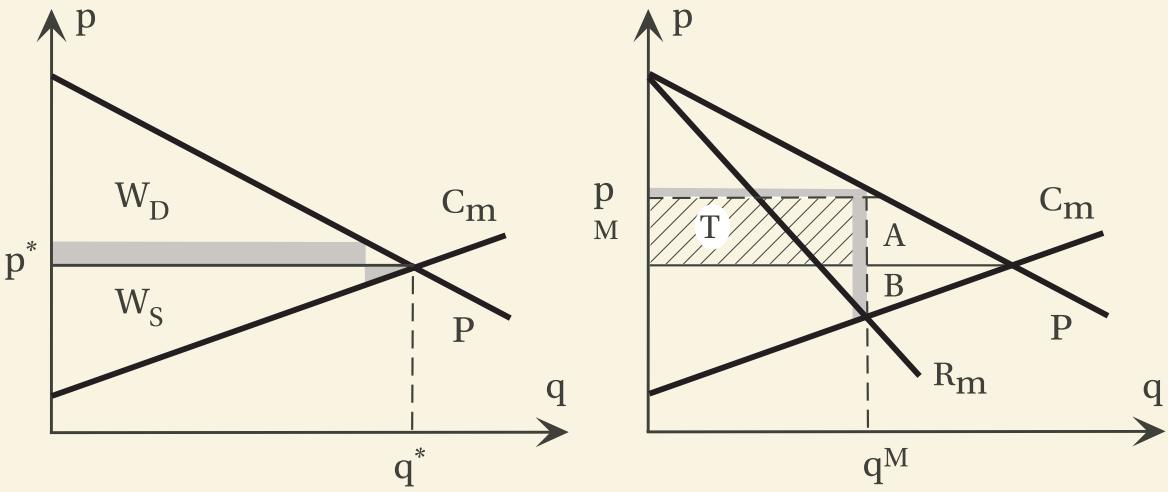


Figure 3.1: Competitive and Monopolistic Markets

3.1.4 Output Distortion

Optimal Choice

Formally, the optimal quantity q^M maximizes the profit Π^M and since this function is concave, its maximum is reached when the derivative becomes nil. The derivative of the

profit is called the marginal profit and since the profit is made of revenues minus cost, the marginal profit is the difference between marginal revenue and marginal cost. The maximum profit is thus reached when marginal profit is nil, that is to say when marginal cost and marginal revenue are equated. In formulas, this first order condition for profit maximization reads

$$\frac{\partial \Pi^M}{\partial q} = 0 \Leftrightarrow R_m = C_m \tag{3.2}$$

As we saw with equation (2.20), the marginal revenue is the price minus the marginal loss of consumer surplus: $R_m = P + W'_D < P$. We see on the right panel of Figure 3.1 that the R_m curve slopes down faster than the demand curve, hence the monopoly quantity q^M is lesser than the competitive one q^* (the latter solves $P = C_m$). Given the negative relation between price and quantity, the monopoly price p^M is larger than the competitive price p^*, as we already saw using economic intuition. To conclude,

> A monopoly maximizes his profits by equating marginal revenue and marginal cost; as a result, he sells less than the efficient quantity at a price greater than marginal cost (the efficient price).

Let us relate the optimality equation (3.2) to the previous graphical analysis. The gain from raising the price, the volume effect, is $q\Delta p$ while the loss of consumers is[6@] $\Delta q = D'\Delta p = \frac{\Delta p}{P'}$ which leads to a profit loss of $\Delta q(P - C_m)$; this is the price effect. Starting from p^* where $P = C_m$, the gain clearly outweighs the loss (which is zero); the optimum is attained when both effects neutralize themselves i.e., when

$$q\Delta p = \Delta q(P - C_m) \Leftrightarrow q\Delta p = \Delta p \frac{P - C_m}{P'} \Leftrightarrow C_m = P + qP' = R_m \tag{3.3}$$

One should take care to avoid mixing the marginal changes shown by the grey areas with areas $T + A$ and B that sum those marginal changes when the price rises from p^* to p^M. The change in profits for the monopoly in passing from p^* to p^M is the difference of areas T and B; we might wonder whether this quantity is positive. The answer is a clear yes provided by a revealed preferences argument: offering $p^M \neq p^*$ reveals that the monopoly has found a way to increase profits for he could always pretend to be myopic and act as a price taker if it was better for him. The excess of profits $T - B$ above those characterizing perfect competition is called an extraordinary profit or a quasi-rent; as we shall see in §17.3.1 on regulation, the presence of quasi-rents tells us that the capital invested into the monopoly has a return higher than the market average.

A monopoly obtains economic profits above the competitive level.

Lerner Index

Using the price elasticity of demand $\epsilon = \left|\frac{P(q)}{qP'(p)}\right|$ (cf. eq. 2.14), Lerner (1934) rewrites equation (3.2) characterizing the optimal monopoly price as

$$R_m = C_m \quad \Leftrightarrow \quad \mathscr{L} \equiv \frac{p - C_m}{p} = 1/\epsilon \tag{3.4}$$

i.e., at the optimum, the Lerner index of market power $\mathscr{L}$ equates the inverse of the demand elasticity.[7@]

This formula extends beyond the monopoly case to just any firm. Whatever the market structure (anything between monopoly and perfect competition), the demand addressed to a firm i can be related to the price p_i she charges, thereby leading the so-called residual demand function $\hat{D}_i$ (which possesses an inverse $\hat{P}_i$). Observe then, by going back to the analysis of optimal monopoly behavior, that the optimal pricing policy of firm i is to equate her Lerner index (cf. eq. 3.4) to the elasticity of her residual demand (i.e., change D by $\hat{D}_i$ and P by $\hat{P}_i$ everywhere). In the monopoly case, the relevant elasticity is that of the aggregate demand while in a competitive market, it is infinite (recall indeed that a competitive firm can sell as much as she wants at the market price). We can thus conclude:

As the market structure evolves from monopoly to perfect competition, the ability of a firm to price above marginal cost is reduced from a maximum down to zero.

The previous characterization is also useful to compare two markets like energy and entertainment. Whenever we own a gas owen, our ability to bake without gas is extremely low, therefore our demand for gas is quite insensitive to its price. On the contrary, if the toll for Pay-TV rises we can switch to one of the many substitute forms of entertainment at our disposal, thus our demand for a monopolized entertainment is sensitive to its price. Applying equation (3.4) tells us that energy firms with market power are able to price way above their marginal cost while entertainment firms are bound to be more competitive. This is why basic services like energy or water are regulated to avoid excessively high prices while hertzian TV is not.

Numerical Illustrations

Let us apply our findings to the numerical example used in the perfect competition section; the inverse demand function is $P(q) = \frac{a-q}{b}$ while the cost function is $C(q) = \frac{cq^2}{2}$. Leav-

ing detailed calculations as an exercise for the reader, equation (3.2) reads $\frac{a-2q}{b} = cq$, so that we find $q^M = \frac{a}{2+bc} < q^* = \frac{na}{n+bc}$ for all $n \geq 1$ (the number of competitive firms). The optimal monopoly price is $p^M = \frac{a(1+bc)}{b(2+bc)}$ and the Lerner index achieves the value $\mathscr{L} = \frac{1}{1+bc}$.

For the case of a constant marginal cost c that we predominantly use in this book, the monopolist's FOC becomes $\frac{a-2q}{b} = c$, so that we find $q^M = \frac{a-bc}{2}$ and $p^M = \frac{a+bc}{2b}$. The efficient behavior[8@] is in that case to price at marginal cost i.e., $p^* = c$ from which we deduce the efficient quantity $q^* = a - bc$. At the monopoly equilibrium, the Lerner index (cf. eq. 3.4) is $\mathscr{L} = \frac{a-bc}{a+bc}$ while profits are $\Pi^M = \frac{1}{4b}(a-bc)^2$ and consumer surplus is $W_D^M = \frac{1}{2b}\left(q^M\right)^2 = \frac{1}{8b}(a-bc)^2$. Finally, the (market) welfare is

$$W^M = \frac{3(a-bc)^2}{8b} < W^* = \frac{(a-bc)^2}{2b} \tag{3.5}$$

the welfare at the Pareto optimum, which is the maximum achievable in this market (for given technology and demand). A relationship useful for empirical estimations is to observe that the deadweight loss[9@] is $W^* - W^M = \frac{1}{2}\Pi^M$. Checking the exactness of these formulas is a fruitful exercise for the reader. To conclude,

> A monopoly does not behave efficiently and generates a loss of welfare.

The output distortion and its consequences in terms of welfare will be commented at length in the next section.

3.2 Inefficiency

As we shall see in this section, the unrestricted behavior of a monopoly is inefficient in the sense that it would be possible in an ideal world to achieve higher surplus with the same technology while meeting the same needs. von Pufendorf (1672) tells the archetypal case of the Dutch East India Company who destroyed spice crops in various parts of India in order to prevent an over–supply.[10@] The purely economical result we work out in this section rationalizes the predating moral view according to which monopoly is "evil" and must be reigned in (cf. §8.2.4). The public attitude towards monopoly is the object of chapter 17 on regulation.

3.2.1 Welfare Loss

As can be checked on the right panel of Figure 3.1, the switch from a perfectly competitive market (or any organization yielding efficiency) to a monopoly is accompanied by a wealth transfer T from consumers to monopoly and by a welfare loss. Indeed, we already

saw with equation (2.22) that any price different from the competitive one p^* generates less trade than the maximum achievable q^* and consequently a welfare loss. With a monopoly, welfare is $W^M = \Pi^M(q^M) + W_D(q^M) < W^*$, the difference called the *deadweight loss* triangle being measured by area $A+B$. The loss of consumer surplus over the $q^* - q^M$ units not consumed is A while B is the opportunity cost (profit loss) of not producing and not selling those units.

The increase of producer surplus $\Pi^M(q^M) - \Pi^*(q^*)$ is called the *monopoly rent* and is measured graphically by area $T - B$. Whether the monopoly is a private firm or a public enterprise, it is ultimately own by consumers, thus the only social cost is the deadweight loss $A + B$ caused by insufficient output. For a long time, economists have sticked to this point of view and have estimated empirically the welfare loss. The first study by Harberger (1954) using manufacturing data from the 1920s finds a welfare loss of $\frac{1}{10}$% of GDP, more recent studies compute a greater figure (several percent of GDP).

§17 studies the government intervention aimed at reducing the distortion brought about by unrestrained monopoly behavior.

3.2.2 X-efficiency

According to Hicks (1935), "the best of all monopoly profits is a quiet life", reflecting the idea that an avid young entrepreneur is bound to become a bourgeois once he has achieved industrial success.[11@] Leibenstein (1966) takes on this simple observation to introduce the loose concept of *X-efficiency*[12@] according to which firms with market power have higher costs than competitive firms. It is argued that in the absence of competitive pressure, shareholders get a constant stream of dividends and employees are protected from losing their jobs. Nobody is therefore poised to improve the competitiveness of the firm (i.e., reduce production costs). Employees, managers and workers, have leeway or slack on how to do their jobs and fail to implement "cost-cutting" processes, innovative technologies or learn how to better serve demand; rather they all engage in rent-seeking conduct. Although this view was never backed by formal theory, the European Commission argued in 1988 that "...the new competitive pressures brought about by the completion of the internal market can be expected to ...produce appreciable gains in internal efficiency...[which will] constitute much of what can be called the dynamic effects of the internal market..."

Several criticisms were addressed to this theory and accepted by its author. Firstly it is a sketchy story not supported by a game theoretical analysis of interactions inside the firm. In other words, it fails to explain why there is a change of behavior (with respect to cost minimization) when the environment changes from monopoly to competition. As

we all know, the lust for profits is universal, thus equally strong for the shareholders of monopolies and competitive firms: why would an investor holding shares in two such companies take a different behavior in the boards of the two companies ? Why would he support "cost-cutting" in the competitive sector and "take-it-easy" in the monopolistic one ?

Secondly, the X-inefficiency is said to arise because of a generalized slack in effort in the firm, the absence of adequate incentives towards cost minimization. As forcefully argued by Stigler (1976), the failure by managers to motivate employees or the failure by owners to motivate managers is not a waste; it is the rational outcome (a Nash equilibrium) of the utility maximization performed by every stakeholder in the firm. One must realize that the firm's technology is not limited to machines and land but also encompasses the internal organization and the entrepreneur's ability; each firm therefore follows the neoclassical theory to produce an output belonging to its technological frontier. The only conclusion we can draw from Leibenstein (1966)'s examples is that some frontiers are higher than others (cf. §2.1.2 on productive efficiency) and above all, that managerial innovations like piece rate remuneration and yardstick competition (e.g., "employee of the month" prize) are very effective.

Raith (2003) explores the issue of product market competition and managerial incentives to rationalize the diffuse idea that "more competition induces a better management". He shows that if "more competition" means a more elastic demand or less differentiation, then market power is reduced; this, in turn, forces some firms to exit (recall that in the long run, there are no economic rents) so that the remaining firms sell more. As we show in §12.2.4, the optimal level of management effort (deemed R&D in that chapter) is proportional to sales; in the present case, more competition has raised the incentives toward effort i.e., better management is observed. The same occurs if the market size grows; although there is entry, each firm sells more and therefore provides more incentives to its managers. Surprisingly, a deregulation which generates entry in a formerly monopolistic market leads to a reduction of incentives (according to this theory), contrary to common wisdom and to what the X-efficiency thesis predicts.

In the end, X-efficiency is best seen as a forerunner of the agency theory studied in Part H; it raised the attention of researchers on the inner logic of firms, on incentives, on the features of contracts used inside firms and how changes in organizational structure occur (cf. §13.1). Another argument (absent from the X-efficiency story) that might explain the higher costs observed in monopolistic industries is the presence of strong asymmetries of information across all market players. It is easier to perform relative evaluation of managers or divisions when the firm participates in a competitive market because there are many similar units to compare with (yardstick); estimates are then

more accurate and burden managers with less risk. Hence, shareholders of monopolistic or oligopolistic industries face a more difficulty task to extract the valuable information in possession of their managers. §21.2.3 exposes this issue at length for the regulation of a public monopoly.

> X-efficiency is concerned with the internal organization of firms and the failure of those working inside to minimize productions cost. The source of X-inefficiency are incomplete contracting, effort discretion and misalignment of objectives.

3.2.3 Extensions

Multiple Products

Although Volkswagen (VW) owns Seat (brand #1) and Audi (brand #2), these brands are competitors on the consumer market. As a firm with market power, Volkswagen must account for the possible substitutability or complementarity of these two brands to set its prices optimally. On top of this pricing issue, the company must also account for the fact that each brand is produced in a different plant (cf. the multi-plant firm problem seen in §2.1.3).

To understand how multiple products affect the pricing decision of firms, consider a monopoly offering products #1 and #2; since both demands depend on both prices, total profit is

$$\Pi(p_1,p_2) = p_1 D^1(p_1,p_2) + p_2 D^2(p_1,p_2) - C^1(D^1(p_1,p_2)) - C^2(D^2(p_1,p_2)) \tag{3.6}$$

so that the first order condition for price p_1 is $D^1 + p_1 D_1^1 + p_2 D_1^2 = C_m^1 D_1^1 + C_m^2 D_1^2$ (we denote $D_j^i = \frac{\partial D^i}{\partial p_j}$). After some algebraic manipulations we obtain

$$\mathscr{L}_1 = \frac{1}{\epsilon_{11}} - \mathscr{L}_2 \frac{\epsilon_{12}\, p_2 D^2}{\epsilon_{11}\, p_1 D^1} \tag{3.7}$$

where $\mathscr{L}_1$ is the Lerner index (cf. eq. 3.4) for price p_1, $\epsilon_{11} = \left| \frac{\Delta D^1}{D^1} / \frac{\Delta p_1}{p_1} \right|$ is the own-price elasticity of demand and $\epsilon_{12} = \frac{\Delta D^2}{D^2} / \frac{\Delta p_1}{p_1}$ is the cross-price elasticity of demand.

The interpretation is quite simple: if the two products are substitutes like automobiles, then $\epsilon_{12} < 0$ so that $\mathscr{L}_1 > \frac{1}{\epsilon_{11}}$ which means that the optimal price is larger than without taking into account the sales externality. If, on the other hand, products are complements like computers and operating systems, then the optimal prices are smaller than if they were set by two independent firms. To conclude,

A monopoly lowers the prices of complement products (cross-subsidization) to boost sales while he raises the prices of substitute products to boost margins.

Monopsony

A monopsony occurs when there is a sole buyer for a product. The state is a legal monopsony for the recruitment of civil servants while examples involving private firms and the labour market are for instance VolksWagen in the city of Wolsfburg, Germany or Michelin in the city of Clermont-Ferrand, France.

As we previously saw in §2.1.5, firms have a decreasing marginal willingness to pay V_m for additional units of factors, thus this is also true for a monopoly and for the labour factor. Conversely, the supply of labour L is an increasing function of wage w due to the leisure-work optimal arbitrage made by of consumers-workers. The inverse supply function is denoted $w(L)$ and the total wage expenditure is $B(L) = Lw(L)$. Efficiency commands to increase employment until the productivity of labour V_m equates the opportunity cost w. The monopsony, on the other hand, maximizes $V(L) - B(L)$ i.e., solves $V_m = B_m = w + Lw' > w$ which involves a quantity L^M lesser than the efficient L^* and a wage lesser than the competitive one w^*. Such an outcome occurs because, to employ more people, the monopsony is forced to raise everybody else's wages, so that the marginal cost of labour is not w but the greater B_m. To conclude,

A monopsony employs an inefficiently low amount of productive factor by offering a price lower than the efficient one; this generates a loss of welfare in the factor market.

3.3 Quality

In this section, we study the relationship of quality towards the exercise of market power.

Quality is an attribute of a product or service that all consumers view as useful or desirable. The resistance of an automobile to a crash-test or the number of airbags are objective quality elements since their adequacy can be assessed by technical tests. The presence of air-conditioning or leather seats are also seen as quality additions, although the utility they bring is more difficult to quantify. In the end, the usefulness of any attribute is subjectively judged by consumers alone which means that quality is a *demand side* attribute of a product or service. Contrary to engineers' belief, supply side attributes such as technical prowess or costly parts do not bestow any quality upon an item.[13@]

3.3.1 Monopoly and Quality

Quality vs. Cost

We assume for the purpose of our study the existence of an index s taking into account all quality characteristics of the product; this way, quality ranges from a minimum to a maximum. Since cheap and obvious quality additions to a product are common knowledge, increasing the quality index requires investments into market analysis to discover what new features (or altogether new products) consumers would like to possess. Then, one has to invest into R&D to develop, test and add such features to the current line of products. Finally, one has to advertise the novelties to establish leadership upon competitors (and reap the rewards through a higher selling price). It is generally accepted that all three steps of product development display decreasing returns to scale i.e., if k denotes the investment into any of these activities, the final quality index is $s = f(k)$ with $f'' < 0 < f'$ and where $f(0)$ is the quality of the standard (minimum quality) product.

A client assesses a product in terms of the value it delivers. We may assume that the latter is $v = qs$ where q is the quantity of product consumed. Since the investment k impacts mostly the quality perception of the product, we may, in a first approximation, assume it has no impact on production cost i.e., the cost of producing q units is $c(q)$ irrespective of the investment k. The cost of delivering final value v is thus $c\left(\frac{v}{f(k)}\right)$. We may then interpret the investment into product development (market research, R&D, advertising) as an effort to reduce cost. Hence, we can study w.l.o.g. R&D in §12 as a cost reduction activity rather than a demand booster activity. We apply this trick because viewing R&D as a *supply side* activity makes it much more amenable to the study of oligopolistic competition.

As we explain in §11, two products of different quality are said to be *vertically differentiated* whereas two products with distinct characteristics that appeal to different segments of the market are said to be *horizontally differentiated*. The monopoly attitude wrt. variety is treated in §6.1.4 and leads to a conclusion endorsed by casual observation:

> Firms with market power offer too much variety (brands, shops) given that the fixed cost of establishing a variety although heterogenous consumers long for many different ideal varieties.

Optimal Quality

Many people intuitively believe that the quality of a commodity or of a service like water distribution can never be too high. They err because in the process of providing quality, we are sacrificing scarce resources which would produce more service to the economy in

some alternative use (either in another public service or in production of private goods). This occurs as a consequence of the decreasing returns to scale of the technology used to improve quality. What is then the behavior of a monopoly with respect to quality selection? Is it too low or too high in comparison with an efficient level? To answer this question, we assume that the quality index s impacts demand $D(s,p)$ in the following manner: $\frac{\partial D}{\partial s} > 0 > \frac{\partial^2 D}{\partial s^2}$: an increase in quality raises the demand but with a fading force.

For a monopoly market structure, and ignoring fixed costs, welfare is

$$W = W_D + W_S = \int_p^{+\infty} D(s,x)\,\mathrm{d}x + \pi$$

Contingent on a price level (whether p^M, p^* or anything else), we have

$$\frac{\partial W}{\partial s} = \int_p^{+\infty} \frac{\partial D(s,x)}{\partial s}\,\mathrm{d}x + \frac{\partial \pi}{\partial s} \geq \frac{\partial \pi}{\partial s}$$

Hence, once the monopolist sets s so as to satisfy $\frac{\partial \pi}{\partial s} = 0$, the welfare can still be increased by raising quality further.[14@]

> Contingent on the market price, a monopoly *under provides* quality which rationalizes the frequently observed *minimum quality* requirements imposed by regulators on firms whose price or production is regulated (cf. §16.3).

The above conclusion is partial because the monopoly and the planner (who maximizes welfare) choose different prices, thereby making impossible an absolute judgment with respect to quality. To go further, we have to study in more detail the willingness to pay $P(s,q)$ which now depends on quality.

When $\frac{\partial^2 P}{\partial q \partial s} < 0$ i.e., the WTP for added quality $\frac{\partial P}{\partial s}$ falls as consumption increases, then quality and quantity are said to be *substitutes*; an example is clothing because if you switch to a higher quality, you need to replace less often your wardrobe, thus your WTP for a fifth pair of shoes is lesser than it used to be. If the reverse inequality $\frac{\partial^2 P}{\partial q \partial s} > 0$ holds, then quality and quantity are said to be *complements* as could be the case for food because the WTP for a better restaurant is higher for people who go out dinning often (than for those who go out rarely).

When quality and quantity are substitutes, the general conclusion reached by Spence (1975) depends on the elasticity of demand ϵ which measures the innate market power of the monopoly. If ϵ is large, the monopoly and efficient quantity are not too different so that the quality ranking is not changed: the monopoly under provides quality. If, on the contrary, ϵ is low (e.g., electricity) then the monopoly provides an excessively high quality.[15@] These findings are reversed when quality and quantity are complements.

Table 3.1 summarizes our findings.

market elasticity is / quality and quantity are	high	low
substitutes	s^M is insufficient	s^M can be excessive
complements	s^M can be excessive	s^M is insufficient

Table 3.1: Quality selection by a monopolist

The previous findings shed light on some practices, for instance the policy of an opera house or cultural business. The private operator caters to the marginal client (an occasional opera buff) who is less knowledgeable than the average client (an aficionado); he thus chooses a familiar repertoire whereas a public operator who caters to the average client will choose a wider selection of works and authors. This is empirically validated when looking at the restricted repertoire of privately promoted classical music shows wrt. government sponsored ones. The same analysis goes on to all the recreational activities (media, culture).

Linear Setting

Mussa and Rosen (1978) offer a very simple treatment of quality choice by a monopolist that serves as a basis for the study of competition and differentiation.

- Quality s ranges from zero to some maximum, normalized to unity.[16@]
- Clients are considering whether to buy a single unit.
- Heterogeneity: the WTP θ for the top quality ranges from zero to some maximum, again normalized to unity (cf. extension later on).
- WTP is uniformly distributed; total market mass is unity.
- A client with type θ has WTP θs for quality $s \in [0;1]$.
- Zero production cost.
- Cost of designing quality s is $\frac{s^2}{2}$.

When offered at price p, the item is sold to all types with $\theta s \geq p$. Demand is thus made of all the people with type between $\frac{p}{s}$ and 1 i.e., $q = 1 - \frac{p}{s}$. Since market WTP is $P(s,q) = s(1-q)$, quantity and quality are substitutes (according to the previous classification). Consumer surplus is $W_D = \int_{p/s}^{1} (\theta s - p)\,\mathrm{d}\theta = \frac{1}{s}\int_0^{s-p} x\,\mathrm{d}x = \frac{1}{2s}(s-p)^2 = \frac{s}{2}q^2$. Profit being $\pi = pq$, the optimal price is $p^M = \frac{s}{2}$ leading to sales of $\frac{1}{2}$, profit $\pi^M = \frac{s}{4}$ and welfare $W^M = \frac{3s}{8}$.

Taking into account the cost of designing quality, we maximize net profit $\Pi(s) = \frac{s}{4} - \frac{s^2}{2}$; the optimal quality choice is then $s^M = \frac{1}{4}$. Observe at this point that, since production cost are zero, efficiency commands to giveaway the item, so that demand is unity and

welfare $W^* = \frac{s}{2}$. The efficient quality maximizing net welfare is $s^* = \frac{1}{2}$, twice the optimal choice of the firm. We have thus shown that the monopolist under provides quality because uniform pricing forces her to forgo clients with a low WTP which then dampens her incentive to invest in quality.

A simple extension is to allow for WTP to be distributed between zero and some upper bound $\bar{\theta}$ while keeping total market mass at unity. It is a matter of algebra to check that demand is then $q = 1 - \frac{p}{\bar{\theta}s}$ so that profit is maximum for $p^M = \frac{\bar{\theta}s}{2}$ which gives a maximum profit of $\pi^M = \frac{\bar{\theta}s}{4}$ and welfare $W^M = \frac{3\bar{\theta}s}{8}$. Letting the cost for quality being $\frac{\beta}{2}s^2$, the optimal choice is $s^M = \frac{\bar{\theta}}{4\beta}$, leading to net profit $\frac{\bar{\theta}^2}{32\beta}$ and net welfare $\frac{\bar{\theta}^2}{16\beta}$. Exactly as before, the efficient quality maximizing net welfare is twice larger than the monopolist's optimum and net welfare is also twice larger.

3.3.2 Transportation †

Frequency of service is a key quality attribute for mass transportation (e.g., airplane, train or bus). We apply the previous theoretical considerations to the specific case of a transport monopolist, say a train operator.

Letting t be the time interval between two departures, the maximum waiting time is $\frac{t}{2}$, so the average is only $\frac{t}{4}$. If a consumer values, on average, δ each hour lost waiting for the train, then his expected opportunity cost of waiting is $\frac{\delta t}{4}$. Letting T stand for the duration of the service period (e.g., 24h) and s for the frequency of carriers (the quality index), we have $t = \frac{T}{s}$, so that the unit cost of time is $\frac{\gamma}{s}$ where $\gamma \equiv \frac{\delta T}{4}$ is the sole parameter we shall need to derive the optimal monopoly conduct.

We assume that the market WTP for immediate service is the standard linear formula $\frac{a-q}{b}$; the WTP contingent to a quality of service s is then $P(q,s) = \frac{a-q}{b} - \frac{\gamma}{s}$ as we subtract the opportunity cost of waiting.[17@] The gross utility of consumers is $U(q,s) \equiv \int_0^q P(x,s)\,\mathrm{d}x = q\frac{a-q/2}{b} - \frac{\gamma q}{s}$. In this setting, two cost dimensions emerge, the traditional one linked to volume and a new one linked to quality (frequency). Indeed, the cost of transporting q customers in s carriers is $C(q,s) = \theta s + cq$ where θ is the cost of moving one carrier and c the cost of moving one passenger.[18@] The monopolist's profit is now easily computed as

$$\pi = qP(q,s) - C(q,s) = q\frac{a-q}{b} - \frac{\gamma q}{s} - cq - \theta s \tag{3.8}$$

while welfare is $W = U - C$. The optimal monopoly quantity, conditional on quality s, solves the FOC

$$\frac{\partial \pi}{\partial q} = 0 \Leftrightarrow R_m = C_m \Rightarrow q^M = \tfrac{1}{2}\left(a - bc - \frac{\gamma}{s}\right) \tag{3.9}$$

while the efficient quantity solves $\frac{\partial U}{\partial q} = \frac{\partial C}{\partial q} \Leftrightarrow P(q,s) = C_m$ yielding $q^* = 2q^M$ as in the

standard monopoly set-up since we are assuming the same quality.

The optimal quality chosen by the monopolist solves the FOC

$$\frac{\partial \pi}{\partial s} = 0 \Leftrightarrow s^2 = \frac{\gamma}{\theta} q \tag{3.10}$$

which, in our simple additive model,[19@] is also equivalent to the FOC for an efficient quality $\frac{\partial U}{\partial s} = \frac{\partial C}{\partial s}$. It turns out that the incentives for quality are the same for the monopolist and the social planner but since the later chooses a greater quantity, he also ends-up choosing a greater quality (check in (3.10)). In turn, this implies $q^* > 2q^M$ by a simple comparison of the FOCs for optimal quantities. The optimal carrier size, $\frac{q}{s} = \frac{\theta s}{\gamma}$ by (3.10), is therefore larger at the efficient combination as compared to the monopoly's choice.

To conclude, let us inquire how changes in carrier cost θ, opportunity cost of time γ or market size $a - bc$[20@] impact the quality choice of the monopoly. Using (3.9), we can rewrite (3.10) as an equality between marginal benefit and marginal cost:

$$a - bc - \frac{\gamma}{s} = 2q^M = \frac{2\theta}{\gamma} s^2 \tag{3.11}$$

Since $s \geq 1$, the LHS is an almost flat curve while the RHS is convex increasing. Neglecting $\frac{\gamma}{s}$ in the LHS, an approximate solution is $s^M \simeq \sqrt{\frac{\gamma(a-bc)}{2\theta}}$. Hence, a smaller train cost ($\theta \searrow$), a greater opportunity cost of time ($\gamma \nearrow$) or a larger natural demand ($a - bc \nearrow$) leads to greater frequency s^M, greater traffic q^M and, due to (3.10), a greater carrier size to achieve these choices. Since $q^* \simeq 2q^M$, (3.10) implies that the efficient frequencies and carrier sizes are approximatively 44% larger than those optimal for the monopoly (multiplied by a factor $\sqrt{2}$).

3.3.3 Monopolistic Taxation

Another application of the monopoly theory of quality choice is the provision of public goods and services by the State using taxation as developed by McGuire and Olson (1996). We are dealing here with *public goods* such as national defense or the security of property rights and *public services* such as infrastructures, education or health (cf. §16.1.1 for precise definitions). These goods and services reduce the uncertainty of economic endeavors and many cost items, thus facilitate trade and investment so that in the end the economy grows faster.[21@] The fundamental parameter is then which objective pursues the State. Analytically, a more or less representative elite taxes the economy; it runs from despotic appropriation to benevolent planning at the hand of a democratically elected government.[22@] We also perform comparative statics over the size of the elite

(without however explaining how it can grow).

Leviathan

Consider first the case of a single decision-maker, the ruler,[23@] holding power over a territory producing taxables riches. Concretely, one may think of a lord or a small clique whose source of power is the absolute control over security forces such as police and/or the army (cf. §16.2.4). The ruler can use part of tax revenues to finance public goods in anticipation that the local economy will grow richer and thus become a larger tax base that can be further plundered. The tax base takes the role of market demand and is therefore denoted $D(p,s)$. The tax rate is p (i.e., market price) while and the amount of public goods s is akin to a quality index (measured in monetary units). That D decreases with p reflects the disincentive that taxation creates on producers who anticipate the lower return of their investments.[24@] That D increases with s reflects the productivity enhancement brought about by public goods.

The first-best choice maximizes welfare $W = D(p,s) - s$.[25@] It is immediate to observe that efficiency calls for the elimination of taxes (set $p = 0$) to avoid distorting private activity and to choose the public goods amount s^* solving $\frac{\partial D}{\partial s}\Big|_{p=0} = 1$. The ruler, on the other hand, seeks to maximize her net wealth $\pi = pD(p,s) - s$ under the financing constraint of public goods $\pi \geq 0$. To clarify the exposition, we assume $D(p,s) = r(p)\Phi(s)$ where $\Phi(s)$ is the aggregate maximum production in a free-market economy with a level s of public goods and $r(p)$ is the net labour supply that producers choose upon anticipating taxation of their returns at the rate p. We assume decreasing returns to scale (DRS) wrt. production with $\Phi'' < 0 < \Phi'$. Regarding the reaction of producers to taxation, we assume

- $r(0) = 1$: without taxes, work incentives are perfect and the aggregate maximum production is achieved.
- $r', r'' < 0$: higher taxes generate fewer production at an increasing rate.
- $r(1) = 0$: if all the tax base is taken away, no taxable output is produced because workers are either gone or dead.

In this very simple setting, the efficient level of public goods solves $\Phi' = 1$ while the ruler maximizes $\pi_0 = pr(p)\Phi(s) - s$ i.e., chooses the tax rate p^0 maximizing the net return $R_0(p) \equiv pr(p) < 1$. Since $R_0(0) = 0 = R_0(1)$, the graph of R_0 is bell-shaped, thus reaches an interior maximum at the solution of the FOC $p = -\frac{r}{r'}$, as shown early on by Khaldun (1377) (it is also known as the Laffer curve). The optimal amount of public goods s^0 then solves $\Phi'(s) = \frac{1}{R_0(p^0)}$.[26@] Since the RHS is greater than unity, there is under-investment (wrt. first-best) because the ruler grabs only a share $R_0(p^0)$ of potential output. The crowd's per-capita utility is $u_0 \equiv (1 - p^0)r(p^0)\Phi(s^0)$.

Clique or Elite

The ruler is now replaced by an elite making up a share α of the population; its income thus comprises taxation revenue and net-of-tax producer surplus i.e., its objective is now $\pi_\alpha \equiv \pi_0 + \alpha(1-p)r(p)\Phi(s)$ while the public finance constraint remains $R_0(p)\Phi(s) \geq s$. Since the maximum of $(1-p)r(p)$ is achieved for zero tax ($p=0$), the maximization of π_α involves an intermediate tax rate p^α between 0 and p^0 i.e., the elite has an *encompassing interest* for the economy and restraints itself from too much taxation.

Observing that $\pi_\alpha = R_\alpha(p)\Phi(s) - s$ where $R_\alpha(p) \equiv \big(p + \alpha(1-p)\big)r(p) > R_0(p)$, the optimal tax solves $R'_\alpha(p) = 0 \Leftrightarrow p = -\frac{r}{r'} - \frac{\alpha}{1-\alpha}$ which shows, again, that the optimal tax decreases with the elite size α. The optimal amount of public goods s^α then solves $\Phi'(s) = \frac{1}{R_\alpha(p^\alpha)}$ which leads to $s^\alpha > s^0$ i.e., now that the elite has an encompassing interest with the rest of the crowd, it favors public goods more than a lone ruler.[27@] The crowd's per-capita utility is now $u_\alpha \equiv (1-p^\alpha)r(p^\alpha)\Phi(s^\alpha)$; it increases in all three dimensions, the Pareto possibility frontier rises (more public goods), their incentives are greater (due to less taxes) and more of the produce remains their own.

For small α, the amount of public goods is low and is therefore easily financed i.e., in the neighborhood of $\alpha = 0$, $p^\alpha r(p^\alpha)\Phi(s^\alpha) > s^\alpha$. However, at the other extreme, we have $p^1 = 0$, thus no resource at all to finance the large amount s^1. This means that for large α, the budget restriction $pr(p)\Phi(s) \geq s$ is binding. Since the technology displays DRS, the average cost $\frac{s}{\Phi(s)}$ is increasing, thus the equation $pr(p) = \frac{s}{\Phi(s)}$ defines a positive relationship $s = h(p)$.[28@] The elite's objective is now $\pi_\alpha = \big(p + \alpha(1-p)\big)r(p)\Phi(s) - pr(p)\Phi(s) = \alpha(1-p)r(p)\Phi(h(p))$. Observe that the optimal choice is now independent of α i.e., the elite has grown so large that it is now "super encompassing" and treats the crowd as if it belonged to the elite. Since the term $(1-p)r(p)$ is decreasing while the term $\Phi(h(p))$ is increasing, an interior solution p^* exists together with the associated level of public goods $s^* = h(p^*)$.

As a matter of example, if we take $r(p) = 1-p^2$ and $\Phi(s) = \sqrt{s}$, we find $p^\alpha = \frac{-\alpha+\sqrt{4\alpha^2-6\alpha+3}}{3(1-\alpha)}$ and $s^\alpha = \frac{R_\alpha(p^\alpha)^2}{4}$ while $p^* = \frac{1}{3}$ so that the super-encompassing minority starts at $\alpha = \frac{1}{2}$ when $p^\alpha = p^*$.[29@]

Monopolization and Rent-seeking

Congleton and Lee (2008) enrich this model of public finance with an alternative mean of generating revenue for the ruler: the sale of monopoly licenses. It is easy to see that the monopolization process of the economy though the sale of an ever increasing number of licenses goes on up the point where the marginal revenue equates the marginal loss of general taxation generated by the lesser economic activity in the monopolized sectors

of the economy.

As we recall in §16.2.4, monopoly licensing was the main revenue channel for mercantilist governments in the pre-industrial area. Its relevance nowadays is weaker but still present. Some developing countries use it more or less explicitly (to turn around formal prohibitions). Even advanced countries recur to it occasionally, for instance when auctioning licenses for mobile phone telecommunications. More recently, cash-strapped Greece sold future income from airports and lottery to private investors in return for immediate cash. Similarly, the city of Chicago, in need of funds, privatized its parkings for a price widely believe to understate the net present value of the assets. Nowadays, the most widespread form of privilege sales are those paid in kind with political support (cf. §16.3).

Chapter 4

Differential Pricing

One speaks of "price discrimination" or "differential pricing" when a firm sells two similar products with close marginal cost of production at different prices. The discriminating characteristic can be the identity of the buyer, the calendar date, whether the customer is a new comer, the total volume bought, whether another item is bought jointly. A few examples will help to set the stage. The famous Coca-Cola brand offers a neat example using May 2010 Spanish prices; you can drink this beverage at the bar, from a can bought at an automatic dispenser or a variety of bottles and bundles found at a nearby store.

Size (liter)	.20	.33	12×.20	12×.33	.5	1	2	2×2	4×2
Price (€)	1.5	1	2.84	5.93	0.8	0.97	1.36	2.36	4.65
€/liter	6	3	2.36	1.5	1.6	0.97	0.68	0.60	0.58

Table 4.1: Retail Prices for Coca-Cola

Next, we've queried a major airline on the internet in November 2005 for two round-trips, Paris↔New-York and Paris↔Madrid.

Transatlantic	price (€)	Continental	price (€)
economy 1 month	393	week-end now	190
economy week end	930	week-end later	103
economy in week	2230	long stay	260
~ flexible	3160	one way now	715
~ business class	5084	in week later	408
~ first class	8064	during this week	1000

Table 4.2: Airline Tickets Prices

The "new economy" offers another striking example of differential pricing. Odlyzko (2004a) estimates in Table 4.3 the revenue generated by a megabyte of digital connection according to the use that is made of it; there is a factor of 25 millions between the cheapest and most expensive channels.

Channel	$ / MB
Cable TV	0.00012
ADSL internet	0.025
Voice Phone	0.08
Dial-up Internet	0.33
Cell phone	3.50
SMS	3000

Table 4.3: Revenue from digital services

In this chapter, we go beyond the basic monopoly pricing exposed by Cournot (1838) to explore the refined use of market power as first analyzed by Dupuit (1844). Differential pricing takes advantage of the heterogeneity of customers' willingness to pay to design personalized prices. The idea is to be nearer to one's customers needs in order to charge them higher prices which is echoed in the old maxim that "one ought to tariff at what the demand will bear". Differential pricing was first systematically analyzed by Pigou (1920) under the heading of *price discrimination* (cf. definition in §4.1.1). The complementary strategy of *differentiation*, studied thoroughly in Part E, is to be different from one's competitors through a diversification in quality and/or variety.

The plan of the chapter is as follows. The first section extensively discusses the concept of discrimination, the conditions for its application and its basic instruments. The next section shows how a firm with market power uses the instruments to extract consumer surplus given its legal and informative environment. The last section tackles differential pricing when the firm is in the weakest position and shows that an intelligent design can lead clients to self select a category or segment upon which traditional market power can be applied.

4.1 Premices

It is noticeable that in most of the economics literature, price competition is understood as uniform or linear[1@] whereby the same price is charged for every unit sold. This definition thus excludes the popular strategies of volume discount or packages. Two reasons for this odd restriction can be advanced: the inertia of economic thought and the natural efficiency of economic relationships. Whereas the former does not warrant our attention, we shall deal with the latter: it is thanks to their market power that firms are able to escape uniform pricing and develop differential pricing.

It is also customary in the literature on differential pricing to concentrate on revenue and thereby ignore cost. This practice would seem at odds with the underlying profit maximization objective of firms but it is almost without loss of generality regarding vari-

able cost since we can integrate them in the revenue function which becomes net (of cost). Next, we observe that a private firm covers its fixed cost out of its producer's surplus and thus does not take those into account for the determination of its pricing strategy. The issue of costs however cannot be ignored when dealing with public or regulated firms. For that reason, chapter 17 on *regulation* will focus on cost and check the truth of another old maxim stating that "every tube must stand on its own". Lastly, our results on differential pricing shall equally apply to any firm with market power once the relevant elasticity of demand is used as we already argued when discussing the Lerner formula (cf. eq. 3.4).

4.1.1 Definition

The vernacular definition of *discrimination* is an "unfair treatment of a person or group on the basis of prejudice" while the use in economics adheres to the original latin meaning which is to separate or distinguish. We shall use the expression *differential pricing* and the verb *discriminate*.

A first observation about differential pricing is that it would never appear in a perfectly competitive markets because the law of one price applies: buyers and sellers participating in a competitive market are bound to accept the market price and can only propose a quantity to trade. Hence, it is necessary to hold a minimum of market power to be able to perform differential pricing. Whether this can be interpreted as anti-competitive is deferred to §9 on anti-competitive practices.

In the previous chapter, we introduced the monopoly as a firm having the power to set the price while consumers retained their ability to set the quantity. This market relationship is quite similar to the "divide and choose" procedure to share a pie where one person cuts the cake in two pieces and lets the other person choose the piece she likes best. This story does not fit our intuitive representation of the monopoly as having "maximum market power"; rather one would expect him to make a "take-it-or-leave-it" offer to the helpless consumer in order to grab the greatest profit. This is what differential pricing is about; finding contract proposals–*tariffs*–that extract a maximum surplus from consumers. Practically, the firm uses the heterogeneity of their tastes to treat them differently.

This economically founded discrimination (as opposed to that founded on moral prejudice) can in theory be applied to human traits such as age, residence or occupation if it does not conflict with higher level laws protecting the fundamental rights of citizens. Even if the firm is prevented from applying such a *direct* discrimination, we shall demonstrate that it is still possible to discriminate by *indirect* means. This is why the concept

of differential pricing should be remembered as “the sophisticated exercise of market power”.

Using the previous example of the novel, it is quite simple to recover the steps leading to the observed pricing behavior. The editor knows that different people have different WTP for the novel. He would like to be able to guess it, simply by looking at the client’s face and then charge him/her that price. The first necessary step for successful discrimination is thus the gathering of *information*. Ideally, it is possible to identify the WTP of each and every single consumer; this is a limiting case never observed in reality but useful for the theory. Practically, discrimination is bound to be imperfect and uses statistics, market studies (sociology) and interviews (psychology) to *segment* the demand[2@] according to attributes like age, sex, occupation or the postal code (a good indicator of wealth, itself a good proxy of WTP).

Next, when discriminatory pricing takes place, some consumers realize they are paying more than others for the same good or service. Although the firm will always try to present the cheap price as a “special” discount over the regular price and not the reverse, these consumers will feel deceived and will rebel against this prejudice.[3@] A first solution is to sue the firm for discriminatory and unfair treatment. Historically, the discrimination widely used by railways companies in the XIX^{th} century for categories of freight or passengers generated public outrage, political reaction and governmental prohibition. Later on, during the XX^{th} century, discrimination based on human traits became unlawful in most countries.[4@] We have thus obtained a second necessary condition for successful discrimination: *legality*. Yet, the legal framework for economic activity leaves firms free to set different prices for goods sold in different places, at different times or for different intended uses. For instance:

- German cars are more expensive in Germany than in Italy.
- Holiday packages are cheaper in September than in August.
- Softwares are cheaper for home use than for professional use.
- Mobile phone is cheaper for professional use than for home use.
- Per capita income tax is cheaper for married couples.

If discrimination is legal, *arbitrage* might nevertheless weakens it considerably since the lucky consumers who can buy the item at a bargain price might try to resale it to consumers whom the monopoly asks a high price. For instance:

- Some Germans go to Milan to buy their German car.
- Some people work in August and leave on vacation in September.
- Some independents buy a home version of a software for professional use (hoping not to be sued for such an illegal practice).

- Some people employ their professional mobile phone for personal use.
- Couples get married to save on their tax bill.

Arbitrage can even be such a threat that discrimination altogether disappear (as predicted by the perfect competition model). Some internet vendors such as bookshops or airlines use computer cookies to detect if a visitor has already bought an item from them in which case they quote higher prices because they know that the visitor is serious about buying more items. It is enough to use a different browser or to erase cookies to perfectly bypass this intent of differential pricing. We shall come back later on the issue of arbitrage explaining especially how consumers intent to bypass discrimination and how firms intent to prevent arbitrage.

4.1.2 Typology

Table 4.4 summarizes our finding regarding the steps to follow in order to successfully perform differential pricing:

- *Differential Pricing*: Extract consumer surplus by segmenting demand
- *Segmentation*: Sort heterogeneity of tastes into homogeneous groups
- *Rent Extraction*: Devise specific tariff for each segment
- *Necessary Conditions*: Information, Legality and No-arbitrage

Table 4.4: Stages of Differential Pricing

Notice that segmentation refers to *interpersonal* price discrimination since the firm discriminates across consumers whereas rent extraction refers to *intrapersonal* price discrimination as the firm discriminates across units for the same consumer.

The practical implementation runs into the difficulty of satisfying completely all three conditions of *information*, *legality* and *no-arbitrage*; some degree of incompleteness is often unavoidable. For instance, car insurance premium would be profitably based on the following binary classes: residence (city/countryside), gender (M/F), age (young/old) and accident records (some/none). This would lead to establish 16 different premiums. Yet the last characteristic is unobservable thus only 8 segments can be constructed. If a projected European directive[5@] is adopted, gender discrimination will be prohibited, thus limiting discrimination to 4 segments. Finally, if most families involve young and mature people, they can always bypass the age discrimination (arbitrate) by sending the relevant member to apply for insurance. It might therefore be the case that the firm can only build two segments upon which (direct) discrimination is feasible i.e., city dwellers are being asked to pay a dearer premium for the same service.[6@]

It would seem that, within a segment where *direct* discrimination is impossible to carry out, there is nothing that the firm can do to boost profits. This is a very misleading intuition because the firm can always resort to *indirect* discrimination. The idea is simply to propose the whole range of discriminatory proposals to everybody at the same time and let consumers pick the one they prefer, a behavior known as *self-selection*. The trick is to design these proposals appropriately in order that each consumer picks the proposal you would want him/her to choose. Those criteria are for instance:

- *Quantity*: Unit price varies with the quantity bought
- *Differentiation*: Several versions of the good are offered
- *Temporality*: Price changes with seasons or with time

Table 4.5: Forms of Indirect Price Discrimination

We can summarize our typology with the help of Table 4.6. We adopt a logical approach to discrimination reflecting the many theoretical developments of the last 20 years; it is a departure from Pigou (1920)'s received typology where perfect discrimination is refereed to as "first degree", indirect discrimination as "second degree" and direct discrimination as "third degree".

- *Perfect*: Each segment is formed by homogeneous consumers
- *Imperfect*: Segments contain heterogeneous consumers
- *Direct*: Discriminate across segments
- *Indirect*: Discriminate inside segments
- *Basic*: Based on unit price only
- *Complex*: Based on packages of price and {quantity or attribute}

Table 4.6: Typology for Price Discrimination

4.1.3 Consumer Surplus Extraction

In this part, we look at all the methods available to extract the surplus of an individual consumer when his demand curve is known to the monopoly. We already know that these schemes have to be more complex than basic monopoly pricing since the latter leaves a net surplus to any of his clients. The three methods are the quantity rebate (or price listing), the club pricing (or two-part tariff) and the minimum purchase obligation (aka "take-it-or-leave-it" or "take or pay" offer). As explained in the introduction to this section, we focus on revenue and thus assume a very simple cost structure with constant marginal cost and zero fixed cost. We can already advance the result of this section:

> Conditional on information, legality and no-arbitrage, quantity rebates, club pricing or minimum purchase obligation are equivalent methods to rip off entirely the surplus of consumer. As a by-product, efficiency is restored when compared to standard monopoly pricing.

Price Listings aka. Quantity Rebates

Let us now introduce the first strategy to extract consumer surplus and demonstrate, after Pigou (1920), its efficiency using the movie rental example of §2.2.2 reproduced on the left panel of Figure 4.1. The client's WTP for the first, second, third ... movie rental enables to draw a staircase. Imagine now that the shop manager askes her to fill a survey in order to be able to guess her inverse demand curve (WTP) $P(q)$ (dashed line). The seller can now make a personalized rental offer to the client: the first movie is proposed for the price $p_1 = P(1)$, the second one will be cheaper and is proposed at the price $p_2 = P(2)$, the third unit is offered for a still lower price $p_3 = P(3)$; this discounting process goes on until the WTP becomes smaller than the marginal cost c of a rental. The price of additional rentals then becomes equal to c.

When offered this personalized price-listing, the client will elicit a number of movie rentals equating her WTP with c.[7@] The purchase is therefore the efficient quantity $q^* \equiv D(c)$ while the profit of the monopoly is the area between the demand curve and the marginal cost curve i.e., the maximum market welfare W^* (cf. definition §2.3.2). For the case of linear demand $D(p) = a - bp$ and constant marginal cost c, $q^* = a - bc$ and $W^* = \frac{1}{2b}(a - bc)^2$ (cf. eq. 3.5).

Observe that the pricing list can be interpreted as a list of quantity rebates: the regular price is p_1, a discount $p_2 - p_1$ is offered on the second unit, a further discount $p_3 - p_2$ is offered on the third unit and so on.

Two-part Tariff aka. Club Pricing

Our second consumer surplus extraction strategy, first described by Hicks (1943) as the quantity compensating variation, involves a change of attitude: instead of renting expensive movies inside a shop with free entrance it is wiser to charge an entry fee (make it a club) and then rent at a bargain price inside.[8@]

To see why this is a good idea in terms of profits observe on the right panel of Figure 4.1 that when the manager sets a per-unit price p, the consumer optimally buys $D(p)$ units and derives a (consumer) surplus $W_D(p)$, the excess of her total WTP over her total expenditure. The consumer participation constraint is thus $f \leq W_D(p)$ and it is optimal for the manager to saturate this constraint i.e., ask his client an entry fee f as high as

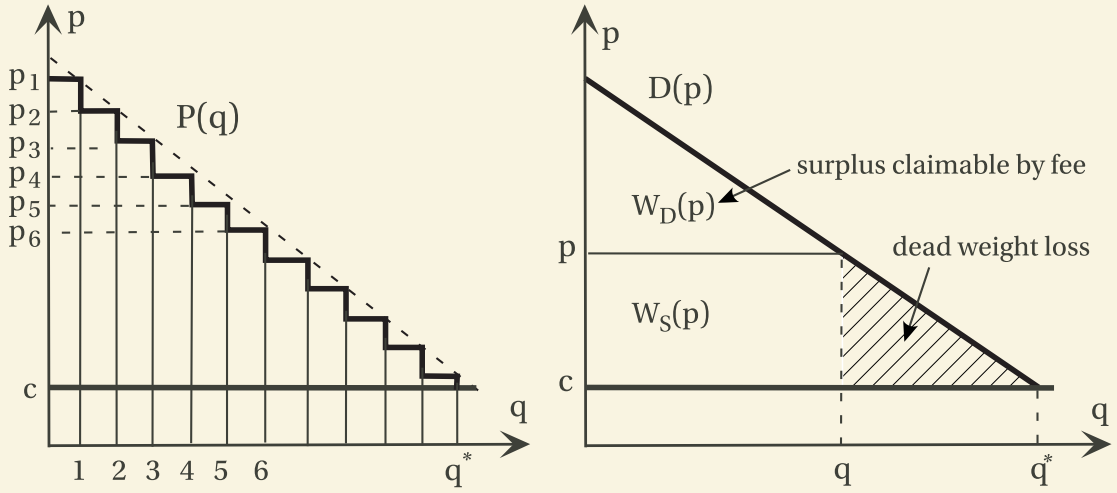


Figure 4.1: Quantity Rebates and Club Pricing

the derived consumer surplus $W_D(p)$. The per-capita profit is then $\pi(f,p) = f+(p-c)D(p) = W_D(p) + pD(p) = W(p)$ by construction. As can be checked on Figure 4.1, meanwhile the price is greater than the marginal cost, there is an inefficiency generating a deadweight loss, which in this particular instance is also a profit loss. Formally, we have $\frac{\partial \pi}{\partial p} = p - c$. The optimal *two-part tariff* is thus to set p equal to the marginal cost c and f equal to the corresponding consumer surplus $W_D(c) = W^*$, which in this case is the maximum market welfare.

The two-part tariff is commonly used by sport clubs or theme parks for kids; a rather expensive annual fee or daily entrance fee is charged but almost nothing is charged for the actual consumption of facilities or attractions. This is because these activities support a high fixed cost but a low marginal cost. Another instance where two-part pricing appears is for bundled goods involving a durable component and a disposable one like {camera + films} or {water filtering jar + mineral recharges} or {razor + blades}; the user must purchase one apparel to enjoy the service but can then freely decide how many disposable units to buy. We develop the analysis of bundling in §5.3.3 and §24.1.

Lastly, most regulated services like telephone, gas, electricity, water are also priced using a two-part tariff. Coase (1946) advocates for the righteousness of this scheme as opposed to marginal cost pricing and public subsidies. The reasons advanced are the same as above, namely high fixed cost and low marginal cost; they and are further explored in §17 on regulation.

Block Tariff An alternative pricing scheme also popular with public services is the *block tariff* consisting of one expensive price p_1 for the units within the block of size q_1 and a cheaper follow-on price p_2. As shown by Gabor (1955), this method enables to

extract the entire consumer surplus.

Indeed, the final consumption elicited by the consumer will either equate his WTP P with p_1 or p_2. If the former case occurs, then we are back to linear pricing which is incapable of extracting all the consumer surplus, thus it must be the case that the consumer demands $D(p_2)$. In order to extract W^*, it is necessary to generate this maximum market welfare, thus necessary to induce efficient sales of q^* which occurs only if $p_2 = c$, the marginal cost of production (recall $D(c) = q^*$). The total revenue under the block tariff is then $p_1 q_1 + c(q^* - q_1)$ while total cost is cq^*, the profit is therefore $\pi = (p_1 - c)q_1$. To equate it with the maximum market welfare W^* there are many solutions (we have a degree of freedom); we can pick some block size q_1 and then set $p_1 = c + \frac{W^*}{q_1}$. If the consumer buys q_1 units at price p_1 plus $q^* - q_1$ units at marginal cost then the first block sales are $cq_1 + W^*$ while the second block ones are $c(q^* - q_1)$, so that total sales are $W^* + cq^*$ leading to the desired profit of W^*.

Since the efficient quantity is q^*, the gross surplus of the consumer is $W^* + cq^*$ and by the very choice of the block tariff parameters, he pays exactly that amount, thus he will accept the block tariff offer and the firm will successfully rip the maximum welfare. The last possibility that must be set aside is that the consumer buys less than q_1 units at the high price p_1 which occurs if $P(0) > p_1 = c + \frac{W^*}{q_1}$. Since $W^* = \frac{1}{2} q^* (P(0) - c)$, the condition we need is $q_1 \leq q^*/2$. Notice finally that choosing $q_1 = 1$ is nothing but a two-part tariff.

The practical advantage of the block pricing method for regulated utilities with respect to a two-part tariff is to accompany the cycles of demand (cf. §25.3 on peak-load pricing). Indeed, during the off-peak period, the client must pay the potentially large subscription of a two-part tariff even though he consumes few units and derives a small surplus; on the other hand, a block scheme enables him to buy the small number of units he needs, at the high price p_1.

“Take-it-or-leave-it” aka. Minimum Purchase Obligation

The third strategy to extract consumer surplus is to bundle price and quantity and make a “take-it-or-leave-it” offer to the consumer.[9@] The fact that a quantity q is offered for a total price (transfer) t means that the consumer has to purchase a minimum number of units. Given our previous derivation of the optimal two-part tariff $(p = c, f = W^*)$, the optimal “take-it-or-leave-it” offer is simply to bundle the efficient quantity q^* with a transfer equal to the consumer’s utility (raw surplus) from consuming q^* i.e., set $t^* = u(q^*) = cq^* + W^*$ (this measure of utility u is defined in eq. (2.17)). Notice that t^* is the total expenditure of the consumer under the optimal price listing and under the optimal two-part tariff. Once again, the optimal bundle leaves the consumer with a zero net surplus while the firm’s profit reaches the maximum welfare.

An equivalent contract is to offer a unit price of $\frac{t}{q}$ bundled with a minimum purchase obligation q, a scheme quite popular with mobile phone operators and manufacturers in the retail sector. Upon having accepted the offer $\left(\frac{t^*}{q^*}, q^*\right)$, the client does not wish to consume more units since her WTP at q^* is exactly c while the unit price is $\frac{t^*}{q^*} = c + \frac{W^*}{q^*}$. Given that her expenditure is $q^* \frac{t^*}{q^*} = t^*$, she will accept the offer.

Historically, Leontief (1946)'s departure point for the "take-it-or-leave-it" method was the comparison of the traditional wage negotiation where only the rate is agreed leaving the firm free to decide later on employment and a more innovative negotiation where a guaranteed wage is accorded forcing the firm to provide a minimum level of employment. He simply observed that the former was akin to a standard monopoly, thus was inefficient while the later was akin to a perfectly discriminating monopoly and demonstrated that the optimal bundle would be (Pareto) efficient.[10@]

4.2 Direct Differential Pricing

In Pigou (1920)'s language, this section addresses first and third degree price discrimination which we deem perfect and imperfect separation of consumers within a market segment. The former case is rarely observed and serves as a benchmark for the latter.

4.2.1 Perfect Discrimination

It is obvious that if a monopoly is legally entitled to discriminate among his clients, holds all the necessary information on their preferences and can avoid arbitrage, he can rip the entire consumer surplus of each client using any of the previous three instruments.[11@] The maximum welfare W^* will be generated but contrarily to the perfect competition case, it will go entirely to the monopoly leaving no (net) surplus to consumers.

Although not the most frequent form of price discrimination, there are real life situations that approach this extreme case; for instance, when an expert seller bargains with a novice buyer over the price of a durable good such as a painting or a car, he can guess his WTP and offers him a rebate on an extremely high starting price that leaves zero net surplus to the buyer. Odlyzko (2004a) explains how the internet and telecommunication (IT) revolution has made perfect price discrimination a near reality. Firstly, firms are able to guess our current WTP using our recorded past purchases and the information we provide them when enrolling in fidelity programs. The *yield management* techniques initially developed by airlines and hotel chains are becoming extremely precise (cf. Netessine and Shumsky (2002) for an introduction). Secondly, IT technologies enable firm to dynamically change their prices at virtually no cost. The limit to this

practice is the annoyance felt by consumers in the face of ever changing tariffs.

The positive side of perfect discrimination wrt. the basic (non discriminating) monopoly is the increase of consumption as it allows some intermediate income households to consume the good or service. Yet, the consumer surplus being nil, the law often forbids (perfect) discrimination to avoid such an extreme sharing of the welfare between producers (here the monopoly) and consumers. Notice also that perfect discrimination leads the monopoly to take efficient decisions in every respect but price. The costly quality of the product, its degree of complementarity with other goods will all be efficiently chosen because perfect discrimination makes the monopoly profit function equal to the objective function of a benevolent social planner.

Let us stress the fact that the full extraction of consumer surplus is feasible only if the monopoly can guess the demand curve of each potential customer and is allowed to make personalized offers in either of the 3 above forms. It is also patent that our demonstration applies to any cost function and to any good or service for which the information and legality constraints are satisfied.

4.2.2 Imperfect Discrimination aka "Segmentation"

Differential pricing as understood by the general public is direct but imperfect price discrimination. We give motivate this practice with historical examples and then derive the optimal prices for each segment and the welfare consequences. We shall also see how the tools of consumer surplus extraction (complex pricing) can be used in the presence of heterogeneity within a segment. Because they provide a further advantage to the firm, these tools are sometimes forbidden, thus forcing the firm to use uniform pricing. Notice that when consumers only need a single unit, the price is by force uniform. This is the case for durable goods like cars or domestic appliances.

Motivation

As recalled in Odlyzko (2004b), differential pricing has been used for centuries in Europe and China in the area of transportation; for instance sea shipping, use of rivers, canals, turnpikes and then railways. Tolls depended on the type of vessel using the facility, the size of the vessel or the merchandise on board. The main defect for users was not so much the level of tariffs but their intricacy. The general rule that slowly emerged in successful ventures was to charge at "what the traffic would bear".

Dupuit (1844) offers an early formal analysis of differential pricing as a mean to improve the profitability of bridges in France. His idea is to drop the single toll and establish instead a discriminatory tariff based either on clothing of the user or on the

time of passage. In the first case, workers pay a low toll provided they wear a cap or any other cloth typical of their statute while other (richer) users pay a higher toll.[12@] As a result of this scheme, the traffic on the bridge is increased as well as the total receipts; we can conclude that a more efficient situation has been reached because new users do no have to take a long detour, old users still pay the same price and owners now make profit from their bridge instead of incurring losses (under the low pricing scheme).

Most often, market segments are exogenously determined by a criteria such as gender, age or geographical residence but as we shall show later on firms prefer to endogenously define segments along the WTP criteria, so as to come closer to perfect discrimination.

Exogenous Segments

For some reasons, the market can be divided into segments upon which it is both possible and legit to apply differentiated prices. We thus treat each segment as a separate market assuming further that no arbitrage among them is possible. The optimal price in each segment therefore satisfies the Lerner formula (3.4) . If there are n segments, we obtain for $i \leq n$

$$\mathscr{L}_i = 1/\epsilon_i \quad \Leftrightarrow \quad \left(1 - \tfrac{1}{\epsilon_i}\right) p_i = C_m(Q) \tag{4.1}$$

where $Q = \sum_{i \leq n}$ is the quantity sold across all segments. This equation means that changes in elasticity are compensated by changes in the optimal price to maintain it equal to the marginal cost. Thus, a segment with a higher elasticity enjoys a lower price because it is more "combative" and forces the monopoly to offer a greater rebate. In the case of two segments, Robinson (1933) speaks of the *strong* and *weak* markets with respectively a higher and lower price wrt. the uniform optimal price (maximizing the sum of market profits).

Equation (4.1) only characterizes differences in the treatment of each market. The vector of profit maximizing quantities $\mathbf{q}^d = (q_i^d)_{i \leq n}$ is found by applying a simple algorithm inspired by the manna trick of §2.1.1.[13@] Assuming to simplify a constant marginal cost c, the monopoly profit is

$$\Pi(\mathbf{q}) = \sum_{i \leq n} q_i \left(P_i(q_i) - c\right) = \sum_{i \leq n} \int_0^{q_i} \left(R_{m,i}(x) - c\right) \mathrm{d}x \tag{4.2}$$

Given the quantities $\mathbf{q} = (q_i)_{i \leq n}$ already sold in each segment (initially zero), the firm ponders producing an additional unit. She computes the revenue $R_{m,i}(q_i)$ she would derive from selling it to segment i and count the cost c as a special zero segment. The allocation rule is then to award the unit to the segment generating highest revenue, as

if each segment was represented by an officer bidding for the additional unit; notice that the unit goes to the zero segment i.e., is not sold, if revenue bids are all below production cost. The analogy with an auction will be used in §22 on auctions.

The analysis can be pursued graphically, considering for instance two markets; it is enough to equate $q_1 + q_2 = R^{-1}_{m,1}(p) + R^{-1}_{m,2}(p)$ with $q = C^{-1}_m(p)$ to obtain the total quantity q^m and the marginal revenue r^m (cf. the right panel of Figure 4.2). Quantities are then given by $q_1 = R^{-1}_{m,1}(r^m)$ and $q_2 = R^{-1}_{m,2}(r^m)$ while prices offered to consumers are $p_1 = P_1(q_1)$ and $p_2 = P_2(q_2)$. Exercise: solve this problem for the linear demands $D^i(p_i) = a_i - b_i p_i$ for $i = 1,2$ and $C_m(q) = cq$.

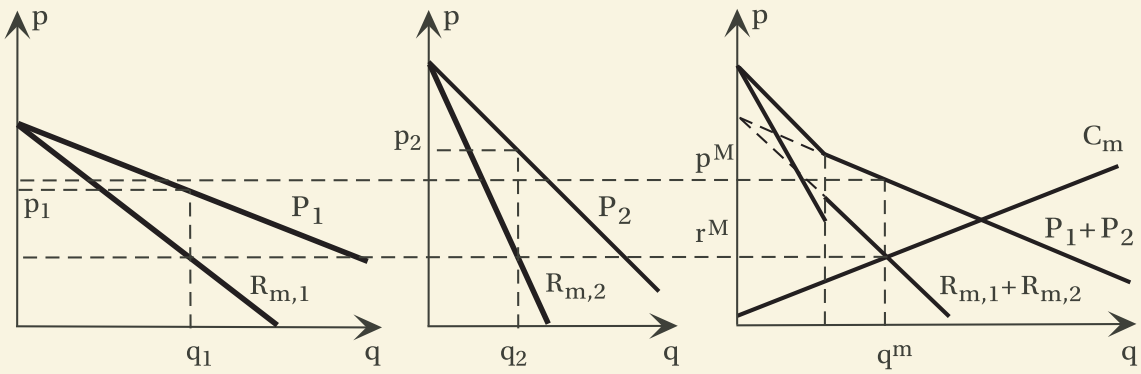


Figure 4.2: Optimal prices in segmented markets

Welfare

Obviously, differential pricing is weakly beneficial for the firm since she has always the option of not using it.[14@] Socially, it will be shown that this practice generates an allocative inefficiency that may be compensated by an increased output (as monopolies price above marginal cost and sell to little).

If demand is linear in all segments with $D_i(p_i) = a_i - b_i p_i$ and marginal cost constant then (4.1) is $c = \frac{a_i - 2q_i}{b_i} \Rightarrow q_i^d = \frac{a_i - b_i c}{2}$ and price $p_i^d = \frac{a+bc}{2}$. Notice now that $q_0^d \equiv \sum_{i \leq n} q_i^d$ is also the optimal quantity under uniform pricing since market demand is $D_0(p) \equiv \sum_{i \leq n} D_i(p) = a_0 - b_0 p$ with $a_0 \equiv \sum_{i \leq n} a_i$ and $b_0 \equiv \sum_{i \leq n} b_i$ (it has the same FOC).[15@] The monopolist thus sells the same total amount whether she price discriminates or not. If she can, she earns more by having strong segments pay more and weak segment pay less. This is inefficient given that opening a second-hand market would allow mutually advantageous re-trade among segments. We thus obtain a necessary (but not sufficient) condition for welfare improvement:

Differential pricing increase welfare only if it increases output.

Schwartz (1990) proves the claim even for decreasing cost, in two steps. Let $q_0^d < q_0^u$ be total output under discriminatory and uniform pricing. We want to show $W^d < W^u$. Given the uniform price $P(q_0^d)$, consumer express demand to maximize utility thus total utility (cf. eq. (2.17)) is greater than if they face the segment prices p_i^d associated with optimal discrimination. As cost only depends on output, welfare, which is utility minus cost, is greater if total output q_0^d is sold uniformly. Next, we observe that at market level, the inequality $q_0^d < q_0^u$ implies $W_D^d < W_D^u$ as consumer surplus rises with output, but also $\pi(q_0^d) < \pi(q_0^u)$ since the latter corresponds to the monopolist's ideal. The sum, which is welfare, must therefore increase.

Applications

Our findings also apply for foreign trade i.e., when a firm from country 2 sells the same good at different prices at home and abroad (country 1). The rationale of this pricing policy is to grasp the heterogeneity of demand at home and abroad. If the home consumers are less reactive to price then $p_2 > p_1$ and the abroad government (country 1) may accuse the firm of dumping (e.g., Japanese cars vs. Europe or USA). German cartels of metals and chemicals used this method in the early XX^{th} century to gain control of foreign markets; they maintained a high domestic price with the help of an import tariff and sold abroad at competitive price, sometimes at a loss.

Another case are DVDs which were clearly priced in a differentiated way between the US and Europe during the 1990s. To avoid parallel imports from the cheap US market to Europe,[16@] a technology had been devised and imbedded into DVD players to force a pre-determined match i.e., a DVD player bought in Europe would only read DVDs bought in Europe (cf. §12.3.4). Because differential pricing also greatly limited the availability of titles in Europe, sales of players were not taking off fast (cf. §24.2.2 on the issue of complementarity) so that retailers started to "hack" DVD players i.e., making them able to play any discs (among which the US ones). This form of arbitrage was so successful that it made its way back to the industrial makers of DVD players who "pre-hacked" the protection system at assembly stage. The machine would not violate the agreement with the DVD forum but the task of the retailer was considerably simplified since it was then enough to open the machine and turn a screw to deactivate the matching technology.

4.2.3 Endogenous Segmentation

If there is a natural segmentation, say between men and women, the firm can take advantage of it to price higher to the group displaying the highest average WTP but the ideal would be to ignore gender and divide the market into weak and strong WTP segments; the differences existing between the original segments would still be reflected into the new segments but in a manner that allows a greater surplus extraction. If, say, women spend twice more on clothing, their discriminatory price is greater than men's but ideally fashion victims are grouped together (may be one third male, two third female) in order to charge them much more. Another example would be air travel for which late buyers, business people, are more eager than early ones, tourists. If it is common knowledge that the price will rise with time, then tourist buy early to enjoy a discount while business people buy late when they learn their exact needs. In absolute terms, this is not direct discrimination given that a tourist can always book a business ticket but the versions are made so as to effectively separate the strong and weak segments and force them into the desired price categories.

Vertical Segmentation

Under endogenous discrimination, profit rise faster (wrt. exogenous discrimination) and welfare is enhanced. We prove this claim within the standard linear model.

Demand $q = a - bp$ is divided into a strong segment $q_1 = a - bp_1$ and a weak segment made of those who enjoy the discount $q_2 = a - q_1 - bp_2$ (obviously, type #1 are barred from enjoying the discount). With marginal cost c, profit is

$$\pi = q_1(p_1 - c) + q_1(p_1 - c) \propto q_1(a - bc - q_1) + q_2(a - bc - q_1 - q_2)$$

thus the two FOCs are $2q_1 = a - bc - q_2$ and $2q_2 = a - bc - q_1$. Their solution is $q_1^d = q_2^d = \frac{a-bc}{3}$ leading to larger sales than under uniform pricing since $q^u = \frac{a-bc}{2}$. Profits are $\pi^u = \frac{(a-bc)^2}{4b}$ and $\pi^d = \frac{(a-bc)^2}{3b}$ while welfare is $W^u = \frac{3(a-bc)^2}{8b}$ and $W^d = \frac{4(a-bc)^2}{9b}$ (recall that the maximum is $\frac{(a-bc)^2}{2b}$). The limit of this process of segmentation by WTP is perfect discrimination which is conducive of full efficiency. The comparison with exogenous discrimination is quite simple. Assume, for instance, the existence of two segments with demands parameters $a_1 = 2a_2 = \frac{2a}{3}$ and $b_1 = b_2 = b/2$, then $p_1^d > p_2^d$ (segment #1 is stronger) and total profit is $\pi = \frac{(2a/3-bc/2)^2}{4b/2} + \frac{(a/3-bc/2)^2}{4b/2} = \pi^u + \frac{a^2}{36b} < \pi^d$.

Innovation

Innovation or the choice of quality is another avenue through which the firm can modify the market perception of differential pricing. We already know that a firm with market power e.g., a monopolist, wishes to to discriminate between a strong and a weak segment. In the standard linear model, we saw in §4.2.2 that total output does not change wrt. uniform pricing which means that the allocation is inefficient.[17@] Now, if we account for the presence of quality and its selection by the firm (to maximize profits), then price discrimination may be welfare improving and even beneficial to both the weak and strong segments. The reason exposed by Ikeda and Toshimitsu (2010) is that price differentiation allows to capture a greater share of the strong segment WTP; the firm's marginal return to quality is thus enhanced and she is lead to choose a higher quality (wrt. uniform pricing). Since this is an efficiency enhancing change, it may compensate for the mis-allocation between strong and weak segments.

To check upon this claim, consider the model of §3.3.1 for two segments with $\alpha_1 = 1$ (strong segment) and $\alpha_2 = \delta \in [0;1]$ (weak segment). If discrimination is impossible, the two segments are pooled together and, thanks to the simplicity of the setting, the new demand is twice that of a segment with parameter $\bar{\alpha} = \frac{2\delta}{1+\delta}$ which is an average between δ and 1. The optimal price is thus $\bar{p} = \frac{\delta s}{1+\delta}$ which is also an average between the prices that are optimal in each segment, $p_2^d = \frac{\delta s}{2}$ and $p_1^d = \frac{s}{2}$. This, in turn, leads to distorted sales $\bar{q}_2 = \frac{\delta}{1+\delta} < q_2^d = \frac{1}{2} = q_1^d < \bar{q}_1 = \frac{1}{1+\delta}$. Lastly, profit is $\bar{\pi} = \frac{\delta s}{1+\delta} < \pi_1^d + \pi_2^d = \frac{(1+\delta)s}{4}$, the discriminatory payoff.[18@] As shown in §4.2.2 and confirmed in this example, total output is the same under uniform and discriminatory pricing (it is unity) but welfare is lower under discrimination.

This analysis is correct for a given quality but once we take into account the ability to adjust quality at the margin, we reach a different conclusion. Indeed, with a cost of producing quality $\frac{s^2}{2}$, the optimal quality level is simply the multiplying factor in the profit i.e., $\frac{\delta}{1+\delta}$ under uniform pricing and the larger $\frac{1+\delta}{4}$ under discriminatory pricing. Taking into account that the net profit in equilibrium is $\frac{s^2}{4}$, choosing a higher quality unmistakably leads to higher profits. We have thus shown that discrimination allows to earn more by investing more into quality. The strongest result of these authors is that price discrimination is also welfare enhancing i.e., the discriminatory firm chooses such a high quality that the welfare gains (for society) from enjoying higher quality outweigh the allocative losses. Under differential pricing, welfare is $\frac{3s}{8}$ in the strong segment and $\frac{3\delta s}{8}$ in the weak one, thus net welfare is $W^d = \frac{3s(1+\delta)}{8} - \frac{s^2}{2}$. Given the optimal quality $s^d = \frac{1+\delta}{4}$, final welfare is $\overline{W}^d = \frac{(1+\delta)^2}{16}$. For the uniform case, net welfare consist of 5 elements, 2 profits, 2 consumers surpluses and cost for quality which simplifies into $W^u = \frac{1+\delta+\delta^2}{2(1+\delta)}s - \frac{s^2}{2}$. As the optimal quality is $\bar{s} = \frac{\delta}{1+\delta}$, we derive $\overline{W}^u = \frac{\delta(1+\delta^2)}{2(1+\delta)^2}$ and it is a

matter of algebra to check that $\overline{W}^d > \overline{W}^u \Leftrightarrow (1-\delta)^4 > 0$. Lastly, it may be verified that both segments benefit from price discrimination.

> For a product whose quality is endogenously chosen by a firm with market power, price discrimination is both privately and socially desirable.

Time Segmentation

Market segmentation opportunities are sometimes determined by external forces such as country borders (and thus different legal systems) or gender but in many instances, firms endogenously fine tune their segments using criteria such as age or time schedule to determine whether the client benefits from a rebate or not. Let us use the happy hour example with T being the time limit where prices return to normal. Afternoon is the weak demand segment (#2) whereas evening is the strong demand segment (#1). We further need to assume that demand at time t is $d(t,p)$ (independent of demand before or after).

When the monopoly moves the threshold T later, he loses $(p_1 - c)d(T,p_1)$ over segment #1 and gains $(p_2 - c)d(T,p_2)$ over segment #2. The optimum is thus found when the opposing forces are equal. Since the demand increases by $d(T,p_2) - d(T,p_1)$ this change is always welfare improving.[19@] We may thus conclude that the monopoly restricts too much the segment with low WTP i.e., happy hours are too short and likewise rebate conditions in transportation are too stringent.

4.2.4 Proviso

In this part, we deal with various reasons why differential pricing may not be used or used unexpectedly.

Positive Discrimination

The law generally forbids discrimination based on human traits but tolerates exceptions that serve the government's objectives in which case we speak of positive discrimination.[20@] Redistribution of income is one important motive for permitting discrimination. For instance, young and old people benefit from preferential tariffs for many services that are under public regulation (e.g., transport) because it is a mean for the government to redistribute income within the population. As an extension, private firms are authorized to discriminate young and old people, as long as, they get better conditions than regular clients (e.g., cheaper tickets in movies theaters). Another curious case of positive gender discrimination is the free entrance for women in night clubs; men welcome this practice

since the extra price they pay is compensated by a greater feminine attendance which after all is the reason why they patronize those clubs.

In the case of insurance, discrimination based on a variety of human traits is occasionally permitted on grounds of fairness because it enables to adjust the individual premium to the expected cost of the individual; this way no group is forced to subsidize other groups. Taxi fares are regulated but in recognition of the painfulness of night driving, they are allowed to be more expensive during the night.

Arbitrage

Arbitrage is a complex reaction to discrimination attempts and is inter-wined with information and legality.

A first strategy to avoid discrimination is to disguise oneself or mask one's intentions by adopting the iron mask of poker players; for instance, in a flea market, the experienced seller is able guess your WTP from your cloths and attitude, thus you better dress casually and pretend to be vaguely interested if you want to make a good deal. For the same reason, some bidders in art auctions prefer to participate by phone rather than appearing in person in the auction hall. Secondly, consumers might try to defraud the scheme by pretending to be eligible for a better price i.e., your partner buys a software suite for home use and then lets you use it for your business or you buy a one-license software and install it on 20 computers. We won't pursue this possibility since economics always assume that laws are exactly enforced (unless we study this specific problem).

For most services, arbitrage is easy to avoid because the contractual relationship is between the firm and a particular consumer, so that the latter cannot resale the service because he is not the owner of the service, he is just entitled to receive the service (e.g., plane ticket). However, for expensive items like a yearly subscription to opera, transferability is often permitted to enable the customer to share the cost with friends and family.

Arbitrage is a reality for goods for the simple reason that the transaction involves a transfer of ownership. The means to circumscribe arbitrage on goods are:

- *Trademarks*: a legal limitation to resale studied in §12.3.3.
- *Warranties*: offered exclusively to the direct buyer of the (durable) good; they decrease the value of resold items.
- *Security Concerns*: authorities limit the number of dealers who can lawfully sale and resale the item but oblige them to guarantee reliability and security of the product (cf. §9.1.2).

One important practical case is the European car market; it used to be quite complex to buy a car in a European country and import it into another one because the vehicle had to pass security inspections. Unless the price difference was very high, it was not worthwhile to spend time and money to perform the arbitrage. With the advent of the internal market and the progressive implementation of the new regulation of 2002 for the car sector, these limitations have disappeared. Nowadays, there are intermediaries who buy new cars in countries where manufacturers prices are cheap in order to resale them in countries where manufacturers prices are dear.

What this example highlights is the existence of a transaction cost associated to the resale operation. Prohibition amounts to an infinite transaction cost while in the case of perfect arbitrage the cost is zero. The transaction cost t is thus protecting the monopoly in its intent to price discriminate among categories of clients. Returning to our analytic model and considering two segments, if $t > p_2 - p_1$, then the arbitrage poses no threat to the monopoly. Otherwise the monopoly cannot freely discriminate because arbitrage would drive the high price down to $p_1 + t$; the optimum is then to set a higher bottom price p_1 and a lower top price $p_2 = p_1 + t$.

Seasonality potentially enables firms to directly discriminate but arbitrage is also present although less active when it comes to services. In many retail sectors, sales on Saturday are notoriously larger than those of the remaining days of the week, thus firms ought to raise their price on Saturdays. One first obvious obstacle is the cost of changing price tags twice a week but another much more compelling is that many goods that can be cheaply stored (e.g., books using shelves or fresh food using a refrigerator) so that *arbitrage* is a meaningful threat at the disposal of consumers to force outlets to maintain constant prices. For a larger time span, arbitrage becomes expensive so that shops can take advantage of the demand peak; this is why prices for luxury items increase one month before Christmas and decrease after new-year's eve.

Regarding services such as travel, consumers can arbitrate and wait the off-peak period to enjoy lower prices. Yet, they suffer an opportunity cost of switching their consumption to the off-peak period (e.g., they must free paid working time to go shopping during week days or free leisure time to go wash the car during the week-end), thus the firm can maintain a price difference. Even so, congestion will frequently appear during the peak period (capacity is exhausted) which means that the pricing of these services goes beyond simple revenue maximization and spills over cost savings and investment policy; this issue is tackled in §25.3 on peak-load pricing.

Complex vs. Uniform pricing

For most services, it is technically possible to record detailed consumption and charge the client according to it. Yet this option is not as frequently used as theory would predict, this because of several drawbacks. Firstly, it requires an extra fixed cost in the form of labour and capital investment to perform metering, billing and monitoring (avoid free riding and defrauding). Secondly, per-unit pricing, by disrupting service provision, tends to lower quality and thus reduce the global WTP for the service. Thirdly, the previous annoyance cannot be eliminated through ex-ante contracting because ex-post contingencies (i) force parties to fine tune the characteristics of the service in real-time and/or (ii) change the value of the service in real-time, making the determination of price difficult. In both cases, haggling and inefficient opportunistic behavior is bound to appear (cf. §14.2 on hold-up). If these costs exceed the extra profit afforded to the firm by differential pricing then only access charging will be observed.[21@]

Another compelling reason why many services only involve access or membership charging is because the optimal pricing scheme exhausts the gains of trade with the client and this amounts to price at marginal cost which is often zero. This claim is well known for the monopoly case (cf. §4.1.3) and since it does not depends on the existence of competition nor on its degree, it readily extends to other market structures. Each client is then characterized by his WTP for the global service (i.e., access and utility maximizing consumption) and it is on this dimension alone that firms compete.

4.3 Indirect Differential Pricing

After presenting some examples of indirect discrimination based on quantity rebates, we proceed to study the design of an optimal scheme of indirect discrimination in two steps. Considering a market made of a weak and a strong segments, we first look at a "no-discrimination" case i.e., the firm offers a *unique* tariff to his whole clientele. In a second step, we show how the addition of a second tariff can improve profits by leading heterogeneous consumers to sort themselves out. Whether the menu of tariffs is offered depends then on transaction cost i.e., whether it is not too costly to create, manage and control its correct implementation, including the backlash of angry consumers who have a tendency to hate blatant discrimination.

As for the pricing tools used by the firm, recall that in a competitive market the consumer is faced with the linear (aka uniform) tariff $T(q) = pq$ where p is set by the market while in a monopolized market, the firm has the ability to go beyond uniform pricing and offer any potentially complex quantity-price relationship (aka non linear

pricing). The latter will often include a subscription, something entirely absent from a competitive market for "free entry" drives it to zero.

4.3.1 Quantity Rebates

Quantity rebates are present everywhere around us; for instance, the price per kilo of many food items diminishes with the size of the container (much more than what the savings on the container would permit) as the example of Table 4.7 shows (prices quoted in November 2005 for exactly the same wine). Some people then buy the compact size, others prefer the regular size and the remaining prefer the family size; in the end everyone has picked the version best suited to him/her in terms of size and price (cf. also Table 4.1).

Bottle size (liter)	1/5	3/8	1/2	3/4
Price (€)	1,5	2,25	2,35	2,65
Price / liter	7,5	6	4,7	3,5
Mg. Price	7,5	4,3	0,8	1,2

Table 4.7: Retail Prices for Wine

Another popular example of quantity rebate is the "buy-one-get-one-free" (BOGO) promotion at your local pizzeria for take away (instead of home delivery). Suppose you'd pay up to 14€ for a first pizza but only 7€ for a second one, then the BOGO package at 18€ will appeal to you. Selling pizzas at the same average price of 9€/piece would trigger only one sale, thereby halving the profit (assuming negligible production cost).[22@]

The BOGO also acts an indirect discrimination scheme because high WTP clients who bother walking to the pizzeria pay the regular price of 18€/piece to get home delivery; consumers sort themselves out in two distinct segments paying different prices for the same basic good, pizza (cf. §4.3.1). Lastly, the BOGO acts an intertemporal discrimination scheme because it is only active from Monday till Thursday when business is low (off-peak period). During the week-end when customers display a higher WTP (this defines the peak period), the restaurant has the opportunity to sell pizzas at the regular price of 18€/piece without generating any hard feeling (cf. §25.3 on peak-load pricing).

Mobile phone price plans like many services use two-part tariffs and can be designed to incorporate quantity rebates i.e., the average price is decreasing with the contracted monthly number of minutes. Table 4.8 show some data from a carrier of a large European country. Thanks to the clever design of these tariffs, an *indirect* discrimination occurs because each person picks the proposal best suited for him/her and ends up paying and consuming varied amounts. Clients have sorted themselves out for the benefit of the firm.

Monthly Fee (€)	18	28	36	43	61	77	90	126
Allowance (hours)	1	2	3	4	6	8	10	15
Av. Price (€ct./mi.)	30	23	20	18	17	16	15	14
Mg. Price (€ct./mi.)	30	17	13	12	15	13	11	12

Table 4.8: Mobile phone price plans

The tariffs presented in Table 4.8 form a menu, the design of which is extremely complex. A first step to understand the characteristics of the optimal menu is to look for the optimal single tariff.

4.3.2 Optimal Single Tariff

In this paragraph, the firm voluntarily limits herself to propose a unique tariff to both segments, for instance because direct discrimination is either prohibited or too costly to implement. The firm must arbitrate between market coverage and mark-up. Indeed, she may cater to all segments with a low subscription fee, an inclusive strategy, or target the strong segment to extract more consumer surplus from it with a high fee, an exclusive strategy. Obviously the optimal strategy depends on the respective sizes of the two segments. Lastly, a "middle of the road" solution that mixes the two previous approaches (and improve upon them) is the three-part tariff consisting in a subscription, a free allotment of units and a per-unit price for additional units.

In the more general case where the market is populated by a continuum of heterogenous clients, ranked by WTP from strongest to weakest, the optimal strategy is to set an exclusion level through the access fee and then apply an aftermarket mark-up (i.e., price above marginal cost) in order to extract surplus from strong clients. One then speaks of price discrimination by "metering" since clients are sorted out by how much they consume in the aftermarket (cf. Blackstone (1975) for an historical case).

Inclusive Tariff

Oi (1971) derives the optimal inclusive (in) two-part tariff when the monopoly faces strong and weak segments such as poor and rich households or home and professional users but is not able to discriminate directly among them. With the help of the left panel on Figure 4.3, we shall demonstrate that pricing at marginal cost is not optimal anymore once there is some heterogeneity in the demand (in contrast with the individual surplus extraction characterized in §4.1.3).

The demand functions of a weak and strong consumer satisfy $D_1 < D_2$. Starting with pricing at marginal cost c, the demands are q_1^* and q_2^* generating maximum consumer

surpluses $W^*_{d,1}$ and $W^*_{d,2}$. To attract both consumers, the monopoly can set the common subscription at $F = W^*_{d,1} < W^*_{d,2}$. Consider now raising the unit price to $p > c$; this generates a sales profit from consumer #1 measured by area G on Figure 4.3 but also a slightly bigger reduction $G+\alpha$ of his surplus (α, like β, is a deadweight loss). In order to keep this client, the monopoly must reduce his subscription by exactly the same amount i.e., $\Delta F = -G-\alpha$. The operation therefore involves a small loss, area A, over consumer #1. Regarding consumer #2, the subscription loss is $G+\alpha$ while the gain from additional profitable sales is $G+\alpha+H$ as seen on the left panel of Figure 4.3; overall, there a benefit H from raising the price above marginal cost (assuming equal size populations for both types).

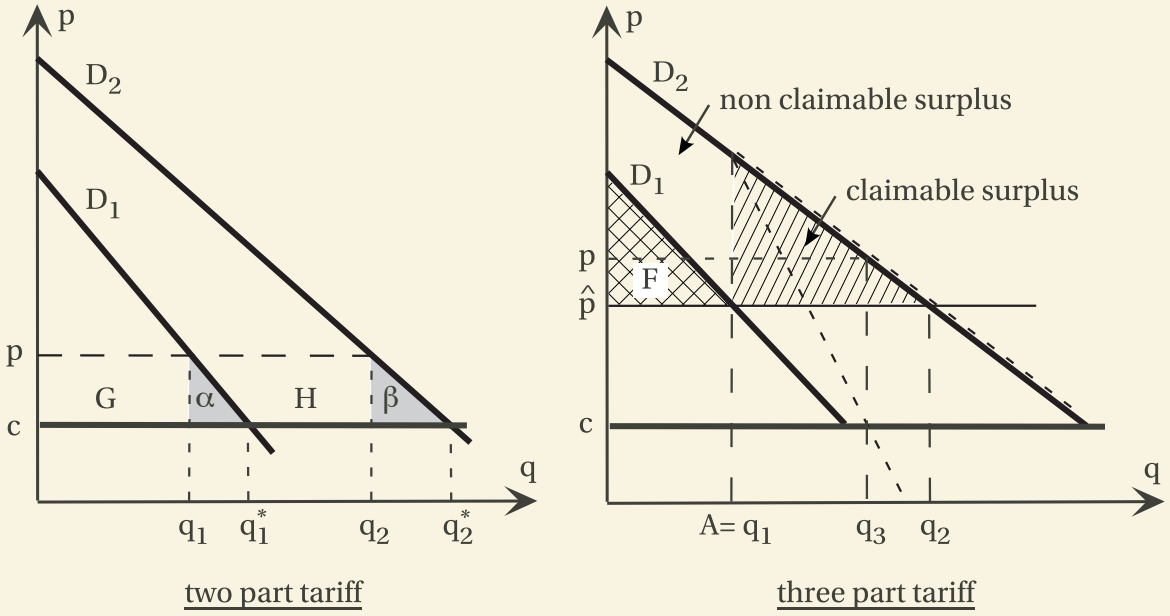


Figure 4.3: Two and Three Part Tariffs

To derive an exact formula, we use the individual demands $D_1(p) = a_1 - b_1 p$, $D_2(p) = a_2 - b_2 p$ and denote $\mu \in]0;1[$ the proportion of people with low WTP in the overall population. We use $\hat{}$ to denote the average of a variable or parameter, hence demand is $\hat{D}(p) = \hat{a} - \hat{b}p$. Within an inclusive strategy, the monopoly optimally sets $F = W_{d,1}(p)$ to attract both types of clients, hence profit is $\Pi^{in}(p) = W_{d,1}(p) + (p-c)\hat{D}(p)$. Since $\frac{\partial \Pi^{in}}{\partial p} = (p-c)\hat{D}'(p) + \hat{D}(p) - D_1(p)$, we have $\left.\frac{\partial \Pi^{in}}{\partial p}\right|_{p=c} > 0$ given that average demand exceeds that of the weakest segment. This is the analytical equivalent to the graphical demonstration given above that the inclusive monopolist inefficiently price above marginal cost to take advantage of the strong segment (cf. also generalization below). The optimal policy solves $0 = \frac{\partial \Pi^{in}}{\partial p} \Leftrightarrow (p-c)\hat{b} = \hat{D}(p) - D_1(p) \Rightarrow p^{in} = \frac{\hat{a}+\hat{b}c-a_1}{2\hat{b}-b_1}$. It is a matter of exercise to check that the basic monopoly price corresponding to free subscription and consumption by

both groups is $p^M = \frac{\hat{a}+\hat{b}c}{2\hat{b}} > p^{in}$.[23@]

Exclusionary Tariff

The exclusionary tariff (ex) that deliberately excludes the weak segment (#1) from consuming is found by setting $F^{ex} = W_{d,2}(p)$ and $p^{ex} = c$ to extract the whole consumer surplus of people within the strong segment (#2). Comparing the two policies, inclusive and exclusive is quite simple since the profits can be decomposed as follows: $\Pi^{in} = \alpha\Pi_1^{in} + (1-\alpha)\Pi_2^{in}$ and $\Pi^{ex} = \alpha\Pi_1^{ex} + (1-\alpha)\Pi_2^{ex}$ with $\Pi_1^{ex} = 0$ given that segment 1 is excluded but $\Pi_2^{ex} > \Pi_2^{in}$ since the profit made on segment 2 is maximized. It is therefore obvious that keeping all consumers as clients is worthwhile only if their proportion in the population α is large enough.

In real situations, there are many categories of clients so that firms rationally exclude the classes with a low valuation for the service (often poor households) in order to extract surplus from more profitable clients. Typically, large industrial buyers are favored by such pricing policies. This is why quantity rebates are sometimes forbidden by authorities to protect small buyers. The same rule applies for monopsonies, as in the case of the supermarket chains who devise such two-part tariffs to extract the surplus of food producers.

Three Part Tariff

Oi (1971) further rationalizes the use of *three part tariffs* consisting of an upfront payment (transfer) t, a free allotment of time $\bar{q}$ and a price p for additional units. These schemes were popular in the lease of mainframe computing power in the 1960s (cf. Waldman (1997) and §24.1) and are now widely used by mobile phone carriers. The data on mobile phone plans given in Table 4.8 also include a price $p = 0.45$ € for extra minutes, which is far above the maximum average price $t/\bar{q}$ of any plan.

We plot on the right panel of Figure 4.3 the two individual demands we previously considered; imagine that the optimal inclusive two-part tariff features the price $\hat{p}$ leading to purchases q_1 and q_2 by the two types of consumers. The fee is $F = W_{d,1}(\hat{p})$ shown by the grid area under D_1. The monopoly can still improve on this scheme by offering his clients a three-part tariff where the allotment is $A = q_1$ and the fee is $t = F + A\hat{p}$. Whatever the choice of the unit price $p \geq \hat{p}$ for extra units (above A), consumer #1 is indifferent among the two plans because he won't consume extra units with the three-part tariff. By construction of this three-part tariff, the monopoly makes the same profit over this client. If consumer #2 agrees to the tariff change, he will consume the allotment and express a residual demand shown by the dotted line parallel to D_2 on the right panel of

Figure 4.3. This means that some of his net consumer surplus under the two-part tariff can be reclaimed through standard monopoly pricing: we equate the residual marginal revenue with marginal cost to obtain an optimal quantity $q_3 < q_2$ and the corresponding price p. As a consequence, client #2 reduces his consumption and the monopoly succeeds to increase profits by recapturing part of consumer #2's surplus (the stripped area).[24@]

Generalization

If, following Littlechild (1975), we allow for more classes of consumers (in- stead of two), it is easiest to consider a continuum indexed by a characteristic $\theta \in [0;1]$ running from most to least interested by the item for sale i.e., the WTP $P_\theta(q)$ of a consumer with characteristic θ satisfies $\frac{\partial P_\theta}{\partial \theta} < 0$. When faced with the unit price p, user θ demands $D_\theta(p)$ solving $P_\theta(q) = p$ and derives consumer surplus $W_{D,\theta}(p)$. It is easy to check that both functions are decreasing with the characteristic θ (cf. §2.16). Hence, if we let μ be the marginal user for which $W_{D,\mu}(p) = f$, then the firm's clients are all the consumers with a characteristic $\theta \leq \mu$ so that choosing f is equivalent to choosing μ. This setting allows to introduce the fixed cost γ of serving a customer. The per-capita profit with type $\theta \leq \mu$ is $\pi_\theta(\mu, p) = W_{D,\mu}(p) + (p-c)D_\theta(p) - \gamma$ whereas welfare is the larger $W_\theta(p) = W_{D,\theta}(p) + (p-c)D_\theta(p) - \gamma$. Letting H denote the law of population distribution among characteristics and $\bar{D}_\mu(p) = \frac{1}{H(\mu)} \int_0^\mu D_\theta(p)\,\mathrm{d}H(\theta) = E[D_\theta | \theta \leq \mu]$ being the average demand of consumers with characteristics lesser or equal to μ, profit reads

$$\pi(\mu, p) = \int_0^\mu \pi_\theta(\mu, p)\,\mathrm{d}H(\theta) = \left(W_{D,\mu}(p) + (p-c)\bar{D}_\mu(p) - \gamma\right) H(\mu) \tag{4.3}$$

while welfare is

$$W(\mu, p) = \int_0^\mu W_\theta(p)\,\mathrm{d}H(\theta) \tag{4.4}$$

Since the consumer surplus satisfies $\frac{\partial W_{D,\theta}(p)}{\partial p} = -D_\theta(p)$ (cf. §2.16), the FOC on price for (4.4) characterizing an efficient allocation is $0 = \frac{\partial W}{\partial p} = \int_0^\mu \frac{\partial W_\theta}{\partial p}\,\mathrm{d}H(\theta) = (p-c)\int_0^\mu D'_\theta(p)\,\mathrm{d}H(\theta) \Rightarrow p = c$, as expected. The situation is different for the monopolist since the FOC on price for (4.3) is $D_\mu(p) - \bar{D}_\mu(p) = (p-c)\bar{D}'_\mu(p)$. Introducing the (average) demand elasticity $\epsilon = \left|\frac{p\bar{D}'_\mu(p)}{\bar{D}_\mu(p)}\right|$, we get

$$\frac{p-c}{p}\epsilon\bar{D}_\mu(p) = \bar{D}_\mu(p) - D_\mu(p) \Leftrightarrow \mathscr{L} = \frac{p-c}{p} = \frac{1}{\epsilon}\left(1 - \frac{D_\mu(p)}{\bar{D}_\mu(p)}\right) \tag{4.5}$$

i.e., the standard Lerner formula (cf. eq. 3.4) except for the presence of the ratio of marginal to average demand which tends to weakens the firm's market power.

If consumers are identical then the margin and the average are identical so that $\mathscr{L}$ is nil which means that profit arises from the entrance fee since the price is optimally set at the marginal cost. If instead consumers are heterogeneous and are sorted by WTP, the marginal one demands less than the average one so that price is inefficiently (socially speaking) greater than marginal cost, yet set optimally (from the point of view of the private firm). The model can be extended to any type of heterogeneity and yield the same formula, only that the relationship between average and margin can go either way.

Overall, we may conclude that the monopoly excludes too many people and charge too much to its clients. Indeed, since $W_\theta - \pi_\theta = W_{D,\theta}(p) - W_{D,\mu}(p) \geq 0$ and the monopoly optimum satisfies $\frac{\partial \pi}{\partial p} = 0 = \frac{\partial \pi}{\partial \mu}$, we have $\frac{\partial W}{\partial p} = \frac{\partial W_D}{\partial p} \propto D_\mu(p) - \bar{D}_\mu(p) < 0$ and $\frac{\partial W}{\partial \mu} = -\frac{\partial W_{D,\mu}}{\partial \mu} H(\mu) > 0$.

Lastly, we characterize the privately and socially optimal clientele size wrt. the fundamentals of the market. The FOC for (4.4) characterizing an efficient allocation is $0 = \frac{\partial W}{\partial \mu} \Rightarrow W_{D,\mu}(c) = \gamma$ i.e., the market is thus efficiently served up to the person whose maximum consumer surplus is equal to the cost of service. The size FOC for (4.3) is expressed in terms of the access fee $f = W_{D,\mu}$ and is

$$\pi_\mu(\mu, p) h(\mu) = -\frac{\partial f}{\partial \mu} H(\mu) \Leftrightarrow f - \gamma = \frac{H(\mu)}{h(\mu)} \frac{f \bar{\epsilon}_\mu}{\mu} - (p - c) D_\mu(p) \tag{4.6}$$

which is negative if the profit to be made in the aftermarket is large and average elasticity of demand to income $\bar{\epsilon}_\mu$ is low.[25@] This happens if there is a large heterogeneity among clients regarding service use but little regarding subscription i.e., everybody wants to enjoy the good but at varying levels of intensity. If so, the monopolist offers a "bargain-then-rip-off" scheme as we have $f < \gamma$ and $p \gg c$.

4.3.3 Optimal Quantity Discrimination aka Non Linear Pricing

The issues dealt with in this section belong to the more general theory of screening developed in §21.2.

Self-Selection

As we saw with mobile phone price plans, firms devise not one but many complex tariffs to cater to the heterogeneity of their clients. Their idea is to propose tariffs called green and red to both professional and home users. If everybody picks the green tariff, it is better to throw out the red one rather than annoying people with it; in that case, we say that the firm *pools* clients because nothing distinguish one type from the other. If, on the other hand, professionals choose the red tariff while home users pick the green one,

then we might as well rename the tariffs “pro” and “home”; in that case we say that the firm *separates* clients as they identify themselves when picking an offer among the two proposals.[26@]

The optimal policy of the firm involves tree steps; firstly, design the best pooling tariff which was done in the previous part; secondly, design the best pair of separating tariffs which shall be done right away; lastly, compare the two strategies and select the optimal one in terms of final profits. This is a complex task which we will only touch; an early reference is Maskin and Riley (1984) while the classical treatment is Wilson (1993).

When designing the “pro” and “home” tariffs, the firm must always take care that professionals prefer the “pro” over the “home” proposal while home users hold the reverse preference. We now see clearly that indirect discrimination is more stringent than direct discrimination because under the latter framework, the firm can *force* professionals to buy the “pro” tariff (or not consume) and *force* home users to buy the “home” tariff (or not consume). The firm has therefore more freedom to design tariffs under direct discrimination. We shall see that indirect discrimination compels the firm to leave an information rent to customers with a high WTP.

Graphical Analysis

The individual demands are the low one D_l from home users and the larger one D_h from professionals. The optimal discriminating two-part tariffs is to charge units at marginal cost c and ask for a subscription $F_i^* = W_{D,i}(q_i^*)$ equal to the consumer surplus corresponding to the efficient demand $q_i^* = D_i(c)$ of group $i = l, h$. This pair of tariffs is illustrated on the left panel of Figure 4.4 where F_l^* is the area of the small triangle while $F_h^* = F_l^* + A_h^* + B_h^*$. An equivalent method is to offer the optimal discriminating bundles (q_i^*, t_i^*) for $i = l, h$ with $t_i^* = q_i^* c + F_i^*$.

If the monopoly cannot force consumer i to buy the two-part tariff (c, F_i^*) but is obliged to propose both tariffs to both consumers then all will go for (c, F_l^*) since the unit price is the same while the subscription is cheaper.[27@] The optimal discriminating bundles also generates the same ordering i.e., everybody prefers (q_l^*, t_l^*) over (q_h^*, t_h^*) so that the monopoly earns F_l^* with both consumers. Yet bundles and two-part tariffs are different from the point of view of a type h consumer; indeed, when picking (c, F_l^*) over (c, F_h^*), he derives a net surplus of $A_h^* + B_h^*$ since he consumes q_h^* units. Now, when picking bundle (q_l^*, t_l^*) over (q_h^*, t_h^*), consumer h derives the smaller net surplus A_h^* because his consumption has been reduced.

This observation indicates a first improvement for the monopoly: offer bundles (q_l^*, t_l^*) and (q_h^*, t_h) with $t_h = t_h^* - A_h^*$ to convince consumer h to accept the second one; indeed, it leaves him a net surplus (also called a rent) of exactly A_h^*. The profit over consumer

h therefore pass from F_l^* to $F_h^* - A_h^* = F_l^* + B_h^*$ i.e., part of the lost surplus has been reclaimed (cf. left panel of Figure 4.4).[28@] To reclaim further surplus from consumer h, it is necessary to scrutinize the optimal quantities.

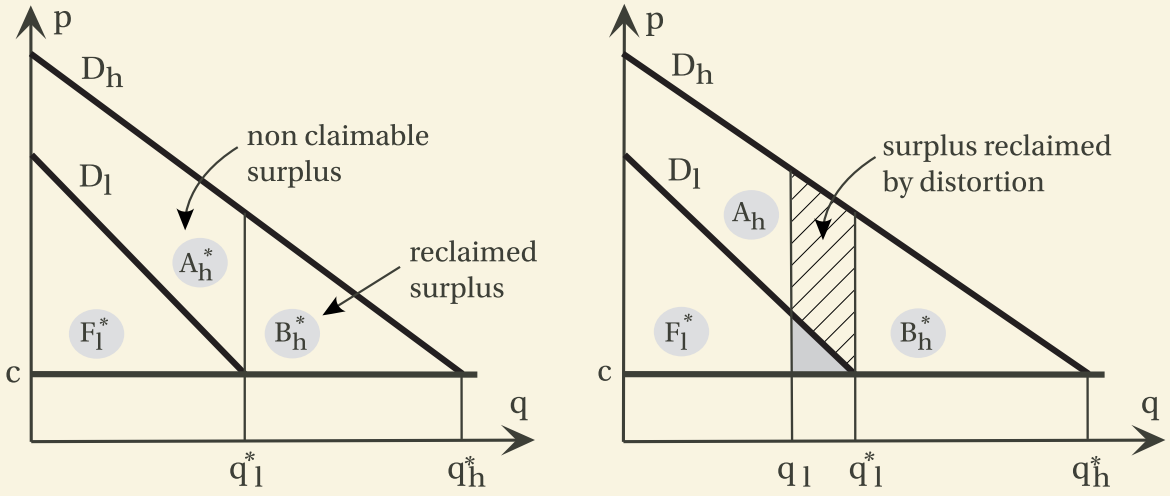


Figure 4.4: Self-Selection among Tariffs

A second improvement is shown in the right panel of Figure 4.4 where q_l is reduced below the efficient level q_l^*; this obviously reduces the surplus of a type l consumer so that the transfer t_l must accordingly be reduced to maintain him indifferent and keep him as a client (recall that he derives zero net surplus from the bundle). Overall there is downfall in the profit made over consumer l because a deadweight loss (grey triangle) has been created and the monopoly must assume it. Yet the drop of profits is small because the distortion we are imposing starts from the efficient level. The great benefit of this distortion is to strongly reduce the attractiveness of (q_l, t_l) for consumer h. His "non claimable surplus" has been cut so that his transfer t_h can be raised. The stripped area represents the final net gain.

In this process, the quantity q_h is kept at the efficient level q_h^* since it generates a surplus B_h^* that can be siphoned by the transfer. The limit to this process of reducing q_l is achieved when the loss made over consumer l cancels out with the gain made over consumer h i.e., when the left side of the grey triangle has the same height as the left side of the stripped area.[29@] Hence, area A_h will never be reduced to zero and this means that consumer h enjoys a positive net surplus. We call it an *information rent* to emphasize that it comes from the impossibility for the monopoly to identify him and fully extract his consumer surplus; somehow, he is able to hid behind the low WTP consumer (l).

More generally, the limit to the distortion process illustrated on Figure 4.4 depends on the proportion of each group in the total population since in our graphical presentation we were considering a one-to-one situation (50% home users vs. 50% professionals).

The same analytical problem will be treated with more rigor in §21.2. The general conclusions that emerge from that theory are:

- For all buyers excepts those with the highest WTP, the quantity consumed is inefficiently low.
- Tariff design is limited by the fact that a hight WTP consumer tries to grab the package set for the person immediately below him in terms of WTP.
- The buyer with lowest WTP derives no net surplus (indifferent with not consuming)

4.3.4 Discrimination by Differentiation (Versioning)

This last section presents shortly the remaining forms of indirect discrimination that are based on the heterogeneity of tastes regarding quality or attributes of items and on the moment at which they are consumed.

An early case of indirect discrimination by quality differentiation is seating in theaters that has been in use over centuries. The cheapest rate allows one to stand to see the act; at a higher price, one can seat on a bench in a "first-come-first-seated" fashion; by paying a premium, one gets a reserved chair and finally, the wealthiest people hold a box all year long.

Regarding network industries, the oldest example is probably the 3 passenger classes introduced by all railways companies during the XIX^{th} century. At the time, the third class was treating passengers very harshly and many advocated that the railways companies should spend the small cost of improving their comfort. Dupuit (1849), commenting these schemes, answers them with an argument of self-selection: if the wood benches in the third class were covered with a cheap cloth, many second class users would switch to the cheaper third class given that the price difference would cease to be worth the quality difference. If this were to happen, the company would need to compensate second class users, for instance by introducing a moving service inside the second class cars. But then the first class users would feel unjustified the premium they pay.

To implement the class system, it is necessary to maintain a large quality difference between any two adjacent classes. For trains, the third class disappeared in developed countries when everybody became able to afford second class while nothing justified raising the first class rate to the sky. Planes offering very long trip are still able to maintain 3 classes thanks to the "Sleeper-Seat" innovation that justify the premium between the first and business classes; the reader has probably noticed that tourist class in these

flights is no more fun than the wood benches in the third class railways of the past, but this is a necessary evil if we want to be offered cheap tickets.

Just like third class in railways was a downgraded second class, many firms today deliberately impair a good or service to create a lower quality version that can be sold at a discount price; these products are called *damaged* or *hobbled* goods. For instance, many commercial softwares have a light version with restricted functionalities or some media products are made artificially smaller (60 mn disc vs. 74 mn). In other cases, printers, computer chips or postal delivery services are artificially slowed. Finally, the introduction of computerized systems in automobiles has lead automakers to design motors whose horsepower can be electronically reduced so as to offer a lighter but cheaper version of the car (which in several countries also qualify for lower tax). Likewise the onboard computer is always present but the display functions are empowered at assembly only if the client has paid for the options. A recent popular case is Apple's 33% price cut of the iPhone in 2008 on the day the iPod Touch was introduced. This apparatus is similar except for the absence of telephony service. This solved the cannibalization problem whereby consumers would abandon the pricey iPhone for the less expensive but almost as useful iPod Touch.

Up to now, we have discussed a *vertical* differentiation based on quality; the other common form of differentiation is *horizontal* and uses the heterogeneity of consumer tastes, for instance that related to gender (men vs. women) or age (young vs. mature). Both types of differentiation are pursued with more detail in §11.

4.3.5 Inter-temporal Discrimination

Dupuit (1844)'s original example of differential pricing remains topical to introduce the time dimension: the wise editor of a novel can print a limited number of luxurious copies for the first edition and sell them at 40€ each; the second edition will come out immediately for the general public and will be priced at 20€. A few months later, the publisher will issue a third edition by subscription at the reduced rate of 10€. The next year, a fourth illustrated edition will come out at only 5€, soon followed by a popular edition on cheap paper at 3€. Finally, a compact small print edition will be issued for 2€. All in all, the publisher will sell many more copies and earn much more than if he were to set a single monopoly price, for instance 20€. This scheme was not well received at the time and even called a scam. Dupuit (1844) countered that, thanks to the greater profits of the editor, the writer can be paid a decent wage and that many people have been able to enjoy his book so that everyone is better off; in other words the pricing scheme is a Pareto improvement upon traditional pricing.

As opposed to flow goods like a beer or a train ticket whose consumption is synonymous of destruction, a durable good lasts for at least two periods. Thus, once the monopoly has sold to some consumers he cannot sell again to them later on. Our purpose is to see how this limitation can be overcome. We shall consider in turn several models of increasing complexity regarding the behavior of both the monopoly and the consumers. We say that customers are *sophisticated* if they anticipate any future price cut and can therefore postpone their purchase (if they wish). Inversely, customers are naive or *myopic* if they have no memory or fail to anticipate changes in the pricing policy.[30@] We also distinguish whether the firm can commit or not to a pricing policy over the two periods. If not, she is bound to set a price optimal given the current condition i.e., the game is solved by backward induction. The more general interaction of commitment and price discrimination is the object of §4.3.6.

Pricing for Myopic Consumers

Intuitvely, the monopoly can sell the good to rich consumers (more precisely those with a high WTP) in the first period and reduce the price in the second period to reach poorer consumers. For instance, he could initially charge the monopoly price p_1^M corresponding to a once–for–all sale, then estimate the residual demand $\hat{D}$ for the second period and charge the updated monopoly price $\hat{p}_2^M$ (thereby fooling the memoryless consumers). The monopoly therefore earns Π_1^M and a lower Π_2^M.

Concretely we assume additive temporal preferences. The demand for one period use of the good is $D(p) = 1 - p$ and the extra value of an additional period ownership is δ times the single period value where $\delta < 1$ is the obsolescence factor of the good common to all consumers. The market WTP in the first period for a once–for–all sale is $P_1(q_1) = (1+\delta)(1-q_1)$. With constant marginal cost (normalized to zero), the monopoly sells $q_1^M = \frac{1}{2}$ at price $p_1^M = \frac{1+\delta}{2}$ to earn $\pi_1^M = \frac{1+\delta}{4}$. The residual demand is $\hat{D}(p_2) = 1 - q_1^M - p_2$ because the rich consumers have bought first (the sale is like an auctioning process); the market WTP is now $P_2(q_2) = \frac{1}{2} - q_2$, hence the optimal second period sales are $q_2^M = \frac{1}{4}$ at price $p_2^M = \frac{1}{4} < p_1^M$ yielding $\pi_2^M = \frac{1}{16}$. Total profit is thus $\frac{5+4\delta}{16}$.

The truly optimal strategy when facing myopic consumers is slightly different although it obeys the same intuition. Given initial sales $\bar{q}_1$ (not necessarily half of the market), there is a residual demand, thus an optimal second period price p_2. We then compute the reduced form second period profit $\Pi_2(\bar{q}_1)$ and maximize the total profit $\Pi_1(\bar{q}_1) + \Pi_2(\bar{q}_1)$. Recalling that profits are always increasing with market size, $\Pi_2(\bar{q}_1)$ must be a decreasing function since more sales today means that tomorrow's market will be smaller. Hence, the marginal profit of the second period is negative, it acts as a marginal cost within the total marginal profit. Now, being faced with a higher overall marginal

cost, the monopoly will optimally reduce its first period sales; therefore the optimal initial price $\hat{p}_1$ is greater than the monopoly price p_1^M but the rebate for the second period is also greater.

In our example, if $\bar{q}_1$ units are sold initially, the residual demand is $\hat{D}(p_2) = 1 - \bar{q}_1 - p_2$, thus optimal sales are $\hat{q}_2(\bar{q}_1) = \frac{1}{2}(1 - \bar{q}_1)$ at price $\hat{p}_2(\bar{q}_1) = \frac{1}{2}(1 - \bar{q}_1)$ leading to second period profit $\hat{\Pi}_2(\bar{q}_1) = \frac{1}{4}(1 - \bar{q}_1)^2$. Recalling that the first period profit being $(1+\delta)(1-q_1)q_1$, total profit is

$$\Pi = q_1 P_1(q_1) + \hat{\Pi}_2(q_1) = (1+\delta)(1-q_1)q_1 + \tfrac{1}{4}(1-q_1)^2 \tag{4.7}$$

The optimal initial sales, maximizing (4.7), are thus $\hat{q}_1 = \frac{\delta}{1+2\delta} < q_1^M$ at the price price $\hat{p}_1 = P_1(\hat{q}_1) = \frac{(1+\delta)^2}{1+2\delta} > \frac{1+\delta}{2} = p_1^M$. Since the monopoly restrict first period sales, it increases the second period ones as $\hat{q}_2(\hat{q}_1) = \frac{1+\delta}{2(1+2\delta)} > q_2^M$. As we previously claimed, the new rebate $\hat{p}_1 - \hat{p}_2 = \frac{1+\delta}{2}$ is greater than the old one $p_1^M - p_2^M = \frac{1+2\delta}{4}$.

Summarizing, if consumers are *myopic* and don't consider the future when taking today's purchasing decision, the monopoly can basically sell the good to rich consumers in the first period and reduce the price in the second period to reach poorer consumers. In fact, the optimal strategy when facing myopic consumers is different because the two markets, today and tomorrow, are related. Recall indeed that periodic profits are increasing with market size and that more sales today means a smaller market tomorrow. Hence, increasing sales marginally today generates a marginal loss on tomorrow's profits. The monopoly has therefore an incentive to reduce today's sales with respect to the standard monopoly computation. He then offers a larger rebate in the second period to capture a large market share.

Leasing

Lowering prices in the second period means lowering benefits, hence a clever attitude is to turn the durable good into a non-durable one by *leasing* (renting) it in every period. This way, rich consumers are forced to pay twice the good while poor are excluded from consumption. Since the market WTP for the first period of use is $1 - q$, the optimal first period volume is $\frac{1}{2}$ at the lease price $\frac{1}{2}$. In the second period, the good loses values thus the WTP is $\delta(1-q)$ so that the optimal second period volume is $\frac{1}{2}$ at the lease price $\frac{\delta}{2}$; total profit is thus $\frac{1+\delta}{2}$. The key to this result is that by leasing the good, the monopolist retains ownership of all produced units and thus is not tempted to expropriate the owners of previous production by flooding the market because he would expropriate himself.

Since the full monopoly outcome is trade restrictive and quite inequitable, it has been ruled out by the US antitrust authorities in several cases against Xerox in the 1960s and IBM in the 1970s (cf. Waldman (1997)). More recently, the 1994 Consent Decree has

prohibited Microsoft from offering long-term contracts (contracts with duration longer than one year) for operating systems. A strategy similar to leasing is the *repurchase agreement* according to which, the monopoly accepts to repurchase all units previously sold if he comes to offer a lower price, thereby permitting old customers to buy again at the new and lower price. Almost identical is the *most favored customer* (MFC) clause which compensates old customers for any price rebate offered to new customers; the only difference is the continuity of ownership of the units previously sold. These two contractual arrangements give the same commitment power to the monopoly to maintain its price constant over time; they are often viewed as anti-competitive by authorities as we explain in §9.2.1.

Pricing to Sophisticated Consumers

We abandon the option to lease the good and come back to sale as ownership transfer. The myopic attitude is hard to sustain as consumers do observe the price cut. Some first period buyers may simply anticipate it and decide to wait in order to enjoy the good, later but cheaper. This anticipation forces the monopoly to reconsider its pricing strategy and adopt a dynamic perspective consistent with its clients expectations. The solution is found by applying backward induction: given what happened in the first period (be it coherent or not) how will the consumers and the monopoly act in the last period?

Although consumers observe the price p_1, they anticipate sales $\bar{q}_1 < D(p_1)$ because some people will delay their acquisition. Consumers are rational and know that the monopoly is rational too, hence the initial sale of $\bar{q}_1$ units will be followed by the optimal price $\hat{p}_2(\bar{q}_1) = \frac{1}{2}(1-\bar{q}_1)$. To advance further, we need to delve into the roots of consumer demand. When initial sales are $\bar{q}_1$, the margin buyer has per-period WTP $1-\bar{q}_1$, hence his utility for early purchase is $(1+\delta)(1-\bar{q}_1)-p_1$ whereas he nets $1-\bar{q}_1-\hat{p}_2$ for late purchase. This is truly the marginal buyer if he is indifferent i.e., if $p_1 = \bar{p}_1(\bar{q}_1) \equiv \left(\frac{1}{2}+\delta\right)(1-\bar{q}_1)$, the unique price consistent with consumers' expectations. Notice that $\bar{p}_1 < P_1$ i.e, consumers by anticipating the future price cut, force the monopoly to distort the initial price downward. The total profit that the monopoly can now anticipate is

$$\Pi^M(\bar{q}_1) = \bar{q}_1\bar{p}_1(\bar{q}_1) + \hat{\Pi}_2(\bar{q}_1) = \left(\tfrac{1}{2}+\delta\right)\bar{q}_1(1-\bar{q}_1) + \tfrac{1}{4}\left(1-\bar{q}_1\right)^2 \tag{4.8}$$

The optimal initial sales, maximizing (4.8), are $q_1^* = \frac{2\delta}{4\delta+1} > \frac{\delta}{1+2\delta} = \hat{q}_1$ i.e., sophistication forces greater initial sales (using a lower price) because clients anticipate the future rebate; the rebate is also lowered since $p_1^* - p_2^* = \delta(1-q_1^*) < (\frac{1}{2}+\delta)(1-\hat{q}_1) = \hat{p}_1 - \hat{p}_2$. Total profit is $\pi^* = \frac{(1+2\delta)^2}{4(4\delta+1)}$.

Commitment

If the firm has the ability to commit to an inter-temporal price strategy, she can avoids some of the loss due to consumers' sophistication by selling once for all the good to half of the population and earn $\frac{1+\delta}{4} > \pi^*$ although this is less than the profit derived against myopic consumers. Commitment does not come cheaply, it must be build. An example of a successful reputation is Disney's strategy of time limited releases; every year a movie is released in video (now DVD) for the consumer market during 6 months with a publicly made promise to stop selling it for 5 full years. Since kids' interest for the feature is bound to disappear after that lapse of time, parents have no choice but to buy early at the high price set by Disney. This profitable one-shot strategy works if the seller keeps its word. Now, as Disney's catalogue of candidate for releases is large, it has a strong incentive to maintain this reputation. Similar issues of dynamic consistency are studied in §10.3 on foreclosure.

Coase conjecture

The above logic of falling prices and patiently waiting consumers leads Coase (1972) to conjecture that

> If the monopolist can sell any amount and cut price quickly, her initial offer would be approximately marginal cost, thereby replicating the competitive outcome.

The monopolist's problem is her inability to commit to maintain a high price as she competes with price-cutting future incarnations of herself. The mathematical proof of this result is highly sensitive to details and in particular to the service capacity which cannot be quickly adjusted upwards or downward. Mcafee and Wiseman (2008) show that under such a plausible capacity constraint, profit cannot be lesser than 30% of the rental monopoly profit (as shown in appendix).

4.3.6 Behavior-Based Discrimination

We leave durable goods and return to flow goods that are repetitively bought.

Purchase History

Many industries, including supermarkets, airlines, and credit cards compile individual consumer transactions, study the purchasing behavior and make tailored offers to individual consumers, via direct (e)mail or other forms of targeted marketing. Over the

internet, cookies, static IP addresses, credit card numbers, and direct user authentication make this strategy ubiquitous. Historic discrimination is thus the conditioning of the current price on the purchase history.

As with all intents to discriminate, firms have to deal with the reaction of consumers who can hide the fact that they bought previously or try to simply escape recognition. The reason why buyers reveal their private information to sellers is to allow them to design a better service which in turn makes a premium price justified. Typically, a small piece of information regarding preferences enables to design at low cost an addition, a customization to the baseline service which is of great value to the customer. For instance, we stick with the same doctors, lawyers, accountants, dentists, butchers because these professionals know our tastes and cater to them.[31@] Expedited checkout, online or in a shop, is available to loyal fully registered clients but comes at the cost of forfeiting "new-client" discounts. Similar customized services are recommendations, reminders, preferences. There are also purely monetary rewards to loyalty with coupons and fidelity discounts (airline miles).

Model

Armstrong (2007) considers the periodic demand is $D(p) = 1 - p$ (individual values are uniformly distributed in $[0;1]$) repeated over two periods. The monopolist sets price p_1 for the first period and prices p_2 and $\hat{p}_2$ for the second period according to whether the sale is new or repeated. Inter-temporal discrimination occurs if $p_2 \neq p_1$ while historic discrimination occurs when $\hat{p}_2 \neq p_2$.

Commitment Let us analyze first the case where the firm can *commit* in advance to its pricing policy. When facing naive customers, the firm optimally sets $p_2 = \frac{1}{2}p_1$ for new buyers (this is the standard monopoly price over the low WTP segment) and $\hat{p}_2 = p_1$ for repeaters (the monopoly wants to sell to all people with strong WTP). We have second period profit $\pi_2 = p_1(1-p_1) + \frac{1}{2}p_1\left(p_1 - \frac{1}{2}p_1\right)$ which is maximum at $p_1 = \frac{2}{3}$. Yet, the real objective sums past and present profits. As $\pi_1 = p_1(1-p_1)$, we have $\pi = \pi_1 + \pi_2 = \frac{1}{4}p_1\left(8 - 7p_1\right)$ which is maximum at $p_1 = \frac{4}{7} > p^M$ and leads to profits $\pi = \frac{4}{7}$.[32@] Consumer surplus is $\frac{1}{2}(1-p_1)^2$ for the first period and $\frac{(1-p_1)(1-\hat{p}_2)^2 + p_1(p_1-p_2)^2}{2}$ in the second one (taking into account the dichotomy between first time and repeating buyers), overall consumers net $W_D = \frac{53}{343} \simeq 0.15$ at the firm's optimum.

When facing sophisticated customers, the optimum is to repeat the one-shot monopoly price $p^M = \frac{1}{2}$ and earn $\pi = \frac{1}{2}$. Indeed, the arbitrage $v - p_1 = v - p_2$ for the indifferent buyer still holds, thus $p_1 = p_2$ so that no time arbitrage occurs. Since $\hat{p}_2$ is chosen at will, it is

optimal to set $\hat{p}_2 = p_1$ for repeaters (as before). It is now clear that the common monopoly price is optimal. Consumer surplus is then $W_D = 0.25$.

Concluding, under commitment, the firm takes advantage of consumers' naivety to increase its profits. Since prices are higher welfare decreases which means that consumer surplus is drastically reduced (from 0.25 to 0.15). In a free market economy, a firm sets price as she wishes in any period, thus inter-temporal discrimination is always possible (though not useful if consumers are sophisticated). Now, historic discrimination may be unavailable for either legal or informative reasons in which case the firm stands to lose.[33@]

Commitment Failure We now move to the case where the firm cannot commit to its pricing policy i.e., cannot help but offer potentially different prices to the two groups of early and late buyers, if she finds it at her advantage to do so. The failure to commit is inconsequential when facing naive customers since they fail to perceive commitments, hence the firm takes advantage of them exactly as when she can commit.

Against sophisticated customers, the last early buyer is $v > p_1$ because people with WTP close to p_1 prefer to wait for the ineluctable price cut. The optimal second period prices are as in the Naive–Commit case, $p_2 = \frac{1}{2}v$ and $\hat{p}_2 = v$. Since the sophisticated cut-off buyer is indifferent between early and late purchase we have $v - p_1 = v - p_2$. Thus, $p_1 = p_2 = \frac{1}{2}\hat{p}_2$, a strategy quite different from that seen before: the price asked to first-time buyers is low but the repeater's price is high as observed in many instances of web sales. Profits are then $\pi_1 = \frac{1}{2}\hat{p}_2(1 - \hat{p}_2)$ and $\pi_2 = \hat{p}_2(1 - \hat{p}_2) + \frac{1}{4}\hat{p}_2^2$. The only difference with the Naive–Commit case (up the change of optimizing variable p_1 by $\hat{p}_2$) is the $\frac{1}{2}$ factor in π_1; the solution is the greater price $\hat{p}_2 = \frac{3}{5} > \frac{4}{7}$ which allows the firm to earn $\pi = \frac{9}{20} < \frac{4}{7}$ (the commitment level). Consumer surplus is $W_D = \frac{139}{1000} \simeq 0.14$. The lack of commitment power thus arms everybody since both the firm and consumer lose; they are trapped into a web of mistrust about price changes that generates a . If historical discrimination is unavailable, then the firm applies the one-shot optimal monopoly price twice.[34@]

Recent advances in marketing techniques together with the advent of the computer age, mean that the commitment problem has become worse. Indeed, the finely tuned customer data that firms often possess allow them to build personalized prices. Such prices are often "secret" rather than public, and it is unlikely that firms can commit to such prices. Moreover, even if firms could commit, the complexity of the linkages between consumer actions and future prices may be too complicated for many consumers to comprehend.

Part C

Strategic Interaction

Chapter 5

Imperfect Competition

Oligopoly with Simultaneous Decisions

This chapter is the building block of the book. We concentrate on duopoly models i.e., the competition among two firms. Although this may seem a limitation when trying to study the abyss existing between perfect competition and monopoly, most results of this chapter extend straightforwardly to any finite number of firms. It is also frequent to find the perfect competition outcome when the number of firms grows large which gives in retrospect a justification for the price-taking assumption in the perfect competition model.

The main models of oligopoly competition view firms competing by choosing prices and quantities (or production capacities). As we shall see, the Cournot (1838) approach is adequate for markets where prices can vary faster than production capacities (e.g., wheat, cement) whereas the Bertrand (1883) approach is adequate when firms can adjust capacities faster than prices (e.g., software). We also study a new form of competition based on contracts whereby firms offer multiple combinations of price and quantity and let clients pick their preferred choice; this is relevant for markets where the demand changes unexpectedly.

Our first section complies with the basic requirement for the study of strategic behavior, the introduction to game theory, the tool permitting the modeling of rational interaction. The following section focuses on the Cournot model of competition through quantities whereas the next one considers the more direct competition through prices together with some comparison and extensions. Lastly, we present succinctly under the heading of contest, some of the scenarios where firms compete outside the market.

5.1 Competition via Quantities: Cournot

The idea that firms compete via the quantities they wish to sell was first analyzed by Cournot (1838) who modelized the competition among two producers of mineral spring water of identical quality with a common nil marginal cost. In today's world, these firms could be Danone and Nestlé with their respective brands of mineral still water Evian and Vittel.

5.1.1 Symmetric Cost

Let us present the basic model in the following manner: farmers 1 and 2 come to the market to sell their harvest (e.g., wheat); their (maximum) supplies are fixed quantities q_1 and q_2. This is the crucial hypothesis of Cournot's model; it can be explained by a technological limitation like a production capacity or the crop enabled by this year's weather. Assuming further that the two goods are homogeneous, a single price must hold in equilibrium.[1@] If the equilibrium price p is such that demand $D(p) = a - bp$ is lesser than the maximum aggregate supply $q_1 + q_2$, then at least one farmer fails to sell all of his units. He can then undercut p to secure maximum sales; as shown in the price competition section, this behavior improves his profit. We therefore conclude that the equilibrium price satisfies $D(p) \geq q_1 + q_2$. In equilibrium of the market, demand equals supply thus we have $D(p) = q_1 + q_2$; hence the market price is given by the willingness to pay of the market through the relation

$$p = P(q_1 + q_2) = \frac{a - q_1 - q_2}{b} \tag{5.1}$$

Like a monopolist, farmer 1 understands that his decision q_1 directly impacts the market price; the novelty of the Cournot model is the presence of the decision q_2 taken by competing farmer 2 inside the price equation (5.1).

Let us assume that both farmers share the same constant marginal cost of production c. Farmer 1, being rational, is able to anticipate that when farmer 2 brings in q_2, its profit as a function of its own quantity q_1 is

$$\pi_1^C(q_1, q_2) = q_1 P(q_1 + q_2) - c q_1 \tag{5.2}$$

His objective being to maximize profits, he will produce the quantity equating marginal revenue to marginal cost, here c. It is interesting to note that the marginal revenue

$$R_{m,1} = P(q_1 + q_2) + q_1 P'(q_1 + q_2) \tag{5.3}$$

is greater than for a monopoly (obtained by merging farmers 1 and 2). Indeed, when farmer 1 wants to expand production, he must give a rebate to his current clients, but these do not constitute the entirety of the demand, hence the cost of the rebate is lesser than for a monopoly who has to give it to the whole demand (check the difference with eq. 3.3).

Using our linear demand, we compute $R_{m,1} = \frac{a-q_2-2q_1}{b}$, thus the best reply of farmer 1 to q_2 is the quantity q_1 that maximizes $\pi_1^C(q_1, q_2)$ i.e., solves

$$R_{m,1} = c \quad \Leftrightarrow \quad q_1 = BR_1^C(q_2) \equiv \frac{a - bc - q_2}{2} \tag{5.4}$$

> In the Cournot duopoly, each firm behaves like a monopolist, only that the demand he faces is the market demand minus the (anticipated) supply of his challenger.

Symmetrically,[2@] the best reply of farmer 2 to q_1 is $q_2 = BR_2^C(q_1) \equiv \frac{a-bc-q_1}{2}$. Both curves are shown on Figure 5.1 and their intersection defines the unique Nash equilibrium since it the only situation where each action (quantity) is a best reply to the other. Recalling the definition of the Nash equilibrium, each farmer takes a decision contingent on what he believes his competitor will bring. The equilibrium occurs when both conjectures are actually correct.

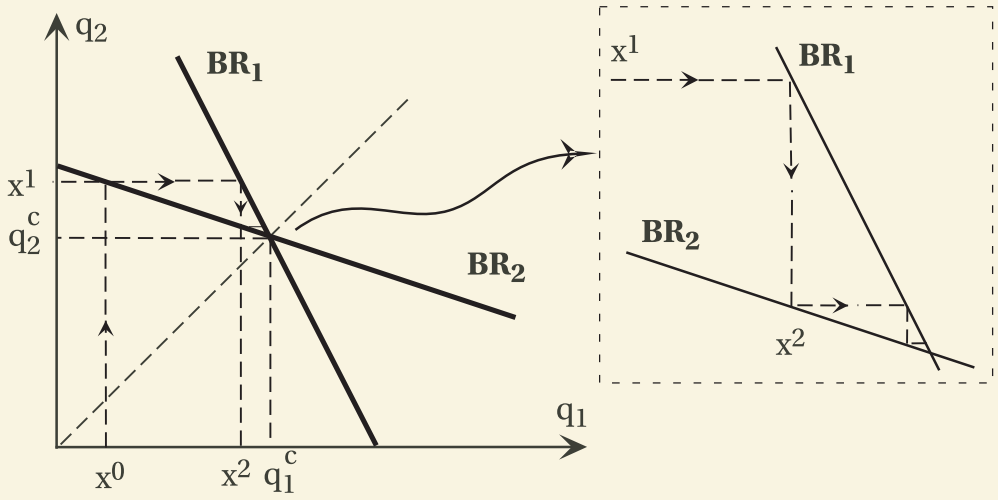


Figure 5.1: The Cournot Web

To understand the process by which an equilibrium is reached imagine that on day 0 farmer 1 brings a small quantity x^0. Then the next day, farmer 2 having observed x^0, choose as a best reply to bring a large quantity $x^1 = BR_2^C(x^0)$. In turn farmer 1 will react on day 2 by bringing $x^2 = BR_1^C(x^1)$. It easy to observe that the quantities converge to the intersection of the two reaction curves. To derive formally this intersection, we solve

simultaneously the two equations $q_1 = BR_1^C(q_2)$ and $q_2 = BR_2^C(q_1)$; substituting one into the other we obtain

$$q_1^C = q_2^C = q^C \equiv \frac{a - bc}{3} \tag{5.5}$$

It is clear that when farmer 1 anticipates that farmer 2 will bring q_2^C, he finds it optimal to bring q_1^C and symmetrically, when farmer 2 anticipates that he will face q_1^C, he is rationally lead to bring q_2^C. Hence both actions reinforce the anticipation held by each farmer. The Nash equilibrium point is said to be *self enforcing*. [3@]

Replacing the result from (5.5) into (5.1), we obtain the equilibrium price:

$$p^C \equiv P(q_1^C + q_2^C) = \frac{a + 2bc}{3b} \tag{5.6}$$

leading to total sales of $2q^C$. Since cost are symmetric, the profit margin is the same for both farmers; using (5.6) and symplifying, we get

$$p^C - c = \frac{a - bc}{3b} = \frac{q^C}{b} \tag{5.7}$$

The last relevant magnitude is the (common) equilibrium profit which is computed using (5.5) and (5.7) as

$$\pi_1^C = \pi_2^C = q^C(p^C - c) = \frac{1}{b}\left(q^C\right)^2 = \frac{(a - bc)^2}{9b} \tag{5.8}$$

The Lerner index of market power (cf. eq. (3.4)) is $\mathscr{L} = \frac{p^C - c}{p^C} = \frac{a - bc}{a + 2bc} > 0$. The consumer surplus is $W_D^C = \frac{1}{2b}\left(2q^C\right)^2 = \frac{2}{9b}(a - bc)^2$ (cf. eq. (2.19)) and welfare (cf. definition in §2.3.2) is thus

$$W^C = W_D^C + \pi_1^C + \pi_2^C = \frac{4(a - bc)^2}{9b} < W^* = \frac{(a - bc)^2}{2b} \tag{5.9}$$

the maximum welfare achieved at the Pareto optimum where willingness to pay is equated to marginal cost.[4@] The welfare loss is to due the application of market power by the duopolists. The ratio of inefficiency is $\frac{W^* - W^C}{W^*} \simeq 11\%$ which is less than the 25% loss occuring when there is a single firm, the monopoly (cf. eq. 3.5).

> The equilibrium of the Cournot duopoly involves an intermediate outcome in terms of efficiency. The equilibrium price, aggregate sales, individual profits and market welfare are all found between their respective monopoly and efficient levels.

In the following sections, we use the results of the basic Cournot model to address a variety of questions.

5.1.2 Asymmetric cost

Theory

There many reasons for the marginal cost of the two firms to differ. In the case of farmers, the fertility of their plots of land is likely to be different. More generally, one firm can have a better technology.

When farmer 1 has a cost advantage with $c_1 < c_2$,

the best reply (5.4) becomes $q_1 = \frac{a-bc_1-q_2}{2}$ and symmetrically $q_2 = \frac{a-bc_2-q_1}{2}$. Solving this linear system, the equilibrium is

$$q_1^C = \frac{a-b(2c_1-c_2)}{3} \quad \text{and} \quad q_2^C = \frac{a-b(2c_2-c_1)}{3} \tag{5.10}$$

The equilibrium price is $p^C = \frac{a+b(c_1+c_2)}{3b}$ and the total quantity sold is $Q = \frac{2a-b(c_1+c_2)}{3b}$. The more competitive firm ends-up selling more than its competitor, the difference being $q_1^C - q_2^C = b(c_2 - c_1)$. The profits of firms remain given by the same formula as before with $\pi_1^C = \frac{1}{b}\left(q_1^C\right)^2$ and $\pi_2^C = \frac{1}{b}\left(q_2^C\right)^2$; they are ranked like sales.

Cost Edge

To enable a comparison with the previous symmetric situation, let us assume that firm 2's marginal cost is $c_2 = c$ while firm 1's is $c_1 = c - \gamma$ where γ is a cost advantage. Applying (5.10), we obtain $q_1^C = \frac{a-bc+2b\gamma}{3}$ and $q_2^C = \frac{a-bc-b\gamma}{3}$ thus firm 1 ends-up selling $q_1^C - q_2^C = \frac{b\gamma}{3}$ more units thanks to her cost advantage; she also earns a premium of $\pi_1^C - \pi_2^C = \frac{1}{3}(q_1^C + q_2^C)\gamma$.

> A marginal cost advantage traduces, within the Cournot model of competition, into additional sales and profit that are both proportional to the edge.

Trade Policy

Even when the world best's technology is available to all, local regulations or taxes can distort market outcomes. An interesting application of competition with asymmetric cost is trade policy whereby a government impinges on the competition between a national firm (#1) and a foreign firm (#2). For instance, it can impose a duty t upon every unit imported in the country. If the two firms have initially the same marginal cost c and sell the common output $\bar{q} = \frac{a-bc}{3b}$ at price $\bar{p} = \frac{a+2bc}{3b}$, then the trade policy effect is to make $c_1 = c$ and $c_2 = c + t$. We can immediately look at the equilibrium for asymmetric costs to derive conclusions. We observe that $q_1^C = \bar{q} + \frac{t}{3}$ while $q_2^C = \bar{q} - \frac{2t}{3}$, hence the national firm increases its sales at the expense of the foreign one (who sells less than before). Although

this policy protects the local firm, it hurts local consumers since the equilibrium price, $p^C = \bar{p} + \frac{t}{3}$, increases with the duty t and the total quantity sold, $Q^C = \bar{Q} - \frac{t}{3}$ decreases with it.

Switching cost

Many of the products we consume are durable and generate a long lasting utility; being accustomed to them, we find it difficult to switch to an alternative even if it is functionally identical.[5@] The following list reveals the extend of the issue.

- Need for compatibility with existing equipment
- Transaction costs of switching suppliers
- Costs of learning to use new brands
- Uncertainty about the quality of untested brands
- Discount coupons and loyalty programs (aka volume discount)
- Brand loyalty and/or psychological inertia[6@]

As reported in Farrell and Klemperer (2007), empirical evidence of switching costs has been found for credit cards, cigarettes, software, supermarkets, air travel, frequent-flyer programs, telephone, TV channels, online broker, electricity supplier, bookstores and automobile insurance.

As we shall see with a straightforward application of the Cournot model, switching costs have a notable impact on competition and are sought after by firms. Assume that for all consumers, there is a switching cost s€; the willingness to pay for a new product $P(q)$ is therefore diminished by this amount which appears as an additional marginal cost for the firm in its profit function. If only one firm, say an entrant (#2), is affected by this phenomenon, its full marginal cost is $c_2 = c + s$ while the incumbent keeps its marginal cost at $c_1 = c$. All of our conclusions on the distortionary effect of a tax can be accommodated to the present setting.

5.1.3 Oligopoly

Symmetric Cost

Extending the duopoly analysis to oligopoly (n active firms) is needed to address the issue of entry and understand why some markets have only 2 or 3 active firms while others are characterized by a two-digit figure.

In the analysis of the duopoly, we used the $Q = q_1 + q_2$ decomposition of the total supply. For the oligopoly, we select one firm i and write $Q = q_i + Q_{-i}$ where $Q_{-i} \equiv \sum_{j \neq i} q_j$

is the total amount brought to the market by i's competitors. The individual profit is $\pi_i(q_i) = q_i P(Q) - C_i(q_i)$ and the best reply solves the FOC

$$P(Q) - C_{m,i}(q_i) = -q_i P'(Q) \quad \Leftrightarrow \quad \mathscr{L}_i = \frac{s_i}{\epsilon} \tag{5.11}$$

where s_i designates the firm's market share. In our linear example, (5.11) simplifies into the previously seen FOC (5.4), so that the best reply is simply $q_i = BR_i^C(Q_{-i})$ i.e., when goods are homogeneous it does not matter who produces the remaining units that the market will absorb. Repeating this operation for all n firms, we obtain n equations for the n unknown quantities yielding a finite number of solutions (very often a unique one).

We now show for the linear demand $D(p) = a - bp$ that if the production technologies are identical (i.e., same marginal cost $c_i = c$ for all $i \leq n$), then a unique symmetric Nash equilibrium exist. In a symmetric equilibrium, the individual quantities selected by firms are all identical to some level q, so that the best reply equation (5.4) needs to be applied to $q_i = q$ and $Q_{-i} = (n-1)q$; we thus obtain

$$q = \frac{a-bc-(n-1)q}{2} \quad \Rightarrow \quad q = q^C \equiv \frac{a-bc}{n+1} \tag{5.12}$$

which is smaller than its competitive equilibrium counterpart $q^* = \frac{a-bc}{n}$ (case where all farmers are price-takers). The total sales simplify into $Q^C \equiv \frac{n}{n+1}(a-bc)$, the equilibrium price is $p^C \equiv \frac{a/b+nc}{n+1}$, while the individual firm profit is $\pi^C \equiv \frac{(a-bc)^2}{b(n+1)^2}$, leading to industry profits $\Pi^C = n\pi^C$, a proportion $\frac{4n}{(n+1)^2}$ of the monopoly level. Lastly, the Lerner index of market power (cf. eq. (3.4)) is $\mathscr{L} = \frac{p^C - c}{p^C} = \frac{a-bc}{a+nbc} > 0$ which is decreasing with the number of active firms.

When n becomes large, the competition in the Cournot set-up brings the price nearby efficiency since both p^C and p^* tends to zero (industry profit also shrink towards zero). Another nice property of the Cournot model is its neutrality with respect to market size: as can be checked from (5.12), if the demand doubles (multiply a and b by 2) then sales and profits also duplicate. We can now state:

> As the number of active firms increases in the Cournot oligopoly, total production increases in equilibrium, the equilibrium price decreases and welfare rises. If the equilibrium is symmetric, then individual sales and profits decrease.

Asymmetric Cost

We end with the derivation of the equilibrium in the presence of cost asymmetries which shall be useful for later chapters. In the linear example, the best-reply of firm i is $2q_i =$

$a - bc_i - Q_{-i}$ for all $i \leq n$. Summing these equations we get

$$2Q = 2\sum_{i\geq 1} q_i = \sum_{i\geq 1}(a - bc_i) - \sum_{i\geq 1} Q_{-i} = n(a-bc) - (n-1)Q$$
$$\Rightarrow \quad Q^C = \tfrac{n}{n+1}(a-bc) \tag{5.13}$$

where $c \equiv \frac{1}{n}\sum_{j\geq 1} c_j$ is the average industry cost. Lastly we use $Q^C_{-i} = Q^C - q^C_i$ inside the best reply equation to derive the individual quantity

$$q^C_i = \frac{a - b\left(nc_i - \sum_{j\neq i} c_j\right)}{n+1} = \frac{a - b((n+1)c_i - nc)}{n+1} \tag{5.14}$$

This step-by-step derivation of the equilibrium has also proven its uniqueness.

Industry Shock or Taxation

Nelson (1957) wonders what happens to the industry (and its individual members) if the government taxes the output, say, to address a negative externality such a pollution, or if a competitive input market suffers a negative shock such as higher wages or dearer energy. These shocks raise marginal cost and lead unambiguously each firm to sell less, whatever the market structure and the form of interaction. There is thus a price hike. Intuition would then suggest a reduction in the profit margin and total earnings. This is correct for the linear demand used in this book (check eq. (5.12)) but need not always be so. Indeed, since the shock applies to the entire industry, it may reduce the intensity of rivalry and may thus end up being beneficial. Obviously, the same paradox may apply in reverse to a positive shock i.e., a reduction of taxes, wages or energy prices may be detrimental to the industry. As shown by Seade (1985) (cf. §7.2.3), when an oligopoly serves an inelastic demand, a negative outside shock is always profitable. An example was the gasoline market of western countries when the oil price rose in the 1970s.

5.1.4 Welfare

Market welfare sums the consumer and producers surpluses $W = W_D + W_S$. It is well known that a social planner wishing to maximize welfare would impose the competitive price p^*. Interestingly, Bergstrom and Varian (1985) characterize the Cournot outcome in a similar way: a social planner would choose the Cournot price if his objective was to maximize $\widehat{W} \equiv W_D + \frac{n}{n-1} W_S$ i.e., if he would favor firms over consumers (twice in the duopoly case).

To prove this claim recall first that $W_S(Q) = \sum_{i\geq 1} \pi_i = QP(Q) - \sum_{i\geq 1} C_i\left(q_i\right)$ so that $dW_S =$

$QP'(Q)dQ + \sum_{i\geq 1}\left(P(Q) - C_{m,i}\left(q_i\right)\right)dq_i$ while we saw in (2.19) that $dW_D = -QP'(Q)dQ$, thus

$$\begin{aligned}(n-1)d\widehat{W} &= (n-1)dW_D + ndW_S = QP'(Q)dQ + n\sum_{i\geq 1}\left(P(Q) - C_{m,i}(q_i)\right)dq_i \\ &= \sum_{i\geq 1}\left(q_iP'(Q)dQ + \left(P(Q) - C_{m,i}(q_i)\right)ndq_i\right) = P'(Q)\sum_{i\geq 1} q_i\left(dQ - ndq_i\right)\end{aligned}$$

using FOC (5.11) for an individual firm $p - C_{m,i}(q_i) = -q_iP'$. In the symmetric Cournot game where the equilibrium is necessarily symmetric, we have $q_i = \frac{Q}{n}$ so that $d\widehat{W} = 0$ since $\frac{Q}{n}$ factorizes out and $dQ = \sum_{i\geq 1} dq_i$ by construction.

This characterization echoes the regulatory capture (cf. §16.3 and §17.3.3) that is observed in most industries: the producer's lobby through its professional association succeeds to increase its weight in the governmental balance between demand and supply by arguing, for instance, that a strong industry can create jobs. This is indeed important but let us not forget that the consumer surplus for that market is a monetary measure; thus, the higher it is, the more consumers are able to spend in other productive sectors or simply save (to finance investment in the rest of the economy). Welfare is the right measure of efficiency because it treats all sectors of the economy equally for their ability to generate wealth to the benefit of all (cf. §2.3.3).

5.2 Competition via Prices: Bertrand

Actual examples of markets for homogeneous goods where firms compete in prices are the online markets for software, music, movies, computers, electronic appliances, or books, to name a few.[7@] The conditions for perfect competition are satisfied thanks to the multitude of price comparison engines and shopping bots that enable potential buyers to compare most prices within minutes in order to buy from the cheapest seller.

The same idea is turned upside down by firms who are large consumers of standardized products such as paper, toner or pens. They set up a periodic procurement auction for the specific items they desire such as "A4 paper of 80g/m^2 in 500 sheets packaging", how many units they need and the maximum price they are ready to pay. When the auction starts, potential providers offer lower and lower prices until one remains the most inexpensive and win the right to provide the batch for the proposed price (cf. §22.1.1 for a detailed analysis of auctions).

As intuition suggests, these markets are extremely competitive because firms are forced to name low prices to keep their customer base; the outcome is then propitious to consumers but unfavorable for firms who end up earning meager profits. In the following section, we study from a theoretical point of view, how this competition develops, who

wins it and how losers might do something to avoid their fate.

5.2.1 Pure Competition: Homogeneous Goods

Bertrand Paradox

The quantity based approach previously used to study oligopoly does not apply when price is the main variable by which firms compete to attract consumers. If two shops located side by side sell the same good or service (e.g., two bars with the same design and range of drinks), then consumers will obviously patronize the one displaying the lowest price. Hence, each firm is lead to undercut her competitor in order to serve the whole market. This process stops only when one of the firms starts losing money (in the economic sense). If the marginal production costs are c_1 and a larger c_2, the equilibrium is $p_2 = c_2$ and $p_1 = c_2 - \epsilon$ where ϵ is the smallest monetary denomination. Indeed firm 2 would make strictly negative profits if it were to match p_1 or even propose less. Likewise, firm 1 who is serving the whole demand and earning $c_2 - c_1 - \epsilon$ per unit sold would lose half of its customers with $p_1 = c_2$ and all of them with a larger price, hence it is better off this way.

An example of such a "cut-throat" behavior is empirically validated by Koerner (2002) in the German market for coffee even though this market is highly concentrated (six firms account for 90% of total sales). To find other examples, it is enough to search google news for the expression "price war" to observe its constant appearance in many different sectors of the economy. This result was called a paradox by Bertrand (1883) because although there are only two firms which are not price-takers, they end up proposing a price nearby the competitive price (exactly when $c_1 = c_2$). Chicago economists in the 1960s turned the Bertrand Paradox into the slogan that "two is enough for competitive outcomes".

Edgeworth Critique

Edgeworth (1925) is neither satisfied by Cournot's model nor Bertrand's critic; he observes that constant marginal cost is only an approximation and is only valid for a small range of production. Taking these elements into consideration, it is easy to see that the undercutting reasoning does not hold anymore. Consider for instance the marginal cost curve displayed on Figure 5.2. This technology enables production at constant cost c up to the capacity k of the plant; for larger quantities, the unit cost increases by an amount δ that represent extra-hour wages or a more expensive energy.

The critique reconsiders the previous analysis as follows: starting from a common

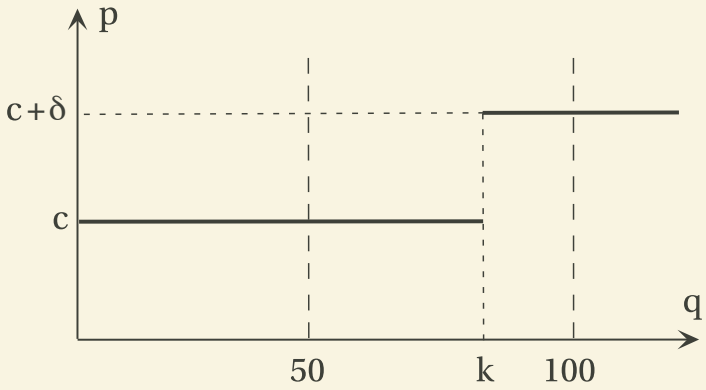


Figure 5.2: Increasing Marginal Cost

price p where firms share the market evenly (50% each), firm 1 may consider undercutting its rival with $p_1 < p$. Initially, 100% of the consumers will rush to firm 1 but because its sales will double, it will refuse a fraction $100 - k$ of the new customers who will then turn back to firm 2. The latter being already the most expensive can then raise its price to compensate for the loss of some of her customers. This process of competition is quite difficult to study but it is clear that firms do make positive profits since their marginal cost is increasing. Indeed, the price equates the cost of last and most costly unit so that all infra-marginal units either make a profit or avoid at least losses.

To illustrate this critique, we can go back to Cournot's original example of competition between to mineral spring waters like Danone's Evian and Nestlé's Vittel. The flow of a spring being limited by geological conditions, the production of bottled water is likewise limited and this rationalizes the Cournot approach to competition. Nowadays, major players in the beverage sector like Coca-Cola, PepsiCo and Nestlé sell a purified and re-mineralized tap water under the names of "BonAqua", "Aquafina" and "Pure Life". Competition is then no more limited by production capacities and the Bertrand model applies. To address quality concerns in Europe regarding bottled tap water, these major players now sell spring waters of similar taste and composition under a common name such as Nestlé's Acquarel, Danone's Dannon (US market) or FontVella (Spain).

Kreps and Scheinkman (1983) and later Boccard and Wauthy (2000) show that if firms choose production capacities (k_1, k_2) and later compete in prices (p_1, p_2) then, in equilibrium, there is no excess capacity (demand equals the sum of capacitites) and the price lies between the Cournot price p^C and the competitive one depending on the level of the extra cost δ. More precisely, $p = \min\{c + \delta, p^C\}$ and capacities are previously chosen so that total supply $q_1 + q_2$ equates demand for that price.[8@] Hence we may conclude that

> Competition via quantities (à la Cournot) is a *synthesis* of a more complex dynamic competition involving the choice of technologies and a competition via prices (à la Bertrand). If building capacity is costly, then Cournot (quantity) is the right model, otherwise Bertrand is a better description of duopoly competition.

Empirical Validation

The above prediction is validated by recent empirical research. Goolsbee and Chevalier (2002) study the competition between online booksellers Amazon and Barnes&Noble (BN); they show that a 1% price increase at Amazon reduces quantity by about 0.5% at Amazon but raises quantity at BN by 3.5%. Given that Amazon sells much more books, it is as if every customer lost by Amazon was buying from BN. On the other hand, raising prices by 1% at BN reduces sales about 4% but increases sales at Amazon by only 0.2% i.e., the lost customers from BN do not switch to Amazon, most of them are lost altogether. Studying online retail competition for books, cameras and printers among firms with groups of loyal consumers of varying size, Kocas (2005) show that cheap retailers compete with more randomized prices to capture switchers (non loyal consumers) while dear retailers prefer charging the higher reservation price with less randomization to maximize their profits from their loyal segments (the size of loyal client base plays the role of capacity).

Another illustration is based on the market for Video Game Consoles in the US[9@] shows that both phenomena can take place sequentially. The three contenders are Sony with the Playstation2 launched in October 2000, Nintendo with the GameCube launched the 15th of November 2001 and Microsoft with the X-Box launched 3 days later.[10@] As one can observe from Table 5.1, the high christmas prices the first year are explained by the bundling of consoles with games and also by the shortage of supply (which may have been voluntary). Shortly after christmas 2001, both Nintendo and Microsoft reduces their prices and it took 5 months for Sony to match the lowest price. This immediately triggered a further price reduction by Microsoft in order for its product to remain the cheapest of the market. But just one week later, Nintendo decided to cut another 50$ from its price to recover its position of "best bargain". Undercutting has been going on until the products were gradually replaced by the next generation of game consoles. Total sales were respectively 35, 14 and 11 million systems for Sony, Microsoft and Nintendo.

Console	12/01	1/02	5/02	1 day	1 week	5/03	1 day	9/03	3/04
Sony	299	299	↓199	199	199	↓179	179	179	179
Nintendo	350	↓199	199	199	↓149	149	149	↓ 99	99
Microsoft	400	↓299	299	↓199	199	199	↓179	179	↓ 149

Table 5.1: Evolution of Prices for Video Game Consoles

The battle for e-book readers displays a similar trend. Sony introduced a crude Japan-only model in 2004 at 368$, then PDF e-readers in 2006 at 349$, in 2007 at 299$, in 2008 at 399$ (later reduced to 359$), in 2009 at 299$. The market took off in November 2007 with Amazon's 399$ Kindle. A new model was introduced in February 2009 at 359$, a modest 10% discount probably much less than the cost savings allowed by 15 months of experience (a long time in the electronic industries). Amazon then lowered the price to 299$ in July and further down to 259$ in October anticipating the release of the competing Nook the following month at the same price. When the Nook price was cut to 199$ in June 2010, Amazon did not wait a day to undercut its competitor with a price 189$. In July, the third version of the Kindle (3G) appeared at the same price. Sony lowered its price to 169$ in July 2010.

5.2.2 The Hotelling model of Differentiation

Differentiation in all its dimensions (horizontal, vertical) is studied thoroughly in chapter 11. We content here to introduce Hotelling (1929)'s simple elucidation of the Bertrand Paradox.

Setting

Upon observing that many people buy from many different shops charging different prices, this author reflects that if a seller starts to increase his price, he will start to lose some customers, but not all of them as perfect competition would predict. Secondly, the perfect competition framework does not say anything about where would go the customers lost by the seller; shall they refrain from consuming or buy from a competitor ? If product are perfectly independent, there won't be any spillover in the sense that the competitor will receive no additional clients while if products are perfect substitutes, as in Bertrand competition, there will be a maximal spillover since the competitor will recover all the clients. The issue is thus to determine the *own price* and *cross price* elasticities of demand. This is precisely what Hotelling (1929)'s construction is about.

He considers a city with a single boulevard; two movie theaters are located at the outskirts, theater A standing on the west side and theater B on the east side; the distance between the two theaters is the unit of distance. Both show the same selection of movies

at respective prices p_A and p_B (all prices are labeled in the same currency). Moviegoers are ready to pay $\bar{p}$ to watch a movie but suffer from walking to any of the theaters and value that loss at t per unit of distance. The alternative situation that Hotelling (1929) had in mind was of a transcontinental train line linking two harbors located on the east and west coast of the US; firms located inland can buy foreign merchandise in any harbor but must pay for the transportation cost back to their hometown.

Model

The length of the city boulevard is set to unity. A moviegoer living at point $x \in [0;1]$ obtains a surplus or utility $u_A(x) = \bar{p} - p_A - tx$ from buying at theater A, a utility $u_B(x) = \bar{p} - p_B - t(1-x)$ from buying at theater B and a zero level of satisfaction ($u_0 = 0$) if staying at home watching TV (alone). Figure 5.3 shows in plain lines the utility levels u_A and u_B as a function of location x (beware of the differing vertical axes). At a given location x, three levels are compared and the highest one indicates the optimal choice. The prices giving rise to the left panel of Figure 5.3 are sufficiently low in order that everybody goes to the movies.[11@] The upper envelope of the three curves u_A, u_B, u_0 is the bold dashed kinked line on the left panel of Figure 5.3. Notice that consumer $\tilde{x}$ for whom $u_A = u_B$ is indifferent about which theater to patronize; the precise address of this indifferent consumer is

$$\tilde{x} = \frac{1}{2} + \frac{p_B - p_A}{2t} \tag{5.15}$$

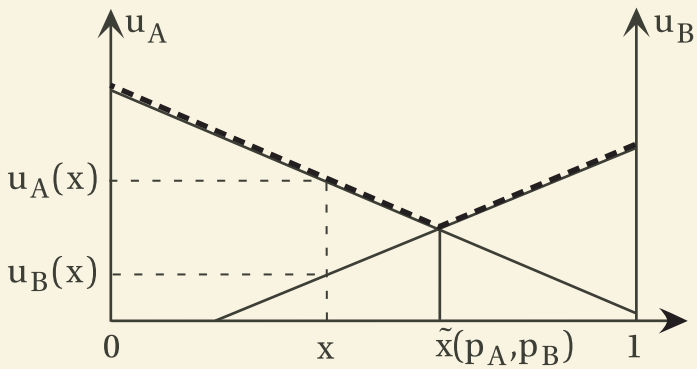


Figure 5.3: Price Competition à la Hotelling

One easily guesses that movie lovers leaving to the left of $\tilde{x}$ go to theater A while those on the right go to theater B. To ease the derivation of demands, we assume a uniform distribution of consumers along the segment $[0;1]$; this way a location x is also the proportion of consumers living to the left of x. The demands are thus $D_A = \tilde{x}$ and

$D_B = 1 - \tilde{x}$. Assuming a marginal cost c for selling a ticket, the profit of theater A is

$$\pi_A(p_A, p_B) = (p_A - c)\,\tilde{x} = (p_A - c)\,\frac{t + p_B - p_A}{2t} \tag{5.16}$$

and a symmetric formula for theater B. The best reply of theater A against her competitor's price p_B maximizes π_A i.e., solves

$$\frac{\partial \pi_A}{\partial p_A} = 0 \quad \Leftrightarrow \quad p_A = \frac{t + c + p_B}{2} \tag{5.17}$$

The best reply of theater B against the price p_A quoted by his opponent is entirely symmetric.[12@] Both are displayed on the right panel of Figure 5.3. Notice that contrary to the Cournot case, the best reply functions are increasing which leads to say that prices are strategic complements i.e., the higher my opponent's price, the higher I shall set mine.

A Nash equilibrium of the pricing game is then a pair of prices that are best replies to each other i.e., at the intersection of the best reply curves on the right panel of Figure 5.3. Because of the symmetry of the best reply functions, the equilibrium is also symmetric i.e., both theaters name the same price. To find this equilibrium price, it is enough to solve (5.17) for $p_A = p_B$ which yields $p^h = t + c$. If the transportation cost was nil, we would fall back on the Bertrand outcome where both theaters price at their marginal cost.[13@] To conclude,

> The existence of the transportation cost for consumers enables competing firms to escape from the price war leading to the Bertrand paradox. They secure profits with clients living nearby; thanks to the heterogeneity of the market, each firm is able to behave as a *local monopoly*.

5.2.3 Duality of Cournot and Bertrand

Singh and Vives (1984)'s duopoly model generalizes the Hotelling model of differentiation to enable a comparison of the Bertrand and Cournot framework of competition.

Price Competition

We now proceed to present the model and derive these results. Shubik and Levitan (1980) show how to build a simple model of consumption (cf. §2.2.1) in which the demand addressed to firm $i = 1, 2$ is

$$q_i = D(p_i, p_j) \equiv a - bp_i + d(p_j - p_i) \tag{5.18}$$

where the coefficients are the market size $a > 0$, the price elasticity $b > 0$ and the degree of differentiation or substitutability d. The two goods are independent for consumers if $d = 0$, substitutes if $d > 0$ and complements if $d < 0$.

Let us concentrate on the most frequent case of substitutability ($d > 0$) and on zero marginal cost to keep formulas simple.[14@] When firms compete in prices, each maximizes $p_i D(p_i, p_j)$. Solving the FOC, we obtain the best reply

$$p_i = BR_i^B(p_j) \equiv \frac{a + d p_j}{2(b+d)} \tag{5.19}$$

for $i = 1,2$; it is displayed by the plain lines on Figure 5.4. The symmetric Bertrand equilibrium solving (5.19) is $p^B = \frac{a}{2b+d}$ and using (5.18) we obtain $q^B = (b+d)p^B$.

Quantity Competition

Let us now study the quantity competition i.e., the case where firms choose quantities and prices are determined by the market clearing conditions. The first step is to invert system (5.18) in order to express prices as functions of the choice variables. We transform the first equation into $(b+d)p_1 = a - q_1 + d p_2$ and the second into $p_2 = \frac{a - q_2 + d p_1}{b+d}$; plugging the latter into the former, we obtain $(b+d)p_1 = a - q_1 + \frac{d}{b+d}\left(a - q_2 + d p_1\right)$ thus $p_1\left((b+d)^2 - d^2\right) = (b+d)(a-q_1) + d(a-q_2)$ and finally $p_1 = \frac{a(b+2d) - (b+d)q_1 - d q_2}{b(b+2d)}$. Letting $\beta = \frac{b+d}{b(b+2d)}, \delta = \frac{d}{b(b+2d)}$ and $\alpha = a(\beta + \delta)$, we can write for $i = 1,2$

$$p_i = P(q_i, q_j) \equiv \alpha - \beta q_i - \delta q_j \tag{5.20}$$

Now that firms compete in quantities, each maximizes $q_i P(q_i, q_j)$. Solving the FOC, we obtain the best reply

$$q_i = \frac{\alpha - \delta q_j}{2\beta} \tag{5.21}$$

for $i = 1,2$. The symmetric Cournot equilibrium found using (5.20) and (5.21) is $q^C = \frac{\alpha}{2\beta + \delta}$ and $p^C = \beta q^C$.

Merging the two approaches

It is not possible yet to mix graphically the two regimes of competition in order to compare them because the Cournot best replies are quantities relationships. However, we can use the demand equation (5.18) to transform the quantity best reply (5.21) into a price relationship; we have

$$2\beta\left(a - (b+d)p_i + d p_j\right) = \alpha - \delta\left(a - (b+d)p_j + d p_i\right)$$

$$\Leftrightarrow \quad p_i = BR_i^C(p_j) \equiv \frac{(b+d)(a+dp_j)}{2(b+d)^2-d^2} > BR_i^B(p_j) \tag{5.22}$$

since $a+dp_j > 0$ (cf. (5.19)) and $\frac{(b+d)}{2(b+d)^2-d^2} > \frac{1}{2(b+d)} \Leftrightarrow 2(b+d)^2 > 2(b+d)^2 - d^2$ is true. The curve corresponding to (5.22) is drawn, together with its symmetric, as a dotted line on Figure 5.4.

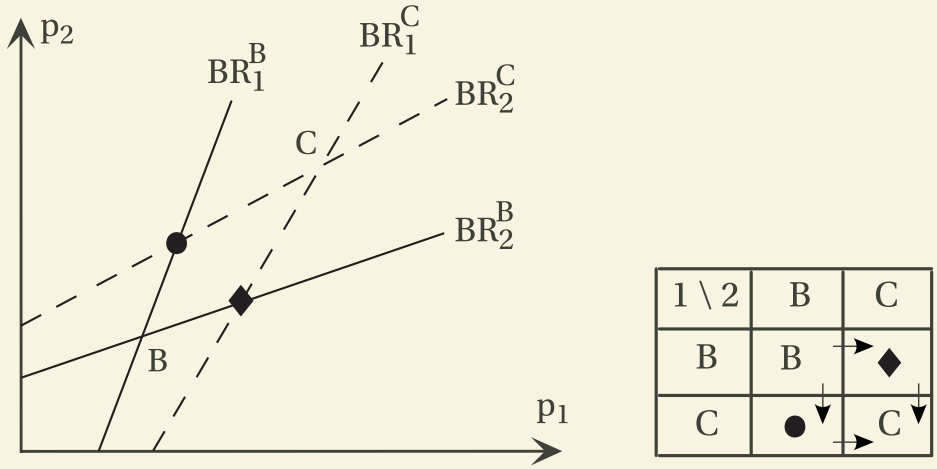


Figure 5.4: Bertrand vs. Cournot

We immediately observe that the (symmetric) Bertrand equilibrium (where best replies cross) involves a lower price than the (symmetric) Cournot equilibrium. To derive a similar ordering of quantities, we simply use the equilibrium relation $q = a-(b+d)p$ to deduce $q^C < q^B$. We can now conclude that

> In a model of duopoly competition for differentiated products, consumer surplus and welfare are greater under Bertrand competition while profits are greater under Cournot competition.

As the substitutability factor d goes to zero, the difference in (5.22) vanishes (and both prices converge to the monopoly price since the two goods become independents). When goods are complements ($d < 0$), the best reply functions are decreasing (check in (5.19) and (5.21)) but their ordering remains the same so that equilibria are ranked as before. We can therefore say that

> The Bertrand framework is always more competitive than the Cournot one, independently of whether the differentiated goods are substitutes or complements.

The intuition underlying this important result is simply that firms perceive a higher elasticity of demand (in absolute value) under Bertrand than under Cournot competition;

indeed in the first case it is $(b+d)\frac{p_1}{q_1}$ while in the second case the inverse elasticity is $\beta\frac{q_1}{p_1}$ leading to a smaller price-elasticity of $\frac{1}{\beta}\frac{p_1}{q_1}$ (independently of the sign of d).[15@]

Choosing the mode of competition †

The second result of Singh and Vives (1984) is that firms prefer to behave in a Cournot fashion (fix quantities) over a Bertrand fashion (fix prices) if they are free to choose. This is shown in a limited version of Grossman (1981)'s contract competition (cf. §6.1.7) where firms can choose one but only one of the following contract with their clients: guarantee a price or a quantity.[16@] This leads to a simple 2×2 matrix game illustrated on the right pane of Figure 5.4 for which we already know the payoffs when the choices are identical i.e., (B,B) or (C,C). Indeed the equilibria are shown on Figure 5.4 at the intersection of best reply curves and we know that firms prefer high equilibrium prices. We only have to study the case of heterogeneous choices to be able to conclude.

In case of heterogeneous choices, the price player, e.g. #1, sets his price p_1 to maximizes p_1q_1 where $p_1=\alpha-\beta q_1-\delta q_2$ due to the commitment of firm 2 over q_2; hence firm 1 maximizes $\frac{1}{\beta}p_1(\alpha-\delta q_2-p_1)$ and sets $p_1=\frac{\alpha-\delta q_2}{2}$. Likewise, the quantity player (#2) sets his quantity q_2 to maximize q_2p_2 where $q_2=a-(b+d)p_2+dp_1$, hence maximizes $\frac{1}{b+d}q_2(q_2-a-dp_1)$ and sets $q_2=\frac{a+dp_1}{2}$. Once again these best reply cannot be drawn directly in the prices space, thus we use equations (5.18) and (5.20) to observe that the price player (#1) plays the Cournot best reply (expressed in prices) $p_1=BR_1^C(p_2)$, while the quantity player (#2) plays the Bertrand best reply $p_2=BR_2^B(p_1)$. The ensuing equilibrium is easily identified on Figure 5.4 as the diamond while the equilibrium where the firms adopt the reversed roles is the large dot. The proof is now trivial: independently of firm 1's commitment, firm 2 benefits from switching its own commitment towards quantity (C) since the final price increases from B to ♦ against a price player and from • to C against a quantity player. This preference is shown by the small arrows in the matrix game of Figure 5.4. The conclusion we can draw from this analysis is that

> Firms are likely to favor investments or technology choices that limit or slow down their price reactiveness i.e., turn them into less aggressive competitors.

5.3 Contract Competition

When firms compete for clients by offering multi dimensional contracts involving price, quantities and quality features, the distribution of consumers among firms obeys the maximum expected utility principle. Taking into account the intrinsic differences between rival suppliers (cf. differentiation), consumer flock towards that offering the great-

est level of satisfaction. This process is a generalization of Bertrand competition for an homogeneous product. In this section, we illustrate this process with several examples.

5.3.1 Supply Function

We first inquire a form of competition lying between the Bertrand and Cournot benchmarks. Recall indeed that in the former, a firm proposes a price and agrees to supply any quantity whereas in the latter, a firm commits a quantity but agrees to whatever price the market dictates. Many electronic markets are organized in a less extreme manner, as firms are required to propose quantity-price contingent plans i.e., blocks of units at specific prices.

A Contractual View

Grossman (1981) shows how Cournot and Bertrand competition can be related within a common framework putting *contracts* at the center stage and how the whole competition process transmits to the entry decision.

Under Cournot competition, the monopoly quantity may be large enough to make the price lower than the entrant's average cost. In that case, the entrant stays out and the incumbent enjoys monopoly rents. In this story, when the incumbent announces he will supply the monopoly quantity, the potential entrant believes that there is no way to steal his customers. How can she hold such a belief? After all, the challenger could advertise a contract guaranteeing a lower price to customers accepting to switch provider; this way it would be possible to seize a portion of the incumbent's market share. Yet, the latter will himself react to this threat, for instance, by signing an exclusive dealing agreement with his customers (an example of which is studied in §9.2.1). In that case, the entrant is right to believe she cannot expand her market share.

Notwithstanding, if such binding exclusivity agreements are impossible to write or too costly to enforce or are simply illegal, then the entrant can effectively get the whole market with her undercutting price. We are therefore back into the Bertrand story and the ensuing price war. Monopoly is not anymore an equilibrium outcome since the entrant will step into the market. To sum-up, Cournot appears to be the right competition framework if binding contracts with customers are available, otherwise Bertrand is the right approach.

Credible Contracts under Certainty

Grossman (1981) then adds Edgeworth's original critique of Bertrand competition: who-

ever pretends to sell a large quantity at a price lower than her own average cost goes bankrupt; in other words, a seller will ration potential clients to sell only profitable units.[17@] The consumers may therefore anticipate that the total quantity supplied by a firm will, in fact, be a function of the lowest price in the market. A typical contract will be an exclusive dealership for a quantity $\bar{q}$ at a given unit price $\bar{p}$ which does not leads to bankruptcy i.e., such that $\bar{p} \geq AC(\bar{q})$ and the guarantee of meeting any lower price that would be posted in the market. A general contract is thus a supply function $\hat{q}(.)$ linking prices to quantities and satisfying the no-bankruptcy condition. The Cournot conjecture is the extreme case where the supply function is flat ($\hat{q}(p) = q, \forall p$) while in the Bertrand case it is vertical ($\hat{p}(q) = p, \forall q$).

Within this framework, Grossman (1981) shows that the outcome of perfect competition is also an outcome of the supply curve competition and conversely, that a supply curve equilibrium yields a perfect competition outcome. Surely, this is an elegant result but of scant use to draw policy implications because too many things can happen; indeed, the game at stake between the firms has too many equilibria. A refinement or more ambitious theory is needed to disentangle the issue.

Uncertainty

The reader certainly recalls that the supply of a perfectly competitive firm is a (supply) function while firms with market power propose a quantity-price pair (in fact either of them since the other one is determined by the residual demand of the firm). The reason for this discrepancy is well known but worth restating here: the competitive firm ignores what the equilibrium price will be, thus she prepares herself for any contingency, for any possibility that might arise and offers one quantity for every single price that may become the equilibrium price. The firm with market power, on the other hand, anticipates the market demand and the supply of her competitors (if any); she is thus able to estimate with exactness her residual demand, so that she only has to pick the optimal price or quantity that maximizes her profit over that demand.

The observation made by Klemperer and Mayer (1989) is that many market situations involve a degree of uncertainty regarding demand, at least at the time where the firm's strategy must be committed. One fundamental reason, studied in Part F, is the internal organization of firms: adjustment cost, communication failures or managerial incentives combine to force top management to set fixed rules of marketing for the lower levels of management; these decisions implicitly determine a supply function. Examples are services such as consulting, law, software programming or any other qualified activity sold by specialized firms; they often quote a per hour price but the real allocation of people on tasks by the general manager depens on the global amount of business. If there

is some slack then more hours are worked than billed but if business is tight, all extras such as travel time or intial negotiations are added to the bill. All in all, the greater the contracted time, the greater the average price, that is to say, the policy designed by the general manager gives rise to a supply function.

Another domain where supply functions naturally appear are the markets for services with unexpected variations of demand such as electricity,[18@] hotels or airplanes. Producers cannot bid ever changing quantity-price pairs that adjust in real time to the changes in demand; rather they must submit a series of strategies that cover all relevant possibilities. In the case of electricity, a firm bids a series of pair (q_1, p_1), $(q_2, p_2), (q_3, p_3), ...$ where the first q_1 units are offered at the unit price p_1, the next q_2 units are offered at the (greater) unit price p_2, the next q_3 units are offered at the (still greater) unit price p_3 and so on (these quantities often match the capacities of plants). The average price for a production of units $q \leq q_1$ is p_1, while it is $\frac{p_1 q_1 + (q-q_1)p_2}{q}$ for $q_1 < q \leq q_2$, $\frac{p_1 q_1 + q_2 p_2 + (q-q_1-q_2)p_3}{q}$ for $q_2 < q \leq q_3$ (and so on). The relation between average price and quantity is a supply function $s(p)$ stating how much the firm is willing to supply for the unit price p.

Klemperer and Mayer (1989) prove that the supply function equilibrium is unique and lies between the Cournot and Bertrand outcomes.

> Demand uncertainty reduces the market power of active firms; they are forced to propose supply functions rather than fixed quantities; as a result, the equilibrium is more competitive.

Formal Analysis †

To study formally the competition between two firms in the presence of uncertainty, it is enough to assume that the market size parameter a in the demand $D(p) = a - bp$, is unknown to the firms at the time where they set their pricing strategy. The common technology of firms displays DRS i.e., the cost function is increasing convex $C(q) = F + \delta q^2/2$.[19@] Firms choose their supply functions s_i and s_j before learning the shock parameter a. Since both supply functions are increasing, their sum is also increasing while the demand is decreasing; hence there is a unique market price p equating demand and supply i.e., solving $a - bp = s_i(p) + s_j(p)$.[20@]

Given the strategy s_j of the opponent, the residual demand for firm i is $a - bp - s_j(p)$ so that her profit is $\pi_i(p) = p(a - bp - s_j(p)) - C\left(a - bp - s_j(p)\right)$ and the FOC characterizing the optimal price is then

$$0 \quad = \quad a - bp - s_j(p) - \left(p - \delta\left(a - bp - s_j(p)\right)\right)\left(b + s_j'\right)$$

$$\Rightarrow \quad s_j'(p) + b = \frac{a - bp - s_j(p)}{p - \delta\left(a - bp - s_j(p)\right)} \tag{5.23}$$

Equation (5.23) determines an optimal price p_a^i that depends on the current shock a together with a quantity $q_a^i \equiv a - bp_a^i - s_j(p_a^i)$. As the parameter a varies, the pair (p_a^i, q_a^i) describes a curve in the price-quantity space which is firm i's optimal supply function $s_i^*(p)$; it catches the idea that for every possible macro-economic contingency, there is an optimal response to the competitor's supply function. Notice that equation (5.23) can be written

$$s_j'(p) + b = \frac{s_i(p)}{p - \delta s_i(p)} \tag{5.24}$$

The same analysis for the other firm yields an optimal supply function $s_j^*(p)$. Let us look for a linear symmetric supply function (SF) equilibrium i.e., where $s_i(p) = s_j(p) = \alpha + \beta p$ thus (5.24) becomes

$$\beta + b = \frac{\alpha + \beta p}{p - \delta\alpha - \delta\beta p} \Leftrightarrow \left[\left(\beta + b\right)\left(1 - \delta\beta\right) - \beta\right] p = \alpha\left(\delta(\beta + b) + 1\right) \tag{5.25}$$

Since the polynomial in (5.25) must be nil for all p, it must be the nil polynomial i.e., all its coefficient must be zero. From the LHS, we derive[21@] $\beta = \frac{1}{2}\left(-b + \sqrt{b^2 + 4b/\delta}\right) > 0$ while the RHS yields the necessary condition $\alpha = 0$ because $\delta(\beta + b) + 1 \neq 0$.

The slope of this equilibrium supply function is intermediate between the Cournot and Bertrand cases which are respectively flat and vertical. Finally, the equilibrium price solves $a - bp = s_i(p) + s_j(p) = 2\alpha + 2\beta p$ so that $p^{SF} = \frac{a}{2\beta + b}$. The Bertand equilibrium price cannot be determined here due to Edgeworth cycles; however when δ goes to zero, marginal cost tend to zero, the SF equilibrium slope β tends to infinity driving the SF equilibrium price towards zero i.e., we are back to a Bertrand-like outcome. The Cournot price can be taken from equation (6.8), it is $p^C = \frac{a(1+b\delta)}{b(3+b\delta)}$ which is greater than the SF equilibrium price.[22@]

5.3.2 Clubs

The economic notion of a club good, as reviewed by Scotchmer (2002) usually refers to public goods and public services jointly consumed.[23@] The Public Economics approach takes a cooperative and pro-active view and concentrates on the formation and stability of the club together with its financing under a series of geographical or jurisdictional constraints.

From our perspective in this book, clubs relate to public utilities (cf. §17.1.2) and private services building on a network of customers such as recreational or cultural clubs.

From an analytical point of view, competing clubs offer contracts to their prospective members with a view to raise their utility and sway their endorsement. We are thus faced with a particular brand of oligopolistic contract competition.

Tradition sees club goods as services naturally displaying a positive externality through *network* effects. Yet, they may also suffer a negative externality, *congestion* when membership grows too much because at any point in time, a club has a limited capacity of (instantaneous) service (cf. §25). From the industrial organization viewpoint, a club is more than a local association (book, wine, chess or sport club), it includes any service delivered through a minimum size facility (e.g., movie theater) or infrastructure (e.g., highway). At the heart of any club formation is the recognition that members confer positive externalities onto each other but also that a minimum (critical) size must be reached for two reasons: to start-up the network effect and to finance of the fixed cost of club creation. By its very nature, a club can serve many users simultaneously at near zero marginal cost; the club service is then "non rival", which is why it resembles a public good (but different as the latter holds the further property of being "non-excludable").

From the point of view of a user, aggregate membership (of the club) is a quality attribute. At low levels, the relationship may be positive because we enjoy better service in community than alone,[24@] but for levels close to capacity, congestion kicks-in and aggregate use undoubtedly becomes a negative quality attribute. Yet, rather than being exogenously chosen by the service supplier (cf. §3.3), aggregate use is endogenously determined by the equilibrium behavior of potential users. The thrust of the problem lies over the congestion range because the manager, whether a welfare or profit maximizer, will always try to motivate use or membership close to capacity level since it raise utility (thus WTP) while cost is barely affected.

To study this issue, Scotchmer (1985b) assumes that competition among clubs is fierce enough to maintain the overall price of the service below the WTP of consumers so that the market is covered in equilibrium. One can then inquire about how many firms will be active, how many clients will they have and how much will these consume. Modeling depends on whether the facility size is stretchable or equivalently whether investment is lumpy. The dichotomy could be between a golf club (fixed size) and a tennis club (variable size) or between a highway and an optic fiber network. In any case, we may define the willingness to pay for the service (WTP) and the willingness to avoid congestion (WTA) which are assumed decreasing and increasing, respectively. Hereafter, total WTP and WTA sum these monetary valuation over club members.

We develop the variable capacity case. Utility $u(k,n)$ depends negatively on membership n (crowding) and positively on the facility size k whose (per period) cost is $C(k)$. We have WTP $= u_k$ and WTA $= -u_n$. Let us derive first the efficient organization.

The welfare objective is to maximize the utility of a representative consumer net of his fair share of cost i.e., $W = u(k,n) - \frac{C(k)}{n}$. The club size FOC yields the Samuelson equation[25@]

$$\frac{\partial W}{\partial k} = 0 \quad \Leftrightarrow \quad C_m = nu_k \quad \Leftrightarrow \quad \text{MC = TWTP} \tag{5.26}$$

while the membership FOC yields[26@]

$$\frac{\partial W}{\partial n} = 0 \quad \Leftrightarrow \quad \frac{C(k)}{n} = -nu_n \quad \Leftrightarrow \quad \text{AC = TWTA} \tag{5.27}$$

The solution of (5.26) is k_n^*; plugging into (5.27) , we derive the efficient membership n^*. To cover the entire unit mass population efficiently, we thus need $\frac{1}{n^*}$ clubs.

Let us now consider m competing clubs independently choosing their size k and membership fee f. Given these, consumers allocate themselves among clubs, seeking the greatest utility. In equilibrium of this process, all utilities are equalized to $\bar{u}$.[27@] If a club has n members, each derives utility $u(k,n) - f = \bar{u}$ and the club's profit is then $\pi = nf - C(k) = n(u(k,n) - \bar{u}) - C(k)$. The capacity FOC is

$$C_m = n\frac{\partial f}{\partial k} = nu_k \tag{5.28}$$

since capacity k has no direct bearing upon utility elsewhere $\bar{u}$. The optimal size, conditional on membership, is the efficient one k_n^* since (5.28) reduces to (5.26). This is an important property of contract competition: once buyers become customers, their relationship with the firm is Pareto efficient because the fee can be used to share welfare so that other variables can be set so as to maximize welfare.[28@] Lastly, the membership FOC is

$$f = n\frac{\partial f}{\partial n} \tag{5.29}$$

If a club wishes to attract more clients, she is bound to increase local congestion thus she must reduce her fee to avoid loosing all of her clients. Furthermore, as competitors receive now less clients, congestion is reduced and satisfaction is increased at these clubs. This forces the deviator to further decrease her fee. This equilibrating process implicitly define the relationship between membership and fee. To uncover it, suppose all but one firm play a candidate equilibrium $(\hat{f}, \hat{k})$ while the remaining one adjusts f to receive n clients (having picked the size k_n^* in anticipation before). Other firms then receive $\frac{1-n}{m-1}$ clients and the market clearing equation is $u\left(\hat{k}, \frac{1-n}{m-1}\right) - \hat{f} = u(k,n) - f$. Differentiating wrt. membership base n yields

$$\frac{-1}{m-1}u_n\left(\hat{k}, \frac{1-n}{m-1}\right) = u_n(k,n) - \frac{\partial f}{\partial n} \quad \Leftrightarrow \quad \frac{\partial f}{\partial n} = \frac{m}{m-1}u_n \tag{5.30}$$

at the symmetric equilibrium where all actions are identical.

Under free entry, profits are gradually eroded and drive the fee tends toward the average cost so that using (5.30), (5.29) becomes

$$\frac{C(k)}{n} = f = -\frac{m}{m-1} n u_n \tag{5.31}$$

The comparison with (5.27) is immediate: the equilibrium fee exceeds the congestion cost by the ratio $\frac{m}{m-1}$; the solution is thus lesser[29@] than n^* meaning that the optimal club size is smaller than required by efficiency. As an ultimate consequence, too many firms enter the market.[30@]

As in the standard Cournot competition, the market power afforded to firms by their peculiar mode of rivalry allows them to earn economic profit and thus attracts too many in the field. Note however that as the cost of operating a club falls (due to technological progress), the efficient club size falls i.e., the efficient number of clubs rises so that the ratio $\frac{m-1}{m}$ converges to unity meaning that an almost correct number of firms will enter. The policy recommendation flowing from the model is to limit entry in facility based services (i.e., prone to congestion) only when the required investment for building a minimum size facility is so large that the efficient supply structure is close to a natural monopoly.

In the golf club case (or amusement park), the service capacity is fixed but it makes sense to allow members to vary usage which can then be priced. The constrained Pareto optimality arising from contract competition allows firms to apply marginal cost pricing. In the end, competition builds on the membership fee only. The analysis then follows along the previous lines. Scotchmer (1985a) shows that the first-best and the free entry equilibrium equations are those characterized above.

5.3.3 Platform Economics

The novel "platform" terminology aims to convey the idea that competition for modern bundled goods differs sensibly from the standard modes of either price or quantity competition.

Introduction

As can be seen from the list of examples below, technological progress creates a new long-term relationship between a firm and its customers. What is at stake is a package made of a fixed asset, the *platform*, together a flow of services.[31@]

Game Console Device and game cartridges (Playstation, Gameboy, Xbox)

Printer Device and ink cartridges (e.g. HP, Lexmark)

Nespresso Coffee machine (build by several makers) and patented Nestlé capsules

iPod Portable device and songs bought from iTunes store

iPad Portable device and digital contents bought from e-retailers

Mobile Phone Device and usage plan

TV Device and distribution plan for channels (cable, satellite, DVB)

Health membership and plan with a health maintenance organization (HMO)

Car Automobile and repair parts (mostly electronic)

Software Initials release vs. upgrades and support services

The key characteristic of a platform is that the asset and the services are complementary in a way that technology makes unique and hard to emulate. This bestows the vendor of the platform an exclusivity for the sale of future services, at least until the customer terminates the entire relationship. Given this lock-in of customers, firms compete vigorously to sign them up with schemes colloquially referred to as "bargain-then-ripoff". One also speak of "fore-market competition" followed by "aftermarket monopolization". Indeed, the asset is often sold at a price close or even below cost to lure clients but these eventually suffer high markups over services given that the "aftermarket" is monopolized. The phenomenon is so strong that it can even happen with a monopolist as shown in §4.3.2. Because firms compete for customer "before the market", the standard conclusions of price theory for competition "in the market" are altered.

A crucial distinction for the analysis of platforms is whether the firm can commit or not to future prices for services. If so, the bilateral relationship between a "committed" firm and a client is Pareto efficient i.e., generates maximum welfare. Profits for the firm and utility for the client are then shared by sliding the initial asset price as a function of competitive pressure in the market and outside opportunities for the client. The most frequent case, however, is when firms are unable to commit to a future pricing behavior; they cannot help but monopolize the aftermarket with high service prices (cf. §14.1.3). Since this behavior reduces the consumer surplus, the overall WTP for the platform falls and so do profits, independently of whether the firm is a monopoly or an oligopolist. This shows that, in a world devoid of risk, firms would like to commit over their future prices. If they so frequently pass on this opportunity, it must be due to uncertainties regarding the evolution of input prices, technological development or future demand. We analyze the two situations in turn before comparing them.

Committed pricing A firm has unit cost γ for the asset and c for services, profit is thus $\pi = (f - \gamma + (p-c)q)D(f,p)$ where q is the service demand of a customer facing price p and

$D(f,p)$ the market share of the firm when she offers the bundle (f,p).[32@] If the firm can commit to future prices it is optimal to set $p=c$ so that profit becomes $\pi=(f-\gamma)D(f,p)$ and the FOC is $(f-\gamma)\frac{\partial D}{\partial f}+D=0\Rightarrow\frac{f-\gamma}{f}=\frac{1}{\epsilon_f}$. We thus observe that when firms compete for customers through the access fee only, profit maximization leads to equate the Lerner Index to the inverse elasticity of demand (cf. eq. (3.4)) with the important proviso that the traditional unit price is replaced by the access fee.

As product differentiation lessens or as the number of firms increases, the elasticity of access demand facing any single firm is likely to increase, causing the equilibrium fixed fee and price-cost margin to fall.

Uncommitted pricing If the firm is unable to commit to future prices, then after-market monopolization occurs in the sense that the firm cannot help but sell her flow services at the standard monopoly price i.e., "rip-off" her customers. The FOC is commonly expressed with the Lerner index as $\frac{p-c}{p}=\frac{1}{\epsilon_p}$. Consumption and consumer surplus are greatly reduced (wrt. the previous case). In anticipation, clients require a very low entry price or even a subsidy. Indeed, the FOC for setting the access fee is $f-\gamma+(p-c)q=1/\frac{\partial lnD}{\partial f}\Rightarrow\frac{f-\gamma}{f}=\frac{1}{\epsilon_f}-\frac{pq}{f}\frac{1}{\epsilon_p}$ (using the other FOC). The Lerner index for the entry fee is a modified inverse elasticity because the impact of after-market monopolization must be subtracted. The margin may be negative meaning that the asset becomes a loss leader .

Examples abound with mobile phone plans including a free high-tech phone bundled with a two-year permanence and minimum billing, Nespresso's affordable gorgeous machines and expensive coffee capsules, HP's cartridge set which is more expensive than the printer or Brita's water jug cheaper than a pack of filtering cartridges.

Comparison The absence of commitment generates a deadweight loss from the monopoly pricing, but competition for customers puts strong downward pressure on access fees that conceivably could result in higher overall market penetration which is socially desirable. Yet, insofar as there are no serious income effect, the committed case yields more utility thus more participation and a larger welfare. It is also unequivocally yield more profit because the per-capita profit is larger.

If the per-capita earning is the same then client's expenses are identical and the "committed" client must be happier as he consumes the efficient amount.[33@] We may then increase the "committed" firm's access fee until the client is equally happy under either scheme; thus the "committed" firm earns more per-capita. However, equal utility implies the same market share and the same sensitivity to the fee[34@] and equality of $\frac{\partial lnD}{\partial f}$ for both kind of firms. In order for the FOCs to be satisfied, it must the case that

the "committed" fee is lower (since $\frac{\partial lnD}{\partial f}$ is decreasing in fee). This means that the client under "commitment" must enjoy a higher utility in equilibrium. As a consequence, fewer people choose the outside option so that market coverage is greater.

5.3.4 Competition and Discrimination

In this section on oligopolistic contract competition, we have expanded the "weapons" from price or output to (potentially complex) contracts. Among those, we find elaborated tariff structures such as volume discount, subscriptions or versioning which are all instances of differential pricing. This topic is thoroughly in chapter 4 for the case of firm holding market power and facing an exogenous population of potential clients. In the present paragraph, we study the interaction between competition and complex (non linear) pricing pricing following Stole (2007) and Armstrong (2007) .

Regarding welfare or the social desirability of price discrimination, we may look at the impact of entry against a discriminating monopoly or at the repercussion of allowing price discrimination in an oligopoly.

Perfect Discrimination

Spence (1976b) argues that when a seller can price discriminate perfectly, she captures her marginal contribution to consumer surplus so that her profit coincides with her marginal contribution to welfare. As a producer, she is lead to choose her product variety, quality and quantity so as to maximize welfare. Entry decisions are then efficient since a firm enters the market if and only if her profit which is her marginal contribution to welfare exceeds the entry cost. This is entirely true for a monopoly but in oligopolistic situations care must be taken.

Regarding pricing, one must understand that perfect discrimination refers in fact to the greater availability of contractual means for buyers and sellers to negotiate their deals (cf. §4.1.3 with the many examples of consumer surplus extraction schemes). A result ubiquitous of many models is that once firms are able to offer rich contracts, these induce an efficient bilateral relationship with each client. In essence, bargaining over the terms of the contract yields an efficient result, the Pareto optimal barter of the Edgeworth box (cf. §2.4.3).[35@] Perfect price discrimination thus allows firms to capture their marginal contribution to welfare. The non cooperative choice of characteristics are then efficient since firms face the same incentives as a social planner.

Entry

As shown by Bhaskar and To (2004), there is too much entry although a potential entrant's profit is his welfare contribution. The hidden distortion is that this welfare contribution is computed for an inefficient bunch of characteristics on the part of the incumbents, so that it is greater than if the optimal reconfiguration was used as we now proceed to show in a particular case.

In the circular city location model of Salop (1979) (cf. §11.1.3), perfect discrimination lead firms to compete in a Bertrand fashion over every single customer. As intuition would suggest, the firm closest to a consumer wins him: efficiency obtains. The equilibrium price is equal to the best that the second-nearest could offer which ends up being the difference in transportation cost. Indeed, consider consumer located at x between firm A located at 0 and firm B located at $\frac{1}{n}$, his utility is either $u_A(p_A) = v - tx - p_A$ or $u_B(p_B) v - t(\frac{1}{n} - x) - p_B$. Now, when $x < \frac{1}{2n}$, the best that B can offer is $u_B(0) = v - t(\frac{1}{n} - x) \leq u_A(p_A) \Leftrightarrow p_A \leq t(\frac{1}{n} - x) - tx$ which is the optimal price offered by A to win the customer over B. The producer surplus is then $2\int_0^{1/2n} \frac{t}{n} - 2tx\,\mathrm{d}x = \frac{t}{2n^2}$. Taking into account the fixed cost of entry F, profit is $\frac{t}{2n^2} - F$ and the free-entry number of firms is $\sqrt{\frac{t}{2F}} = \sqrt{2}n^*$, where n^* is the efficient level characterized in §6.1.4. Since free entry without price discrimination leads to even more entry ($\bar{n} = 2n^*$), we may say that price discrimination reduces excessive entry but comes short of bringing full efficiency.

Imperfect Discrimination

As shown in §4.2.2, if firm discriminate among segments but are bound to use uniform pricing within segments then each segment is a standard oligopolistic market where goods may be homogeneous or differentiated.

If firms have the same technology and demand in segment $\#j$ is the homogeneous $Q_j = a_j - b_j p_j$, the oligopolistic equilibrium is $Q_j^* = \frac{n}{n+1}(a_j - b_j c)$, thus total output is $Q_0^* \equiv \sum_j Q_j^* = \frac{n}{n+1}(a_0 - b_0 c)$ (with obvious notation). Observe then that in the absence of discrimination, total demand is $Q_0 = a_0 - b_0 p_0$ and the equilibrium is Q_0^*. Thus, discrimination does not change total output in an homogeneous market, only the identity of buyers which is inefficient since low prices in elastic segments mean that people will low WTP buy the good while people with higher WTP fail to do so if they belong to low elasticity segments (where the price is set higher).

We may thus conclude that price discrimination tends to reduce welfare as it generates misallocation among consumers. More generally, when demand curves are nonlinear or some markets would not be served under uniform pricing, price discrimination may increase welfare as it may trigger a greater covering of the market.[36@]

The effect on output of allowing discrimination in a duopoly market is not clear. Holmes (1989) shows that if the ratio of cross price elasticity to monopolist's elasticity is greater in the strong segment then total output increases because output falls less in the strong segment than it rises in the weak one. If the reduction in the strong segment is sufficiently small relative to the weak one, then welfare will also rise. Even less can be said of how equilibrium profit in the industry changes with the introduction of price discrimination (cf. sec 3.3 in Stole (2007)).

Chapter 6

Strategic Moves

Oligopoly with Sequential Decisions

The simultaneous competition featured in the previous chapter, either through prices, quantities or contracts is only the last stage of the global rivalry that takes place among oligopolists i.e., once firms have chosen their technology and entered the market. In this chapter, we broaden our viewpoint and analyze entry, production and sales as sequential decisions whose rationality turns them into *strategic moves*.

Our first section is devoted to the analysis of *entry* under a variety of market structures and conducts but with a strong symmetry hypothesis: all firms, incumbents and potential entrants alike, have access to the same technology. We derive the long term equilibrium in each case and assess its efficiency.

We then consider the opportunity for a firm to make a commitment such as producing and dumping a given quantity on the market using the Stackelberg (1934)'s leadership model. The motive for such a decision was seen in the previous chapter: a cost advantage translates into a profit advantage; thus, each firm will try to invest into a better production system or a better distribution network to gain that edge over the rest of the herd. The next sections then analyze variations around this issue with forward contracting (i.e., separation of production and sales) and vertical relationships as a mean to commit to a specific behavior.

6.1 Dynamic of Entry and Exit

The reduction of concentration through the entry of new firms in markets is key to foster competition and innovation among all contenders. This whole process is desirable as it eventually advances economic growth and general welfare. We study in this section the conditions under which a potential challenger takes the irrevocable entry decision i.e., spending the sunk cost that enables him to become an effective competitor in the mar-

ket. The study of this long term market dynamic under a variety of market structures and conducts will serve as a reference for §10.2 where we tackle the possibility that incumbent firms establish barriers to entry. Recall indeed that the less concentrated is the market, the more intense is the competition and the lesser are the profits which ultimately means that entry harms incumbents.

The welfare concept we have adopted in equation (2.23) is contingent on the current actors of the market. Indeed, consumer surplus derives from the demand which itself sums the individual demands of economic agents standing ready to buy in that market. Similarly, the producer surplus sums the individual surpluses of active firms, those which are ready to produce and sell. When considering entry, we must extend our definition of welfare to account for the resources that the entrant spends (or save) to enter, the most important being the sunk cost that must be incurred to gain access to the market or the fixed cost that has to be disbursed in every period. Recall indeed that none of these is not captured by our producer surplus definition.[1@]

The cost function we adopt in this section derives from a decreasing returns to scale (DRS) production technology. This more complex specification is needed to treat several market structures simultaneously; it nevertheless contains, as a limiting case, the constant returns to scale (CRS) we use in most of the book.

6.1.1 Entry under Perfect Competition

Firms choose to enter a market for an homogeneous good at a fixed cost F that accounts for building a minimally efficient plant or buying a governmental license. Although unrealistic, we assume that, whatever the number of firms present in this market, they are price-takers i.e., there is perfect competition.

Let us consider a common technology for all firms characterized by decreasing returns to scale and postive fixed cost: $C(q) = F + cq + \delta q^2/2$. A first method uses the theoretical result according to which the long-term supply of an active firm is her minimum efficient scale which itself determines the long-term price (cf. §2.1.4). The former solves $AC = C_m$ and is $\bar{q} = \sqrt{2F/\delta}$ so that the latter is $\bar{p} = C_m(\bar{q}) = c + \sqrt{2\delta F}$. The limiting number of firms is then the ratio of market demand at the long-term price $D(\bar{p}) = a - b\bar{p}$ by the individual long-term supply $\bar{q}$. Solving this equation yields[2@]

$$n^* = \frac{a-bc}{\sqrt{2F/\delta}} - b\delta \tag{6.1}$$

> The sustainable number of firms in a perfectly competitive market is determined by the minimum efficiency scale and the demand.

The alternative method is to derive the full equilibrium in order to get the (same) condition for entry. The individual supply of a price-taker producer solves $p = C_m = c + \delta q \Leftrightarrow q = \frac{p-c}{\delta} \equiv s(p)$; his profit is

$$\begin{aligned} \pi(p) &= ps(p) - C(s(p)) = \left(p - c - s(p)\tfrac{\delta}{2}\right) s(p) - F \\ &= \tfrac{p-c}{2} s(p) - F \quad \text{since } \tfrac{s(p)\delta}{2} = \tfrac{p-c}{2} \\ &= \tfrac{1}{2\delta}(p-c)^2 - F \end{aligned} \tag{6.2}$$

The aggregate supply when there are n active firms is $S_n(p) = n\frac{p-c}{\delta}$. The competitive equilibrium price, equating demand $D(p) = a - bp$ and supply, is $p_n^* = \frac{nc+a\delta}{n+b\delta}$; individual sales are

$$q_n^* = s(p_n^*) = \tfrac{a-bc}{n+b\delta} \tag{6.3}$$

and profits are

$$\pi_n^* \equiv \pi(p^*) = \tfrac{\delta}{2}\left(\tfrac{a-bc}{n+b\delta}\right)^2 - F \tag{6.4}$$

Since profit is computed using economic costs (i.e., including the opportunity cost of every factor but excluding sunk cost), the relevant benchmark for a firm manager to decide whether to be active or not is zero. The limiting number of firms $n^* \geq 1$ solves $\pi_n^* \geq 0 > \pi_{n+1}^*$ and is the integer part of n^* computed in (6.1).

The comparative statics are in line with intuition: a greater market size a favors entry, as well as lower entry cost F, lower price reactiveness b and lower marginal cost factor c. When the fixed cost F vanishes, the long-term number of firms increase unboundedly i.e., DRS commands many small firms. If the technology approaches constant returns to scale ($\delta \to +\infty$), then the long-term number of firms tends to unity i.e., CRS commands a single large price-taking monopoly in order to minimize total fixed cost.

In connection with our extended definition of welfare as $W(p^*, q^*) - nF$ in the presence of n firms, one might wonder whether the efficient number of firms would not be that which maximizes this expression and not, as we previously did, that which equates (individual) producer surplus to fixed entry cost i.e., solves $\pi = F$. It occurs that both answers coincide. To check this claim, we study the effect of an additional firm on welfare $W(p^*, q^*)$. Firstly, there is the direct effect of having an additional firm whose producer surplus π contributes positively to welfare. Then, there is the indirect effect on welfare through the effect on the market equilibrium; in the case of perfect competition, the equilibrium price maximizes welfare which means that the change of price induced by entry has no first order effect on welfare i.e., the indirect effect is nil.[3@] The overall effect of entry on extended welfare is thus $\pi - F$ which is the profit of the entrant.

6.1.2 Entry under Bertrand Competition

As always with Bertrand competition, we reach a paradox: fierce price competition yields a monopoly outcome although there is free entry. The key to understand this puzzle is to note that an incumbent monopoly already paid the fixed cost F while a potential competitor must ponder the opportunity to spend F. If there is Bertrand competition in the market and firms have identical constant marginal cost c, then the equilibrium price after entry drops to c. As a consequence, both firms earn zero producer surplus; this is dramatic for the entrant because it blocks him from recouping the initial fixed cost F. Hence, because competition is fierce no one dares to enter and the incumbent can enjoy monopoly profits!

If, however, the potential entrant discovers a better technology, he will have a lower marginal cost $\underline{c}$, will enter and force the incumbent to exit. Indeed, we saw in §5.2.1 that the cheapest of the two firms in presence wins the entire market and earns a margin equal to the cost difference, here $c-\underline{c}$ for the challenger. If the resulting profit $(c-\underline{c})D(c)$ is larger than the fixed cost of entry F, then entry is unstoppable and will occur. The old monopoly is now excluded from the market which means that the entrant has become the new monopoly; this situation will last until a better technology is discovered and the same story repeats itself.

The condition for unstoppable entry $(c-\underline{c})D(c) > F$ is equivalent to an improvement in marginal cost greater than $\frac{F}{D(c)} \simeq \frac{F}{2q^M}$, since the monopoly quantity is roughly half the competitive one for linear demand and constant marginal cost. As a matter of comparison, the profit margin for the monopoly $p^M - c$ has to be greater than $\frac{F}{q^M}$ to make his presence worthwhile in this market. Thus, the cost improvement must roughly be at least half of the current monopoly profit margin.

From this analysis, we conclude that markets where price is the strategic variable and where production capacity is cheap to build (constant marginal cost) should be highly unstable; they are good examples of the Schumpeterian process of creation-destruction. When production capacities matter or equivalently, when marginal costs are increasing, things becomes more difficult to analyze as we already commented in §5.2 and a convenient competition framework is the Cournot one that we now proceed to study.

6.1.3 Entry under Cournot Competition

In this section, potential competitors share the cost function $C(q) = F + cq + \delta q^2/2$. If there is Cournot competition in the consumer market, the profit of one firm is $\pi_i(q_{-i}, q_i) =$

$q_i P(q_{-i}+q_i) - C(q_i)$ and the FOC of maximization is

$$R_m = C_m \Leftrightarrow \frac{a-q_{-i}-2q_i}{b} = c + \delta q_i \tag{6.5}$$

leading to the best reply function $q_i = BR(q_{-i}) \equiv \frac{a-cb-q_{-i}}{2+\delta b}$. The symmetric Nash equilibrium among n active firms solves

$$q = BR\big((n-1)q\big) \Leftrightarrow q(2+\delta b) = a - bc - (n-1)q \tag{6.6}$$

and is

$$q_n^C \equiv \frac{a-bc}{\delta b+n+1} < \frac{a-bc}{\delta b+n} = q_n^* \tag{6.7}$$

seen in equation (6.3).

We immediately deduce that total supply nq_n^C is lower than in the competitive equilibrium (for the same number of firms), thus the equilibrium price is greater: $p^C > p^*$. More precisely,

$$p_n^C \equiv \frac{a(\delta b+1)+nbc}{b(\delta b+n+1)} \tag{6.8}$$

Finally, using equation (6.4) we obtain the individual profit:

$$\pi_n^C = \frac{1}{2\delta}(p_n^C - c)^2 - F > \frac{1}{2\delta}(p_n^* - c)^2 - F = \pi_n^* \tag{6.9}$$

using equation (6.2). We can therefore conclude after von Weizsäcker (1980) that the limiting number of firms

$$n^C = \frac{\delta + 1/b}{\sqrt{2\delta F}}\left(a - bc - b\sqrt{2\delta F}\right) \tag{6.10}$$

is larger than n^* since the lesser competition yields higher individual profits and allows for more firms to be active than would be efficient. Even in the limiting case of constant returns to scale ($\delta \to 0$), the Cournot price p_n^C tends to $\frac{a/b+nc}{n+1}$ and remains above the marginal cost c, so that firms keep earning extraordinary profits or rents.

> A market characterized by Cournot competition tends to be flooded by a too large number of firms charging too high prices.

Notice that two opposite forces are at work; on the one hand, entry is welfare enhancing through the increase of the entrant's profits but on the other hand, entry is welfare reducing through a "business stealing" effect: in reaction to entry, all incumbent firms strategically reduce their production.

The intuition held for a long time in the economic profession according to which "free entry is conducive of a higher degree of efficiency" is in fact wrong. It is misleading to

dissociate the issue of pricing from that of entry because both of them matter to assess the efficiency of a market. True, the entry of a new competitor increase rivalry in the market and leads to lower prices and greater consumer welfare but this is only one part of market welfare, our efficiency concept. There is an indirect effect of entry that was neglected in early studies, the reaction of incumbent firms to entry. If the pricing behavior of incumbents is such that, the reaction to new entry is to increase price or reduce individual production, then entry surely reduces welfare because the reduction of profits for the incumbents outweighs the gain in consumer surplus. Somehow, the entry of "bad" economic agents such as firms with market power can be a source of additional inefficiency.

6.1.4 Hotelling Competition

The Hotelling duopoly model of differentiation (cf. §5.2.2) is extended to oligopoly in §11.1.3. This circular city model of differentiation is amenable to the computation of entry. When n firms are active, the equilibrium price is $\frac{t}{n}$ and each firm serves her fair share $\frac{1}{n}$ of the market earning profit $\pi_n = \frac{t}{n^2} - F$ where F is the entry cost. The limiting number of firms is then $\bar{n} = \sqrt{\frac{t}{F}}$.

Welfare, however, commands to minimize total cost made of fixed cost nF and transportation cost since the market is closed and each consumer buys a single unit. Now, there is a distance $\frac{1}{n}$ between any two firms so that the maximum distance travelled is $\frac{1}{2n}$ (indifferent consumer located at mid-distance). Since people are uniformly distributed, the average distance travelled is half the maximum i.e., $\frac{1}{4n}$. The total transportation cost is thus $C = \frac{t}{4n}$ and the FOC for minimizing $nF + C$ has solution $n^* = \sqrt{\frac{t}{4F}} = \bar{n}/2$.

Consider lastly the case of a cartel among the n shops (cf. §9.1). To guarantee market coverage, the monopoly sets the price $p^M = \bar{p} - \frac{t}{2n}$ that leaves exactly his opportunity cost to the consumer living at mid-distance (i.e., $\frac{1}{2n}$) between any two contiguous shops. The monopoly thus cares for the maximum cost $\frac{t}{2n}$. To decide on the optimal number of shops (or brands), the monopoly compares the fixed cost of setting up one more shop F with the increase in price $\frac{t}{2n} - \frac{t}{2(n+1)}$ that he will be able to apply. The condition for opening one more shop is $n(n+1) < \frac{t}{2F}$. The approximate optimal number is thus $n^M = \sqrt{\frac{t}{2F}}$.

> We may thus conclude that, in the Hotelling framework of horizontally differentiation, a monopoly opens some 41% more shops than what efficiency commands but that open access competition leads to an even stronger 100% excess opening. Once we make the alternative interpretation of transportation cost as opportunity cost for characteristics (cf. §11.2.1), we may say that competition leads to variety proliferation that even

monopolization would fail to stop.

6.1.5 Generalization †

In a more general setting of imperfect competition, Mankiw and Whinston (1986) show that entry is excessive if, in a symmetric equilibrium, the individual quantity decreases with entry. As we just saw, Cournot competition is one such case. To prove this result formally, denote q_n and Q_n the individual and total sales in a symmetric equilibrium with n active firms. The individual firm equilibrium profit is $\pi_n = q_n P(Q_n) - C(q_n)$, hence

$$\frac{\partial \pi_n}{\partial n} = \frac{\partial \pi_n}{\partial q}\frac{\partial q_n}{\partial n} + \frac{\partial \pi_n}{\partial Q}\frac{\partial Q_n}{\partial n} = (P - C_m)\frac{\partial q_n}{\partial n} + q_n P' \frac{\partial Q_n}{\partial n} \tag{6.11}$$

Market welfare in the presence of n firms is $W_n \equiv n\pi_n + W_D(Q_n)$. We use equation (2.20) to write $\frac{\partial W_D}{\partial Q} = -Q_n P'(Q_n) = -nq_n P'(Q_n)$ since $q_n = Q_n/n$ in a symmetric equilibrium, thus

$$\frac{\partial W_n}{\partial n} = \pi_n + n\frac{\partial \pi_n}{\partial n} + \frac{\partial W_D}{\partial Q}\frac{\partial Q_n}{\partial n} = \pi_n + n(P - C_m)\frac{\partial q_n}{\partial n} < \pi_n \tag{6.12}$$

using (6.11) since we assumed $\frac{\partial q_n}{\partial n} < 0$ and a rational firm adjusts its production (may be using rationing) to maintain the price above its marginal cost.

We see in equation (6.12) the two effects alluded before, the (positive) entrant's profits and the (negative) "business stealing" effect. Entry will occur until $\pi_n = 0$ is satisfied[4@] while it ought to stop sooner when $\frac{\partial W}{\partial n} = 0$ is satisfied. All in all, there is excess entry.

The same idea can be pursued even when products are not perfect substitutes i.e., in the presence of product diversity. It is however necessary to adopt a convenient formulation of consumer surplus like the one proposed by Spence (1976a). Letting $\vec{q} \equiv (q_i)_{i \le n}$ be the production of the n firms, the gross consumer surplus (net of price paid to sellers) is $W_D(\vec{q}) = \Phi\left(\Sigma_i \xi(q_i)\right)$ where Φ and ξ are concave increasing. Total welfare in a symmetric equilibrium is now $W = \Phi(n\xi(q_n)) - nC(q_n)$. Surplus maximization leads consumers to demand a quantity q_i of variety i such that $p_i = \frac{\partial W_D}{\partial q_i} = \Phi'\xi'$, hence the profit of a firm in the symmetric equilibrium is $\pi_n = \Phi'\xi' q_n - C$. Now we get

$$\begin{aligned} \frac{\partial W_n}{\partial n} &= \Phi'\xi - C + (\Phi'\xi' - C_m)n\frac{\partial q_n}{\partial n} \\ &= \pi_n + (\Phi'\xi' - C_m)n\frac{\partial q_n}{\partial n} + \Phi'(\xi - \xi' q_n) \end{aligned} \tag{6.13}$$

We observe that (6.13) is similar to (6.12) but with an added term, the last one, whose sign is positive by the concavity of ξ. It simply means that because consumer like variety,

entry contributes not only to competitiveness but also to increase the desired variety, thus is not as bad as in the case of homogeneous products. In this generalization, no definite conclusion can be reached unless we make some further assumptions regarding the pricing behavior. Roughly speaking, if firms are close to be price-takers then diversity is insufficient while it is excessive when firms have significant market power.

6.1.6 Market Integration

The political acceptation of an economic integration process such as the European one is strongly dependent on the effect it may have on national market structures and national welfare. One element related to the present theory of oligopoly competition is whether we shall see more or less active firms in the long run once integration has taken place.

Quantity Competition

To resolve this question consider two national markets for the same homogeneous good where national firms compete in quantities and share the same production technology displaying decreasing returns to scale (DRS) and a positive fixed cost of entry. Before studying integration, we have to look at the autarchic long-run Cournot equilibrium in each country. Autarchy is the relevant hypothesis because import taxes or import quotas in each country prevent foreign firms from participating in national markets.

The long-run Cournot market structure in a single country is a number n of *national* active firms. When integration takes place, each market has suddenly $2n$ active firms so that the equilibrium price falls in both markets. If we assume that demands are identical in both countries (i.e., economies are of similar size), the profit of an individual firm is $2\pi^C - F$ since it is now active in the new economic area consisting of both countries.[5@] The entry condition being now $\pi^C \geq F/2$, the long run market structure of the integrated economic area is that of a (new) single country where the fixed cost of entry would be twice smaller. The integrated economic area therefore offers more room for firms but not twice more i.e., the long-term number of firms m satisfies $n < m < 2n$.

In the example we have been using to study oligopoly in §5.1, the long term number of firms in a single country is the solution of $\pi^C = F$ where, for simplicity, we neglect the integer problem. Using equation (5.12), we derive $n = \frac{a-bc}{\sqrt{bF}} - 1$. The post-integration stable number of firms solves, on the other hand, $\pi^C = F/2$ which yields

$$m = \sqrt{2}\frac{a-bc}{\sqrt{bF}} - 1 = n + (\sqrt{2}-1)\frac{a-bc}{\sqrt{bF}} < n + (\sqrt{2}-1)n = n\sqrt{2}$$

which is more or less 70% of $2n$, the initial number of firms in the integrated economic

area. Hence, some 30% of the original firms will have to exit the integrated market.[6@]

Perfect Competition

Let us compare this result to what happens when both national markets are perfectly competitive. We know that previous to integration, the autarchic long-term equilibrium in country A features firms which are all selling at their minimum efficient scale $\bar{q}_A$ (the level of production minimizing the average cost) while the price is $\bar{p}_A = C_{m,A}(\bar{q}_A) = AC_A(\bar{q}_\alpha)$. The long-term number of active firms is then given by the equality of demand and supply i.e., $n_A = D_A(\bar{p}_A)/\bar{q}_A$. The same goes on in country B. When the two economies merge, either everything remains the same because prices were already equal or there is a comparative advantage for one country, say A, if $\bar{p}_A < \bar{p}_B$. In that case, the production technology of country B become obsolete and all firms using it have to close gradually since the long-term price is $\bar{p}_A$ which is unsustainable for them. To meet the demand of both countries with the better technology (A), new firms will enter the market; there is obviously no condition of nationality for them since our simple model has neither administrative nor transportation cost to limit entry within the integrated area.[7@]

Summing up, economic integration of perfectly competitive market enables all consumers to take advantage of the best available production technology while the supply side of one country may need to reorganize itself to meet the standard set by the other (currently better) country. The fact that the better technology leads to greater sales[8@] means, as a consequence, that more inputs will be used in the integrated economic area as compared to what was previously used in the autarchic countries. The geographic origin and composition of these inputs (e.g, unqualified labour, qualified labour or capital) cannot be assessed in such a crude model. However, the origin of the comparative advantage we alluded to before can give a hint of which input is most needed in the economy; the most frequent sources of technological advantage are the wage level, the cost of capital and the level of human capital of national workers (productivity).

Conclusion

Whatever their initial market structures, the integration of two economic areas has a globally beneficial effect since the long-term number of active firms increases in both countries which means that the price will fall everywhere thereby generating a greater consumer surplus and also a greater total welfare. This is so in the Cournot model because the fixed cost of each remaining firm is spread over larger sales thanks to the duplication of the market size. To sum-up:

> Economic integration is good for all sectors even-though some firms belonging to inefficient oligopolistic markets will have to close down; their factors, capital and labour then find employment in the more competitive sectors of the economy where the integration process creates an additional demand for factors that favors either the creation of new firms or the expansion of existing ones.

There is few doubts that the early economic integration of the US states during the 20^{th} century contributed meaningfully to the might of the US economy. A likewise observation can be made of the European integration after the second world war.

6.1.7 Contestability

Baumol et al. (1982)'s theory of *contestable markets* present an ambitious theory of industrial organization which breaks away from the SCP paradigm but also with the core methodology of this book, *strategic behavior*.

We first summarize the theory before succinctly presenting its formal aspects. Next, we turn to its theoretical shortcomings and finally to the lasting impact it has shown in the policy arena.

Basic Tenet

Let us start with an historical clearcut example. The Ballpoint Pen was patented in 1938. In 1945, Milton Reynolds bought one and copied the idea bypassing the patented capillarity mechanism with a slightly less efficient gravity one. By doing so, he beat the official distributor to the US market and sold his "Rocket" pen as a luxury item, generating large profits. Within one year, all incumbent pen makers had entered the market pushing the price down from 12$ to 3$. With even more challengers, the price dropped to 40 cents in 1948, so that Reynolds cleverly sold all his assets in 1951, cashing on his initial success.[9@]

Contestability is most best understood as an extended Bertand paradox (cf. 5.2.1): if an incumbent prices so as to obtain an 8% rate of return on investment whereas the market average is 6%, then a potential entrant can actually enter the market at a lower price, capture all demand and still make a 7% return that keeps beating the market. This situation will last until the incumbent is able to react with an even lower price to recapture some market share. Even if he ends-up being expelled from the market, the transitory entrant has managed to earn an above average return for some period of time; this, in itself, warrants the initial entry. The incumbent being equally rational, cannot

ignore the menace. In other words, meanwhile she does not adjust her pricing strategy so as to yield the average return on investment, "hit and run" will continue to take place.

To understand the potential extent of *contestability theory*, it is worth recalling that perfect competition does not cover the case where a technology displays economies of scale or scope (cf. §2.3.1). Thus, the welfare theorems do not apply for natural monopolies and fail to provide us with an efficient benchmark for policy prescription. Filling this void has been hotly debated since the 1930s. Tenants of SCP call for regulation as exposed in §17.1 while the Chicago school seeks to axe all barriers to entry i.e., create the conditions for *free entry* in order to unleash the forces of *potential competition*.[10@] The theory of *contestable markets* formalizes this latter point of view by generalizing perfect competition to cover natural monopolies and more generally, industries displaying economies of scale. The key features of a contestable market are free entry and the existence of at least one potential competitor. The equilibrium in such a contestable market is then extremely efficient:

- If technology calls for a natural monopoly, there is one active firm and she prices her bundle of products in a Ramsey-Boiteux fashion[11@] so that second best efficiency is achieved.
- If technology calls for two or more active firms, then the number of firms minimizes industry cost; furthermore these firms price all their goods at marginal cost so that first-best efficiency is achieved i.e., "two is enough for competitive outcomes".

The policy implication of these results is obvious: if a market is contestable, no regulation at all is needed because the hidden forces of potential competition literally compel incumbents to behave efficiently. If the market is initially regulated, it should be deregulated. As we explain hereafter, the contestability theory is too thin to support such a call but other models of industrial organization have shown that administrative barriers to entry that protect incumbents have negative welfare consequences and should thus be removed.

Theory and Critique

The main academic achievement of contestability theory is the characterization of efficient cost structures for a multi-product production technology; as reported by Baumol and Willig (1986), it has been used in many econometric studies of regulated oligopolistic industries. Since this topic is intensive in mathematics, we skip it and focus instead on the implications of contestability for competition and market performance which are almost formula free and can therefore be exposed here. A *contestable market* satisfies the following hypothesis:

- H1: The best incumbent's technology is available to anyone.
- H2: The profitability of entry is evaluated at the current price.
- H3: Fixed costs are not sunk.

(H1) implies that the product sold by active firms is homogeneous since they all share the same production technology. It also implies that incumbents derive no quality advantage from the positive experience of actual clients, nor goodwill from their historical presence in the market. (H2), the Bain-Sylos postulate, means that a potential entrant completely disregard the reaction that his entry will trigger among incumbents.[12@] Lastly, (H3) means that although a fixed investment may be necessary to enter, the entrant can hit-and-run the market at no cost; his capital investment is not specific to this market and keeps its full value for use in another market.

The next fundamental ingredient to the theory is the ubiquitous presence of a *potential competitor*, in the background, ready to "hit and run".[13@] When (H1-H3) are satisfied, it is quite easy to see that incumbents, whatever their number, are forced to price the product at their long-term average cost i.e., as if they were perfect competitors. No regulation is needed to guarantee efficiency since firms, including natural monopolies, will (auto) regulate themselves towards that goal. Contestability is thus a logically straightforward theory building on clear-cut assumptions. As with all economic theories, the latters are not entirely fulfilled in real markets:

- H1: A potential competitor need not exist because the incumbent's technology is not freely available; rather technologies are protected by patents, regulatory capture through lobby groups, or simply experience and know-how.
- H2: The Bain-Sylos postulate is turned upside down by firms' quickness to react to a changing environment. It is customary to see a major player reacts within days or weeks with a price cut, an improved warranty, a special gift or an advertising campaign to counter the entry or launching of a new competing product or service.[14@]
- H3: There are virtually no investment that can be freely moved from a market to another, hence some of the fixed costs are sunk even if a small share.

Weak empirical contrasting is however not an impediment to the success of a theory. The perfect competition paradigm (PC) is a polar framework that never fully applies to real markets; its fame derives from its ability to serve as a benchmark. Indeed, economists have been able to relax the strong hypothesis of PC to construct models of imperfect competition and compare them to the PC benchmark. So, we know better the contribution of each PC hypothesis to the welfare theorems and this is also useful

for policy recommendations. Such an extension process has not taken place within the theory of contestability.[15@] Whereas nearly competitive markets have nearly efficient outcomes, Baumol (1982) confess that nearly contestable markets **might** have nearly efficient outcomes. Contestability theory is thus *inflexible* and not amenable to change and progress.[16@] Lastly, contestability assumes away strategic behavior (H2) in order to achieve logical simplicity. By going contrariwise to the general development of scientific thinking in the field of economics, it has put itself outside the paradigm of modern economics.[17@] Nonetheless, this vision has made its way into the policy arena and for that reason cannot be altogether ignored. Note finally that the modern use of the terminology points at models of entry barriers based on game theoretical reasoning.

Impact

The sharp conclusions of contestability have been adopted by lawyers defending firms standing accused in antitrust cases starting with its architect William Baumol, a long time consultant for AT&T.[18@] The following quote by economics professor Richard Schmalen illustrates this. Upon being hired by Microsoft in the antitrust case over the tying of the "Internet Explorer" browser to the "Windows" operating system, he declares that "*the extremely high market share of Microsoft does not translate into market power because the company faces the threat of potential competition. Of great concern to Microsoft is the competition from new and emerging technologies, some of which are currently visible and others of which certainly are not. This places enormous pressure on Microsoft to price competitively and innovate aggressively*" (cf. §24.1).

The "invisible technologies" alluded to in the previous statement are a reality; the history of the software industry is replete with case of sudden death of an incumbent upon appearance of a better product launched by a challenger. Microsoft which itself successfully outplayed competitors is therefore right to be anxious about this treat. Yet the response does not seem to be aggressive pricing[19@] but rather the preemptive buying of small firms that have developed and patented new and interesting innovations.[20@]

As we indicate in §24.3.5, Microsoft achieved a quasi monopoly with its software "Microsoft Excel" and "Microsoft Word" at the end of the 1990s; however, after the uninterrupted price fall of the 1990s it did not raised its price as we would expect from a traditional monopoly. That is an indication of the seriousness of the threat alleged by Schmalensee. However, Microsoft products remain more expensive than the competitive fringe and have a rather short live span due to the Microsoft policy of issuing new versions that force users to upgrade (the backward compatibility is of low quality). This strategy can be then seen as the optimal policy of a monopolist selling a durable good (cf. §4.3.5), all the more since "Office XP" has been marketed as a leasing product in several

countries instead of being sold.

The very idea of contestability is also related to Markides and Geroski (2004)'s "fast second" notion whereby the ability to copy a first mover may save on cost of development while losing little in terms of image and thus be overall advantageous (cf. Lee (2009)). An example is how MySpace was gobbled up by Facebook and Twitter in 2009.

6.2 Stackelberg Leadership

We now study how a firm can weigh in a profitable manner on the terminal stage of price or quantity competition by taking an observable and sometimes costly decision which irrevocably changes her behavior.

6.2.1 Intuition

Stackelberg (1934) observing the competition among mineral water producers in Germany notices that one firms seems to be a leader whose changes of prices systematically trigger responses by all other contenders. To explain and rationalize such a sequentiality of moves, this author modifies Cournot (1838)'s model of quantity competition.

Keeping in line with our previous farmers' story, imagine that farmer 1 takes a *leading* action by announcing publicly the quantity $\bar{q}_1$ it will bring to the market (e.g., publication in the local newspaper). Farmer 2 who forgot to do the same becomes a *follower* and its best response to the announced $\bar{q}_1$ is to bring $q_2 = BR_2^C(\bar{q}_1)$. It is crucial to understand that farmer 2 cannot credibly lead farmer 1 or anyone else to believe that he will bring a quantity different from $BR_2^C(\bar{q}_1)$.

The behavior of farmer 1, in this new context, differs qualitatively from that of farmer 2 as he can anticipate that his announcement $\bar{q}_1$ will be followed by $BR_2^C(\bar{q}_1)$. A first and obvious result is that farmer 1 will be better off than in the Cournot equilibrium. Indeed he can choose q_1^C in which case farmer 2 replies with q_2^C; thus farmer 1 obtains at least the Cournot profit Π_1^C. We now study his incentive to deviate from the Cournot value.

Recall first that farmer 2's reaction is decreasing with the commitment $\bar{q}_1$ (this is the basic property of Cournot competition). In the Cournot framework, when farmer 1 increases his sales by, say 50 kg, market sales grows by 50 kg forcing the market price to fall by, say 8€. The optimal level is then when the additional profit made on the 50 kg just equates the margin loss of 8€ over his total sales. In the commitment framework, when farmer 1 increases his sales by 50 kg, farmer 2 reduces his own sales by, say 25 kg in response; hence, market sales grows by 25 kg only and the price falls by a smaller amount, say 5€. This is attractive for farmer 1 because the profit made on the 50 kg

is as before while the margin loss over his total sales is 5 instead of 8 per kg, thus the additional sales are profitable under the commitment while they were not under Cournot competition.

The usual problem with increasing sales for firms with market power is that such a move depresses the price and thus the profit margin. The value of a commitment to large sales is to force the challenger to reduce his own sales in a "desperate" attempt to detain the price tumble provoked by the commitment. The leader is thus shifting part of the problem on the challenger and this is why increasing sales can be profitable when it would not be under a regime of simultaneous Cournot competition.

6.2.2 Formal Analysis

Formally, the merit of the commitment is seen by exploring in detail the marginal revenue of farmer 1. Indeed, the commitment affects only the behavior of the other farmer, thus the demand side and not the supply side (cost function). In the Cournot framework, the quantity sold by farmer 2 is, from farmer 1's point of view, a parameter insensitive to his own sales decision while in the commitment framework, the quantity sold by farmer 2 is a response to the commitment, we thus have

$$\bar{R}_{m,1} = P + P'\bar{q}_1 + P'\frac{\partial q_2}{\partial \bar{q}_1} = R_{m,1} + P'\frac{\partial q_2}{\partial \bar{q}_1}$$

because he *knows* that his decision $\bar{q}_1$ will affect the decision q_2 of farmer 2. As $q_2 = BR_2^C(\bar{q}_1)$, we deduce $\frac{\partial q_2}{\partial \bar{q}_1} < 0$ so that the marginal revenue of additional production is greater than under the Cournot competition; hence there is a motive for increasing production which implies that the other farmer will have to reduce his own. Nevertheless it is not clear whether total sales will increase or not.

In our linear example, the relation between sales $\bar{q}_1$ and price becomes

$$p = \frac{a - \bar{q}_1 - BR_2^C(\bar{q}_1)}{b} = \frac{a + bc - \bar{q}_1}{2b} \tag{6.14}$$

using (5.4): the demand addressed to farmer 1 is twice more elastic than in the Cournot game; this leads farmer 1 to produce a greater quantity (as we already saw). Equivalently, the profit to take into consideration for farmer 1 is not $\Pi_1\left(\bar{q}_1, q_2\right)$ where a value of q_2 has to be guessed but

$$\Pi_1\left(\bar{q}_1, BR_2^C(\bar{q}_1)\right) = \bar{q}_1(p - c) = \tfrac{1}{2}\bar{q}_1\left(a - bc - \bar{q}_1\right) \tag{6.15}$$

using (6.14). The optimal choice, q_1^S, maximizes (6.15); it solves the first order condition

$a - bc - 2\bar{q}_1 = 0$ and is

$$q_1^S \equiv \frac{a-bc}{2} > q_1^C = \frac{a-bc}{3}$$

which in turn leads farmer 2 to respond with

$$q_2^S \equiv BR_2^C(q_1^S) = \frac{a-bc}{4} < q_2^C = \frac{a-bc}{3}.$$

It is a matter of algebraic computations to check that the Stackelberg equilibrium price is $p^S = \frac{a+3bc}{4b} < p^C = \frac{a+2bc}{3b}$, the Cournot price so that efficiency unequivocally increases (total consumption increases). Regarding equilibrium profits we obtain

$$\Pi_2^S = \frac{1}{16}\frac{(a-bc)^2}{b} < \Pi_2^C = \frac{1}{9}\frac{(a-bc)^2}{b} = \Pi_1^C < \Pi_1^S = \frac{1}{8}\frac{(a-bc)^2}{b}$$

so that there is indeed a first mover advantage but also

$$\Pi_1^S + \Pi_2^S = \frac{3}{16}\frac{(a-bc)^2}{b} < \frac{2}{9}\frac{(a-bc)^2}{b} = \Pi_1^C + \Pi_2^C$$

meaning that industry profits decrease by $\frac{2/9-3/16}{3/16} \simeq 18\%$ which does not come as a surprise since the price is lower.

To conclude, the flooding of the market by farmer 1 reduces the cake to be shared by 18% but enables him to grab two thirds instead of one half, so that he ends up better by 12% but farmer 2 loses 44% of his original profit.

6.2.3 The value of Commitment

Time Consistency

The Stackelberg model of leadership (and the first mover advantage result) although appealing has a serious drawback because it is *time inconsistent*: at the moment where farmer 1 has to take his quantity decision he is tempted to renege his word (the announcement he made previously). Indeed, when facing q_2^S, his best reply is $q_1' \equiv BR_1^C(q_2^S) = 3\frac{a-bc}{8}$ which is between q_1^C and q_1^S (check it on Figure 5.1 by drawing q_1^S, then the best reply q_2^S and finally the best reply q_1' to q_2^S). In simple words, if 60 is the Cournot quantity, farmer 1 announces 80 to force farmer 2 to reduce his own sales from 60 down to 40 but once this is achieved, farmer 1 would like to sell only 70! The reason for that is quite simple: $(60,60)$ is the only stable pair which means that $(80,40)$ is unstable in the sense that at least one of the farmer prefers a different choice. Now, since 40 is optimal vis-à-vis 80, it must be the case that 80 is not optimal vis-à-vis 40 and this is exactly what is happening here; in front of 40, the optimal response is 70.

Legal System

This paradox (and its solution) can be understood with the help of Figure 6.1 below. Farmer 1 plays first by announcing a quantity. We see on the lower branch that q_1^C is followed by the optimal response q_2^C. If now farmer 1 announces the large quantity q_1^S then we expect farmer 2 to respond with the smaller q_2^S. After observing the irrevocable choice q_2^S, farmer 1 would like to revise his quantity down to q_1'. What can possibly happen in that case? It could be the case that Farmer 1 is sued by angry rationed consumers and condemned by the judge (J) for lying in the newspaper. If the penalty is large[21@] than then it is optimal to keep the promise and bring the quantity q_1^S to the market, so that in turn it is optimal for farmer 2 to bring q_2^S.

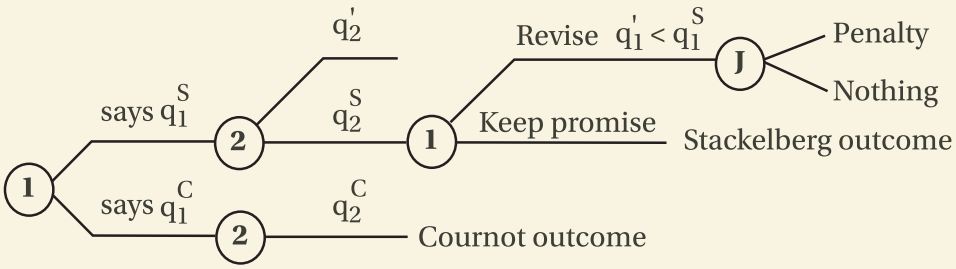


Figure 6.1: Sequence of Announcements

Consider now a twist of our basic story where the advertising says in very small characters that "the announced quantity is non binding obligation". The judge, when called to act, would turn down the consumers because the advertisement only referred to an intention, so that no malevolence is involved. Reassured by this judgment, farmer 1 would rationally choose q_1' over q_1^S Anticipating that the announcement was just some kind of advertising (economists say "cheap talk") farmer 2 would therefore bring $q_2' = BR_2^C(q_1')$. We have just started the process of optimal revisions that bring the two farmers back to the Cournot equilibrium.

Interpretation

Our little digression within the judicial system tells us that the Stackelberg model applies only to situations where the first mover can make sure that he would suffer if he were to change his mind later; in other words, the first mover must be able to tie his hands to convince the follower that what he said will indeed happen. The public announcement is a device to commit to something that would otherwise be untenable.

Notice that the existence of the judicial system is necessary to enforce the Stackelberg equilibrium for otherwise the equilibrium is the Cournot one. The paradox is that in equilibrium, farmer 1 always keeps his promise so that the judge is never call to act.

Society is thus paying judges for no actual work. It is precisely for this reason that the judicial system has to be publicly financed; being inactive it cannot sell its service and cannot therefore recoup its cost.

Nonetheless, trade which is the basis of market economies relies on morality, trust and an effective commercial law. It is quite simple, using Figure 6.1 in a more general context, to understand that the inefficiency and corruption of judicial systems in developing countries is undermining the whole trading system; this issue of "good governance" is now recognized as a major impediment to growth and actively promoted by international organizations like the OECD or the World Bank.

6.2.4 Business Strategies †

The idea of strategic commitment "à la Stackelberg" is extended by Fudenberg and Tirole (1984) and Bulow et al. (1985) to study the global effect of an investment k_1 in capacity or advertising by firm 1. This firm competes on the consumer market against firm 2. The strategic variables are x_1 and x_2 (e.g., prices or quantities). The profit of firm 2 depends obviously on the strategic choices in the consumer market but also on the previous choice k_1 of her opponent so that we write it $\Pi_2(k_1, x_2, x_1)$.

In the last stage, both firms are rational, thus choose x_1 and x_2 to form a Nash equilibrium of the situation where they fight, a situation that is potentially affected by k_1. Hence the equilibrium is $(x_1(k_1), x_2(k_1))$ such that $x_i(k_1)$ maximizes $\Pi_i(k_1, x_i, x_j(k_1))$ for $i = 1,2$. We have

$$\frac{d\Pi_1}{dk_1} = \underbrace{\frac{\partial \Pi_1}{\partial k_1}}_{\text{direct}} + \underbrace{\frac{\partial \Pi_1}{\partial x_2}\frac{\partial x_2}{\partial k_1}}_{\text{strategic}} + \underbrace{\frac{\partial \Pi_1}{\partial x_1}}_{0}\frac{\partial x_1}{\partial k_1} \tag{6.16}$$

and

$$\frac{d\Pi_2}{dk_1} = \underbrace{\frac{\partial \Pi_2}{\partial k_1}}_{\text{direct}} + \underbrace{\frac{\partial \Pi_2}{\partial x_1}\frac{\partial x_1}{\partial k_1}}_{\text{strategic}} + \underbrace{\frac{\partial \Pi_2}{\partial x_2}}_{0}\frac{\partial x_2}{\partial k_1} \tag{6.17}$$

Regarding the direct effect we can distinguish two cases:

- if k_1 is spend on advertising to increase fidelity of consumers then $\frac{\partial \Pi_2}{\partial k_1} < 0$ as the potential market for firm 2 is reduced.
- if k_1 is a technological investment aimed at decreasing firm 1's cost or improving quality it has no direct effect on firm 2 thus $\frac{\partial \Pi_2}{\partial k_1} = 0$.

The indirect effect is called *strategic* because it takes time to act: the decision of today, k_1, will force firm 1 to change her behavior tomorrow. We say that the investment

k_1 makes firm 1 look *tough* (resp. *soft*) when the strategic effect is bad (resp. good) for firm 2 i.e., if $\frac{\partial \Pi_2}{\partial x_1}\frac{\partial x_1}{\partial k_1} < 0$ (resp > 0).

We now ponder the indirect effect of k_1 over firm 1's own attitude by looking at the decomposition of $\frac{d\Pi_1}{dk_1}$ into a direct effect $\frac{\partial \Pi_1}{\partial k_1}$ and an indirect one $\frac{\partial \Pi_1}{\partial x_2}\frac{\partial x_2}{\partial k_1}$ to observe that whenever the strategic term is negative (resp. positive), firm 1 has a tendency to underinvest (resp. over-invest).

We assume that the strategic variables x_1 and x_2 are both prices or both quantity, more generally that $\frac{\partial \Pi_1}{\partial x_2}$ and $\frac{\partial \Pi_2}{\partial x_1}$ have the same sign. Using $\frac{\partial x_2}{\partial k_1} = \frac{\partial x_2}{\partial x_1} \times \frac{\partial x_1}{\partial k_1}$ we obtain

$$sign\left(\frac{\partial \Pi_1}{\partial x_2}\frac{\partial x_2}{\partial k_1}\right) = sign\left(\frac{\partial \Pi_2}{\partial x_1}\frac{\partial x_1}{\partial k_1}\right) \times sign\left(\frac{\partial x_2}{\partial x_1}\right)$$

We say that the strategic variables x_1 and x_2 are strategic *substitutes* when $\frac{\partial x_2}{\partial x_1} < 0$, as it is the case for Cournot competition. We speak of strategic *complements* whenever $\frac{\partial x_2}{\partial x_1} > 0$, as it is the case for Bertrand competition.

We can now draw, using the colorful language of Fudenberg and Tirole (1984), a clear taxonomy of the strategic effect of the investment k_1 for the leading firm 1:

- If k_1 makes firm 1 look tough and any increase in her strategic variable forces firm 2 to reduce her own then firm 1 desires to overinvest to gain an additional advantage (beyond the inner value of the investment) i.e., be a "top dog".
- If k_1 makes firm 1 look tough but compete with strategic complements then it prefers to underinvest to avoid tough competition later; it mimics a "puppy dog".
- If k_1 makes firm 1 look soft and it competes with strategic substitutes then firm 1 desires to underinvest or look like an "hungry dog" in order to show her readiness to fight in the last period.
- If k_1 makes firm 1 look soft but compete with strategic complements it prefers to overinvest to reduce her future aggressiveness, it mimics a "fact cat".

x_1 and x_2 are strategic / k_1 makes firm 1	Substitutes $\frac{\partial x_2}{\partial x_1} < 0$	Complements $\frac{\partial x_2}{\partial x_1} > 0$
Tough ($\frac{d\Pi_2}{dk_1} < 0$)	overinvest (top dog)	underinvest (puppy dog)
Soft ($\frac{d\Pi_2}{dk_1} > 0$)	underinvest (hungry dog)	overinvest (fat cat)

Table 6.1: Taxonomy of Business Strategies

Beware that this taxonomy is limited to oligopoly and excludes monopoly and perfectly competition. Indeed, the idea developed here rests on the indirect effect that a particular behavior has over a competitor's market strategy.

6.3 Forward Market

A simple way to commit to a large production in the future is to make some anticipated sales. This can be done using a *forward contract* which is a promise to deliver a fixed quantity at a future date in exchange of a monetary payment. Since forward trading is a pervasive instrument to protect oneself from risk, we briefly recall its functioning before studying the relation between forward contracting and Stackelberg leadership.

6.3.1 Introduction

Forward trading is either organized in a dedicated forward market, an exchange[22@] or through bilateral transactions deemed "over the counter" (OTC). A party to a forward transaction adopts either a "long" position if she promises to buy or a "short" position if she promises to deliver (sell).[23@] Since the main use of forward trading is to hedge the risk of a price variation in tomorrow's spot market, one might as well use a purely financial instrument called a "contract for difference" (CdF) that pays the (algebraic) difference between today's agreed price and tomorrow's price on the spot market (for an agreed quantity); that way a seller or buyer can guarantee himself a fixed price for the good he will transact.[24@]

When a trader who is not a producer holds a short position, a promise to deliver, he has no other choice but to buy the units in the spot market at the equilibrium price, be it high or low. Likewise, when a trader who is not a consumer has a long position, he has no use for the physical unit he is entitled to receive, thus he prefers to sell it in the spot market rather than keep a useless item, this whatever the standing price. Matters are different for producers holding forward positions since they can actually produce the units they must deliver or save those they will receive by virtue of their long position. We shall demonstrate later that a producer gains no advantage from this feature because the equilibrium is the same whether he behaves as a pure trader or not.

6.3.2 Analysis

Although forward markets have long been studied in finance, they were seen from a theoretical point of view solely as "hedging" instruments, a method to spread risk among

many agents and therefore reduce its impact on individuals. Allaz and Vila (1993) offer a new rationale by showing that forward markets are also useful to reduce the market power of large participants in spot markets.

Let us demonstrate this result using the Cournot model of quantity competition for an homogeneous good (cf. §5.1). The spot market where the product is sold and bought is preceded by a forward market where anyone, firm, buyer or broker,[25@] can buy or sell forward contracts; an algebraic position f in the forward market generates an algebraic gain $(\tau - p)f$ to its holder where p is the spot market price and τ is the forward market price.[26@] Hence $f = 1$ is a unit long position, a promise to deliver later while $f = -1$ is a unit short position, a promise to buy later.

We assume without loss of generality zero production cost and denote f_i the position of firm $i = 1,2$ in the forward market. When $f_1 + f_2 > 0$ i.e., producers are globally long, some counter-party traders are entitled to receive physical units of the good; since spot market buyers are the only ones who value these units, they will end up buying them in the spot market. As a consequence, their overall demand $D(p)$ will be met by the balance of forward contracts $f_1 + f_2$ and by the sales of producers in the spot market $x_1 + x_2$. If we denote $q_i = x_i + f_i$ the total amount to be produced by firm $i = 1,2$, then the spot price satisfies $D(p) = q_1 + q_2 \Leftrightarrow p = P(q_i + q_j)$ so that the total profit of firm i made up of spot sales and forward sales, reads $\pi_i = P(q_i + q_j)(q_i - f_i) + \tau f_i$.

The marginal revenue of an additional unit produced (independently of where it will be sold) is $\frac{\partial \pi_i}{\partial q_i} = P + (q_i - f_i)P'$. It is important to observe the positive effect of forward sales (short position or $f_i > 0$) on marginal revenue. The negative term is smaller than in the Cournot model (no forwards) because the rebate that must be offered to current customers in order to gain an additional customer applies only to spot market sales, not the entirety of production (check the difference with equation (2.15)). Using our usual linear demand $D(p) = a - bp$, the best reply solving $\frac{\partial \pi_i}{\partial q_i} = 0$ is easily computed as $q_i = \frac{a - q_j + f_i}{2}$ which is increasing with f_i thereby confirming our intuition that forward sales enable to commit to a larger overall production. Using the symmetric equation for the other firm, we can solve the equilibrium system and derive equilibrium quantities $\bar{q}_i = \frac{a - f_j + 2f_i}{3}$ for $i = 1,2$; after simplifications, we find the equilibrium spot price $\bar{p} = \frac{a - f_j - f_i}{3b}$ and the total sales $\bar{Q} = \frac{2a + f_j + f_i}{3}$.

As always in financial markets there are pure traders (brokers) in the forward market willing to exploit any arbitrage possibility. If the aggregate position of producers is long ($f_1 + f_2 > 0$), they expect a profit over the forward units if $\tau > p$, thus their counterparts expect a profit if $\tau < p$. We deduce that brokers are interested to take the short position, pay τ now only if $\tau < p$. In that case they compete à la Bertrand over this opportunity which means that they will offer an higher and higher forward price τ; in equilibrium

there is equality. A similar reasoning holds when the aggregate position of producers is short, hence the spot price is perfectly anticipated by arbitrageurs so that $\tau = \bar{p}$ holds.

In the initial stage where the producers have to decide on their forward sales, they take into account the future equilibrium of the spot market. Spot sales and forward sales are therefore made at the same price so that total profit is production times spot price:

$$\Pi_i(f_i, f_j) = \bar{p}\bar{q}_i = \tfrac{1}{9b}(a - f_j - f_i)(a - f_j + 2f_i) \tag{6.18}$$

The best reply to f_j is easily computed by maximizing (6.18). The solution to $\frac{\partial \Pi_i}{\partial f_i} = 0$ is $f_i = \frac{a-f_j}{4}$. Since individual forward sales are limited by the market size a, each firm will sell forward ($f_i > 0$). The underlying reason is that quantities are strategic substitutes (cf. §6.2.4), thus the (positive) strategic effect of anticipated sales is stronger than the (negative) direct effect of lowering the spot price.

The equilibrium of the forward market is found by solving for the system of best replies; we obtain $\hat{f}_i = \hat{f}_j = \frac{a}{5}$ out of which we compute $\hat{q}_i = \hat{q}_j = \frac{2a}{5}$ i.e., each firm sells $\frac{a}{5}$ forward and then another $\frac{a}{5}$ in the spot market. The equilibrium price at which everything is bought is $\hat{p} = \frac{a}{5b} < p^C = \frac{a}{3b}$, the Cournot price (for zero cost), hence total sales increase as well as efficiency (recall from §2.3.2 that higher consumption is synonym of higher welfare).

To summarize, the existence of a forward market with perfect foresight and arbitrage intensifies quantity competition and brings about more efficiency. Because forward sales are beneficial to a single firm, they enter into a race for selling in advance; this situation is very much like the prisoner's dilemma: it would be better to avoid forward sales, but if you are the only one to do so, it becomes a worthwhile option! (cf. §2.4.1)

6.3.3 Comparisons

The relation of forward sales with Stackelberg leadership is now quite simple to see (using the physical forward sales contract). If firm 1 is the only one who can sell forward then she chooses $f_1 = \frac{a}{4}$ (the best reply to $f_2 = 0$), so that firms behave as in the Stackelberg game. Indeed, it is a matter of simple algebraic computation to check that $\bar{q}_1 = q_1^S$ and $\bar{q}_2 = q_2^S$.

We have thus found a new way to commit to sales of $q_1^S = \frac{a}{2}$. Firm 1 can sell immediately half of it and then compete in the spot market to (optimally) sell the other half. Thus, we can say that the commitment to bring a large quantity to the market or the forward sale are identical tools **if** only one firm has access to them. However, as soon as the two competitors can both commit or both sell forward, there is a difference among the two instruments. In the case of double commitment, it is as if there was no

commitment at all, so that the Cournot outcome emerges again. On the contrary, the case of mutual forward sales represent a stronger commitment than the mere promise to fuel the market; both firms actively engage into this practice and get trapped into an unwanted flooding of the market that ultimately reduces their profits but increases the overall efficiency.

We have to conclude with a word of caution spelled out by Mahenc and Salanié (2004): the social desirability of forward markets is crucially dependent on the mode of competition in the spot market. As we explained above, forward sales always end up in the hands of buyers, thus reduce the demand appearing in the spot market and therefore depress the equilibrium spot price. When firms compete in quantities which are strategic substitutes, it is more important to increase sales using forward contracts than sustain the spot price. On the other hand, when firms compete in prices which are strategic complements, the reverse property holds: firms prefer to act so as to sustain high prices even if this means selling less. They therefore take a long position (buy some of their own production) but do not go as far as selling their entire production in advance.[27@] This might seems unnatural on the part of a producer. When such long positions were taken in 1978 by large coffee producing countries, observers thought it was an attempt to squeeze the traders holding the short positions; indeed, the latter might be forced to accept very high prices in the spot market in order to meet their obligation to deliver the physical good (since they will hardly be able to procure it elsewhere). The present theory shows that this might well have been a strategic behavior aimed at taking advantage of the countries market power where that term is understood in the classical sense and not as an exploitative behavior (cf. sec. 4 in Mahenc and Salanié (2004)).

6.4 Vertical Relationship

In this section, we look at several strategic commitments. The first two examples look at the authority relationship with managerial compensation and manufacturer-retailer trade. We show that a relation governed by contract (aka “arm’s length relationship”), can provide an advantage over integration where the relationship is governed by authority. Obviously, we leave aside the possible benefits of integration which are treated in §14. Lastly, we show how firms can use debt leverage to commit to a greater aggressiveness in the market.

6.4.1 Managerial Compensation

The rational owner of a firm strives for profit maximization but she has to rely on a manager to achieve this goal. The latter then does not seem to care much for profits but rather for the size of the firm he runs (and the perks associated with the position). §13.1 shows how manager's objectives can be re-aligned with the owner's views. §15.1, on the other hand, presents some arguments as to why excessive firm growth is inevitable.

The originality of Vickers (1985) and Fershtman and Judd (1987) is to argue that, may be, it is in the interest of owners to have their managers maximize sales or revenue instead of profits. The reason is that most firms are neither monopolies nor perfect competitors (price-takers), hence their behavior influences the behavior of their competitors. In a context of Cournot competition, there is a first mover advantage to flood the market because it forces opponents to reduce their own sales. To achieve that advantage, it is enough for the owner of the firm to write a *managerial compensation* linked to sales (or revenue).[28@] Because such a bright idea is immediately adopted by all firms, they become trapped in a prisoner's dilemma. Indeed, the new equilibrium is characterized by higher sales, a lower price, a higher welfare but also lower industry profits.

To check this claim, assume that the traditional compensation package made of a base wage $\underline{w}$ and a share λ of profits (bonus) is enough to align managers incentives with the maximization of profit. Consider now an alternative compensation scheme[29@] that additionally pays a share μ of the firm's revenue R:

$$w \equiv \underline{w} + \lambda\pi + \mu R = \underline{w} + (\lambda + \mu)(R - \alpha C)$$

where $\alpha \equiv \frac{\lambda}{\lambda+\mu}$ since $\pi = R - C$. It is obvious that the manager will act (choose a quantity) so as to maximize $R - \alpha C$. The owner can therefore tune α to obtain his desired strategic effect and use the other parameters to offer (on average) his opportunity cost to the manager (saturate his participation constraint).

Let us now study the effect of this strategic tuning upon market strategies. Each firm owner signs a contract with her manager; later on, managers learn their marginal cost and the market demand $D(p) = a - bp$. The analysis of Cournot competition for asymmetric costs can be directly used to solve for the competition among managers. Indeed, the choice of the coefficient α_i by owner i for the compensation scheme of her manager amounts to change the *true* marginal cost c_i into a *virtual* one $\alpha_i c_i$ upon which the manager of firm i will base his behavior (he maximizes $R_i - \alpha_i C_i$).

Manager i's best reply function is equation (5.4): $BR_i^C(q_j) = \frac{1}{2}(a - bc_i - q_j)$; it is drawn as a downward sloping plain line on Figure 6.2. Observe that the lesser the cost c_i, the greater the response. Now, when the owner of firm i introduces a coefficient $\alpha_i < 1$

in his manager's contract, the whole best reply curve moves up to $\widetilde{BR}_i$ (dashed line) and the new equilibrium (circle) features a larger quantity for the strategic owner and a smaller one for her competitor. Since profits are increasing with the quantity sold, the strategic owner obtains a greater equilibrium payoff. The intuition of the result is simple: by linking wage to revenue, the owner turns his manager into an aggressive player behaving *as if* his marginal cost was lower.

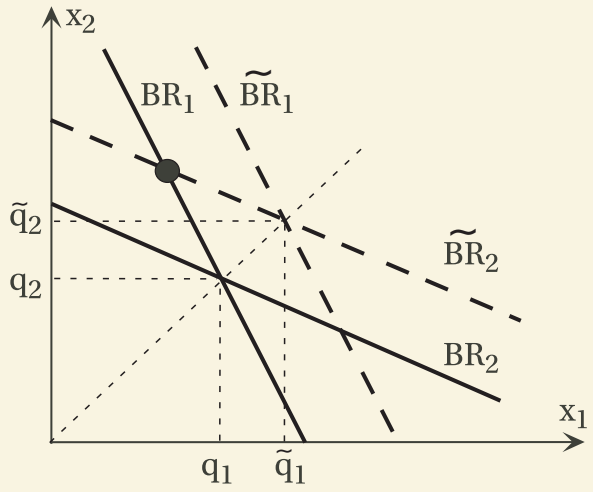


Figure 6.2: Strategic Effect of Managerial Compensation

Being equally rational, owner j will also offer a distorted compensation scheme to her own manager in order to make him more aggressive. To compute the equilibrium of the Cournot game among "hypnotized" managers, we assume that firms share the same technology i.e., have the same true marginal cost c; in equilibrium quantities are given by equation (5.10): $q_i = \frac{1}{3}\big(a - bc(2\alpha_i - \alpha_j)\big)$, but profits use the true cost hence

$$\pi_i = (p-c)q_i = \tfrac{a-bc-q_i-q_j}{b} q_i = \tfrac{1}{9b}\big(a - bc(3-\alpha_i-\alpha_j)\big)\big(a - bc(2\alpha_i - \alpha_j)\big)$$

for $i = 1,2$ and $j \neq i$.

The objective of owners is to maximize profits minus the opportunity cost of their respective managers which is a constant; hence owner i maximizes π_i as a function of α_i and her optimal choice is easily computed as $\alpha_i = \frac{3}{2} - \frac{a+\alpha_j}{4bc}$.[30@] Using the symmetric formula for owner j, we deduce the symmetric equilibrium $\alpha_i = \alpha_j = \frac{6}{5} - \frac{a}{5bc}$. The overall sales are $q_i = q_j = \frac{2}{5}(a-bc)$ just as in the forward market equilibrium[31@] so that the same efficiency conclusions are valid.

The same result obtains if the wage's bonus is based on sales q (physical quantity).

As can be seen on Figure 6.2, the equilibrium involving the strategic use of managerial incentives catches both owners into a more aggressive competition. The new equilibrium, at the intersection of the dashed lines, involves greater quantities, thus lower price and also a greater efficiency.

Notice finally that there is no point to design distorted compensation schemes for the manager of a monopoly or of a competitive firm. Indeed, these are the cases where the competitors fail to react to the firm's behavior either because there are no competitors or because they are ignored by price-takers. Hence, owners will look forward to align managers incentives with their own, the maximization of profit. To conclude, the issue of distorting manager's incentive is akin to the strategic behavior characteristic of oligopolistic market structures.

6.4.2 Vertical Integration

We have just seen that it is in the interest of a firm owner to maintain a certain degree of misalignment of incentives with her manager. The possible vertical integration between an upstream and a downstream firm entails the same tradeoffs: if firms integrate, their interests will be perfectly aligned while if they remain independent, the upstream firm will be able to influence the behavior of the downstream firm through the wholesale price.

In the case of the automobile industry, it is customary to see manufacturers enter exclusivity agreements with distributors whereby the dealer agrees to sell only the brand of the manufacturer in return for a local monopoly. The distributors of the various brands in turn compete for the final consumers. For such markets, Bonanno and Vickers (1988) show that it is in the interest of manufacturers to deal with independent dealers rather than integrate with them (absent any other benefit or cost of integration).

To check this claim, consider a manufacturing duopoly whose downstream retailers compete in prices; it has been shown in §5.2.3 that these strategic variables are complements i.e, if p_1 rises (for whatever reason) then dealer 2 will optimally respond by raising his own price. Imagine now that manufacturer 1 raises her wholesale price w_1; this will raise the marginal cost of dealer 1 who will have to raise his own retail price p_1 to keep maximizing profits (as shown in eq. (5.19)); hence dealer 2 will raise his price in response.

If the upstream manufacturer can aspire her dealer's profits $\pi_1 = (p_1 - c_1 - w_1)D_1$ using a subscription fee F_1, then she earns the payoff $U_1 = F_1 + w_1 D_1 = \pi_1 + w_1 D_1 = (p_1 - c_1)D_1$ which happens to be the profit of an integrated firm. We claim that it is a good idea to raise w_1 above zero because the direct effect is nil (envelope theorem) while the strategic

(indirect) effect has just been shown to be positive. Formally, the retailer maximizes π_1, thus $\frac{\partial \pi_1}{\partial p_1} = 0$ implying that $\frac{\partial U_1}{\partial p_1} = w_1 \frac{\partial D_1}{\partial p_1} \leq 0$ ($= 0$ if $w_1 = 0$). Since $\frac{\partial U_1}{\partial p_2} = (p_1 - c_1 - w_1)\frac{\partial D_1}{\partial p_2} > 0$, we obtain $\left.\frac{dU_1}{dw_1}\right|_{w_1=0} = \left.\frac{\partial U_1}{\partial p_1}\right|_0 \frac{\partial p_1}{\partial w_1} + \left.\frac{\partial U_1}{\partial p_2}\right|_0 \frac{\partial p_2}{\partial w_1} > 0$ given that both retail prices increase with w_1. Raising the wholesale price increases retail prices but also profits as the two competing goods are strategic complements.[32@] It is obvious that the competing manufacturer will use the same device raising further the consumers prices and the industry profits. The limit of this process occurs when the marginal loss due to forgone sales (direct effect of $\frac{\partial D_1}{\partial p_1}$) equates the marginal benefit due to increased retail prices.

If, on the other hand, the downstream competition is in quantities, the clever strategy for the manufacturer is to subsidize her retailer by lowering the wholesale price to the maximum.[33@] Indeed, quantities being strategic substitutes, retailers earn more whenever their marginal cost is lower. This time, the subscription fee will be very large since it sums the surplus earned by the retailer to the total subsidization amount. As we already saw with forward markets and managerial compensation, the subsidization strategy is individually rational for a manufacturer but catastrophic at the industry level since lower retail marginal costs mean higher sales and therefore lower industry profits. This occurrence of the prisoner's dilemma is entirely due to the fact that quantities are strategic substitutes.

6.4.3 Strategic Debt Commitment

In another example of strategic commitment, Brander and Lewis (1986) blend the financial strategy and the product market strategy of a firm to offer a novel explanation for the use of debt (in favor of equity) in the capital structure.

Intuition

The argument draws on the "asset substitution" effect of debt explained in §23.2.1: debt finance together with limited liability turns a firm manager into a risk lover who tends to overestimate (resp. underestimate) the positive (resp. negative) consequences of strong (resp. low) consumer demand. As a consequence, the manager behaves more aggressively in the market thereby forcing his opponents to retreat in a manner similar to the case of manager's incentives. In short then, debt makes a firm look tough to her opponents and we might say that issuing debt ex-ante embodies a beneficial strategic commitment similar to a Stackelberg first mover advantage.

Nevertheless, the use of debt is no costless because of moral hazard. Indeed, when the firm goes bankrupt (in a state of very low demand), the debt-holders fail to be repaid entirely; anticipating this fate, they require a higher interest rate for the funds they lend.

It is therefore doubtful that firms will emit debt to take advantage of the aforementioned strategic effect. To lift this suspicion, we note that when the firm is unlevered (zero debt), taking on a little debt has no financial effect because bankruptcy is not a concern (it is highly improbable). However, any level of debt changes the firm's market behavior so that the strategic effect is always significant. Hence, it pays to strategically increase debt above zero even though the optimal leverage might be quite limited.

As often with strategic commitments, the prisoner's dilemma looms large because if a strategic move can bestow an advantage, it tends to be adopted by all contenders. Hence, all rival firms increase their leverage and become more aggressive competitors. In equilibrium, the price falls, the overall efficiency of the market increases[34@] and profits shrink.

Model †

We now formally demonstrate these assertions. The profit of one firm is $\pi(x,y,\sigma)$ where x is her own decision, y her competitor's and $\sigma>0$ is a macro-economic indicator whose law is given by the cumulative distribution function H. Without loss of generality we assume $\frac{\partial\pi}{\partial\sigma}>0$ i.e., a higher σ indicates a better economic environment. The strategic market variable is chosen so that $\frac{\partial\pi}{\partial y}<0$ i.e., it can be quantity, advertising or the negative of price. The crucial element of the analysis is whether $\frac{\partial^2\pi}{\partial x\partial\sigma}$ is positive or negative. Positiveness could arise from cheaper inputs that reduce the marginal cost or greater willingness to pay of consumers or positive demand shock (check in the Cournot formula §5.4 the effect of parameters over the best reply). Negativeness might occur if firms compete in advertising since there is not much point to advertise when everybody wants to buy the good (the marginal effect of advertising is dampened by a positive demand shock). We perform the analysis for the most probable case of $\frac{\partial^2\pi}{\partial x\partial\sigma}>0$ assuming for ease of presentation that x and y are quantities.

After eliciting a debt level D and collecting the proceeds B of her sale to investors, shareholders are left caring for the expected value of profits net of debt repayment, denoted S. Whenever $\pi<D$ is realized, there is bankruptcy because the debt obligation cannot be met entirely. In that case, the firm is closed and the shareholders walk out safely because they are protected by limited liability.[35@] Defining θ as the unique solution to $\pi(x,y,\theta)=D$, we have $S=\int_\theta^\infty\big(\pi(x,y,\sigma)-D\big)\,\mathrm{d}H(\sigma)$ which satisfies

$$\frac{\partial S}{\partial x}=\int_\theta^\infty\frac{\partial\pi}{\partial x}\,\mathrm{d}H(\sigma)+\frac{\partial\theta}{\partial x}\Big(\underbrace{\pi(x,y,\theta)-D}_{=0}\Big) \tag{6.19}$$

Since D does not appear inside the integral and, by construction, $\frac{\partial\theta}{\partial D} > 0$, we have

$$\frac{\partial^2 S}{\partial x \partial D} = \frac{\partial^2 \pi}{\partial x \partial \sigma}\bigg|_{\sigma=\theta} \times \frac{\partial \theta}{\partial D} > 0 \tag{6.20}$$

In the market competition, shareholders choose the strategic variable x to be a best reply against the choice y, they expect from their competitor i.e., x solves $\frac{\partial S}{\partial x} = 0$. As made clear by (6.20), an increase of debt has the unambiguous effect of increasing the marginal value $\frac{\partial S}{\partial x}$, thus the entire best reply function shifts up with debt: $\frac{\partial x}{\partial D} > 0$.[36@] The new equilibrium can be analyzed with the help of Fig. 6.2 pertaining to the study of manager's incentives. Whereas the unleveraged equilibrium is symmetric, the new one (shown by the circle) features a larger quantity for the strategic player ($\frac{\partial x}{\partial D} > 0$) and a smaller one for his competitor ($\frac{\partial y}{\partial D} < 0$). Hence, when a firm increases her debt, she earns more in the market.

Whether any debt is emitted at all depends on the balance between its pros and cons. As we abstract from bankruptcy costs and tax advantages of debt, the cash flow is always distributed in its entirety to either shareholders or debtholders.[37@] Hence the total payoff to firm owners $B+S$ is simply the expected profit $V \equiv \int_0^\infty \pi(x,y,\sigma)\,\mathrm{d}H(\sigma)$. The effect of a debt increase is indirect and goes through the change in quantities, it is

$$\frac{\partial V}{\partial D} = \frac{\partial x}{\partial D}\int_\theta^\infty \frac{\partial \pi}{\partial x}\,\mathrm{d}H(\sigma) + \frac{\partial x}{\partial D}\int_0^\theta \frac{\partial \pi}{\partial x}\,\mathrm{d}H(\sigma) + \frac{\partial y}{\partial D}\int_0^\infty \frac{\partial \pi}{\partial y}\,\mathrm{d}H(\sigma) \tag{6.21}$$

The first term in (6.21) is zero by the envelope theorem since the firm elicits x to maximize profits (cf. (6.19)). Given that this first term is an average, $\frac{\partial \pi}{\partial x}$ must be strictly negative at $\sigma = \theta$; now as $\frac{\partial \pi}{\partial x}$ is increasing in σ, the second term is negative and represents the loss in debt value. Lastly, the third term is the positive strategic effect upon the competitor of strengthening one's behavior in the product market (recall that $\frac{\partial y}{\partial D} < 0$ and $\frac{\partial \pi}{\partial y} < 0$). It remains to observe that the second term is zero when debt is initially nil ($\theta = 0$ in that case), thus $\frac{\partial V}{\partial D}\Big|_{D=0} > 0$ which means that, in equilibrium, both firms emit strategic debt. As can be checked on Figure 6.2, the final Nash quantities are greater, thus the strategic use of debt is also welfare enhancing since it increases the intensity of competition in the consumer market.

The analysis so far has been conducted with quantity as the strategic variable. Since prices and quantities tend to yield opposite conclusions, it is worthwhile to take a closer look at the case of prices. If we keep interpreting σ as good news for marginal profit, we have $\frac{\partial^2 \pi}{\partial p \partial \sigma} > 0 \Leftrightarrow \frac{\partial^2 \pi}{\partial x \partial \sigma} < 0$ following our convention $x = -p$. We then obtain $\frac{\partial x}{\partial D} < 0 \Leftrightarrow \frac{\partial p}{\partial D} > 0$ i.e., higher leverage motivates the firm to raise her price. Since prices p and ρ (the competitor's price) are strategic complements, we have $\frac{\partial \rho}{\partial p} > 0$ thus $\frac{\partial \rho}{\partial D} = \frac{\partial \rho}{\partial p}\frac{\partial p}{\partial D} > 0 \Leftrightarrow$

$\frac{\partial y}{\partial D} < 0$ following our convention $y = -\rho$. Using the fact that $\frac{\partial \pi}{\partial \rho} > 0 \Leftrightarrow \frac{\partial \pi}{\partial y} < 0$, the analysis performed over equation (6.21) remains identical for prices: the first term is zero, the second one is zero at the unleveraged benchmark while the third one, being the product of two positive terms, is positive. We may conclude that the strategic effect of raising leverage is beneficial when firms compete in prices.

Chapter 7

Economic Rivalry: Contest and Conflict

In this chapter, we leave market-mediated competition to address the more general notion of **rivalry**, when economic agents, driven by self interest, interact in a strategic manner. This chapter, using the tools of game theory, will provide us with models useful for Industrial Organization and beyond, for all microeconomics.[1@] We distinguish *organized* from *spontaneous* rivalry under the headings of *contest* and *conflict*.

The plan of the chapter is not linear. After an introduction linking rivalry to more traditional economic concepts, we study purely wasteful conflict under a guise of assumptions to test predictions. Then, we move to the arbitrage between production and appropriation, the so-called productive conflicts. In the third section, we formalize a number of stylized model of strategic interaction pertaining to political economy which from our IO point of view are all the relationships taking place outside markets. We also analyze various refinement to understand the impact of risk, group play or information asymmetries. Lastly, we study patent races and long lasting conflicts known as attrition.

7.1 Introduction

7.1.1 Typology

Economic rivalry is best introduced with a few quotes:

Edgeworth (1881) The first principle of economics is that every agent is actuated only by self-interest. The workings of this principle may be viewed under two aspects, according as the agent acts without or with, the consent of others affected by his actions. In wide senses, the first species of action may be called war; the second, contract.

Pareto (1906) The efforts of men are utilized in two different ways, they are directed to the production or transformation of economic goods, or else to the appropriation of goods produced by others.

Lerner (1972) An economic transaction is a solved political problem. Economics has gained the title of queen of the social sciences by choosing solved political problems as its domain.

Williamson (1985) Homo economicus engages in a full set of ex ante and ex post efforts to lie, cheat, steal, mislead, disguise, obfuscate, feign, distort, and confuse.

Edgeworth (1881) stresses the dichotomy between violent and peaceful pursuit of happiness whereas Pareto (1906) assumes a modern legal framework where violence has been eradicated. He then emphasizes the dichotomy between productive and unproductive enrichment. Following the latter, we disregard coercion and violence so that we fully adhere to *individual rationality* in the sense that economic agents freely enter into contracts over items that have recognizable property rights (cf. §8.1 for a justification). Lerner (1972)'s quote now makes full sense: there is a complex economic life beyond the overtly simple markets and its idiosyncrasies must be understood. Lastly, Williamson (1985) reminds us that there are many legal or at least non violent means to pursue one's objective. These are a potential source of inefficiencies that contracting can tame, if used properly.

Our typology of lawful appropriation for spontaneous and organized rivalry is *conflict* and *contest*, examples of which are shown in Tables 7.1 and 7.2.[2@]

This typology is far from perfect. For instance, advertising is not a face-to-face contest between two firms since consumers are involved. The advertising mechanism builds on the fact that no customer is bind forever to his current provider; this means that he can switch at will if the promotional effort of a competitor is convincing enough.

A patent race is not a peaceful conflict revolving on appropriation. Firms seek the 20 years legal protection bestowed by a patent because of its market value which originates in the service it renders to consumers; it is thus to some extent socially useful although this benefit must be mitigated by the cost of developing the innovation (cf. §12.2.3).

Rent-seeking is an implicitly organized contest since its very existence arises from a decision taken by the State. To avoid rent-seeking, it is enough to award the license at random (using a lottery) but by doing so, the State bureaucrat foregoes a useful interaction with lobbyists. It does not matter whether he seeks a bribe or relevant information in order to make an efficient choice.

Procurement mixes contest and conflict because a contender will make both a productive investment to develop a good project and a power (lobby) investment to gain influence over the jury (cf. §14.4.1).

Advertising and marketing as non price market competition (cf. §11.5).

Rent-seeking: vying for State awarded economic rents (either by a bureaucrat or a politician). Applies in international trade policy, industry regulation, price support schemes, privatization or subsidies (cf. §16.3).

Innovation and *Patent* races see firms invest into R&D to develop drugs, technical devices or ideas (software) in order to either clinch a leadership or receive a legal monopoly for 20 years (cf. ch.12).

Litigation in court over ill defined property rights (cf. §14.3), especially intellectual ones (cf. §12.3). This wasteful activity is a key ingredient of transaction cost economics (cf. §13.3.3).

Influence over tax rates or wages indirectly through the media or directly through strikes to pressure a decision maker, be it a court, a government or a firm.

Political competition for elected offices by swaying electors (public choice theory).

Attrition: battle to control a new technology or standard (cf. Bertrand Paradox in §7.4.2).

Externality: congestion (cf. §25) or depletion of a common resource that force other to bear extra cost (indirect conflict).

Table 7.1: Examples of Economic Conflicts

Apart from most auctions where payment of bids is contingent on winning the item, contests and conflicts tend to display an "all-pay" structure whereby everybody, winners and losers alike, forfeit the resources expanded to win the prize. This feature tends to make the aggregate effort or investment large, possibly greater than the prize. In productive contests, this is a desirable property while in the case of unproductive conflicts, this is a source of inefficiency.

7.1.2 Origins of Economic Rivalry

A conflict arises when a valuable item has weak property rights, so that it becomes cheaper for some agents to engage in a struggle over possession rather than buy it. The means so employed can be legal (e.g., litigation, influence, pressure), can border on illegality (e.g., white collar crime, creative accounting), involve deception (e.g., fraud, theft, debt defaulting, lying about relevant characteristics) or, at the limit, use violence (coercion, extortion, outright aggression and war). Since the legal system is not able to perfectly protect us from these risks, we are forced to spend resources at protecting ourselves; as shall be shown later on, this is a major source of inefficiency and waste.

Contests are more specific to the economy since they may be viewed as optimal responses to market failures. In the oligopoly models of this book as well as in the perfect competition paradigm, the interaction of firms is mediated by the market through prices. There are however many reasons why the (anonymous) market may fail to perform the

Auctions: the seller puts an item for sale with the hope of attracting many bidders and earning the largest possible income (cf. ch.22).

Beauty Contests: an implicit auction where contenders expand effort in a more or less licit manner to impress the jury and win the prize e.g., olympic games locations, license awarding (cf. §22.1.2).

Procurement: Firms or public bodies design a set of requirements and let contenders strive to achieve them at the lowest cost (cf. §14.4.1).

Tournaments: Firms and organizations use relative performance tools to decide on promotions or design reward schemes with the varied aims of ranking (cf. screening in 21.2) or motivating employees (cf. incentive theory in §20).

Regulation of public services providers: tournaments are used to gather discriminating information regarding firms and be able to reward and punish them (cf. yardstick competition §17.3.2).

Education: a filter to assign people to jobs or schools (cf. screening in §21.2).

Sport: competitions where the objective function of participants goes beyond the maximization of monetary rewards to include fame and other forms of social recognition.

Scientific Prizes: intended to foster effort, imagination and ultimately advance human knowledge (cf. §12 and Bays and Jansen (2009)).

Table 7.2: Examples of Contests

desired exchange. To name a few, property rights are difficult to define or enforce, the product specifications are not standard, the traded item is indivisible or there is asymmetric information regarding the value and cost of the item. These transaction costs may lead the parties to engage in direct contracting (which is the object of Part F on integration). An implicit assumption of face-to-face relationships is that the parties know they can achieve a mutually beneficial trade together. In many situations, however, each prospective trader has to search for the ideal partner; it thus makes sense to organize a *contest*; the rules and prizes will be designed to elicit the private information of contenders (adverse selection) and/or motivate them toward maximum effort (incentives). An organized contest is thus strategically similar to a conflict but it is *productive* as it serves the promoter's objective, which is often aligned with society's own values.[3@]

7.1.3 On Rivalry and Efficiency

According to von Weizsäcker (1980), economic competition can take place at the levels of consumption, production and innovation to which we respectively associate the concepts of allocative, productive, and dynamic efficiency.

Direct competition (aka rivalry) in consumption is generally prohibited by property rights (e.g., A cannot lawfully compete with B to occupy B's house) but competition for

the exchange of the property rights relating to consumption is lawful (e.g., A can compete with others to buy B's house). Restricting competition at the level of consumption (by upholding property rights) gives incentives for productive activities and hence for enhancing the means of consumption. Analogously, intellectual property rights (cf. §12.3) prohibit some forms of competition in production in order to stimulate competition in innovation, which in turn expands the means of production. We cannot then talk of more or less competition, but of more competition at one level and less at another. The limits set by law upon economic rivalry aim to uphold the *Holy Trinity* of efficiency:

❶ **Productive Efficiency** parties invest resources into welfare enhancing activities (seller minimize production cost, buyer maximize value of use).

❷ **Allocative Efficiency** parties create value through exchange (from low WTP towards high WTP).

❸ **Dynamic Efficiency** resources are efficiently invested into economic sectors using the return on capital criteria.

According to Usher (1987), of all forms of rivalry, theft epitomizes an antithesis of the *Holy Trinity*, we choose to call the *Wicked Trinity*:

① Parties waste resources into offense and defense, not value creation.[4@]

② Value is reduced either by the destruction of items or their inadequate ownership; at best there is a wealth transfer.

③ Economic sectors prone to theft are under-developed as they yield poor returns, thus creating a deadweight loss.

Legal forms of rivalry partake some but not all of these wicked ways. A negative externality or a free rider problem features only ③, inadequate investment because the absence of property rights prevents the correct allocation of cost and benefits (cf. §2.4.4).

Over-exploitation of a common pool resource (CPR) features ① and ②, the latter because excessive extraction today reduces the available stock for tomorrow and the former because, in anticipation of this fate, all users enter a race to the bottom with excessive investment.

Advertising and other profit seeking strategies feature ① and generate ③ as products become dearer and sell less. The opposite of ② may nevertheless appear if advertising is informative or when the profit seeking strategy has a productive component.

Monopoly statutes and publicly awarded rents motivate ① (investment into acquisition or to bar a challenger). Monopolistic pricing then generates ③ directly as the standard dead weight loss but also indirectly because the extraordinary returns allowed by monopolization foster too much investment in the sector.

Tax incidence in public finance studies how the tax base reacts to the tax code and is thus about ① (inefficient investments in tax avoidance) and ③ (distortion of economic activity).

7.2 Wasteful Conflict

Our focus is on conflicts such as rent-seeking (cf. §16.3), advertising, lobbying or influence; note however that beauty contests share strategic similarities. We content ourselves to develop some game theoretical models addressing the issue of rent dissipation in a partial equilibrium framework i.e., how much of the prize is wasted by the efforts to win it. After presenting the symmetric case, we look at a variety of asymmetries that dramatically reduce the wealth dissipation.

In contrast to productive activities in which inputs are combined *cooperatively* in the production technology, the inputs to appropriation are combined *adversarially* in the conflict technology. Game theoretic models of contest and conflict either use a perfectly or imperfectly discriminating contest technology i.e., either the strongest side wins for sure or she simply has an edge over the rest of the flock. In this section, we adopt the latter setting. Auctions, by their very definition, adhere to the former (cf. §22).

This section encompasses four parts. In the first, we consider the simplest and most famous model of contest with exogenous prize and identical contestants. Then, we expand this basic staple by allowing many forms of asymmetry among participants. Next, we take a different path by endogenizing the prize to show that the contest paradigm covers many models of strategic interaction. Lastly, we show how some conflicting situations can be brought toward efficiency using taxation.

7.2.1 Wealth Dissipation

Many partial equilibrium economic models of rivalry, whether spontaneous or organized, rely on strategic contests whereby a prize (exogenous or endogenous) is alloted among contenders. In the *sharing* interpretation, each receives a percentage of the prize value whereas in the *lottery* interpretation, a unique agent, the winner, receives the entire prize. Under risk neutrality, both visions are equivalent in terms of induced behavior. Like most of the literature, we follow Tullock (1980)'s lottery approach for ease of exposition.[5@]

The basic payoff structure involves n participants (denoted i, j), each choosing an investment k_i (in either financial or human capital). We denote v_i the WTP of player i for the prize and by p_i his probability of winning the contest (alternatively, his share

of the prize). Assuming risk neutrality, a party cares only for her expected profit which is $\pi_i = p_i v_i - k_i$. Most game theoretic models of imperfectly discriminating contests or conflicts use the separable contest success function (CSF) with[6@]

$$p_i = \frac{h_i(k_i)}{\sum_j h_j(k_j)} \tag{7.1}$$

The *influence technology* $h_i(.)$ transforms a monetary investment k_i into an amount of leverage $e_i = h_i(k_i)$ that can be meaningfully compared to the challengers' choices. As customary in economics for activities performed by humans, we assume decreasing returns to scale ($h_i'' < 0 < h_i'$) so that p_i is also increasing concave in own investment.[7@] The payoff of contender i in the contest is

$$\pi_i \quad = \quad \frac{h_i(k_i)}{\sum_j h_j(k_j)} v_i - k_i \quad = \quad \frac{e_i}{\sum_j e_j} - c_i(e_i) \tag{7.2}$$

where $c_i(.) \equiv \frac{1}{v_i} h_i^{-1}(.)$ is convex increasing. The interplay between *investment* and *leverage* shows that the canonical model of rent-seeking is one where participants seek a unitary prize allocated by a pure lottery but with differentiated cost of participation.[8@] It often makes sense to assume a common influence technology with the power function while allowing for private valuations i.e., $c_i(e) = \frac{e^\alpha}{\alpha v_i}$ with $\alpha \geq 1$. Observe then that a higher prize valuation is akin to a greater influence ability.

By convention, for any list $(x_j)_{j \leq n}$, we define $x_0 \equiv \Sigma_j x_j$ as the aggregate contribution and $x_{-i} \equiv \Sigma_{j \neq i} x_j$, the aggregate influence of i's challengers, the FOC of optimal individual behavior is easily computed as

$$0 = \frac{\partial \pi_i}{\partial e_i} = \frac{e_{-i}}{e_0^2} v_i - e_i^{\alpha - 1} \quad \Rightarrow \quad e_0^2 e_i^{\alpha - 1} = e_{-i} v_i \tag{7.3}$$

In many rent-seeking situations, it makes sense to assume a constant and common valuation v of the prize because it is determined by the State and has a well known market value which contestants cannot alter. In that case, the equilibrium is symmetric with $e_i = e$ so that (7.3) becomes $n^2 e^{\alpha+1} = (n-1)ve$. The individual investment is then $c(e) = \frac{1}{\alpha} e^\alpha = \frac{n-1}{\alpha n^2} v$ while the aggregate cost, $nc(e)$, is a proportion $\beta \equiv \frac{n-1}{\alpha n}$ of the prize v; it is called the index of *rent dissipation*. The equilibrium expected individual profit is $\pi^* = \frac{v}{n}\left(1 - \frac{n-1}{\alpha n}\right) > 0$ since $\alpha \geq 1$. This shows that each contender has an incentive to participate, even after factoring the cost of influence. A larger prize v triggers more influence activity but as the rent dissipation index is lesser than unity, participants still gain.

The original literature on rent-seeking and regulatory capture assumed $\alpha = 1$ and n

large so that the rent dissipation was to be almost complete. As we show hereafter, many realistic features contribute to reduce this dissipation. Furthermore, the lottery models are ad-hoc and fail to provide a micro-foundation of the interplay between contenders and the contest organizer.

7.2.2 Asymmetry

The basic model of rent-seeking is extended to allow for heterogenous contenders. The general conclusion will be a softening of the previous striking conclusion regarding rent dissipation. An example of asymmetric situation is the regulation of a monopoly as it involves firms and consumers who differ in their valuation of the rent to be or not to be awarded. Indeed, as we saw in 3.2, the consumer surplus rises more when passing from monopoly to efficiency than the producer surplus upon moving in the opposite direction. It is thus important to assess the impact of asymmetric valuations.

Asymmetric Valuations

Hillman and Riley (1989) focus on the pure lottery contest technology ($\alpha = 1$). The FOC (7.3) now reads $e_0^2 = (e_0 - e_i)v_i \Rightarrow e_i = e_0(1 - e_0/v_i)$ i.e., the greater the WTP, the greater the influence expanded. Summing over the participants, we get $e_0 = ne_0 - e_0^2/\hat{v}_n$ where the harmonic mean is defined by $\frac{1}{\hat{v}_n} \equiv \frac{1}{n}\Sigma_j \frac{1}{v_j}$. We deduce the total effort $\hat{e}_0 = \frac{n-1}{n}\hat{v}_n$ i.e., dissipation is reduced (w.r.t symmetry) since $\hat{v}_n < \bar{v}_n \equiv \frac{1}{n}\Sigma_j v_j$ is always true.[9@] For later use, we derive the individual investments of the bilateral case. When $n = 2$, $\hat{v} = \frac{2v_i v_j}{v_i + v_j}$ and $\hat{e}_0 = \frac{\hat{v}}{2}$, thus individual investments are

$$e_i = \frac{v_j v_i^2}{(v_i + v_j)^2} \tag{7.4}$$

which satisfies the intuition according to which the more motivated participant (greater valuation) invests more in order to produce a greater influence on the outcome.

Lastly, we must check how many contenders are active since the asymmetry of values generates an asymmetry in profits. We have $\frac{\partial \pi_{n+1}}{\partial e_{n+1}} = \frac{v_{n+1}}{e_0} - 1$. For $e_{n+1} = 0$, $e_0 = \Sigma_{j \le n+1} e_j = \Sigma_{j \le n} e_j = \frac{n-1}{n}\hat{v}_n$ at the equilibrium of the contest with n active participants. The ultimate entrant is thus n such that $nv_{n+1} < (n-1)\hat{v}_n$. As a matter of example, we consider geometric ranking with $v_i = v_0\gamma^i$ for $0 < \gamma < 1$. The solution γ_n to the entry equation $\gamma(1 - \gamma^n) = (n-1)(1-\gamma)$ means that for $\gamma < \gamma_n$ at most n contestants participate. Table 7.3 displays numerical results and the corresponding dissipation (computed at the maximum WTP).

n	1	2	3	4	5	8	∞
γ_n (%)	0	62	81	89	93	98	100
β (%)	0	38	53	62	69	78	100

Table 7.3: Contest Performance

Asymmetric Burden

A typical case of asymmetric valuations occurs with lobbying over a policy measure (e.g., subsidy or tax avoidance) whose potential recipient is specifically group #1. If the government is restricted to budget neutrality, the cost v of the policy, if implemeted, is spread equally over the remaining groups; hence $v_1 = v$ whereas $v_i = \frac{v}{n-1}$ for $i > 1$ (payment avoidance). Applying (7.3) to groups #1 and #i for $\alpha = 1$, we have $e_{-1}v_1 = e_0^2 = e_{-i}v_i$. Using $e_{-i} = e_1 + e_{-1} - e_i$ and the fact that $e_i = \frac{e_{-1}}{n-1}$ in equilibrium, we get $e_{-1}\frac{v_1}{v_i} = e_1 + e_{-1}\frac{n-2}{n-1}$, thus $e_1 = \frac{n^2-3n+3}{n-1}e_{-1}$. Plugging back into $e_{-1}v = e_0^2$, we finally obtain $e_{-1} = \left(\frac{n-1}{(n-1)^2+1}\right)^2 v$ so that the wealth dissipation is $\beta = \frac{e_1+e_{-1}}{v} = \frac{n-1}{(n-1)^2+1} \simeq \frac{1}{n-1}$. Lastly, the probability that the measure passes the voting hurdle is $p_1 = \frac{e_1}{e_1+e_{-1}} = 1-\beta$. This result rationalizes the observation by Pareto (1906) that "A policy measure benefitting one group but paid by all is hardly opposed".

Commitment

Dixit (1987) studies Stackelberg leadership in rent-seeking games and shows that among identical participants, everyone would like to become the leader, a result reminiscent of quantity competition.

Consider then a symmetric duopoly rent-seeking game and assume that one party, say #1, can pre-commit her influence effort k_1. She will overinvest if the indirect effect of her influence is positive i.e., when $\frac{\partial \pi_1}{\partial k_2}\frac{\partial k_2}{\partial k_1} > 0$.[10@] By construction, $p_1 + p_2 = 1$ and since an investment is always useful it must be true that $\frac{\partial p_1}{\partial k_2} < 0$, hence $\frac{\partial \pi_1}{\partial k_2} = \frac{\partial p_1}{\partial k_2} < 0$ which means that overinvestment requires strategic substitutability $\frac{\partial k_2}{\partial k_1} < 0 \Leftrightarrow 0 > \frac{\partial^2 \pi_2}{\partial k_2 \partial k_1} \propto \frac{\partial^2 p_2}{\partial k_2 \partial k_1}$ since the impact of k_1 on π_2 runs only through the winning probability p_2.

To be able to conclude, we must consider explicitly the separable CSF (7.1) to derive $\frac{\partial p_2}{\partial k_1} = \frac{h_2 \dot{h}_1}{(h_1+h_2)^2} \Rightarrow \frac{\partial^2 p_2}{\partial k_1 \partial k_2} = \frac{\dot{h}_2 h_1 (h_2 - h_1)}{(h_1+h_2)^3} \propto h_2 - h_1$. Hence, the leader is motivated to overinvest when she is the most influential participant. In conclusion, the strongest contestant would like to pre-commit in order to extend further her leadership. Observe that this conclusion carries to the oligopoly context.

We can recover this result analytically in the duopoly with pure lottery. The Cournot equilibrium is given by (7.4) but as we need the best reply, we have to solve the FOC $\frac{v_i k_j}{(k_i+k_j)^2} = 1$ i.e., $k_i = \sqrt{v_i k_j} - k_j$. The Stackelberg leader maximizes $\hat{\pi}_i = \frac{v_i}{\sqrt{v_j}}\sqrt{k_i} - k_i$ hence

solves $\frac{v_i}{2\sqrt{k_i}\sqrt{v_j}} = 1 \Rightarrow k_i^L = \frac{v_i^2}{4v_j}$ and the follower chooses $k_j^F = \frac{v_i}{4v_j}(2v_j - v_i)$. We check that $v_i > v_j \Leftrightarrow k_i^L > k_j^F$. Observing that total investments are $k^C = \frac{v_i v_j}{v_i + v_j}$ and $k^S = \frac{1}{2}v_i$, we have $k^S > k^C \Leftrightarrow v_i > v_j$ so that the ability to pre-commit is socially wasteful if and only if it is handed over to the most interested player.

Litigation

In many conflicting situations brought to justice for settlement, an element of fault or misconduct impinges on the outcome. We may cite divorce, patent infringement, wrongful contract termination, insufficient performance or failure to deliver (a good, a service). Ideally, tort or innocence should be attributed with certainty to whom it belongs. Likewise, the outcome should only depend on litigation efforts if there is an equal degree of fault and should only depend upon the degree of fault if there is an equal litigation effort. Modeling these features after Hirshleifer and Osborne (2001), we show that litigation can be seen through the lens of the canonical conflict.

Letting $x_i \in [0;1]$ indicate party i's degree of fault and k_i her litigation effort, a sensible way to meet these criteria is for the ratio of winning probabilities to be $\frac{p_i}{p_j} = \left(\frac{x_j}{x_i}\right)^\beta \left(\frac{k_i}{k_j}\right)^\gamma$. The positive parameter β and γ make each ratio more or less influential. For instance, the codified legal systems (e.g., French and Germanic) leave little leeway to attorneys so that $\beta \gg \gamma$ whereas in common law systems (e.g., US and UK) the reverse would hold. Since $p_j = 1 - p_i$, the influence function is $h_i(k_i) \propto k_i^\gamma x_i^{-\beta} \Rightarrow c_i(e_i) \propto (e_i x_i^\beta)^{1/\gamma}$ so that we are back to the canonical model with individual valuation $v_i \propto x_i^{\beta/\gamma}$ which tends to infinity when the agent becomes the deserving owner of the item ($x_i \to 0$). Using (7.4) for $\gamma = 1$, we deduce that the more deserving agent (smaller fault factor) invests more i.e., being right motivates to fight harder.

If the law contemplates this option, conflicting parties can settle their difference out-of-court with a side payment (withdrawing their lawsuit and counter-lawsuit). This way, they avoid the costly legal expenses of an actual trial, the negative media exposure and the inefficiencies associated with a rigid court judgement. However, parties often disagree wrt. their odds of winning the trial making their demands incompatible. As they fail to reach an agreement, trial becomes inevitable. In §7.3.5, we model this failure of the Coase theorem.[11@]

Incumbent Advantage

Fisher (1985) argues that most rent-seeking contests have an incumbent whose renewal is almost assured. As we show hereafter, asymmetrically empowered rent-seekers jointly spend less in wasteful rent-seeking activities than if they are symmetrically empowered.

The result also applies to rent-seekers within an organization, public (e.g., agency) or private (e.g., firm) where one stands at a higher rank on the hierarchy ladder.

Assume that contender #1 is an incumbent whose rent-seeking investment is $\lambda > 1$ times more productive than other people's money. The asymmetric winning probabilities are $p_1 = \frac{\lambda k_1}{\lambda k_1 + k_{-1}}$ and $p_i = \frac{k_i}{\lambda e_1 + e_{-1}}$ for $i > 1$. For the case of linear cost $\alpha = 1$, expected profits are

$$\pi_1 = \lambda \left(\frac{k_1}{\lambda k_1 + k_{-1}} v - \frac{k_1}{\lambda} \right) \quad \text{and} \quad \pi_i = \frac{k_i}{\lambda k_i + k_{-i}} v - k_i \text{ for } i > 1$$

The FOCs are then

$$\lambda v k_{-1} = k_0^2 = v k_{-i} \text{ for } i > 1 \tag{7.5}$$

In a semi-symmetric equilibrium, we have $k_i = k$ for $i > 1$. The FOC system (7.5) now reads $\lambda(n-1)k = (n-2)k + k_1 \Rightarrow \frac{k_1}{k} = (n-1)\lambda - n + 2$. Then, the left FOC yields $\lambda v(n-1)k = (k_1 + (n-1)k)^2 = k^2((n-1)\lambda + 1)^2 \Rightarrow k^* = \frac{(n-1)\lambda v}{((n-1)\lambda+1)^2}$. Lastly, the degree of rent dissipation is total cost over the prize value i.e.

$$\beta = \frac{k_1^*/\lambda + (n-1)k^*}{v} = \frac{k^*}{v}\left(\frac{(n-1)\lambda - n + 2}{\lambda} + n - 1\right) = \frac{2\lambda - \frac{n-2}{n-1}}{\left(\lambda + \frac{1}{n-1}\right)^2}$$

so that $\beta \simeq \frac{2\lambda - 1}{\lambda^2}$ when n is large or $\beta \simeq \frac{2}{\lambda}$ when λ is large.

In a typical case with three challengers ($n = 4$) and where the incumbent holds as much effectiveness as they do altogether ($\lambda = 3$), then rent dissipation is 50% instead of 75% if players were equally influential.

7.2.3 Canonical Rivalry

Up to now, the prize was exogenously given; this suits well the idea of a contest where the organizer determines the prize characteristics ex-ante. Many forms of rivalry (all conflicts and some contests) however involve an endogenous prize determined by the conjoint actions of contenders. In the following generalization, the participants' inputs will simultaneously determine the size and distribution of the prize.

This is the case for instance when economic agents work in team and pool their efforts to produce a result which is then divided in proportion of individual contributions. The exploitation of a common pool resource (cf. §18.2.3) presents a similar situation since aggregate effort determine total catch or extraction while nature makes sure that, on average, each receives a share of output in proportion to his input. Still another application of this general setting is the Cournot model for an homogeneous good given that the price determination pools the efforts of agents, here the individual outputs of firms (cf. §5.1).

The canonical model of rivalry takes the canonical model of rent-seeking (7.2) and assumes that investments $\mathbf{k} = (k_i)_{i\leq n}$ are perfect substitutes whose aggregate $k_0 \equiv \sum_{i\leq n} k_i$ determines the prize $V(k_0)$. We assume that investment has always a marginal value i.e., $V_m > 0$ but that the unitary value $v(k_0) \equiv \frac{V(k_0)}{k_0}$ is decreasing (a property close to DRS). Given proportional sharing of the prize i.e., $p_i = \frac{k_i}{k_0}$, individual payoff is

$$\pi_i = \frac{k_i}{k_0} V(k_0) - k_i = v(k_0) k_i - k_i \tag{7.6}$$

In the Cournot model with common marginal cost c, $V(q_0) = \frac{q_0 P(q_0)}{c}$ is the normalized market revenue generated by the aggregate output of firms (cf. §5.1). In the Commons model, investments are pooled to extract valuable output from the community owned stock (cf. §18.1.4).

The collective benefit is $\pi_0(k_0) \equiv \sum_{i\leq n} \pi_i = V(k_0) - k_0$. Observe that in this particularly simple model, each participant gets his fair share since we have $\pi_i = p_i \pi_0$ (by construction). The aggregate profit is maximum for k_0^* solving $V_m = 1$, yielding π_0^*. In the Nash equilibrium, an individual firm solves the FOC

$$1 = k_i v_m + v = k_i \frac{V_m - v}{k_0} + v = p_i V_m + (1 - p_i) v \tag{7.7}$$

i.e., marginal cost equal to a weighted average of marginal and average value.

We immediately observe that the equilibrium must be symmetric because (7.7) yields $p_i = \frac{V_m - v}{1 - v}(k_0)$ which is independent of the identity. It must therefore be true that $p_i = \frac{1}{n}$ so that (7.7) becomes

$$1 = \frac{1}{n} V_m(k_0) + \frac{n-1}{n} \frac{V(k_0)}{k_0} \tag{7.8}$$

i.e., the marginal cost (LHS) is an average between marginal revenue and average revenue with weights determined solely by the market structure. Equation (7.8) thus ranges from the monopolistic case for $n = 1$ to the competitive limit with free entry as n grow large. Observe finally that the monopoly implements the collective optimum i.e., the inefficiency brought about by non-cooperative interaction can only be eliminated by forming a cartel. Alternatively, the collectivity may use a tax instrument to achieve the same efficient outcome as shown in §7.2.4.

Exogenous Shock†

Kotchen and Salant (2009) inquire the impact of a cost increase (or a value decrease) upon aggregates in a contest where technology is common. Removing the cost normalization used above, individual profit becomes $\pi_i = (v(k_0) - c) k_i$. The FOC for profit

maximization remains (7.7) but with c on the LHS. Since it is symmetric, we have $nc = \beta(k_0) \equiv nv(k_0) + k_0 v'(k_0)$. Symmetry is also used to expressed the SOC in a leaner way:

$$0 \geq 2v' + k_i v'' \propto 2nv'(k_0) + k_0 v''(k_0) \Leftrightarrow \gamma \leq 2n$$

where $\gamma \equiv \frac{k_0 v''(k_0)}{-v'(k_0)}$ is the elasticity of the slope of v. We get a unique equilibrium if the RHS of the FOC is decreasing i.e., if

$$0 \geq n\beta' = (n+1)v'(k_0) + k_0 v''(k_0) \Leftrightarrow \gamma \leq n+1$$

Under this condition, $\frac{dk_0}{dc} < 0$ and since $k_i = \frac{k_0}{n}$, we also have $\frac{dk_i}{dc} < 0$. To tackle the effect of an exogenous cost shock, we compute

$$\frac{d\pi_i}{dc} = \frac{\partial \pi_i}{\partial k_i}\frac{dk_i}{dc} + (n-1)\frac{\partial \pi_i}{\partial k_j}\frac{dk_j}{dc} - 1 = 0 + (n-1)v'(k_0^*)\frac{dk_j}{dc} - 1$$

and since $\frac{dk_j}{dc} = \frac{dk_j}{k_0}\frac{dk_0}{dc} = \frac{1}{n}\frac{1}{\beta'}$, we obtain

$$\frac{d\pi_i}{dc} = \frac{n-1}{n}\frac{v'}{\beta'} - 1 > 0 \Leftrightarrow (n-1)v'(k_0^*) > n\beta'(k_0^*) \Leftrightarrow \gamma > 2 \Leftrightarrow 2v'(k_0) + k_0 v''(k_0) > 0$$

which is equivalent to V being convex (recall that $V(k) = kv(k)$).

Thus, when there are sufficiently many firms (n not too low) and value is convex at the equilibrium, we may have the paradoxical result that a cost increase benefits all members of the industry (case where $2 < \gamma < n+1$ holds true). In the pure Cournot model studied by Seade (1985), $\Phi(x) = x$, thus $v(.) = P(.)$ and $\gamma > 2$ is equivalent to local convexity of market revenue $q_0 P(q_0)$ around the equilibrium. In our leading linear demand example, market revenue is concave thus the paradox does not occur. However, it will arise if the demand displays a low constant elasticity ϵ since in that case we have $\gamma = 1 + 1/\epsilon > 2 \Leftrightarrow \epsilon < 1$.

7.2.4 Enforcing Agreements

Consider a strategic interaction where the Nash equilibrium is deemed collectively inadequate because society's members are trapped in a prisoner dilemma (cf. §2.4.1). Think for instance of the over-exploitation of a natural resource under open-access wrt. private property management (cf. §18.1.4) or the oligopolistic equilibrium of an industry wrt. the monopolization by a cartel (cf. §9.1).

We shall prove that when firms or agents are only mildly asymmetrical, the ideal out-

come can be reached if the collectivity can institute a tax system and redistribute part or all of its proceed without incurring too large a transaction cost. Setting adequately this Pigouvian tax, each member remains free to pick his strategic decision.[12@] The opposite situation where the non-cooperative equilibrium features an under achievement is solved in a symmetrical fashion with a subsidy instead of a tax.

Theory

Consider an economic interaction summarized by the game $\left(\pi_j(\mathbf{q})\right)_{j\leq n}$ where $\mathbf{q}=(q_j)_{j\leq n}$ indicate strategies. We designate the Nash equilibrium with exponant symbol $\hat{}$ while * stands for an ideal situation for which we assume aggregate payoff dominance: $\Delta\equiv\pi_0^*-\hat{\pi}_0>0$ where $\hat{\pi}_0\equiv\sum_{j\leq n}\hat{\pi}_j$ and $\pi_0^*\equiv\sum_{j\leq n}\pi_j^*$.

The unit taxes are $\boldsymbol{\tau}=(\tau_j)_{j\leq n}$, the revenue collected $T\equiv\sum_{j\leq n}\tau_j q_j$ is redistributed according to shares $\boldsymbol{\lambda}=(\lambda_j)_{j\leq n}$ satisfying $\lambda_j\geq 0$ and $\sum_{i\leq n}\lambda_i\leq 1$ to guarantee that the scheme does not run a deficit. The modified game has payoff $\tilde{\pi}_i(\mathbf{q})\equiv\pi_i(\mathbf{q})-\tau_i q_i+\lambda_i T$. The FOC characterizing the player's best reply is $0=\frac{\partial\tilde{\pi}_i}{\partial q_i}=\frac{\partial\pi_i}{\partial q_i}-(1-\lambda_i)\tau_i$. The unit tax τ_i needs to solve the previous FOC for q_i^* i.e., $\tau_i=\frac{\beta_i}{1-\lambda_i}$ where $\beta_i\equiv\left.\frac{\partial\pi_i}{\partial q_i}\right|_{\mathbf{q}^*}$. Notice that a larger fraction λ_i of tax revenue leads to a larger tax rate τ_i.

If $\sum_{i\leq n}\lambda_i=1$ (i.e., there is full refund), then total profit in equilibrium is $\tilde{\pi}_0=\pi_0^*-\sum_{i\leq n}\tau_i q_i^*+\sum_{i\leq n}\lambda_i T=\pi_0^*-T+T=\pi_0^*$. If the starting situation is entirely symmetric in terms of preferences $\left(\pi_j(.)\right)_{j\leq n}$, ideal output $\mathbf{q}^*$ and redistributive shares $\boldsymbol{\lambda}$, then tax rates are also symmetric. In that case, individual profit is just the fair share of total profit and it is obvious that all players gain by Δ/n. By continuity, the result must carry to a mix of incomplete refunding and mild asymmetries, the individual rationality (IR) condition for participation being

$$\begin{aligned}\hat{\pi}_i &\leq \tilde{\pi}_i=\pi_i^*-\tau_i q_i^*+\lambda_i T=\pi_i^*-\frac{\beta_i q_i^*}{1-\lambda_i}+\lambda_i\sum_{j\leq n}\frac{\beta_j q_j^*}{1-\lambda_j}\\ &\Leftrightarrow \lambda_i\sum_{j\neq i}\frac{\beta_j q_j^*}{1-\lambda_j}\geq\beta_i q_i^*-\pi_i^*+\hat{\pi}_i\end{aligned}$$

or in aggregate $\sum_{i\leq n}\lambda_i\sum_{j\neq i}\frac{\beta_j q_j^*}{1-\lambda_j}\geq\sum_{i\leq n}\beta_i q_i^*-\Delta$. If some of these inequalities are violated, one needs to compensate involved players in an ad-hoc fashion which is likely to be impractical and/or unfeasible.

If firms are identical, it makes sense to consider a common share λ and the (IR) condition becomes $\hat{\pi}\leq\pi^*+\frac{1-\lambda n}{1-\lambda}\beta q^*\Leftrightarrow\frac{1-\lambda n}{1-\lambda}\leq\alpha\equiv\frac{\pi^*-\hat{\pi}}{\beta q^*}=\frac{\Delta}{\beta Q^*}\Rightarrow\lambda\geq\bar{\lambda}\equiv\frac{1-\alpha}{n-\alpha}$. For any $\lambda\in[\bar{\lambda};\frac{1}{n}]$, firms end up better off. The maximum untax revenue that can used to different purposes (double-dividend) is simply the prize for cooperation $\Delta\,(=(1-\bar{\lambda}n)\,T$ in this case).

Cartel

In the cartel case (cf. §9.1), the ideal is the monopoly outcome and the underlying interaction is Cournot competition for the market demand $D(p)$ with individual cost functions $C_i(q_i)$. The cartel's optimum level of sales $q_0^* = \sum_{i \leq n} q_i^*$ solves $P(q_0^*) + q_0^* P'(q_0^*) = \hat{C}_m(q_0^*) = C_{m,i}(q_i^*) \equiv c^*$, thus

$$\beta_i \equiv \left.\frac{\partial \pi_i}{\partial q_i}\right|_{q_i^*} = P(q_0^*) + q_i^* P'(q_0^*) - C_{m,i}(q_i^*) = \left(q_i^* - q_0^*\right) P'(q_0^*) = \frac{q_0^* - q_i^*}{q_0^*}(p^* - c^*)$$

If firms are identical then $\beta = \frac{n-1}{n}(p^* - c^*) = \frac{(n-1)\pi_0^*}{n q_0^*}$ so that $\alpha = \frac{\Delta}{\beta q_0^*} = \frac{n}{n-1}\left(1 - \frac{\hat{\pi}_0}{\pi_0^*}\right) < 1$. In the case where marginal cost is constant, the ratio of industry profits is $\frac{4n}{(n+1)^2}$ (cf. §5.12), thus $\bar{\lambda} = \frac{3n+1}{2n+n^2+n^3}$ which is quite small.

Commons

Contray to what happens in the cartel case where consumers and welfare are hurt by the enforcement mechanism, the latter may be used in problems of congestion and overexploitation of commons to bring an inefficient equilibrium towards the efficient outcome. As before, the ideal situation is the profit maximizing exploitation of the common that would be undertaken if it were under sole ownership whereas the default situation is the open access equilibrium.

Formally, this is like the Cartel since both share the same underlying model of canonical rivalry (cf. §7.2.3). Hence the lower limit for refunding only depends on the ratio of industry profits between cartelization and competition. If we further apply this result to the logistic case (cf. §18.2.3), then there is a perfect identity with the Cournot model, hence the same lower limit for refunding.

7.2.5 Productive Conflict

Investing into a conflict or a contest to improve one's odds of success is, of necessity, a relinquishment of immediate consumption or production, aka. the "guns vs. butter" tradeoff.[13@] The contested prize then becomes endogenous because it is created by the very parties seeking its appropriation. For instance, labor tournaments commonly reward behavior that contributes to increase productivity within the firm, thus the distributable profit. Likewise, employees within large scale organization routinely engage into lobbying because their output and position is insecure as it can be taken away by a challenger in a back room deal with the hierarchy. The dilemma is then to allocate optimally one's resources between production and appropriation.

Paradox of Power

To account for the production vs. appropriation trade-off, we need to approach rivalry from a general equilibrium perspective as originally done by Haavelmo (1954).[14@] To ease the approach, Hirshleifer (1988) models *Production & Conflict* (PC) in a partial equilibrium framework where decision makers must allocate their scarce human and/or financial capital between creation and capture.[15@]

Each economic agent has a maximum welfare contribution $\bar{q}_i$ reflecting differentiated productive abilities as well as endowments. Each can invest k_i to build influence at the rate $e_i = k_i/\lambda_i$ where a greater λ_i reflect a lower influence ability (productivity).[16@] The net welfare contribution is thus $q_i = \bar{q}_i - \lambda_i e_i$. The influence technology is the standard one with $p_i = \frac{e_i}{e_0}$. Normalizing maximum GDP to unity ($\bar{q}_0 = 1$), the profit of player i is

$$\pi_i = p_i q_0 = \frac{e_i}{e_0} - \hat{\lambda} e_i \text{ with } \hat{\lambda} \equiv \frac{1}{e_0} \Sigma_j \lambda_j e_j \tag{7.9}$$

This is the rent-seeking setting except that the linear cost has an endogeneous slope. Yet, the usual intuition remains correct: an agent better at appropriation invests more into rent-seeking. We use the influence e_i as strategic variable instead of the investment k_i. The FOC of optimal power investment is $0 = q_0 \frac{\partial p_i}{\partial e_i} + p_i \frac{\partial q_0}{\partial e_i}$. Since $\frac{\partial p_i}{\partial e_i} = \frac{1-p_i}{e_0}$ and $\frac{\partial q_0}{\partial e_i} = -\lambda_i$, we obtain

$$\frac{q_0}{e_0} = \lambda_i \frac{p_i}{1-p_i} \tag{7.10}$$

i.e., the product of individual productivity by the hazard rate of victory is constant over all participants.[17@] Given that the RHS ratio is increasing in p_i while the LHS is constant, a better influence technology ($\lambda_i < \lambda_j$) leads to a greater chance of winning ($p_i > p_j$) i.e., a larger influence ($e_i > e_j$). Since differential capture productivities impact the results, one may wonder if differential creative productivities reinforce or counter these. It so happens, in this simple model, that the two productivities play opposite roles i.e., being twice better at capture is like being twice worse at creation.[18@] The two activities are thus complementary and we obtain the *exploitation paradox*:

> In the production vs. appropriation dilemma, agents more creative or less able at capture, earn smaller shares of the overall output.

This paradox explains the prevalence of rent-seeking when binding distribution agreements are unenforceable.[19@] In the case of influence within organizations, the *exploitation paradox* implies that those most able at managing power and human relationships enjoy better jobs and perks than engineers or technicians. Furthermore, once someone has developed an edge into an activity, she is likely to improve further her productivity

through learning and will therefore completely specialize in it.

The previous paradox was conditional on a positive influence activity by all agents in order for the FOC to apply. Under that proviso, differential endowments do not seem to matter. To say more about that we need to solve for the equilibrium. Observe that the FOC for (7.10) reads $(1-p_i)q_0 = \lambda_i p_i e_0 = \lambda_i e_i = \bar{q}_i - q_i$ so that summing over all contenders, we obtain $(n-1)q_0 = 1 - q_0$ (recall $\bar{q}_0 = 1$), thus $q_0^* = \frac{1}{n}$ which means that aggregate expenditure on influence, dissipation, is a fraction $\frac{n-1}{n}$ of maximum potential output as in the lottery model of rent-seeking. From (7.10), we have $q_0^*(e_0 - e_i) = \lambda_i e_i e_0 \Rightarrow e_i = \frac{q_0^* e_0}{q_0^* + \lambda_i e_0}$ and summing again, we derive $e_0 = \Sigma_j \frac{q_0^* e_0}{q_0^* + \lambda_j e_0} \Rightarrow 1 = \Sigma_j \frac{1}{1 + \lambda_j e_0 / q_0^*}$ which has a unique solution e_0^* determining indiviudal degrees of influence.[20@] With two contenders, (7.10) can be worked out to yield the more readily interpretable $\frac{e_j}{e_i} = \sqrt{\lambda_i / \lambda_j}$.

We see that at least half of the potential production (value) is lost in wasteful rent-seeking with a potential full dissipation limit .[21@] We can now draw some conclusions regarding final payoffs in the duopoly case. If parties have the same capture abilities, they share equally the value produced, this independently of their initial wealth differences (i.e., $\lambda_i = \lambda_j$ and $\bar{q}_i \neq \bar{q}_j \Rightarrow \pi_i^* = \pi_j^*$). This is so because productive inputs are pooled before the contest decides on the division. If one of the agents now improves her capture productivity, she ends up investing more into capture, capturing more and earning more. However the inequality will remain stable in subsequent periods. The case of a windfall income is altogether different. If δ_i accrue to party i, other parties grabs a share $1 - p_i^*$ of it. We thus obtain the *paradox of power*:

> In the production vs. appropriation dilemma, wealth is equalized inso far as parties have the same abilities at creation or capture. Any windfall income for one party is necessarily shared with the other in the proportion given by the ratio of creative (or capture) productivities, so that income inequality is reduced.

In the alternative interpretation where λ is a creative productivity, endowments are inputs for final welfare creation. In equilibrium, endowments are wasted at 50% rate into influence. It is quite obvious that if an agent can produce a non appropriable output with a productivity greater than $\frac{1}{2}$, she will forfeit influence activity and risky (though valuable) production to concentrate on secure (although obsolete) means of profit. This can explain why developing countries are autarchic in nature since exports require licenses and more risk taking than local production and retail. This argument can be extended to innovation: in the presence of a weak protection of property rights (e.g., illegal file sharing through peer-to-peer networks), a firm will pass over an innovation if it is too easily copied or illegally distributed.

Settlement

Meanwhile parties engaged in lengthy negotiations (or conflict) have not settled their differences, they must forego alternatives activities, for instance valuable investment opportunities.[22@] More generally, haggling over a distributive issue reduces the amount to be distributed. We may thus assume that, either the managers reach an agreement (λ_i, λ_j) for sharing the prize q or they engage into conflict over the smaller prize θq where $\theta < 1$ measures the transaction cost of squabbling. Under this specification, influence activity is indeed an investment such as building trust with the CEO that can be used later on if no agreement is found with a competing manager.

Since their disagreement payoff is θq while the agreement payoff is q, the value of cooperation is $\delta = (1-\theta)q$ which is shared evenly (cf. §2.4.3). To ease calculations, we assume equal productivities with $\lambda_A = \lambda_B = 1$. The profit is now $\pi_i = p_i \theta q + \frac{1}{2}(1-\theta)q$. The FOCs leads to

$$\frac{e_A}{p_A p_B}\left(\theta p_A + \tfrac{1}{2}(1-\theta)\right) = q = \frac{e_B}{p_A p_B}\left(\theta p_B + \tfrac{1}{2}(1-\theta)\right) \tag{7.11}$$

If $e_A > e_B$ then $p_A > p_B$ by the definitions of the probabilities but then this would imply $e_A < e_B$ in (7.11), a contradiction. The solution is thus symmetric with $e_A = e_B$ and $p_A = p_B = \frac{1}{2}$. Plugging back into the FOC and using the same method as above, we derive $e^* = \frac{\theta}{2(\theta+1)}$ which is increasing with θ whereas final production $q^* = \frac{1}{\theta+1}$ is decreasing with θ. As conflict becomes more wealth-destructive ($\theta \searrow$), parties get less incentives to seek rent and thus share a greater expected prize. Adopting a warlike language, we may summarize this result as:

> Immediate settlement is thus more attractive, as the weapons at disposition of each party become more deadly.

7.3 Political Economy

In this section, we apply the basic rent-seeking model to a range of political economy issues, among which lobbying, political pressure, collective action or the determination of the size of firms.

7.3.1 Lobbying

Since legislators, governments and bureaucrats have the power to enact and implement regulations or legislations, special interest groups vie to influence them in order to sway the outcome towards their interests (cf. §16.3.2). Becker (1983)[23@] analyzes the inefficiencies generated by monetary transfers between SIGs and the government over the

choice of a tax-subsidy program which is opposed by tax-payers and supported by recipients lobbies. The impact of the State's technology is also assessed. Since the gain to one side is the other side's loss, this conflict is a zero sum game as opposed to the constant sum games seen up to now (they revolve on the division of a valuable item).[24@]

Letting denote Q the monetary amount at disposal of the government, the relation to the disposable income of recipients S and the tax-payers contribution T is respectively $S = F_2(Q)$ and $Q = F_1(T)$ where both functions display decreasing returns to scale (like a monopoly's marginal revenue) because the tax collection and distribution systems generate deadweight costs (cf. marginal cost of public funds in §17.1.2). Pressure determines the public decision through the influence function $Q = \Upsilon(k_t, k_s)$ which displays decreasing returns to pressure i.e., $\Upsilon_{ii} < 0 < \Upsilon_i$ for $i = t, s$. As the sign of Υ_{ts} cannot be determined from intuition, we consider $\Upsilon(k_t, k_s) = k_s - \frac{k_s^2}{2} - k_t + \frac{k_t^2}{2} + \gamma k_s k_t$ with γ small but of no particular sign. We let also $F_1(T) = \theta_t T$ with $\theta_t \leq 1$ and $F_2(Q) = Q/\theta_s$ with $\theta_s \geq 1$. This particular labeling choice simplifies the derivation of the equilibrium but one must take care that a higher θ_s worsens the distribution technology while a higher θ_t improves it. The utilities are then

$$\begin{cases} \pi_t = -T - k_t = -\frac{\Upsilon(k_t,k_s)}{\theta_t} - k_t \\ \pi_s = S - k_s = \frac{\Upsilon(k_t,k_s)}{\theta_s} - k_s \end{cases} \quad \Rightarrow \quad \text{FOCs:} \begin{cases} \theta_t = -\Upsilon_t = 1 - k_t - \gamma k_s \\ \theta_s = \Upsilon_s = 1 - k_s + \gamma k_t \end{cases}$$

$$\Rightarrow \quad \text{best reply:} \begin{cases} k_t = 1 - \theta_t - \gamma k_s \\ k_s = 1 - \theta_s + \gamma k_t \end{cases} \quad \Rightarrow \quad \text{equilibrium:} \begin{cases} k_t^* = \frac{1-\theta_t-\gamma(1-\theta_s)}{1+\gamma^2} \\ k_s^* = \frac{1-\theta_s+\gamma(1-\theta_t)}{1+\gamma^2} \end{cases}$$

Notice the "reverse slope" property of the best-replies which is due to the zero-sum nature of the game at hand. The equilibrium transfer is then $Q^* = \frac{1}{1+\gamma^2}\left(\gamma(1-\theta_s\theta_t) - (\theta_s - \theta_t)\frac{\theta_s+\theta_t}{2}\right)$ while indirect utilities are

$$\begin{cases} \pi_t^* = -Q^* - k_t^* & \propto \quad \theta_s^2 - 2 + \theta_t(2-\theta_t) - 2\gamma\theta_s(1-\theta_t) \\ \pi_s^* = Q^* - k_s^* & \propto \quad \theta_t^2 - 2 + \theta_s(2-\theta_s) - 2\gamma\theta_t(\theta_s - 1) \end{cases}$$

If recipients reduce their internal free riding, they are able to apply pressure k_s at lower cost, as if $\pi_s = S - \lambda k_s$ for $\lambda < 1$ but this is equivalent to set $\pi_s = \frac{1}{\lambda\theta_s}\Upsilon(k_t, k_s) - k_s$ as if θ_s was smaller. The effect is to increase own pressure k_s^* at the rate $\frac{1}{1+\gamma^2}$ and decrease that of opponents at the much smaller rate $\frac{\gamma}{1+\gamma^2}$. The government is thus pressured more and the outcome changes in favor of recipients, though not much. The analysis is symmetric for taxpayers.[25@] Hence we may conclude that

The political effectiveness of a group is determined by its relative efficiency, not the absolute one.

We know study how changes in the efficiency of the State machinery impact lobbies. A larger θ parameter amounts to a worsening. We have $\frac{\partial Q^*}{\partial \theta_s} \propto -\theta_s - \theta_t\gamma < 0$ and $\frac{\partial Q^*}{\partial \theta_t} \propto \theta_t - \theta_s\gamma > 0$ i.e., both a greater distribution efficiency and taxation efficiency increase the final transfer. This means that the large redistribution programs we observe are in fact the " cheapest" to implement. In other words, a higher marginal deadweight cost (worse State machinery on either side) reduces the size of the program. This is because recipients have little incentive to make costly political investments in support of redistribution when it is carried out by highly distorting policy instruments while taxpayers, on the contrary, have a strong incentive to oppose them. This does not mean that everybody wants to scrap a program based on an inefficient transfer technology.

Regarding this issue, political competition tends to promote efficient means of taxation because both groups prefer a program with low deadweight cost of taxation. That taxpayers prefer an efficient taxation system is clear ($\frac{\partial \pi_t}{\partial \theta_t} \propto 1-\theta_t+\gamma\theta_s > 0$) but since there is complementarity, recipients like it too because the low effort of taxpayers enables them to invest little ($\frac{\partial \pi_s}{\partial \theta_t} \propto \theta_t - \gamma(\theta_s - 1) > 0$ for small γ). There is no such clear cut conclusion with the distribution system. Not surprinsingly, recipients are in favor of an efficient system ($\frac{\partial \pi_s}{\partial \theta_s} \propto 1-\theta_s-\gamma\theta_t < 0$) but taxpayers will oppose it because it attracts large political investments from recipients, thus forcing them to respond accordingly with a larger investment ($\frac{\partial \pi_t}{\partial \theta_s} \propto \theta_s - \gamma(1-\theta_t) > 0$ for small γ). This may explain why the tax system appear rather efficient (at least in advanced countries) whereas subsidies and other redistribution program are almost always seen as money "thrown by the window".

Note finally, that the general inefficiency of rent-seeking for political favors derives from the partial equilibrium assumption that any investment into lobbying is a waste. Whenever we cease to view the State as an inert black box but rather as a monopolist who auctions favors, efficiency is restored (cf. Martimort and Stole (2003) on common agency).[26@]

7.3.2 Capture and Pressure

In most countries, overtly corrupt politicians are defeated by honest ones. Dal Bo and Di Tella (2003) show however that threats and smear campaigns applied by lobbies may capture them and produce similar outcomes, though for quite different reasons. As consequence, honest politicians must endure negative pressure and earn less than their market equivalent wage which is conducive of entry of untalented people into the activity.

Consider then an honest politician who must choose between policies supported by lobbies A and B. It is easier for exposition to associate A with efficiency (i.e., supported by the majority of the population) and B with one special interest group (i.e., supported by a minority). Choosing policy A over B has two consequences for the politician. On the one side, she will gain electoral support and increase her probability of remaining in power but on the other side, she will have to endure the offensive pressure of the disgruntled lobby B. At this point, lobby A will apply defensive pressure to counter-act B's influence.[27@] Letting e_i be the pressure applied by group i when policy A is chosen for $i = A, B$, the net offensive pressure endured by the politician is $p^A = e_B - e_A$. In this simplistic model, p^B is normalized to zero since lobby B is happy with policy B whereas lobby A cannot overturn it.

The lottery (contest success function) is then introduced through a macro-economic shock $\theta \geq 0$ that affects negatively the economy before the politician is required to make a decision. For policy $i = A, B$, the probability of losing the next election is the increasing function $G_i(\theta)$ i.e., it is harder to win an election in times of depression. The electoral support gathered with implementing policy A over B means that the wedge $G_B - G_A$ is positive; as attested by empirical studies, we can further assume it is increasing with the size of the shock i.e., in crisis times, there is strong electoral support for reform (choosing the efficient policy over the "business as usual" one).

Normalizing the utility of remaining in power to unity, the expected utility of the politician is $u_i(\theta) = 1 - G_i(\theta) - p^i$ for choice $i = A, B$ (assuming net pressure to be positive). We have

$$u_A(\theta) > u_B(\theta) \Leftrightarrow G_B(\theta) - G_A(\theta) > p^A \Leftrightarrow \theta > \hat{\theta}(e_B - e_A)$$

where $\hat{\theta}$ is the increasing inverse of $G_B - G_A$. This means that the politician implements the efficient policy only after a large enough shock. The probability of a small shock is then the success probability of lobby B; its profit is thus $\pi_B = v_B \Pr\left(\theta < \hat{\theta}(e_B - e_A)\right) - \frac{1}{2}\lambda_B e_B^2$ where λ_B is the inverse productivity of the pressure technology. A symmetric formula holds for π_A.

In the absence of offensive pressure from lobby B (when policy A is chosen), lobby A enjoys the maximum probability of success, thus it does not engage into defensive pressure. Clearly, given this nil defensive pressure, it pays to be offensive for B since $\hat{\theta}$ is increasing and we assume decreasing returns to scale in the production of negative influence. As in the Hotelling model, the defensive side will respond with some counter pressure.[28@] An equilibrium is found because both pressure technologies displays decreasing returns to scale.[29@]

Since both sides invest into pressure, there will be a positive amount of negative pressure in equilibrium and thus the efficient policy is delayed whenever the shock is

small because the honest politician fears retaliation from the lobby. However, in times of crisis, policy changes hurting incumbent lobbies are realized although they strongly oppose them. Note also that the politician being subject to pressure, he earns less than in a world safe of smear attacks. This inability of public office to pay adequately its servants may explain the failure to attract the most productive people (in a classical sense) and why those with special talent for influence end-up there.

7.3.3 Collective Action

Individuals rarely engage into lobbying or rent-seeking for their own sake, rather they do it on behalf of the group they belong to or represent (cf. §2.4.4). Olson (1965) argues that small groups are more effective because their members have greater incentives to invest effort into the group activities. Indeed, the per-capita stakes are higher and shirking has more deleterious consequences on the group's chances of success. Hence, group members exert a greater effort and the group succeeds more often than a comparatively larger one. It has been argued that this free rider reasoning is only correct if the prize has a private nature (i.e., to be divided among group members) whereas if the prize is enjoyed as a public good by members, then larger group produce a greater amount of effort. Esteban and Ray (2001) address this wisdom in a setting where group benefit are equally divided among members.

Let $e_{i,j}$ and $v_{i,j}$ respectively denote effort and valuation of the rent by member j of group i. The size of group i is m_i, its aggregate valuation for the prize is $v_i = \sum_{j \leq m_i} v_{i,j}$ and its total effort is $e_i = \sum_{j \leq m_i} e_{i,j}$. Lastly, $e = \sum_{i \leq n} e_i$ denotes total effort by contending groups. The probability that group i wins is taken to be $p_i = \frac{e_i}{e}$. It is clear that in a small group, $e_{i,j}$ is an important share of e_i, thus an individual contribution hinges on the success probability. By setting the individual payoff to be their fair share $v_{i,j} = \frac{v_i}{m_i}$, we ascribe high per-capita stakes to small groups and also maximize the free rider problem since individuals get a poor share of the returns on their efforts. The expected utility is $\pi_j = p_i v_{i,j} - c(e_j)$ so that the FOC of optimal effort is $\frac{d\pi_j}{de_{i,j}} = 0 \Leftrightarrow \frac{e-e_i}{e^2} v_j = e_{i,j}^{\alpha-1}$. In equilibrium, all members of a group perform identically thus $e_{i,j} = \frac{e_i}{m_i}$ and the condition becomes $e_i^{\alpha-1} e^2 = (e - e_i) v_j m_i^{\alpha-1}$. We assume next that groups are of equal size m and that the prize is commonly valued (by everyone) at v. From $e = ne_i$, we obtain $e_i^\alpha = \frac{n-1}{n^2} v m^{\alpha-1}$, hence $e_{i,j}^\alpha = \frac{e_i^\alpha}{m^\alpha}$ and total cost is $nm\frac{e^\alpha}{\alpha} = \frac{(n-1)}{\alpha n} v$ which is less than the prize value to a single individual.

When the prize is fully private such as a protective quota or a monopoly rent, its social value W is to be divided among the members of the winning group, thus $v = \frac{W}{m}$. As soon as a group represents dozen of members as in the typical trade or industry association,

the rent dissipation is limited to a few percentage points. If, on the contrary, the prize has a public dimension for the group as in the case of a protective law, the location of a public facility or the specific characteristics of a public project, then v is how much each individual winner values the public policy toward himself. The aggregate value for the group is thus mv and quite often the aggregate value for other groups is positive since the project also serves them, although not as they originally wished. The social value W is thus again in excess of mv and the previous conclusion continues to hold.

7.3.4 Optimal size of the Firm

Müller and Warneryd (2001) offer a rent-seeking explanation for the observation that partnerships never grow too big i.e., go public above some revenue threshold. A recurrent problem for a partnership is free riding among the partners as each is tempted to appropriate more than his fair share of the firm's profits through accounting fraud and other siphoning techniques (cf. §13). If the firm goes public, the managers still tempted to appropriate the firm's wealth will have to go through two stages of bargaining, first as the board of directors against the board of shareholders and then among themselves; because the prize to be divided among them is already smaller than under partnership, their rent-seeking incentives are diluted. Although owners also dissipate resources on their side, it is enough that they be less numerous than managers for total dissipation to be reduced by the incorporation.

Let us illustrate these claims formally. If the firm's profit is a prize v to be divided among the m partners, their fair share is $\frac{v}{m}$. In the lottery model of rent-seeking, each partner burns $\frac{m-1}{m^2}v$ in influence activity, his expected return is only $\frac{v}{m^2}$ while the index of wealth dissipation can be written as $\beta = \frac{m-1}{m}$. If the firm goes public and end-ups with n shareholders (or outsiders); former partners are now managers (or insiders) and must compete against the outsiders for the prize before competing among themselves to divide whatever they got.

Let t_i (resp. s_k) denote the effort of an outsider (resp. insider) and $T = \sum_j t_j$ (resp. $S = \sum_j m_j$) the aggregate effort of the group that defines the share $\frac{T}{S+T}$ (resp. $\frac{S}{S+T}$) of the prize accruing to that group. An outsider (resp. insider) expects a share $\frac{1}{n^2}$ (resp. $\frac{1}{m^2}$) of the prize won by his group, thus his overall expected payoff is $w_i = \frac{1}{n^2}\frac{Sv}{S+T} - t_i$ (resp. $u_k = \frac{1}{m^2}\frac{Tv}{S+T} - s_k$). The FOC system is

$$\frac{v}{m^2}\frac{T}{(S+T)^2} - 1 = \frac{\partial u_k}{\partial m_k} = 0 = \frac{\partial w_i}{\partial t_i} = \frac{v}{n^2}\frac{S}{(S+T)^2} - 1 \quad \Rightarrow \quad v = \frac{n^2(S+T)^2}{S} = \frac{m^2(S+T)^2}{T}$$

hence the solution is $T = \frac{m^2 v}{(m^2+n^2)^2}$ and $S = \frac{n^2 v}{(m^2+n^2)^2}$ (with $S+T = \frac{v}{m^2+n^2}$ and $\frac{T}{S+T} = \frac{m^2}{m^2+n^2}$). The wealth dissipation sums for both insiders and outsiders what they spent in the two

stages i.e.,

$$\hat{\beta} = \frac{S}{v} + \frac{m-1}{m}\frac{S}{S+T} + \frac{T}{v} + \frac{n-1}{n}\frac{T}{S+T}$$

so that

$$\hat{\beta} - \beta = \frac{1}{m^2+n^2} + \frac{T}{S+T}\left(\frac{n-1}{n} - \frac{m-1}{m}\right) = \frac{n(m+1)-m^2}{(m^2+n^2)n}$$

thus it is basically enough to have less outside owners (n) than inside managers (m) for dissipation to shrink with incorporation.

7.3.5 Coase Theorem

In §8.1.3, we recall the conditions under which a laissez-faire policy can be efficient to solve entitlements problems. We show here that this conclusion must be amended when we take into account the likely conflict over the allocation of property rights and the possible ambiguities in this process.

Transaction Costs

In an ideal world, items have well defined property rights and no transaction cost hinders exchange so that the final owner of a contentious item is the one with the highest WTP for it, no matter who initially owned it. Yet, as final payoffs depend on the threat point i.e., the initial ownership, we can expect parties to effort themselves to sway the awarding in their favor.

For instance, the owner of a patent for extracting cheaply tar sand will probably never use it because he will be bribed enormously by oil extractors from all over the world to keep it in the drawer. He will thus act (produce or consume) as if the patent did not exist and full efficiency for the energy industry will be achieved (they avoid a destructive competition). Yet the profit made from threatening to use the patent is a very strong incentive to acquire it in the first place. Likewise, the final days of a war before parties sign the armistice are the most ferocious since each army tries desperately to conquer additional land before the cease fire enters into force. The initial allocation of property rights should therefore be pressure-proof to discourage socially wasteful lobbying.

Robson and Skaperdas (2008) formalize this issue in a simple partial equilibrium framework and show that reducing transaction cost need not always improve efficiency (a Laffer curve result). If S (seller) and B (buyer) contend for the property of an item they value respectively c and $v > c$, then efficiency calls for B to be the final owner. Upon expanding influence investments k_S and k_B, B is awarded ownership with probability $p_B = \frac{k_B}{k_B+k_S}$ in which case he keeps the item. With probability $p_S = 1 - p_B$, S gets the item and sell it to B for the price $\frac{v+c}{2}$ (cf. §2.4.3). Transaction costs are introduced by assuming

that trade fails with probability $\sigma < 1$. Profits are thus

$$\begin{aligned}\pi_B = &\quad p_B v + p_S(1-\sigma)\tfrac{v-c}{2} - k_B &&= p_B V_B - k_B + (1-\sigma)\tfrac{v-c}{2}\\ \pi_S = &\quad p_S\big((1-\sigma)\tfrac{v+c}{2} + \sigma c\big) - k_S &&= p_S V_S - k_S\end{aligned}$$

where $V_B \equiv v - (1-\sigma)\frac{v-c}{2}$ and $V_S \equiv (1-\sigma)\frac{v+c}{2} + \sigma c$ satisfy $V_B + V_S = v + c$. Applying (7.4), the equilibrium influence investments are $k_i^* = \frac{V_i^2 V_j}{(V_i+V_j)^2}$ for $i = S, B$ while $p_B^* = \frac{V_B}{V_B+V_S}$. The welfare loss is found by summing transaction costs (the efficiency loss of erroneous final ownership) to rent dissipation i.e.,

$$\beta(\sigma) \equiv p_S^*\sigma(v-c) + k_B^* + k_S^* = \tfrac{V_S}{V_B+V_S}\left(\sigma(v-c) + V_B\right) = \tfrac{(w-\sigma\delta)(w+3\sigma\delta)}{4w} \tag{7.12}$$

where $w \equiv v + c$ and $\delta \equiv v - c$ denote total and differential value of the item for the parties.

This function of the trade impediment σ is bell shaped with a maximum for $\sigma = \frac{w}{3\delta}$. Minimum welfare loss is thus achieved either when there are no transaction cost ($\sigma = 0$) or infinitely many so ($\sigma = 1$). When trade is sure to take place ($\sigma \simeq 0$), parties become symmetric ($V_B = V_S \simeq \frac{w}{2}$) and thus invest heavily; however this is the only source of welfare loss. When trade is almost impossible ($\sigma \simeq 1$), the item becomes more valuable for the buyer than for the seller ($V_B \simeq v, V_S \simeq c$), hence their lobbying efforts become different (cf. §7.2.2) and are on aggregate lower; rent-seeking has thus allocated the item rather efficiently. However missed trade opportunities must be added to the welfare loss.

To conclude, we might say that for a low transaction cost ($\sigma \simeq 0$), reducing it further is welfare improving but when trade impediment is of medium importance $\frac{w}{2\delta} > \sigma > \frac{w}{3\delta}$, a reduction of transaction cost actually worsens the overall level of efficiency of the exchange process because the gain in recovered trade opportunities is more than offset by the increased rent dissipation in lobbying.

Ambiguous Property Rights

Ambiguity in the allocation of property rights among two parties by the court is another source of transaction cost. One way to model this issue is to assume that after investing k_i and k_j, party i wins ownership with probability $p_i = \frac{\lambda_i k_i}{\lambda_i k_i + \lambda_j k_j}$ where λ_i indicates the weight of legal precedent in his favor i.e., $\lambda_i = 0, 1$ corresponds to certainty (no ambiguity) whereas $\lambda_i = \frac{1}{2}$ indicates maximum uncertainty (and we are back into the pure lottery model).[30@]

The prize for firm i in this context is the profit difference V_i between winning and losing the allocation of property rights.[31@] Up to a constant, each party maximizes $\pi_i = p_i V_i - k_i$ for $i = A, B$. Making again the change of strategic variable, from investment k_i

to leverage $e_i = \lambda_i k_i$, we observe that party i maximizes $\frac{e_i}{e_i+e_j}\lambda_i V_i - e_i$, thus we can apply (7.4) to deduce that in equilibrium $k_i^* = \frac{\lambda_i V_i^2 \lambda_j V_j}{(\lambda_i V_i + \lambda_j V_j)^2}$ and $p_i^* = \frac{\lambda_i V_i}{\lambda_i V_i + \lambda_j V_j}$.

We therefore conclude that the party with more at stake invests more (i.e., $V_i > V_j \Rightarrow k_i^* > k_j^*$). Notice further that if she enjoys legal bias is in her favor ($\lambda_i > 1/2$) then upon increasing this bias ($\lambda_i \nearrow$), both investments fall so that a greater efficiency is achieved[32@] so that we can claim: [33@]

> It is efficient to assign property rights to the party with more at stake.

Since the party awarded exclusive property rights will often use them inefficiently (because he/she ignores the externality produced on the loser), there is room for either an ex-ante or ex-post negotiation towards an agreement where the property rights are put to joint use. Let W^* be maximum joint profit.[34@] In our complete information setting, the agreement will be reached immediately (ex-ante) but under the menace of trial. Now, since investments affect the trial outcome, their equilibrium level will be positive.

Indeed, if the parties fail to agree and go to trial, each expects $p_i V_i$. Let $\delta \equiv W^* - p_i V_i - p_j V_j$ the benefit of cooperation. Ex-ante, expected payoffs are $\pi_i = \frac{\delta}{2} + p_i V_i - k_i \propto W^* - p_j V_j + p_i V_i - 2k_i \propto \frac{\lambda_i k_i V_i - \lambda_j k_j V_j}{\lambda_i k_i + \lambda_j k_j} - 2k_i$. The FOC are thus symmetric and so is the equilibrium with $k_i = k_j = \hat{k} = \lambda_i \lambda_j \frac{V_i + V_j}{2}$ (and $\hat{p} = \frac{1}{2}$). Note that, with or without renegotiation, investments depend positively on $\lambda_i \lambda_j$, thus wasteful influence always increases with ambiguity ($\lambda_i \to \frac{1}{2}$).

It is immediate to see that when $V_i > V_j$, $k_i^* + k_j^* < 2\hat{k} \Leftrightarrow \lambda_i^2 V_i > \lambda_j^2 V_j$. The condition holds if the ratio of precedent weights $\frac{\lambda_i}{\lambda_j}$ is large enough when compared to ratio of values $\frac{V_j}{V_i}$ (which is lesser than unity). Thus, if a party has more at stake and is advantaged by law, she invests so much in influence that the aggregate investments are greater than if Pareto-efficient negotiation was not available. Finally, if, on top of the previous condition, the overall inefficiency is not too large (i.e., δ small) then we may have $\pi_i^* = p_i^* V_i - k_i^* > \hat{\pi}_i = \frac{\delta}{2} + \frac{V_i}{2} - \hat{k}$ i.e., a party prefers to go to trial (and stick to an inefficient outcome) rather than agree to an efficient use of the disputed item. This optimal behavior thus rationalizes some behavior observed in real life.

7.4 Patent Races and Attrition

This section considers dynamic conflicts.

7.4.1 Patent Race

In §12, we treat research and development (R&D) as a strategic investment aimed at improving the competitive position of the firm. In the present section, we study "patent races" whereby firms invest in R&D in order to make a valuable discovery and be the first to get a patent protection. We demonstrate a close equivalence with the standard model of rent-seeking although the conclusions derived from it are less dramatic because the patented discovery has a social value.

Let us assume that all firms seek the same discovery, that the market value v of the patent is fixed and known to all and that it starts to accrue to the patent holder as soon as the discovery is made (no delay in the processing of the patent application). Loury (1979) shows that the instantaneous probability of making the discovery is very much like the winning probability seen in the contest technology. The expected benefit of investing is thus similar to the rent-seeking profit up to the presence of the interest rate that account for the dynamic nature of patent races.

We proceed to prove this claim. For each contender, the time of discovery τ is a random variable with law $\Pr(\tau \leq t) = G(t)$ and density $g(t) = G'(t)$. A fundamental concept is the hazard rate $\frac{g(t)}{1-G(t)}$ which is the probability that the innovation is ready at time $t+dt$ knowing it was not yet ready at time t. The model assumes this hazard rate to be time independent and a function $h(k)$ of investement k; thus $\Pr(\tau \leq t) = 1 - e^{-h(k)t}$ and the density is $\Pr(\tau = t) = h(x)e^{-h(x)t}$.[35@]

Let us denote $z_i = h(k_i), z_{-i} = \sum_{j \neq i} h(k_j)$, τ_i the discovery time of firm i and τ_{-i} the discovery time of the earliest other firm. If $\tau_{-i} = t$, firm i can enjoy the prize whenever $\tau_i = s \leq t$. Letting r be the interest rate, the conditional expected benefit is then[36@]

$$\pi_i(t) = v\int_0^t \Pr(\tau_i = s)e^{-rs}\,\mathrm{d}s = v\int_0^t z_i e^{-z_i s}e^{-rs}\,\mathrm{d}s = v\frac{z_i}{r+z_i}\left(1-e^{-(r+z_i)t}\right)$$

Assuming that the random processes governing firms discoveries are independent, we have

$$\Pr(\tau_{-i} \leq t) = 1 - \Pr\left(\tau_j > t, \forall j \neq i\right) = 1 - \prod_{j\neq i} e^{-z_j t} = 1 - e^{-z_{-i}t}$$

so that $\Pr(\tau_{-i} = t) = z_{-i}e^{-z_{-i}t}$. The unconditional expected benefit is

$$\begin{aligned}\Pi_i &= \int_0^{+\infty} z_{-i}e^{-z_{-i}t}\pi_i(t)dt = v\frac{z_i z_{-i}}{r+z_i}\left(\int_0^{+\infty} e^{-z_{-i}t}dt - \int_0^{+\infty} e^{-(r+z_i+z_{-i})t}dt\right)\\ &= v\frac{z_i z_{-i}}{r+z_i}\left(\frac{1}{z_{-i}} - \frac{1}{r+z_i+z_{-i}}\right) = \frac{z_i}{r+z_i+z_{-i}}v\end{aligned} \tag{7.13}$$

so that the probability of winning the patent race is the contest formula (7.2) up to the

interest rate r.

7.4.2 Attrition

A war of attrition is a battle among firms to control a market, a standard or a new technology; it ends after all but one contender accept defeat rather than at a fixed time or when a player is able to claim victory as in a patent race (cf. ch. 7 on rivalry). It can explain an *Industry Shakeout*, a phase of rapid decline in the number of producers and simultaneous expansion of aggregate output. Indeed, once contenders start to drop out, the remaining incumbents serve larger market shares and can thus take advantage of scale economies to become even more competitive and expel some more firms. The market size also grows because the reduced variety of products for sale increases the potential life cycle of the product, reduces the uncertainty over its core characteristics.

A recent example is the portable game console market where Nintendo remained in 2000 the sole vendor after battling its competitors for 11 years (Atari, Sega, Tiger Electronics, NeoGeo) and selling more than 120 million units. Nintendo nevertheless faces new threats; Nokia in 2004 and Sony in 2005 launched portable game consoles. Lately Apple's iPod Touch has proven an unexpected comptetitor.

The adequate game theoretical framework to analyze this kind of fight is the duel explained below.

Value of Staying Modeling this struggle can lead to a peculiar equilibrium where one firm enters the market and the other one stays out. The obvious problem is to identify the winner. A more meaningful equilibrium is one where the two firms follow the same pattern of action which is a probability α of exit at each period of trade. The winner's prize is the monopoly profit Π (sum of discounted future profits) while the loss is the fixed (sunk) cost F of developing the product. The value of staying in the market V is the sum of two expected terms, one for the case where the firm becomes a monopoly because the contender exited and one for the case where duopoly goes on for one more period. The first term is $\alpha\Pi$ while the second term is the discounted value today of staying in the market tomorrow, $\frac{V}{1+r}$ where r is the interest rate of the firm's owner. Summing and removing the fixed cost, we obtain $V = \alpha\Pi + (1-\alpha)\frac{V}{1+r} - F$, thus $V = \frac{(\alpha\Pi - F)(1+r)}{1+\alpha}$. In equilibrium, each firm is indifferent between stay and exit[37@] and since the value of exit is zero, V must be nil; we therefore deduce that the probability of exiting the race is $\alpha = \frac{F}{\Pi}$ which is independent of the discount factor; what matters are sunk cost and future profits.

Duel A good illustration of predation is the competition between Sony and Nintendo in the development of a new game console. The longer a firm spends on development, the better its product but the first to release has an advantage since its customers will remain captive (lock-in due to switching cost as studied in §24.2.3).

A firm that releases its product first, at time t, captures the market share $h(t)$. The remaining market share is left for the other firm. If they release at the same time, the total market is shared evenly. The profit function is thus

$$\pi_1(t_1,t_2)=\begin{cases} h(t_1)-r\,t_1 & \text{if} \quad t_1<t_2 \\ \frac{1}{2}-r\,t_1 & \text{if} \quad t_1=t_2 \\ 1-h(t_2)-r\,t_1 & \text{if} \quad t_1>t_2 \end{cases}$$

where h increases from $h(0)=0$ to $h(T)=1$ and r is the interest rate (time is costly). Let $\bar{t}$ be such that $h(\bar{t})=\frac{1}{2}$.

When $t_2<\bar{t}$ (quick opponent), a choice $t_1\leq t_2$ implies $h(t_1)\leq h(t_2)<\frac{1}{2}$ while $t_1>t_2$ yields a share $1-h(t_2)\geq\frac{1}{2}$. Thus $t_1=t_2^+$ (slightly more than t_2) is the best reply (r makes the firm impatient). When $t_2>\bar{t}$ then $t_1>t_2\Rightarrow 1-h(t_2)<\frac{1}{2}$ while t_2^- (slightly less than t_2) yields $h(t_2^-)\geq\frac{1}{2}$; it is therefore the best reply. Against $\bar{t}$ there is no need to rush or to wait, hence the best reply is also $\bar{t}$.

The best reply function is $BR_1(t_2)=\begin{cases} t_2^+ & \text{if} \quad t_2\leq\bar{t} \\ t_2 & \text{if} \quad t_2=\bar{t} \\ t_2^- & \text{if} \quad t_2>\bar{t} \end{cases}$; very much like in the Bertrand paradox. We see immediately that that no time greater than $\bar{t}$ can appear in equilibrium because it would be undercut. Likewise a lesser time would be "overpriced" by the opponent so that both timing would increase. It is only against $\bar{t}$ that $\bar{t}$ is the best reply; the equilibrium is therefore the symmetric choice $\bar{t}$ that generates equal market shares.

7.4.3 Performance based Compensation

We compare here two related forms of contest, relative performance evaluation (RPE) and joint performance evaluation (JPE) which are discussed at more length in §13.1.2.

Static Model

A principal hires two agents to work in her business. Her control technology being imperfect, she can either receive a common signal that everybody is working well, with probability σ, or a differentiated signal for each agent (with complementary probability). The personal signal of an agent is good with probability q_k where $k=0$ (shirk) or 1 (work) is the previously chosen effort. We assume $q_1>q_0$ i.e., there is a positive correlation

between effort and signal. Notice that agents' production technologies are independent. Wage can be contingent on signals but must be positive because no financial penalties are allowed by the law governing labor relationships. Given the auditing technology, a wage scheme for a worker is a vector $\vec{w} = (w_{11}, w_{10}, w_{01}, w_{00})$ where $w_{uv} \geq 0$ is the wage if the own signal is u while the colleague's one is v. The expected wage for agent i when efforts are respectively k and l is thus

$$\begin{aligned}\pi_{kl}(\vec{w}) = & \left(\sigma + (1-\sigma) q_k q_l\right) w_{11} \quad + (1-\sigma) q_k (1-q_l) w_{10} \\ & + (1-\sigma) q_l (1-q_k) w_{01} \quad + (1-\sigma)(1-q_l)(1-q_k) w_{00}\end{aligned}$$

When the principal receives a good signal for worker j, she can either pay more to worker i to reward the team (JPE) or less to reward the best of the two agents (RPE) or change nothing. Analytically, JPE is to set $(w_{11}, w_{01}) > (w_{10}, w_{00})$ while adopting a RPE amount to choose wages so as to reverse the inequality. Notice that JPE generates a positive externality since $\pi_{k1}(\vec{w}) > \pi_{k0}(\vec{w})$ for $k = 0, 1$ while RPE generates a negative one. Letting the cost of effort be c, a contract $\vec{w}$ induces effort for both agents as a Nash equilibrium if $\pi_{11}(\vec{w}) - c \geq \pi_{01}(\vec{w}) \Leftrightarrow$

$$\frac{c}{(1-\sigma)(q_1 - q_0)} \leq q_1 (w_{11} - w_{01}) + (1 - q_1)(w_{10} - w_{00}) \tag{7.14}$$

The aim of the principal is to minimize the wages while motivating hard work, thus she aims at minimizing $\pi_{11}(\vec{w})$ under the incentive constraint (7.14). It is clear that setting $w_{01} = w_{00} = 0$ will be optimal. Then (7.14) simplifies to $\frac{c}{(1-\sigma)(q_1 - q_0)} \leq q_1 w_{11} + (1 - q_1) w_{10}$ and the expected utility becomes

$$\pi_{11}(\vec{w}) = (1-\sigma) q_1 (q_1 w_{11} + (1 - q_1) w_{10}) + \sigma w_{11} = \frac{c}{(1-\sigma)(q_1 - q_0)} + \sigma w_{11}$$

so that setting $w_{11} = 0$ is also optimal. Minimizing wage commands to solve (7.14) with equality; we thus obtain then $w_{10}^S = \frac{c}{(1-\sigma)(q_1 - q_0)(1 - q_1)}$.[38@]

> The optimal contract is therefore an extreme form of RPE with $\vec{w}^S = (0, w_{10}^S, 0, 0)$. The key to this result is the presence of noise (σ): a good signal is more informative of effort if the partner's signal is bad. Thus, paying a bonus solely in that asymmetrical situation (which is the essence of RPE) generates a cheap motivation to work.

Dynamic Model

Consider now a dynamic setting of infinite repetition. The discount factor of agents is $\delta < 1$ (alternatively $1 - \delta$ could be the probability of going bankrupt for the principal).

The wage scheme chosen initially applies to all subsequent periods. We study the implementation of effort for both agents. By shirking today, an agent gets $\pi_{01}(\vec{w})$ today and since the other agent can change his effort in the future (possible retaliation), he can guarantee himself at least $\min\{\pi_{00}(\vec{w}), \pi_{01}(\vec{w})\}$ in each future period. Under JPE where $\pi_{00}(\vec{w}) < \pi_{01}(\vec{w})$ must hold, the incentive constraint for hard-work today, given that the other agent is working, is

$$\pi_{11}(\vec{w}) - c \geq (1-\delta)\pi_{01}(\vec{w}) + \delta\pi_{00}(\vec{w}) \tag{7.15}$$

which is strictly slacker than (7.14). On the contrary, in a RPE scheme $\pi_{00}(\vec{w}) > \pi_{01}(\vec{w})$ must hold and this gives rise to an incentive constraint which is simplifies into the static (7.14). This observation means that JPE gives more possibilities to punish a shirker in a dynamic setting because the other agent can retaliate by shirking also in the future.

Let us study the minimum of $\pi_{11}(\vec{w})$ under the incentive constraint (7.15) and the use of a JPE scheme. The latter reads

$$(q_1 + \delta q_0)(w_{11} - w_{10} - w_{01} - w_{00}) + w_{10} + \delta w_{01} - (1+\delta) w_{00} \geq \frac{c}{(1-\sigma)(q_1 - q_0)}$$

thus we can set $w_{00} = 0$ (negative in (7.15) while positive in objective) and $w_{01} = 0$ (same weight as w_{10} in objective, smaller in (7.15)). Since both the objective and the constraint are linear in w_{11} and w_{10}, only one can be positive, but given the restriction to use a JPE scheme it must be w_{11}. Hence the solution is $\vec{w}^J = (w_{11}^J, 0, 0, 0)$ where $w_{11}^J \equiv \frac{c}{(1-\sigma)(q_1-q_0)(q_1+\delta q_0)}$.

If we use a RPE scheme, we already know that the optimum is $\vec{w}^S$; comparing the final cost of both schemes shows that $\pi_{11}(\vec{w})^J < \pi_{11}(\vec{w}^S) \Leftrightarrow \delta > \frac{\sigma}{q_1 q_0 (1-\sigma)}$, hence JPE dominates RPE when the relationship is long lasting (or agents are patient).

It remains to show that retaliation to shirking after observing shirking once is a credible behavior. Since (7.15) is binding at $\vec{w}^J$, $\delta \leq 1$ and $\pi_{01}(\vec{w}^J) > \pi_{00}(\vec{w}^J)$, it must be true that $\pi_{11}(\vec{w}^J) - c < \pi_{01}(\vec{w}^J)$ meaning that shirking is optimal when facing work. From equality in (7.15), we deduce

$$\pi_{11}(\vec{w}) - \pi_{01}(\vec{w}^J) = \delta\left(\pi_{00}(\vec{w}) - \pi_{01}(\vec{w})\right) + c < c$$

Lasltly, noticing that

$$\pi_{11}(\vec{w}^J) + \pi_{00}(\vec{w}^J) - \pi_{10}(\vec{w}^J) - \pi_{01}(\vec{w}^J) = (1-\sigma)(q_1 - q_0)^2 w_{11}^S > 0$$

we obtain $\pi_{10}(\vec{w}^J) - \pi_{00}(\vec{w}^J) < \pi_{11}(\vec{w}^J) - \pi_{01}(\vec{w}^J) < c$ i.e., $\pi_{10}(\vec{w}^J) - c < \pi_{00}(\vec{w}^J)$ meaning that

(shirk,shirk) is a Nash equilibrium of the one stage game. This proves that under $\vec{w}^J$, the retaliation payoff for the future we considered in (7.15) was correct. Furthermore, $\vec{w}^J$ is collusion-proof in the sense that the two agents get more by working than if they adopt any other pattern of behavior i.e., $\pi_{11}(\vec{w}^J) - c > \pi_{00}(\vec{w}^J)$ and $2(\pi_{11}(\vec{w}^J) - c) > \pi_{10}(\vec{w}^J) - c + \pi_{01}(\vec{w}^J)$.

Our theoretical study has thus shown that RPE is better suited to static or short term relations and JPE to long lasting ones absent any consideration of positive externality among agents.

Part D

Antitrust Issues

Chapter 8

Legal Framework

Market competition, contracts, vertical agreements and more generally all relationships between firms do not take place in the limbo, they are in fact mediated by a large legal apparatus guaranteeing each participant, he shan't be coerced into the exchange and that promises shall be kept. That is to say, the State machinery (justice, police) protects private property and enforces contracts smoothly.

Except when stated otherwise, we assume in this book that the *rule of law* is indeed the reference situation for all economic agents. The object of this chapter is then to make sense of this framework. In the first (still sketchy) section, we work out the conditions and the limits to this ideal. The next sections present the current legal framework of the European and US economies. Due to their differing pace of economic and political development, we emphasize the political construction of Europe and the antitrust genesis in the US.

8.1 Rule of Law

Advanced economies live under the rule of law, by which we roughly mean that citizens and firms enjoy effective rights such as freedom (of speech, of religion,...) and political participation. Economic theory then projects this reality into a *market nirvana* whereby exhaustive property rights exist and are swiftly enforced by a costless, error-free judicial system working in tandem with a serviceable police. Obviously, such an idealized world rests on a State strong enough to uphold our political and economic rights. Without losing this proviso from sight, we place ourselves under the realm of the law and disregard any form of coercion. We can thus keep our adherence to *individual rationality* i.e., economic agents freely enter into contracts over items that have recognizable property rights.

In accordance with the contract-based approach to economic relations taken in this book, the transposition of "political freedom" for citizens to economic agents (firms, en-

trepreneurs) revolves around "property rights". Indeed, firms can be effective economic actors only if they can trust each other and do not fear opportunistic or fraudulent behaviors. To guarantee such an environment, the State (cf. §16.1.1) deploys a legal framework leaning on two pillars, *litigation* and *regulation* whose respective enforcers are the judicial system (justice and police) and the bureaucracy, itself encompassing ministries, public commissions and agencies (cf. §13.1.3 on bureaucracy).

8.1.1 Property Rights

Property rights are held either privately (individual, firm, very small number of people), communally (large group of people) or publicly (State) over items such as physical or financial assets, natural resources or knowledge.[1@] Following Grafton et al. (2004), we may identify several characteristics delineating property rights over an item:

Exclusivity Ability to exclude others from either using or benefiting from a flow of benefits originating with the item.

Transferability Ability to transfer or alienate, at will, the item or its flow of benefits.

Duration Time over which the right remains in existence.

Enforceability Extent to which the right is recognized in law (e.g., certificate of ownership). It measures how much protection is available from encroachment i.e., relates to the implementation of exclusivity.

Divisibility Ability of the holder of the right to divide up the asset or the flow of benefits from the asset.

Flexibility Limitations and obligations over the use of the rights not covered by the other characteristics.[2@]

Under these conditions, Adam Smith's invisible hand (e.g.g, greed) pushes owners to perform profitable exchanges and in the end, items end up owned by those most able to put them at a productive use. Demsetz (1967) observes that property rights represent a social institution that creates incentives to efficiently use assets and further to maintain and invest in them. Their enforcement come at the hand of

- courts (judicial)
- administrative agencies (bureaucracy)
- customs and norms (requires trading within kinship or brotherhood)
- repeated market interaction (reputation building)
- coercion (ultimate tool to force an individual to behave)

Property rules form the legal basis for voluntary (market) exchange of rights because a right holder has the exclusive use of the asset and can exclude infringement. In contrast, liability rules only award damages for infringement and thus form the basis for court-ordered non-consensual transactions. If transaction costs are low, many efficient transactions are performed under property rules. If these cost rise then property rules create a market failure and a change to liability rules may become advisable as it allows more efficiency enhancing transactions where the price is decided by the judge. The downside is the errors made by the judge that may leave one party worse off. In general, liability rules cannot create efficient long run incentives because of the constraint that what one party pays the other must receive cannot give both sides the adequate incentives.

Origins of the Law

Hobbes (1651)'s gloomy vision is an important reference in economics. The author claims that in ancient times, humanity lived in a "state of nature" free from ruler or government but also without rules governing ownership of scarce resources. Competing claims over these surely resulted in the widespread use of violence. Society was plunged into a war of all against all where the resulting life would be "nasty, brutish, and short." Although this recollection may at first look seem convincing, there is scant historical evidence of this "state of nature". On the contrary, it is a well accepted fact that our ancestors lived in small bands whose social coherence depended to a considerable extent upon inherited behavior patterns.[3@]

North et al. (2006) propose a framework based on economic reasoning for understanding the evolution of humanity. In a nutshell, human societies evolved from primitive to feudal. Then, most transformed into nation-states and among those, some successfully moved to free-market-democracies. Internal and external violence is ever present in this evolution but as countries grew richer, the State apparatus succeeded to monopolize the use of violence and thereby reduce internal strife and foster growth (cf. §16.1.1). Yet, most facets of social life remained tightly controlled by the State. In the ultimate transition, general freedoms are bestowed by the State upon the entire population leading to unprecedented levels of innovation, growth, material wealth and satisfaction for the citizens (longer and better life) as attested by Usher (2003).

Wallis (2011) argues that one of the crucial change in human relationships brought about by the rule of law is that people and firms now *hold rights* whereas they used to *enjoy privileges*. A right is anonymous and can usually be transferred (sold) whereas a privilege is intimately linked to its receiver and thus cannot be passed, except to heirs in some circumstances. Furthermore, the privilege is bestowed by a ruler or a powerful

economic agent onto someone in payment for past service or in the expectation that the favor will be returned. A right, on the contrary, is a service that the community confers free of charge to all its members without discrimination; it is universal whereas the privilege is particular.

Acemoglu (2010) adds political struggles and social conflict to complete this panorama and explains in a convincing and intuitive manner why politically oppressive and economically inefficient regimes are so enduring (cf. §8.1.3).

Emergence of Private Property A pillar of modern legal systems is private property which often emerged prior to the establishment of the law (cf. Gintis (2007)). This phenomenon can be rationalized as an equilibrium behavior in game theoretic models of rivalry or conflict over ownership of an item (typically land) between an incumbent and a challenger.

The endowment effect or loss aversion states that one is ready to sacrifice more to defend a holding than to acquire it. There are many reasons why it holds. Firstly, the holder knows better the item than a potential acquirer who thus needs to apply a risk premium when estimating the item's value. Next, the holder may have invested into the asset so that its value has increased for him wrt. its original market price (cf. asset specificity).[4@] Lastly, there is a behavioral explanation, empirically validated according to which the incumbent can be identified as "rightful" for the very fact of being first holder. This observable feature serves as a signal which helps society minimize conflict over ownership.[5@]

The transcription of these facets of loss aversion into models of asymmetric conflicts are asymmetric valuation (the incumbent values the item more because of complementary and specific investments), asymmetric burden (the incumbent is more effective at defending the item than the challenger at contesting it) and commitment (the incumbent has committed long-lived resources into defending his position). As we show in §7.2.2, all features point in the same direction, namely that the incumbent expends more resources than the challenger so that the former remains in place with a high probability i.e., there is a natural respect for private property.

Limits to Law

However strong the rule of the law may be, a rational self-interested individual will always strike a balance between lawful and unlawful means of acquiring wealth (or between moral and wicked ones if his education makes him a perfectly law-abiding citizen). Because the technology for identifying and catching thieves is far from perfect and the death penalty cannot be used for minor offenses (what deterrence would be left then for

major offenses?), the expected penalty from wrongful behavior is always lesser than the expected reward, especially for those who have nothing to lose. The prevalence of violence and fraud even in the most advanced societies is proof of this fact. This grey area where agents seek to discover the limits of the law is treated in chapter 7. Let us only say, applying marginalize thinking, that the State will stop investing into the repressive machinery when at the margin, a dollar of taxed wealth fails to increase GDP by a dollar i.e., it is optimal to allow a small amount of unlawful activity. For instance, small tax frauds are not pursued by IRS inspectors.

8.1.2 Regulation vs. Litigation

The main theme of this paragraph is that litigation is optimal when the issues at stake are well defined and can be included into contracts.[6@] When market failures such as externalities or asymmetric information among parties are present, litigation loses its bite to compel firms to behave in a socially optimal manner. Although the bright-lines rules of regulation are inefficiently rigid (aka "red tape"), they may fare better to reduce the impact of socially wasteful conducts. This is why the State imposes so many quantitative regulations such as firm qualification, product certification, minimum quality standards or occupational licenses for individuals. Much of these compulsory informative certificates are delivered by private intermediaries in the shadow of the State who makes sure that they are honest i.e., neither an impersonator of a true certifier nor a malevolent one.

Contract Enforcement

The two technologies used by the State to uphold contracts are litigation and regulation (cf. §16.1.1). In this section, we describe them precisely and give some elements of comparison using the similarity with the "make or buy" " quandary that figures prominently in §13.3.3 on vertical integration and transaction costs.

A commonality between litigation and regulation is their mode of action since both are based on deterrence: whenever an economic agent is found guilty of infringing a legal rule or a private contract, the State can use its coercive power to penalize him (with a fine or a prison term). The rational agent will then abide by the rule if the expected penalty is larger than the gain from misbehaving, that is to say, if both the probability of being (identified, tried and) condemned and the effective fine are large enough (this strategic interaction is detailed by way of an example in §2.4.2 on sequential games). Litigation and regulation however differ in two important dimensions, namely the timing of the State intervention and its odds of success. The optimality of a mechanism vis-à-vis the other will then depend on how it fares on those two dimensions.

Litigation is an *ex-post* mechanism because it is called to the rescue after the parties to a contract (sometimes an implicit one) have started to perform, precisely when one party believes the other has misbehaved. The main enforcer of litigation is the public judicial system but it does not preclude private arbitration to avoid the cost of actually going to court (cf. §2.4.3)[7@] . If complete contracts could be written at no cost, then every single contingency would be covered so that economic transactions would develop smoothly without ever resorting to the judicial system. Indeed, no one would complain because no one would break the rules and this is so because everyone knows that such a behavior would be exposed and punished. As shown by Coase (1960), even the traditional sources of market failure, externalities and asymmetric information, could be dealt with successfully in that ideal framework (cf. next section).

Regulation is an *ex-ante* mechanism whereby the State forces firms to comply with some rigid but clear rules that mostly aim at guaranteeing minimum quality standards. A specialized bureau (either public or private) delivers certificates of compliance and perform inspections to make sure that the rule is followed by all. Offenders can be fined or even jailed. Occupational licenses are one example that applies to individuals who want to enter a regulated profession. Because a regulation forces producers to invest into assets such as human capital or specialized machineries which are only imperfectly related to the quality of their output, there is a productive inefficiency for high quality firms. Indeed, for such firms, the regulation is a source of fixed cost (deemed "red tape") that fails to motivate a higher quality choice (wrt. what they would elicit in the absence of regulation). In a world of complete contracting, regulation is worse than litigation but as we shall now demonstrate, regulation can fare better in a world where contracts are necessarily incomplete.

Comparison

A first weakness of litigation as a technology of legal enforcement is its proclivity to suffer demand shocks. Indeed, two large firms can trade millions of units in a single transaction; if a problem occurs, a single judge is called to decide.[8@] On the contrary, when a firm sells goods or services to millions of consumers, a simple defect in the production process can lead to thousands of lawsuits which can easily overload the judicial system. Although the involved firm will be surely chastised, all other firms in the country will be safe from prosecution (because of congestion) and start to behave as if anything was permitted. Alternatively, the State, in order to avoid congestion of the judicial system, refuses to consider but a handful of lawsuits which means that rules are poorly enforced and again the incentives towards firms are greatly diluted. Class-action, the grouping of individual demands, is a recent answer to this issue but the fact that trials can last for

a decade works against litigation. Regulation, in those instances, is far more economical since it only requires the control of a small number of producer by a limited number of inspectors to obtain the same desired level of quality compliance.

Litigation's second weakness is its reliance over clearly delineated property rights. As we argued in this chapter, many important aspects of a transaction cannot be integrated into a contract. For instance, gas emitted by thousands of sources such as cars or industrial plants are, as a whole, responsible for the deterioration of air quality in cities but it cannot be proven that a particular source caused a particular disease, hence litigation has no bite to reduce this costly negative externality.[9@] The problem here is that emissions are not associated to property rights; a regulation will improve upon litigation by forcing each industry to take preventive measures in order to reduce overall emissions.

Finally, the last weakness of litigation relates to causality or the establishment of accountability. Consider a train conductor; in case of accident, it is difficult to sort bad luck from undue care and therefore to prove responsibility. This means that firms providing low level of quality would not be chastised with sufficient frequency, thus the deterrence effect of litigation would fail and accidents would not stop. The imposition of a minimum quality standard (with controls) or the mandatory display of information reduces the uncertainty regarding quality of the item and eliminates low quality.[10@]

To conclude, regulation is an ex-ante rigid process enforced by specialized bureaucrats whereas litigation is an ex-post flexible process enforced by generalists judges. The latter tend to dominate the former when the involved issue can be delineated by property rights, when it has a deterministic causal relationship with the defendant and if it involves firms only (small numbers). This is why the State imposes so many quantitative regulations such as firm qualification, product certification, minimum quality standards or occupational licenses for individuals (cf. §16.2 & §9.1.2). An analytical model comparing the respective merits of centralized vs decentralized trading mechanisms is presented in §14.4.2.

8.1.3 Coase Theorem

We develop the original, a popular and lastly a political version of the theorem in relation to property rights.

Intuition

Coase (1960) reminds us of something generally taken for granted, namely that a clear delineation of property rights is a pre-condition to exchange. In the modern globalized

and urbanized world, economic agents create countless (positive or negative) externalities in the course of their productive activities. These are a source of inefficiency because they are not marketed and thus have a zero price that does not match their economic cost or valuation. The fundamental contribution of this author is to stress that it is precisely because the carrier of an externality is not amenable to a proper definition that the latter is not marketed in the first place; there are transaction cost because it is hard to define, specify, measure and enforce the property rights related to an externality.

Hence, transaction costs are what ultimately prevent voluntary bargaining from attaining Pareto-efficient outcomes. In order to minimize these costs, economic institutions such as markets (driven by prices) or command-and-control bureaucracies have developed. The same logic is at work in Coase (1937)'s analysis of the "make or buy" dilemma of firms (cf. §13.3.3). When transaction costs are important, some efficient exchanges are impeded. There are then obvious incentives to come up with innovations that reduce transaction costs such as better institutions. Money, for instance, eliminates the transaction cost of barter which requires a double coincidence of wants.

Popular version

Stigler (1966), a prominent member of the Chicago School, popularizes the previous findings as the Coase Theorem in the following manner: "the initial allocation of legal entitlements does not matter from an efficiency perspective, so long as they can be exchanged in a perfectly competitive market or, alternatively, without suffering transaction cost". Critics of this exposition, such as Usher (1998) and Dixit and Olson (2000), deem this statement a tautology. Indeed, the tongue-in-cheek version is simply that "no big bills are ever left on the sidewalk", a mere restatement of Smith (1776)'s invisible hand principle, pointing out "the power of competitive markets to allocate resources efficiently" (cf. our straightforward proof in §2.4.3 on bargaining). Yet, following Machiavelli (1532), we may turn the theorem on its head and say that "nobody will never pass up an opportunity to deceive a fool" i.e., to gain a one-sided advantage by exploiting whatever is at his reach. This duality is developed in chapter 7 on rivalry. Its starting point is precisely the absence of perfectly defined rights for many aspects of economic activity. This reality motivates people or firms to appropriate old wealth rather than work hard to create new worth.

In terms of policy implications, the Chicago version of the Coase theorem logically implies that a Pigouvian tax or a quota is likely to stumble upon transaction costs as much as "laisser-faire" would do. Hence, the latter option is advisable as it saves on implementation cost such as passing laws, creating bureaucracies and more generally public spending and its associated distortions (cf. §17.1.2).

Political Coase Theorem

The extension of the Coase Theorem to the political sphere would suggest that political and economic transactions create a strong tendency towards policies and institutions that achieve the best outcomes given the varying needs and requirements of societies, irrespective of which social group holds political power. Carrying this idea to its limit, we would come to see our current institutions as fully optimized. Furthermore, war would never take place as it systematically destroys wealth. Such conclusions are grossly incompatible with the historical record of wasteful economic policies and murderous wars observed across humankind.

Acemoglu (2003) sees strong empirical and theoretical grounds for the alternative "social conflict" theory. In essence, societies choose inefficient policies and institutions because political elites have the (mostly violent) means to carry out their preferred choices which are rarely aligned with those of society as a whole (cf. §16.2.4). Even though a policy or institutional innovation has the potential to create aggregate wealth, it is not undertaken because there is no way to compensate all those who would lose from the change. The inexistence or inability to set-up redistributive channels lead groups with political power to oppose efficiency enhancing proposals. Hence, transaction costs are at the root of inefficient institutions and policies. A similar phenomenon takes place in economic interactions between firms when they lack credible redistributive mechanisms.

8.2 European Union

We first draw an historical sketch of the political unfolding of the European Union. Then, we single out the relevant features for the economic analysis: the articles of law and the rules of the game between all parties, firms, national governments, European Commission and Court of Justice.

8.2.1 Historical Development

Roots

In 1945, after centuries of almost permanent warfare, European leaders finally understand that the path to peace and prosperity is to form an economic and politic league. The founding father, Jean Monnet, is a convinced federalist, specialist in economic planning with good connections in the US. As French planning commissioner in charge of coordinating the reconstruction effort at the end of WWII and European coordinator of the Marshall plan, he foresees the need to link the French and German economies to make

future wars materially impossible. Since mining (coal, iron ore) and steel production are the basic inputs for these recovering economies, he proposes a pooling of resources in these sectors between the two countries and any other willing to join.

1951: European Coal and Steel Community (ECSC)

The treaty of Paris,[11@] signed by Belgium, France, (West) Germany, Italy, Luxembourg and the Netherlands, establishes the ECSC which is founded on a common market, common objectives and common institutions. The stated objectives are to contribute, in harmony with the member states economies, to economic growth, full occupation and improvement of the standard of living. On the technical side, the goals are to expand production, organize it rationally, increase productivity, safeguard employment, lower prices while duly remunerating capital and finally promote international trade.

The institutions are the *High Authority*, an executive supranational body, whose mission is to implement the treaty's objectives, a *Court of Justice* ensuring that the law is observed in the interpretation and implementation of the treaty, an *Assembly* representing the parliaments of the member states (with very little power) and the *Council* of member states representatives which authorizes some activities of the high authority and generally harmonizes it with member states economic policies.

This treaty focuses on the commercial policy for coal and steel; it already contains most of the current European competition rules and gives the high authority powers to enforce them against firms but also against member states if necessary. It is clear for the drafters that the treaty establishes a "consumers' union" that will limit the natural tendency to output restriction of "business cartels". Apart from the aspect pertaining to the functioning of markets, most decision making in the treaty is based on unanimity in order to protect the peculiarities of each member state.

1957: European Economic Community (EEC)

The ECSC's fast success is not followed by political integration as attested by the failure of setting-up a defense community in 1954. To avoid a loss of momentum, the belgian foreign affairs secretary, Paul Henri Spaak, devises a further economic integration that culminates with the Rome treaty establishing the EEC. It basically extends the previous treaty to all sectors of the economy in order to build a *common market*; the executive body, renamed "Commission", loses some of its prerogatives in favor of the Council especially in the conflicting areas such as agriculture where the treaty remains vague.

The updated objectives are to instate a *regime of undistorted competition*, remove duties and quotas on goods circulating within the community, eliminate the obstacles

to the free movement of people, services and capital, coordinate national policies and approximate national laws to the extent required for the functioning of the common market. A period of twelve years is set for the achievement of these objectives together with binding timetables and numerical benchmarks. Regarding competition law, a 1962 council directive gives practical directions for its implementation.

A key novelty is the establishment of a Common Agricultural Policy (CAP) whose aims are to increase productivity, guarantee revenues of farmers, stabilize markets, guarantee security of supply and low final prices. The weight of the sector in the member state economies as well as the disparities of regions is underlined. The policy tools include price control or support, production and marketing subsidies or storage but none should generate discrimination among producers or buyers within the community. State aid is broadly authorized for inefficient farms or regions. The implementation of the CAP focuses on creating a common market (thereby with a unique community price), on giving preference to community products and on establishing financial solidarity within the community; it takes place only in 1962 after a lengthy process of reconciliation of the differing positions held by Germany and France.[12@] The CAP basically uses prices support and direct subsidies. It sets import tariffs so as to raise world market prices to a so-called EU target price and stands as an unlimited buyer to maintain community prices above a so-called intervention level.

The last innovation in the treaty is a transport policy which is set-up with the broad aims of removing price discrimination for community trade and barriers to entry on national markets (i.e., same rules for everyone); it includes road, rail and waterways but excludes maritime and air transport. Like in other areas, the council can unanimously derogate from the previous competitive rules to protect a region negatively affected.

1986: Single European Act (SEA)

In 1965, a grave crisis over the CAP leads France to a 6 month boycott of the council, thereby effectively blocking EU institutions and policy. General de Gaulle refuses the passage into the second phase of integration scheduled in 1966 whereby every member state will lose his veto power in the council. France finally rips off the “Luxembourg compromise” calling for unanimous decisions in cases concerning a country’s vital interests. Although the customs unions is achieved ahead of schedule in 1968, the entry of vastly different countries in the community slows down the integration process regarding the removal of non-tariff trade restrictions such as bureaucratic regulations.

In 1986, under the leadership of the Commission head Jacques Delors, member countries sign the Single European Act, a rejuvenating act, looking to achieve the four freedoms of movement (people, goods, services and capital) by adopting qualified majority

voting instead of unanimity for all actions necessary to implement the European Single Market aka *internal market*. The "acquis communautaire" gathering all EU law accumulated so far is mentioned as a foundation for future legal developments. The parliament is also associated to many decisions instead of being merely consulted while the ability of the council to overturn the commission's ruling is being restricted.

1992: European Union (EU)

The Maastricht Treaty creates the EU and introduces new forms of cooperation between member states, on defense, justice and home affairs. It also defines the so-called four freedoms of goods movements, people, services and capital and implements an architecture based on three pillars: the existing European Community, a common policy for security and foreign relations and a cooperation for internal policy and justice.

On the other hand, two principles limiting the powers of the Commission, already visible in the case law of the ECJ, are formally stated. According to the proportionality principle, the EU acts within the limits of its powers and of the objectives assigned to it. According to the subsidiarity principle, the EU is to act only if a common objective cannot be implemented in a satisfactory manner by member states acting independently so that, for reasons of scale or externalities, the objective is better attained by a concerted european action.

The EU treaty also lays down the provisions for an Economic and Monetary Union (EMU) that is achieved in 1999 with the transfer of the monetary policy sovereignty of 12 countries to the European Central Bank (ECB); national currencies disappear in 2002 with the introduction of the Euro whose symbol is €.

Regarding political developments, the EU Treaty strengthens the powers of the Parliament, the only European institution directly elected by its citizens and creates a citizenship of the Union which permits EU citizens to vote and stand for election in local and European elections in the Union.

Recent Developments

With the Amsterdam Treaty of 1997, the Union underwent further reforms. The growth of cross-border problems such as crime, asylum seekers etc. led the EU countries to integrate large sections of what had previously been national prerogatives (border controls, immigration and asylum policy). The common external and security policy remained organized on an intergovernmental basis, but in future the European Council was to be allowed to adopt common strategies, actions and positions. In addition the post of "High Representative for the Common External and Security Policy" was created. This was to

give the policy a “public face”. The Amsterdam Treaty entered into force on 1 May 1999.

The enlargement of the EU to include ten new members in 2004 represents an historical step towards reunification of the continent and the promotion of peace, security, stability and prosperity. However, wealth differences between the old and new EU members remain substantial so that catching up with the average will be a process calling for a considerable effort and political sensitivity.

Lastly, the 2008 Lisbon Treaty, superseding the “EU Constitution”, seeks to safeguard the union’s ability to function and take decisions effectively by widening the use of super-majority voting for decision taking.

8.2.2 Institutions of the EU

The European Institutions are the Commission, the Council, the Court of Justice and the Parliament.

Commission

The Commission is the government of the EU, it is a supranational body holding federal powers, has the sole authority to take initiatives to create new Community law and monitors compliance by the member states with EU law. The commissioners are nominated by agreement among member states and are to act independently while member states are required to abstain from pressuring them.

The commission makes hearings that help motivate its decisions are taken by simple majority but generally endeavours to reach a consensus; it can enact regulations which are compulsory orders, directives whose objectives are to be implemented by member states the way they see fit and opinions which are purely informative. The commission must report yearly in front of the assembly representing the parliaments of the member states.

Councils

The first treaty created a single council composed of one delegate by member state along with voting weights reflecting to some extent economic might and population. Its task is to To attain the objectives of the treaty, the council, upon proposal of the commission, enact regulations which the commission implements. The complexity of tasks has lead to a significant expansion.

The *European Council*, composed of the Heads of State and the President of the European Commission, defines the general political objectives and directives of the EU; it

meets at least twice a year.

The *Council of the European Union* or Council of Ministers, is the effective legislative body of the EU. Its members are the specialized ministers of the member states, depending on their particular portfolio. It adopts all the essential legal acts and concludes international agreements, mostly through qualified majority voting.[13@] However, in practice the EU endeavors to avoid majority decisions so as not to have to take decisions against the will of individual member states. In the event of disagreement, proposals are therefore generally withdrawn or negotiations continued until unanimity is reached. Important decisions such as the accession of a new state or the transfer of new powers to the EU always require unanimity.

Court of Justice

The Court of Justice is responsible for the interpretation and application of the Treaties establishing the European Communities and of the provisions laid down by the competent Community institutions. Its members, in odd number, are appointed in a fashion similar to those of the commission. To expand the court's capabilities, a Court of First Instance is created in 1989, turning the original into an appeal court. The Court of Justice has dealt with 13500 cases between 1951 and 2004 mostly involving environmental law, consumer protection, agriculture, tax, social policy and institutional law. The Court of First Instance has dealt with 4100 cases between 1989 and 2004 mostly involving competition law, State aid, trademarks and agriculture.

Parliament

In the first treaty, the assembly was composed of members of national parliaments and had quite limited power. With the widening of the union from the economic sphere to the political one, the parliament has seen its role increase and since 1979 it is directly elected by the citizens of the Union.

In the EU legislative process, the Parliament either has a consultative opinion (the right to be heard) or the same decision making authority as the Council of Ministers (co-decision) depending on the subject concerned. Parliament adopts the annual budget with the Council of Ministers and controls its implementation. It is also a supervisory body in the sense that it confirms the appointment of the Commission and can even force that body to resign.

8.2.3 Economic Principles

The 1951 treaty sets out clearly that the community is to work as a free market economy without distortions (art.2-4). However, it does not surrender fully to economic liberalism since it allows a series of potentially distorting instruments provided their cost-benefit analysis is positive for the attainment of the general community objectives.

Pro-Competitive Principles

Anti-competitive practices such as predatory pricing with a view to build monopoly and price discrimination are prohibited per se (art.60). Offenders can be fined up to twice the value of illegal sales. Other illegal discriminations are that of workers based on citizenship (art.69) and tariffs for transport (art.70).

The commission has full power to prohibit formal agreements, concerted actions tending to distort free competition through price fixing, limitation of production, innovation or investment, the sharing of markets, customers or inputs sources. Offenders can be fined up to 10% of their annual turnover (art.65).

Mergers and acquisitions of firms active in the coal and steel sectors must be notified to the commission who can ask the parties any useful information and should take into account the market structure in its analysis of the case which can be favorably resolved by asking the parties to take the remedies it sees fit (art.66). Small scale operations are automatically authorized ("de minimis" rule);[14@] otherwise clearance is given if the operation does not confer the merged entity the power to control prices or production or distribution or reduce the competitiveness of the involved markets or obtain a dominant position in the access to inputs or outputs. Failure to notify large mergers is fined up to 10% of the merged assets. A last feature (oddly placed in an article on concentration) is the ability to intervene a firm, public or private, abusing its dominant position (in the sense that its behavior is contrary to the objectives of the treaty).

The power of member states to subsidize national firms through direct aid or wages (aka State aid) is being limited (art.67-68). Lastly, the commission is to give support to R&D whose results are to be shared with the community (art.55).

Distortionary Principles

We review here articles safeguarding the interests of member states against unrestricted economic liberalism inside the community but also defensive articles aiming at protecting the community as a whole against foreign competition.

In case of crisis such as a negative demand shock (art.58) or a negative supply shock (art.59), the commission can set up quotas and even temporary compensate high cost

firms (art.62). However indirect instruments such as minimum and maximum prices are preferable (art.57, 61). Agreements can be exempted of the per-se prohibition) for specialization or pooling of resources if this improves production processes or if the agreement does not distort competition meaningfully (art.65). The council can set minimum and maximum custom rates (art.72). In case a foreign firm (outside community) commits dumping (excessively low price), flooding (excessively large imports) or abuses of its dominant position, the commission can act to protect the community (art.74).

Changes brought by the EC treaty

The 1957 treaty extends the rules valid for the coal and sectors to all sectors of the economy. The free circulation of goods applies not only to goods produced in the community but also to those having lawfully entered any member state (art.10); this the origin of the *exhaustion principle*. The elimination of custom duties (art.12-17) and quotas (art.30-37) inside the community is organized. A common customs policy with the rest of the world is also set-up with the aim of establishing bilateral agreements to remove barriers to trade (art.18-29). Temporary safeguards are planned to attend difficulties in any member state. The large number of articles and paragraphs indicates the toughness of negotiations and thus the disparities of national situations. The agricultural policy is based on a list of products escaping the competition rules, a list which can be increased by a qualified majority vote of the council.

8.2.4 Current Competition Law

Origins

As recalled by Roover (1951), monopolies and artificial restrictions were outlawed in Roman times (cf. Justinian Code c. 530). Price fixing and discrimination was likewise seen immoral since the price was to be set either by the State or by the interplay of demand and supply (cf. concept of *just price*). Even the granting of monopolies by kings is the subject of criticism from intellectuals whenever they involve basic goods or services. These moralistic views have passed into the classic writings of Smith (1776) and from then on, economic theory has rationalized their relation to the concept of market efficiency.

Goals

The EU treaty[15@] states in its principles that member states are to adopt an economic policy conducted in accordance with the principle of an *open market economy with free competition*. The European Community competition law is formed by articles 81–89 and

is enforced by the EC. Specifically, art.81 prohibits agreements and concerted practices between firms (i.e., , , restraints), art.82 prohibits the abuse of dominant position, art.83 calls for the setting up of directives implementing the previous goals, art.84–85 cover the transitional period, art.86 details how competition rules apply to public service providers while art.87–89 covers State aid. Additionally, the merger regulation covers horizontal concentration that was not originally contemplated in the original 1957 treaty.

The EU competition policy has two complementary economic goals. The liberalization of monopolistic sectors (art.86–89) is a *pro-competitive* activity that aims at introducing competition where it does not currently exists and seeks to foster the integration of European markets into a single unified market, the so-called "internal market". Liberalization is treated in §16.4 while public service obligations are dealt with in §17.3.4. The control of *State aid* belong to this branch because it potentially distort the conditions of competition; the long term goal is to eliminate any public funding that is not warranted by public service obligations.

The second goal is to maintain the conditions of "free competition" i.e., a competition based on the merits of the contenders' offerings. This endeavor thus seeks to avoid *anti-competitive* behaviors like collusive agreements, restrictive practices and abuses of dominant positions. In this respect, articles 81 and 82 respectively deal with collective and individual abuse of dominant position. This activity called *antitrust* is developed in Part 10 (cf. §8.3 for the etymology).

Realm of application

A necessary condition for the application of the EU competition law is the existence of a direct or indirect, actual or potential *Effect on Trade* between member states of the EU. Otherwise national legislation is called for. Furthermore, the EC applies the *Effects Doctrine* stating that domestic competition laws are applicable to firms irrespective of their nationality, when their behavior or transactions produce an effect within the domestic territory (domestic means here either the EU or a member state). For instance, the merger regulation applies to companies located outside EU territory when it is foreseeable that a proposed concentration will have an immediate and substantial effect in the EU.

Most of the economic definitions we shall use in this book are similar to those from the EC's glossary or the OECD; this is a reflection of the fact that the landscape where firms compete in Europe is effectively shaped by the EC and European Court of Justice (ECJ). Three key concepts are worth emphasizing: firms, competition and market power.

Firstly, the realm of application of the EU treaty with respect to economic activity is not limited to private profit-making firms; the relevant actor is the undertaking, any

entity engaged in economic activity bearing economic risk regardless of its legal status, the way in which is financed or its motive. That said, we shall use the more usual terminology “firm”.

Secondly, *competition* refers to a situation where sellers of a product (or service) independently strive for the patronage of buyers in order to achieve their particular business objective. Sellers and buyers may be individuals or undertakings. Competitive rivalry between firms may take place in terms of price, quality, service or combinations of these and other factors which customers may value.

The third key definition is that of *market power*; it is the ability for a firm to price above marginal cost and is computed with the help of a structural analysis of the market involving market shares, substitutability, barriers to entry , growth and the rate of innovation. Furthermore, it may involve qualitative criteria, such as the financial resources, the vertical integration or the product range. Market power is likely to be present in markets with a small number of large sellers dominating a (competitive) fringe of smaller sellers. The former realize their interdependence in taking strategic decisions on price, output and quality while the latter adapt to their behavior to that of the larger players.

Exemptions

The EU law provides for some exemptions to the restrivitve agreements prohibited by art.81 but in any case, i.e., irrespective of the size of the firms and their market shares, no hardcore restriction must be present. These are

- resale price maintenance: a maximum resale price is allowed but fixed price is not.
- market partitioning by territory or by customer: exclusive or selective distribution system are allowed but unsolicited orders must be served.
- exclusive and selective distribution systems can not be combined.
- spare parts manufacturers can sell to end users, independent repairers or service providers on top of the main original equipment manufacturer.

To benefit from exemption, several criteria may apply. They require the computation of the market share on the relevant–product and geographic–market (cf. §15.3.1). Firstly, agreements that do not impinge meaningfully on trade between member states make art.81 not applicable altogether; we find here agreements between small firms as defined in the de minimis notice (i.e., market share $\leq 15\%$) and agency agreements where the principal bears all financial or investment related risk. Next, for sizeable firms (market share $\leq 30\%$), the block exemption regulation (BER) articulates art.81(3) which allows agreements confering sufficient benefits to outweigh their anti-competitive effects.

The BER covers vertical agreements such as supply and distribution with a special rule for car distribution (cf. Table 15.18). Also covered are technology transfers, horizontal cooperation (R&D, specialization), insurance (standardization and information collection), transport and telecommunications.

8.3 US Antitrust Laws

The current European economic landscape has been heavily influenced by the US experience which we now proceed to describe following its historical developments using the present tense given that many of the legal acts involved are still in force. A classic reference is Posner (2001) (cf. FTC for an online introduction).

8.3.1 Historical Development

Economics is deeply rooted in US culture. Indeed, the US Constitution of 1776 and the Bill of Rights of 1791 guarantee, among other rights, the economic rights to acquire, use, transfer, and dispose of property (including intellectual) and also the right to choose one's work.

In the late XIXth century, the American industry becomes increasingly dominated by large firms which flourish in the virtually regulation-free climate of the age. Members of industries where access to markets is naturally restricted, such as railroads, are free to divide those markets between themselves and charge monopoly prices. Simultaneously, firms in steel, oil and coal industries devise a new ownership structure called "trust" to monopolize their respective industries in a way quite similar to that employed by german cartels. The main leaders simply convince (or coerce) the shareholders of all the companies in their industry to hand over their voting rights to a newly created trust company in exchange for dividends. The trustee then strategically manages all the major competitors of that industry so as to operate as a virtual monopoly and rip the corresponding profits. These practices of collective price settings, eviction of new competitors and unrepentant use of corruption generate public outrage and enormous political pressure to halt the unfettered trade abuses which the trusts and monopolists represent for the democratic society (cf. Glaeser and Shleifer (2003)).

The Sherman Act of 1890, quasi unanimously voted by the congress, establishes the US antitrust policy. In retrospective, Bork (1967) analyzes the act as an intent to advance consumer welfare by minimizing the restrictions to output. The Sherman Act outlaws all contracts, combinations and conspiracies that unreasonably restrain or monopolize interstate (e.g. between Texas and California) and foreign trade. This act provides crim-

inal penalties when enforced by the government. Violation can result in substantial fines and, for individual transgressors, prison terms. These provisions are enforced primarily by the Antitrust Division of the Justice Department. In the 1990 update of the Sherman Act, individual violators can be fined up to $350,000 and sentenced to up to 3 years in federal prison for each offense; corporations can be fined up to $10 million for each offense.

In an early judgment of 1895, the US Supreme Court maintains a strict interpretation of the Sherman Act and refuses to apply it to a cartel of sugar refiners producers controlling 98% of market on the grounds that the law applies to commerce and not to manufacture. The US president elected in 1902, Theodore Roosevelt, wants economic laws to balance the rights of companies and those of the citizens. Since many industries are dominated by a single company or a combination of companies controlled by a trust, he instructs the Justice Department to sue the "bad" trusts, those which fail to act in the public interest. In 1904, the Supreme court rules that the Sherman Act can be applied to holding companies. It deems an "illegal restrain to trade" the merger of Great Northern and Northern Pacific railway companies into the newly founded Northern Securities (controlled by J.P. Morgan). In 1911, American Tobacco is declared an illegal monopoly and is broken up into separate companies. The same year, the Supreme court rules that J.D. Rockefeller's Standard Oil should be broken up into 33 companies.[16@]

Some ambiguities in the Sherman Act leads the US congress to complement it.[17@] The Clayton Act of 1914 prohibits certain specific business practices such as exclusive dealing arrangements, tying together multiple products, mergers or acquisitions that are likely to lessen competition or interlocking directorates (trusts formed by companies with common members on their respective boards of directors). The Federal Trade Commission Act of 1914 prohibits unfair methods of competition in interstate commerce and also creates the Federal Trade Commission (FTC) to enforce antitrust law by temporarily halting suspected anti-competitive practices. The Justice Department remains the investigator and, if necessary, the prosecutor of offending companies. The Clayton Act carries only civil penalties unlike the criminal penalties of the Sherman Act but it also permits an individual or firm who has been injured by an illegal practice to sue the perpetrator in the federal courts for damages 3 times as high as the loss plus attorneys' fees. This feature has proved to be an effective deterrent to antitrust infractions.

The Robinson-Patman Act of 1936 amends the Clayton Act to prohibit price discrimination between purchasers of commodities of like grade and quality which are likely to result in substantial injury to buyers' competition or to sellers' competition. The only justifications are cost differences and reaction in good faith to meet the competition. The Celler-Kefauver Act of 1950 amends the Clayton Act on anti-merger measures. The

Hart-Scott-Rodino Antitrust Improvements Act of 1976 makes it easier for regulators to investigate mergers for antitrust violations. To facilitate the task of firms who want to merge and need to check whether they will be allow or not, the Department of Justice issues the Merger Guidelines in 1968 with revisions in 1982, 1984 and 1992 (cf. DoJ (1997)).

For specialized industries, the Communications Act of 1934 regulates interstate and international communications by radio, television, wire, satellite and cable. The Federal Communication Commission (FCC) is the enforcer of the act. Likewise, the Federal Energy Regulatory Commission (FERC) established in 1977 originates in the Federal Power Commission established in 1920 to coordinate hydroelectric projects; the FERC mission is to regulate interstate commerce of energy (oil, gas and electricity).

Regarding finances, the Federal Reserve Act of 1913 establishes the Federal Reserve to conduct the monetary policy, maintain the stability of the financial system and to regulate banking institutions. The Securities Act of 1933 and the Securities Exchange Act of 1934 establish the Securities and Exchange Commission (SEC); its duties are to secure the disclosure of adequate financial information for firms publicly offering securities for investment and to establish fair rules of the games for brokers, dealers, and exchanges (financial trading posts). The Commodity Futures Trading Commission (CFTC), established in 1974, protects market users from fraud, manipulation, and abusive practices related to the sale of commodity and financial futures and options (cf. also a list of antitrust agencies worldwide).

8.3.2 Modern Application

For legal scholars, antitrust laws are an integral part of the US economic system; their easily understandable statutory provisions embody fundamental principles that have been refined by a century of litigation in courts. Most antitrust laws are interpreted under the *rule of reason* i.e., the potential anticompetitive harm of a challenged practice is weighed against its business justification and its potential pro-competitive benefit. A judgment with respect to the reasonableness of the practice is made. On the other hand, the most blatantly anticompetitive conducts such as price fixing among competitors are judged *per se* illegal.[18@] It is enough to establish that the defendant has engaged in the proscribed practice; illegality follows as a matter of law, no matter how slight the anticompetitive effect, how small the market share of the defendants, or how proper their motives.

Horizontal restraints to trade among competitors are the most serious of antitrust infractions because in a market economy, it is the duty of rival firms to compete with

respect to prices, products, and services. Price fixing–an horizontal agreement affecting prices directly or indirectly[19@] –is one of them and is per se illegal. Other horizontal restraints per se illegal are the allocation of markets or customers and the concerted refusal to deal.

Vertical arrangements and restraints involve relationships among suppliers and customers on different distributional levels; as they do not implicate direct competitors, courts are more more lenient towards them. The previous horizontal and vertical restraints are antitrust offenses of a *behavioral* nature as opposed to those affecting the market in a *structural* way.

8.3.3 Comparison of the EU & US

Legal scholars give credit to the legal systems and practices of France, Germany and Italy in the drafting of the first European treaties (Paris and Rome), only acknowledging a background US influence. However, when we look at the body of law that matters for the study of economic conduct, one cannot avoid to see the resemblance with the US institutions described in the next section.

While the first European Treaty (1951) was limited in scope to a single sector of the economy, it was audacious in terms of political integration with the enactment of two supranational bodies, the (executive) *High Authority* and the *Court of Justice*. This is an unusual feature for countries whose mode of relation (apart from war) has traditionally been cooperation based on consensus i.e., unanimous agreement giving veto power to everyone. The root of this boldness is the understanding that the integration of economies damaged by war and before it, by the great depression, has to overcome the anti-competitive conduct of national monopolies in the basic input industries. The treaty's designers understood that a strong and effective "US-like" competition policy would be necessary for that task. This is why the treaty gave the High Authority broad powers very much like a US federal commission, so much so, it was later renamed the *Commission*. Likewise, one cannot miss the similarity of the Court of Justice with the US Supreme Court as the ultimate interpreter of the law (European Treaty or US Act).

The 1957 EC Treaty's great achievement is to extend the successful coal & steel formula to most areas of the economy by putting them under the existing federal umbrella. At that point, the concept of *European Commission* is far more ambitious than a US federal commission whose scope is delimited by the Act creating it. The downside is obviously the bureaucratic challenge of carrying so many different activities under the same hood.

In the US, the supreme court has often been called to either confirm or infirm the

government and the federal commissions. Similarly, but less intensively, the European Court of Justice has produced a body of case-law upon which the Commission has updated its directives (orders in US parlance).

Chapter 9

Anti-Competitive Practices

The ideal of perfect competition assumes independent suppliers, each of whom is subject to the competitive pressure exerted by rivals. It is therefore a priority task of competition law to preserve this feature and prohibit agreements or practices which might reduce such competitive pressure.

We distinguish horizontal from vertical anti-competitive agreements. The first class regards the collective abuse of dominant position aka *collusion* and *cartelization*. Such a mutual understanding aims at coordinating the policies of the members to increase their overall market power. As we show in the first section, collusive agreements reduce market efficiency and are therefore rightly qualified as anti-competitive. We study their inherent instability and also review some historic evidence.

The second class of anti-competitive agreements regards the relationship with the clients or suppliers. We deal with three categories. Best-price clauses can potentially limit competition and help enforce collusion by removing the ability to steal customers from a competitor. Vertical restraints are price or non-price limitations on distributors with a view to limit downstream competition. Lastly, the legal aspects of differential pricing are mentioned. We conclude with a few figures relative to antitrust activity.

9.1 Cartel and Collusion

A *cartel* or trust in US parlance[1@] is a public agreement among a group of firms while *collusion* refers to the secret version of the same agreement. Smith (1776)'s description of the subject is still topical: *People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices. It is impossible indeed to prevent such meetings, by any law which either could be executed, or would be consistent with liberty and justice. But though the law cannot hinder people of the same trade from sometimes assembling together, it ought to do nothing to facilitate such assemblies; much less to render them*

necessary (§I.10.82).

9.1.1 Inefficiency and Instability of Price Fixing

Inefficiency

We study the effect upon efficiency of an agreement among competing firms to coordinate their pricing policies, independently of whether this happens through collusion or the establishment of a cartel. When firms enter into an agreement, they substitute their individual freedom for a collective constraint; it might at first sight appear prejudicial for them but this commitment generates additional market power and thereby greater profits, which can be later shared.

There is firstly a technical advantage to cooperate because the production technologies can be exploited to reduce overall costs. Given a joint plan of production, if the marginal cost among members A and B of the cartel are not equal then a saving can be realized by switching some production from the dear to the the cheap plant. This makes the cartel a stronger competitor than the sum of the original member's strengths.

The second benefit of cooperation is a strategic advantage for market competition. Coordination amounts to reduce the number of competitors which, ceteris paribus, increases industry profits; there is thus a possibility for the cartel to increase his absolute share of profits. As explained with more details in §15.2 on the merger paradox, if firms compete on quantities, the coalition of firms entering the agreement must be very large to be attractive to all their prospective members. The most frequent form of cartel indeed gathers all major industry members except may be the smallest and youngest ones so that the adequate model is that of a monopoly (the cartel) facing a competitive fringe of price-taker firms.

What is then the optimal behavior of a cartel formed by n firms? Their objective is to maximize the sum of profits

$$\Pi \equiv \sum_{i\leq n} \pi_i = \sum_{i\leq n} q_i P(Q) - C_i(q_i) = QP(Q) - \sum_{i\leq n} C_i(q_i) \tag{9.1}$$

where P is the inverse of the residual demand received by the cartel and $Q \equiv \sum_{i\leq n} q_i$. When the cartel gathers all the industry, this is the market WTP.

The first step to maximize Π is to find out the cartel's cost function i.e., given a total output objective Q, find the cost minimizing distribution $(q_i)_{i\leq n}$ among members. The solution is the multi-plant cost function $C(Q)$ (cf. eq. 2.8) solving $C_{m,1}(q_1) = ... = C_{m,n}(q_n)$.[2@] Now, (9.1) reduces to maximizes $\Pi = QP(Q) - C(Q)$ which is a standard monopoly problem. As will be made clear in the next paragraph, each member is called to produce less than

the "no-cartel" situation, which means that the cartel formation leads unambiguously to an aggregate output reduction.

It is now easy to assess the welfare consequences of a successful cartel formation.

> Whatever might have been the intensity of competition prior to the agreement, the appearance of a cartel reduces competition, thus aggregate production which means that an *inefficient* change of market structure occured; the outcome is even worse in terms of consumer surplus.

Instability

As originally observed by Cournot (1838), anti-competitive agreements are naturally difficult to sustain because each member has an incentive to cheat or free ride on his partners by flooding the market. Recall indeed that the rationale behind the optimal monopoly behavior is the understanding that additional sales can only come at the expense of an additional rebate to existing clients. This logic is less stringent for a member of the cartel because the rebate would be only for his own clients so that his personal incentive to increase sales is stronger than that of the cartel (considered as a single large firm). To draw further implications we use the following equation:

$$C_m = \hat{C}_m = \hat{R}_m < R_m \tag{9.2}$$

where $\hat{}$ denote cartel variables.

The cartel's optimum is chosen so as to satisfy the second equality in (9.2); furthermore, technical efficiency requires the first equality in (9.2) i.e., the production of individual members generate the same marginal cost. Now, our discovery that a cartel member faces a greater marginal revenue enables to conclude that it is worthwhile for each member to increase production, since his marginal revenue is greater than his marginal cost when producing his share of the cartel's optimum. We have just uncovered the inner instability of cartels which is frequently illustrated by the frequent breaches of OPEC quotas agreements as shown in the next section.

To see this result formally, assume that independent firms compete in quantities (Cournot) as it is the case for petroleum. The cartel revenue is $\hat{R}(Q) = Q \times P(Q)$ where Q is the cartel's production while a member's revenue is $R(q) = q \times P(Q) = q \times P(q+\bar{Q})$ where q is his share of the total production and $\bar{Q} \equiv Q - q$ is the aggregate production of the remaining members. If we now look at marginal revenues, we obtain $\hat{R}_m(Q) = P(Q) + QP'(Q)$ and the greater $R_m(q) = P(Q) + qP'(Q)$ (since $Q > q$ and $P' < 0$). The member's marginal profit, when producing his share q of the cartel's optimum Q is thus $R_m(q) - C_m > \hat{R}_m(Q) - \hat{C}_m = 0$

i.e., he ought to increase production beyond what the cartel recommends, more precisely he ought to produce the Cournot best reply to $\bar{Q}$.

In the example where $C(q) = cq^2/2$ and $D(p) = a - bp$ we obtain $\hat{C}(Q) = \frac{c}{2n}Q^2$ and the collusive quantity solving $\hat{R}_m(Q) = C_m(Q)$ is $Q = \frac{an}{2n+bc}$ (this is the monopoly quantity q^M where the cost coefficient c is changed into c/n). The individual quantity is $q = \frac{a}{2n+bc}$ and the optimal deviation is $\hat{q} = \frac{a-(n-1)q^{col}}{2+cb} = \frac{n+1+bc}{2+bc}q$ which is greater than q as soon as $n \geq 2$.

> Price fixing agreements, either through a cartel or collusion, reduce competitiveness and the efficiency of the market. They are inherently unstable because each participant has an incentive to "double-cross" his partners.

It is clear that to maintain the stability of a cartel or a collusive agreement, each member must face the threat of punishment in case he does not fulfill his role. Firms use price wars to retaliate when a member deviates from the agreement. In this respect, mafias are successful cartels because they possess a very powerful enforcement mechanism, the death penalty.

9.1.2 History of Cartels

Origin

A *cartel* (from the german word Kartel) is a formal agreement designed to limit or eliminate competition between participants, with the objective of increasing profits. Members can decide to fix prices, limit their output, share markets, allocate customers, adjudge territories or even manipulate auctions for public work procurement (cf. §22.1.2).

Historically, cartels developed in Germany during the 19^{th} century as associations of firms acting in the same industrial sectors; there were more than 500 at the beginning of the 20^{th} century. Competition on world markets quickly lead to the creation of international cartels; more than 150 were active between the two world wars. Nussbaum (1986) estimates that international cartels controlled approximately 40% of world trade between 1929 and 1937. More recently, the Japanese and Korean industries have used similar associations, respectively called zaibatsu and chaebol (literally "money clan"). Yet experience and theory have shown that in the long run the distortion on competition is too strong and arms society (final consumers and firms which are clients of the cartels). For that reason, almost all countries prohibit cartels.

Figure 9.1 presents a sample of cartels whose demise came from entry, cheating, technological change and antitrust activity (cf. Levenstein and Suslow (2006)).

Industry	Year	N	Years
Parcel Post	1851	5	29
Diamonds	1870	?	60
Ocean shipping	1870	2-8	51
Oil	1871	19-50	2
Railroad-Oil	1871	3-4	7
Railroad	1875	3-15	4
Potash	1877	3-30	9
Bromine	1885	7-15	6
Sugar	1887	8-19	7
Cement	1922	4	40
Steel	1926	4-8	7
Mercury	1928	2	25
Tea	1929	349	4
Rayon	1932	2	8
Beer	1933	500	9
Electrical Eq.	1950	40	8

Table 9.1: Examples of Cartels

OPEC

The most famous example of a legal cartel is the OPEC association of countries producing large quantities of petroleum and natural gas. Each involved government owns a public monopoly for the production of fossil energy but once they go on the world market to sell their production, they are supposed to compete. The cartel was formed in 1960 after the US government imposed a quota adverse to the interest of oil exporters not belonging to the american continent. OPEC's has currently 11 members and meet approximatively three times a year to revise the ceiling allocations for production of crude oil. These ceiling are voluntary export restraints agreed by all but each member is nevertheless entirely free to overshoot them. At the start of the 1970s, OPEC's share of world production was in excess of one half; it is quite clear on Figure 9.1, computed from OPEC (2004), that the strategy of setting binding ceiling allocations in order to reduce world supply was effective to raise the price during a decade. The unexpected consequence the recession it created together with a depressed oil demand and a severe fall of the market price.

In Table 9.2, computed from OPEC (2004), the bottom row displays the ratio of total OPEC production over its own total ceiling allocations;[3@] overshooting occurs when demand is strong and conversely, the quotas fail to bind in periods of low prices and depressed demand. The country rows display a measure of quota respect by each member; less (resp. more) than 100 means that the member behaves better (resp. worse) than the

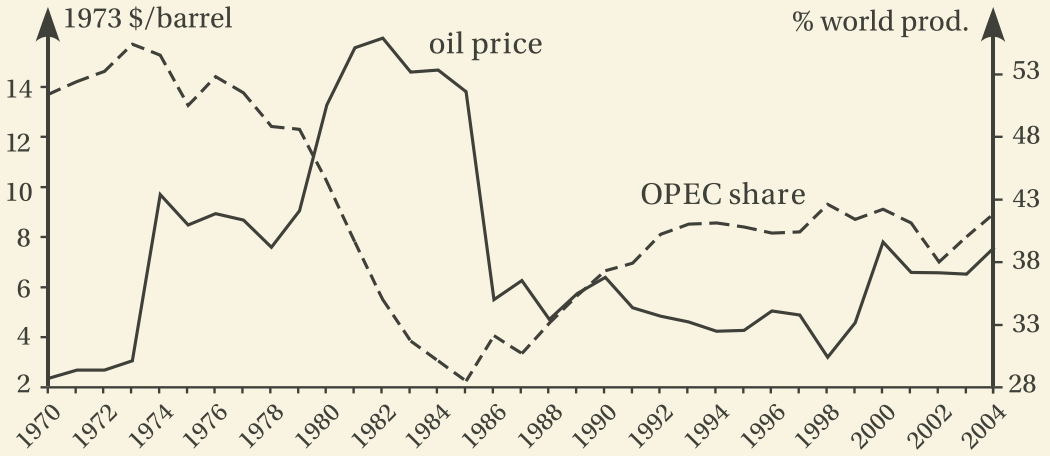


Figure 9.1: OPEC output policy

average.[4@] Notice that Saudi Arabia is systematically below par (except during the first gulf war) because it is the swing member in charge of smoothing the overall production.

Using the HHI index of concentration developed in §15.3, we may state that oil reserves are concentrated, with HHI of 5963 in 2003 if we consider OPEC as an effective cartel, but only 1022 if each country is taken separately. Oil production is rather less concentrated, with an HHI of 1900 if OPEC countries are joined, or 571 taking each country individually. Oil consumption is slightly more concentrated than production if we consider each country individually, with an HHI of 876.

Prod/Quota	*83*	*84*	*85*	*86*	*87*	*88*	*89*	*90*	*91*	*92*	*93*	*94*	*95*	*96*	*97*	*98*	*99*	*00*	*01*	*02*	*03*	*04*
Algeria	94	104	109	93	99	92	92	88	103	94	101	100	100	107	112	93	99	100	100	103	116	150
Indonesia	98	107	106	97	99	89	87	88	106	92	101	100	100	99	99	93	111	98	96	98	86	77
Iran	105	92	102	81	99	95	94	93	112	101	96	100	100	99	99	97	104	100	100	100	100	95
Kuwait	103	120	111	118	100	109	108	53	13	122	95	100	100	99	99	96	99	100	100	98	103	104
Libya	105	97	108	121	100	94	96	105	104	97	99	100	101	99	100	98	102	101	102	101	105	107
Nigeria	98	116	124	104	104	94	111	99	109	105	103	98	99	99	100	97	92	100	101	99	103	104
Qatar	92	118	111	94	75	70	90	101	104	106	104	100	103	103	106	153	99	100	101	99	102	106
Emirates	107	105	114	110	134	127	142	149	92	94	101	100	99	99	99	97	99	100	100	98	101	98
Venezuela	111	110	108	97	102	92	94	102	108	103	100	100	101	100	101	124	101	101	100	109	90	95
Saudi Arabia	93	88	78	102	93	107	98	111	107	99	102	100	100	100	99	97	99	100	100	98	101	100
OPEC	97	92	93	108	100	110	108	108	95	106	99	100	100	101	101	98	99	96	100	102	105	109

Figure 9.2: Quota overshooting by OPEC members

Another international cartel is the International Air Transport Association (IATA) which used to coordinate fares of international air travel. Sectorial deregulation in the US and Europe and the new Open Skies agreement has severely cut back its ability to maintain high prices especially within integrated economic areas as shown with more detail in §17.3.3.

Guilds and Licensed Occupations

Since medieval times, guilds of craftsmen and merchants have been examples of private but regulated monopolies who held special privileges (cf. Smith (1776), I.10.72). Some were organized as cartels while others were monopoly bestowed by the State (or the King). Medieval Kings gave charters to towns allowing them to hold fairs (and tax entrance) or to be the sole producer of a commodity (within the country).

Nowadays, most professions where learning and training is certified by a diploma are organized as "associations" and are lightly regulated by the State (cf. §16.2).[5@] prices may be negotiated with the government or may be set freely by the guild but their most distinctive feature is to be more or less identical in the entire profession. Fundamental to the well being of the association is the ability to limit entry in order to sustain monopoly prices. Examples include doctors, architects, engineers, accountants or lawyers.

Their weight of licensed occupations in the economy has grown from 4.5% of the US workforce in 1950 to 20% in 2000. This is a large increase but not so much if one takes into account the transformation of the economy into services where many hold higher education diploma (college enrollment increased seven fold in the same period or pass from 1.4% to 5.4% of the population). Law and Kim (2005)'s econometric studies concludes that the rise of regulation for many new activities during the progressive era in the US was an organized response to asymmetric information problems. As growth in scientific knowledge was accompanied by specialization, individuals became less knowledgeable about the goods and services they purchased in the marketplace. Moreover, as society became urbanized, market exchanges became more anonymous so that traditional mechanisms for overcoming the asymmetric information problem were less useful. However, the view we purport in this book is that each of these clique captured its regulator and has maintained a too tight supply of services over the last 50 years.

As far as economics are concerned, unions are also guilds. Hollywood hosts two famous guilds, that of actors and writers. The associations representing artists (e.g., ASCAP or SGAE or SACEM) and managing the royalties over the reproduction of their works are also legal monopolies whose statute is increasingly questioned as being anti-competitive and stifling entry and innovation in these industries. Trade associations use public announcements to coordinate their actions; this can be a list of prices, a delegation of power to their union representative, production quotas, agreements on product standards or limits to entry in the sector (numerus clausus). A very important instrument to avoid the price-cutting behavior alluded before is to establish market shares which are most homogeneous in order that each member be a true monopoly facing a partial demand as elastic as the total one; this way the incentive to price at monopoly level be preserved. This kind of market segmentation by geographic origin or sector (service

vs. industry) is practicable only if there are no uncorrelated demand fluctuations among segments that would force to revise too often the quotas or segments definitions.

To summarize, a cartel is inherently fragile for the following reasons:

> *Enforcement*: the coordinated behavior is not self-enforcing because of the free riding incentives, thus it must be enforced by the threat of punishment in case a member does not fulfill his role.
>
> *Sharing*: a successful cartel derives some surplus or quasi-rent; we can expect some quibbling over its division among members.
>
> *Entry*: a successful cartel, by generating above average returns, induces the entry of new competitors in the market which threatens the cartel's stability.
>
> *Retaliation*: the other side of the market suffering from a successful cartel will undoubtedly try to retaliate.

9.1.3 Collusion

Whenever cartels are prohibited, firms willing to form a cartel can instead collude i.e., adopt the same behavior aimed at raising prices and restricting output but secretly to avoid prosecution by antitrust authorities.[6@] We view *collusion* as a broader concept than cartel because, although firms might be forced to act individually and are deterred from explicitly coordination by law, they might nevertheless recognize their interdependence in the market competition and try to behave in ways that signals to others their willingness to *tacitly* collude in order to jointly exercise market power.

Sustaining collusion is obviously more difficult than sustaining a cartel since everything has to be kept secret, most notably the enforcement and sharing issues alluded to before. For instance, the EC fined Japanese video games maker Nintendo and seven of its official EU distributors for having agreed between 1991 and 1998 to maintain price differences across the EU. Each distributor was to prevent so-called parallel trade i.e., exports from its territory to another country by investigating the use of shipments by its clients. Exporters were punished by smaller shipments or boycott. As we show in the next paragraph, best-price clauses are strategic devices that facilitate collusion in a dynamic setting; antitrust autorities refer to them as facilitating practices. Detecting secret price cuts is the primary issue to enforce a collusion. One case where collusion is aided rather than deterred is public procurement since it is frequent that all bids, not only the winning one, are made public thereby enabling the identification of a cheater.

McAfee and McMillan (1992) argue that sharing is so controversial among collud-

ers that it often leads a discontented member to denounce the illegal agreement to authorities. This is illustrated in a number of decisions by the EC relative to cartels and collusion (cf. update):

- In response to increased competition, the auction houses Christie's and Sotheby's colluded to fix commission fees between 1993 and 2000. In 2002, the EC granted immunity to Christie's for coming forward with evidence of collusion while it fined Sotheby's 20 m€ , (6% of its worldwide turnover), including a 40% reduction for its cooperation in the investigation.
- 478 m€ fines to a cartel of plasterboard producers from Benelux, Germany, France and the UK between 1992 and 1998. The incriminated firms had agreed to restrict competition on these markets in line with their interests, exchanged information on their sales volumes and informed one another of price increases.
- 124 m€ fines to eight Austrian banks for their participation in a wide-ranging price cartel.
- 125 m€ fines to one European and one Japanese firm producing food for animals; a third company was granted immunity from fines because it revealed the cartel's existence and provided the EC with decisive evidence.
- 25 m€ fines to seven Dutch companies for fixing prices of industrial and medical gas in the Netherlands.
- 20 m€ fines to one Japanese and two Korean firms producers of food flavour enhancers for price-fixing and customer allocation between 1988 and 1998. A fourth Japanese firm was granted immunity from fines because it revealed the cartel's existence and provided the EC with decisive evidence.

Famous price-fixing issues in the US are

- 1982: IBM won a decade-long battle when charges of illegal monopoly against it were dropped.
- 1983: AT&T was broken into a single long-distance service and the seven "baby Bells" local telephone companies.
- 1998: Eastman Chemical paid an $11 million fine for participating in an international price-fixing conspiracy in the food preservatives industry.
- 1999: the vitamins cartel involving U.S., Swiss, German, Canadian and Japanese firms was fined $850 million; a number of top executives went to jail.

Table 9.2 shows that recent actions against price fixing deal mostly with intermediate goods and services.

Industry	Years
Aluminum Phosphide	1
Bromine Products	3
Cable-Stayed Bridges	1
Carbon Cathode Block	2
Cartonboard	5
Cement	11
Citric Acid	4
Ferrosilicon Products	2
Ferry Operators (Adriatic Sea)	7
Ferry Operators (Channel)	1
Fine Arts	6
Graphite Electrodes	5
Isostatic Graphite	5
Laminated Plastic Tubes	9
Lysine	3
Maltol	6
Marine Construction Services	4
Marine TransportationServices	5

Industry	Years
Plastic Dinnerware	1
Shipping (Central WestAfrican)	20
Shipping (Far Eastern)	4
Shipping (France-Africa)	17
Shipping (NorthAtlantic)	2
Sodium Erythorbate	2
Sodium Gluconate	2
Sorbates	17
Stainless Steel	1
Steel Beam	6
Steel Heating Pipes	4
Steel Tube	5
Sugar	4
Tampico Fiber	5
Thermal Fax Paper	1
Vitamins	9
Wastewater Construction	8

Table 9.2: International Price-Fixing US DOJ and EC Actions

9.2 Vertical Agreements

9.2.1 Best Price Clauses

In 1979, the US Federal Trade Commission (FTC) sued four manufacturers of chemical additives for gasoline because they had adopted marketing practices conducive of collusion, the so-called Ethyl case of Post-sale restraint. The defendants reportedly signed "most favored nation" agreements, gave 30-day notices of price increases before they became effective and applied uniform delivered pricing. The price notices usually resulted in price increases by all four companies on the same day. From the FTC's point of view, coordination among the firms, usually in advance of impending price changes, eliminated uncertainty about each others' willingness to follow an upward price lead. Price signaling through the press was a coordination instrument to maintain uniform pricing and reduce competition sharply in an already oligopolistic industry. Although no conspiracy was involved, the FTC decided in 1983 to bar the defending firms from announcing price changes before their effective date, using price formulas that systematically match

a competitor's price, providing sales information to any of its competitors except in connection with a legitimate sale between them and using "most favored nation" clauses in contracts. The US Court of Appeals reversed the decision the next year.

In our formal analysis, we shall distinguish two clauses, one static and one dynamic. A *meet or release* (MOR) clause sets the current price at the minimum of competitors current prices or releases the client from his buying obligation (if the transaction takes place before the good or service is consumed). A *most favored customer* (MFC) clause guarantees actual customers they will benefit from a future price reduction offered to other customers (of the same firm).[7@]

The usefulness of the MFC clause for a firm is to reduce the temptation to lower prices in order to attract new customers. We saw in §4.3.5 that it could be used by a *durable* good monopoly to mitigate inter-temporal discrimination. Let us show how it turns price competition into the more gentle quantity competition when the items for sale are durable goods such as domestic appliances (TV, fridge) or durable services such as travel tickets. We first observe that, in this context, a sales contract is binding because there is no possibility to return the item to the producer or resale it at full price.[8@]

Consider firms competing on price in a market for an homogeneous good and imagine there is a common initial Bertrand equilibrium price $\bar{p}$; it generates total sales $\bar{Q} = D(\bar{p})$ and individual sales $\bar{q}_j$ for all firms $j = 1,..,n$. Imagine now that firm i sells her product with a MFC clause. Whenever firm i offers a discount price p_i upon the current $\bar{p}$, she gains the $x \equiv D(p_i) - D(\bar{p})$ new customers attracted by the price decline but cannot steal any from her competitors since previous sales are binding. At the same time, the MFC forces her to offer the same rebate to all her former clients so that all units are sold at price $p_i < \bar{p}$, her profit is thus

$$\begin{aligned}\pi_i &= p_i\left(\bar{q}_i + x\right) - c_i(\bar{q}_i + x) = (\bar{q}_i + x)P\left(\bar{Q} + x\right) - c_i(\bar{q}_i + x)\\ &= q_i P\left(\bar{Q}_{-i} + q_i\right) - c_i(q_i)\end{aligned}$$

for any $q_i \geq \bar{q}_i$. The incentive to deviate from the current situation is given by $\frac{\partial \pi_i}{\partial q_i}$ computed at $\bar{q}_i$; hence it is only if the current situation is the Cournot equilibrium that this incentive is nil for all firms. More precisely, if the current price $\bar{p}$ is larger than the Cournot price, then at least one firm has an incentive to offer a discount while if the current price is lesser than the Cournot one and firms share evenly the market, no one has an incentive to offer a discount.

If we now consider *perishable* goods like fruits and vegetables then sales contracts are generally not binding i.e., if a better price is available at delivery time, the buyer can renege his engagement to buy and take advantage of the better opportunity. In that case, the previous result holds only if an additional MOR clause protects the firm victim

of a price-cut from losing any of her clients (they are automatically compensated, thus remain faithful).

Holt and Scheffman (1987) show that MFC and MOR clauses permit to sustain supra-competitive prices in a market for an homogeneous perishable good. Firms publish prices in advance of the actual trade period, consumers address their prospective demand to the minimum of these listed prices; lastly, firms can offer discounts at the trading time. In this framework, any price between the competitive one and the Cournot one is an equilibrium that will not be under-cut by a discount at trading time. As we saw above, agreeing on a price higher than the Cournot one is not an equilibrium since it will surely provoke a later discount. On the contrary, any lesser price is immune to an upward deviation because consumers stick to the lower listed price and to a downward deviation since the incentive identified above tell us that this move reduces the deviant's profit. In case the good is durable, the same supra-competitive outcome holds using an MFC clause only.

Schnitzer (1994) goes further to disentangle the respective role of the two clauses and gives a theoretical foundation for the widely help suspicion that MOR clauses are retaliation devices used to enforce collusive agreement. She consider a market for homogeneous durable goods with new consumers appearing at each period.[9@] In this dynamic setting, the MOR clause makes the detection of cheating on a price-fixing agreement very effective since cartel members get informed by the consumers themselves. Then, it is possible to punish the deviant by offering a (retaliation) rebate to force him to repay this amount to all of his customers, and make a huge loss. It is then possible to sustain the *monopoly* price in all periods except the final ones. Indeed, deviating upward is simply dominated while deviating downward brings a one-shot benefit that is annihilated by the later retaliation using the MOR clause.[10@]

9.2.2 Vertical Restraints

Vertical restraints (or obligations) are contractual clauses designed by upstream firms, most often manufacturers of branded goods like "Adidas" or "Gucci", for their retailers. Some restraints have to do with prices and quantities while other deal with differentiation (cf. ch.11) like the granting of an exclusive territory (horizontal differentiation) or the specification of a quality of service to be provided to clients (vertical differentiation).

When manufacturers and retailers maintain arm's length relationship using incomplete contracts (especially linear pricing), the latter have some freedom to apply their market power. Cournot (1838) shows that a "double marginalization" appears preventing the industry from fully exerting its market power. As we show in §14.1.4 on double

marginalization, integration is a way to restore profits but it also comes with costs. Vertical restraints therefore appear as a lighter alternative.

The resale price maintenance (RPM) stipulates limits to the price that the downstream firm can charge to consumers.[11@] This practice is illegal per-se in most major developed countries although the theoretical literature has never reached a clear-cut conclusion about its efficiency property (harmful vs. beneficial). A reflection of this fact is that the EC classifies it as an hard-core restriction (cannot benefit from a block exemption) but not as an illegal practice per-se. Similarly, a 1997 judgment from the US supreme court has deemed legal a maximum price in the gasoline retail market. The first case in Europe occured in 1964 when the EC declared illegal the granting of exclusive dealership rights by Grundig, a German manufacturer of household appliances to its French subsidiary. The ruling was uphelp by the ECJ in 1966 and expanded the definition of measures affecting trade to include "potential effects".

The strongest arguments against RPMs is that they help enforce collusion at all levels. Whenever a producer uses the same minimum RPM for all his retailers in a given geographic area or market, the retailers become a *de-facto* cartel since they cannot anymore lower the retail price of the product; intra-brand collusion is taking place. For collusion among manufacturers (inter-brand), the argument builds on the following observation: under a RPM, retailers cannot adjust their retail prices to local conjuncture (e.g., macro-economic shocks affecting retail costs or demand). Thus, if a group of collusive producers use RPMs, retail prices will be quite stable. If now one producer breaks the collusive agreement by offering a rebate to his retailers, those will want to pass the rebate to their customer in order to boost sales. Hence, the retail price will fall and since it is a public information, the cheating manufacturer will be easily identified (and punished by fellow conspirators).

Exclusive territories (e.g., one official shop for the brand per city) is another mean to avoid destructive *intra-brand* downstream competition. In the absence of such a restriction, retailers would compete harshly, sell at marginal cost and make zero profits. In that case, the manufacturer would not be able to use franchise fees to aspire consumer surplus from the retailers. If the manufacturer reverts to linear prices then the double marginalization problem reappears. A local monopoly guaranteed by a contractual agreement is the easiest way to restore the retailer's margin that can later be shared with the manufacturer through the franchise fee. The vertical restrictions are generally tolerated on the ground that they permit fair margins for retailers that enable them to provide (finance) a service to consumers that is commensurate with the brand image e.g., a test stand for perfume with an attendant.

A restraint aimed at softening *inter-brand* competition is the exclusivity deal between

a producer and a retailer whereby the later agrees not to sell competing brand in his shop. This completely makes sense for upstream firms like "Adidas" or "Gucci" to avoid having their products along those of "Nike" or "Dior", respectively, on the same shelve. We study in §10.2.1 how the anti-competitive potential of these exclusive deals with respect to the entry of challengers.

9.2.3 Price Discrimination

As we already explained in §4, price discrimination for freight transportation developed hand in hand with the railway system. In Britain and the US, this practice caused unease in many sectors of the economy. The US congress had to pass the Interstate Commerce Act in 1887 to regulate this activity, imposing fair and transparent rates, prohibiting personal or distance related discrimination. The Robinson-Patman act of 1936 later prohibited any price discrimination lessening competition. The landmark Morton Salt ruling of the US supreme court in 1948 made clear that different prices could only be based on different costs, otherwise charging lower prices to large supermarkets would injure groceries (and lessen food retail competition).

These limitations have since then lost their bite and in Europe, no mention of it is made in any of the articles relating to economic laws.

> It may therefore be said that price discrimination is a perfectly legal business behavior inso far as it does not build directly on characteristics such as sex, race or religion that are explicitly protected by laws of higher order.

9.2.4 Antitrust Activity

Europe

Merger scrutinity in Europe is reviewed in §15.1.3. The EC reports that national competition authorities analyzed some 180 antitrust cases in 2005 (cf. EC (2005)), evenly distributed between collusion (art.81) and abuse of dominant position (art.82). With respect to the Commission itself, Table 9.3 gathers some recent trends.

The fines imposed to cartels by the EC (cf. guidelines)and corrected for court judgments are reported in Table 9.4. The ten highest cartel fines since 2000 regard elevators and escalators, vitamins, gas insulated switchgear, synthetic rubber, plasterboard, hydrogen peroxide and perborate, methacrylates, fittings, carbonless paper and lasltly, industrial bags. It is quite clear from this list that cartelization occurs mostly in intermediate goods industries. The fight against cartels rests upon the leniency program

Year	2000	2001	2002	2003	2004	2005
Opening						
Complaints	112	116	129	94	85	55
EC Initiative	84	74	91	97	52	39
Other	101	94	101	71	21	11
Total	297	284	321	262	158	105
Closing						
Informally	362	324	330	295	363	207
Formally	38	54	33	24	28	37
Total	400	378	363	319	391	244

Table 9.3: Dominant Position

giving immunity from fines for the first firm that provides evidence of a cartel to the EC, and a substantial reduction in fines for any subsequent applicant. Under the 1996 first leniency programme, the EC received an average 13 applications per year. With the 2002 revision of the programme, the yearly average surged to 41. In 2005, the Commission received 17 applications for immunity of which 6 were granted and 11 applications for a reduction of fines.

Year	2002	2003	2004	2005	2006
Total fines	905	401	390	683	1846
Decisions	9	5	6	5	7

Table 9.4: Cartel Fines

State Aid in Europe

Before commenting on the EC activity relative to State aid, we provide some basic figures from Eurostat regarding the amounts involved. In 2005, the member states at the EU25 level granted 64b€ of State aid and 40b€ of railways subsidies (which are not accounted as State aid because it is considered a public service); this transfer amounts to 1% of Europe's GDP. The financing of State aid uses equally grants and tax exemptions (plus guarantees and equity participations). Tables 9.5-9.4 gather some geographical and historical figures for the smaller EU15 set of countries where State aid has been halved in GDP terms since 1990.

Country	DE	FR	IT	UK	ES	SW	NL	BE
State Aid	20.3	9.7	6.4	4.5	3.8	3.1	2.0	1.2
Railways	8	10	6	6.5	1.3	1.2	2.4	2.0

Table 9.5: Public Aid in 2005 (b€)

Year	95	96	97	98	99	00	01	02	03	04	05
State Aid	78	77	97	65	56	60	61	68	57	59	59
Railways						33	42	41	39	39	39

Table 9.6: Evolution of EU15 Public Aid (b€)

There are significant differences between member states regarding the sectors to which they direct aid, however at the EU-25 level, 17b€ went to agriculture and fisheries, 4b€ to coal, 2b€ to transport (mostly maritime), 2b€ to manufacture and 2b€ to services. The remaining 38b€ represent horizontal objectives, i.e. not granted to specific sectors. This form of aid is considered better suited to address market failures and thus less distortionary than sectoral and ad hoc aid. Employment received 3.6b€, regional development 8.6b€, R&D 5.4b€, SMEs 4.5b€ while the remaining 12.6b€ went to environment and energy saving.

The trend in State aid cases investigated by the EC is reported in Table 9.7. A comparison with Table 9.3 shows a different scale of complaints, all the more that complaints received in agriculture, fisheries, transport and coal are excluded from available statistics. This feature is a direct consequence of the volume of State aid. As far as member states are concerned, Italy, Germany, France, Poland, Spain and the UK provide the bulk of cases.

Year	2000	2001	2002	2003	2004	2005
Notified	883	892	822	657	661	690
Un-notified	134	152	154	101	84	92
Complaints	94	152	192	177	221	204
Total	1111	1196	1168	935	966	986

Table 9.7: State Aid Scrutiny in Europe

US Antitrust Activity

It is often said that antitrust in the US by the FTC and the department of justice (DOJ) was too active in the 1960s and 1970s, too passive in the 1980s and adequate in the 1990s, reflecting the changes in power between democrats and republicans. Academics rather argue for a paradigm shift with the absorption of Chicago School perspectives into the mainstream of antitrust policy in the 1970s and 1980s as illustrated by the data of Table 9.8 (cf. Crandall and Winston (2003)).

Period	Hor. restraints	Ver. restraints	Dominance	Discrimination
1961-1968	2.6	6.4	2.6	64.7
1968-1976	1.5	13.9	3.3	5.1
1977-1980	5.5	7	1.2	2
1981-1988	7	0.6	0.4	0.6
1989-1992	6.2	1	0	0
1993-2000	7.6	2	1.4	0.1

Table 9.8: Antitrust Non Merger Cases per Year in the US

FTC activity based on discrimination cases was critiqued as incompatible with antitrust enforcement; it was almost abandoned from 1970 on. Horizontal restraints enforcement expanded with a focus on professional associations and trade associations. DOJ became involved with criminal prosecution of supplier collusion; the number of criminal cases rose from 13 to 28 and 80 per year during the 1960s, 1970s and 1980s, before falling back to 60 per year during the last decade (but with a dramatic increase in penalties recovered).

Chapter 10

Barriers to Entry

This chapter studies a specific abuse of dominant position, the erection of barriers that either directly reduce the degree of competitiveness of a market or impede the socially beneficial entry of new challengers.

We first review the standard theory, in use before the development of game theoretical models of entry barriers. Then, we tackle preemption and foreclosure which are strategies of entry deterrence only differing by their timing. We also study strategies of predation, signaling and attrition which deal with exclusionary abuse of dominant position. Lastly, we present the barriers erected by the State, voluntary or not, under the heading of Product Market Regulation.

10.1 Standard Theory

Concept

The notion of *barrier to entry* is due to Wallace (1936) and was formalized by Bain (1956) as "anything that allows incumbent firms to earn above-normal profits without the threat of entry". The EC (as well as other antitrust authorities) makes a clear distinction between the entry barriers resulting from a particular market structure and those resulting from the behavior of incumbent firms. Indeed, the historical presence in a market bestow incumbents with comparative advantages that act as *natural* barriers to the entry of new competitors. Those listed by Wilson (2002) are:

- *Scale economies*: a significant market share is needed for survival.
- *Cost advantage*: accumulation of capital, knowledge or experience.
- *Key input*: access at low cost to a crucial input or asset.
- *Cost of capital*: entrants must pay a risk premium to financial markets.

Hence, an incumbent firm has often a perfectly legal dominant position in the market; what is unlawful is to abuse it. But why should we expect such a behavior? As we saw

in §6.1, entry almost always decreases the profits of an individual incumbent firm and this effect is even stronger for a monopoly. Anticipating this possibility, incumbents may be tempted to strategically erect barriers (or heighten natural barriers) in order to discourage entry.

Examples

Let us first review some examples of barrier to entry in a market. *Preemption* is to claim or preserve a monopoly position by means other than building a cost advantage. The incumbent commits costly actions that irreversibly strengthen its options to later exclude competitors. This behavior is not illegal per-se, it is only some of the later behavior that might be.

In 1994, the US-based biotech company Agracetus won a European patent on all genetically engineered soybean varieties and seeds and all methods of transformation, this despite the opposition of Monsanto and other producers of fertilizers and seeds who argued for lack of novelty or inventiveness. After being overruled, Monsanto bought Agracetus two years later in a clear strategic *preemption* of the desired patent. This move has successfully insulated Monsanto from competition; in 2001, this company provided more than one-half of the soya crop worldwide. Oddly enough, the patent was revoked in 2007 for the reasons initially argued against it.

Other examples of preemption, deemed anti-competitive, are exclusive contracts with upstream or downstream partners and clients, brand proliferation,[1@] excessive production capacity, buying patents but not using them (cf. §12.2.3), advertising, increasing quality beyond the optimal level.

A frequent concern for the EC is *Foreclosure*,[2@] the restriction of market access to potential competitors either upstream or downstream by absolute refusal to deal, degradation of the quality of access, preemption of important sources of raw material supply or preemption of distribution channels through exclusivity contracts. Even a government can help to erect a barrier through licensing requirements and other regulations that slow the entry of (foreign) competitors. Examples of *Foreclosure* are historically numerous, they deal with access to a stadium, a railroad bridge, a train station, a harbor, a power transmission network, a local telecommunication network, and a computer reservation system.

Bain-Sylos Postulate

As recorded by Modigliani (1958) in his review of the books by Bain (1956) and Sylos-Labini (1957), the small number of active firms in an oligopoly might well be due to

entry barriers of technological or administrative nature; this is what the Cournot model implicitly assumes in that it takes the number of firms as a parameter. Endogeneizing entry is made possible by the Bain-Sylos postulate. It is worth noticing that at a time where game theory was not available to systematize the study of sequential interactions, this postulate enabled to derive meaningful predictions, matching factual observations.

The *Bain-Sylos* postulate states that *an incumbent won't change his production or price after entry*.[3@] This behavioral assumption turns the incumbent into a Stackelberg leader (cf. §6.2). The second step of the analysis is to compute the *limit price* which, in some cases, is the minimum of the entrant's average cost. By using that price, the incumbent convinces the entrant that she will make losses upon entering the market. The entrant being rational and believing the postulate will therefore choose not to enter when observing the limit price. Alternatively, a limit quantity would play the same signaling role. Notice that the higher the fixed cost of entry, the higher the limit price i.e., the easier it is to block entry.

Now, the incumbent can decide to use or not the limit price. If the fixed cost of entry is extremely large, the incumbent can continue to use the monopoly price because it is low enough to deter entry; this situation is called "blockaded entry". For a lower fixed cost, the incumbent optimally chooses to use the limit price to impede entry. Finally, if the fixed cost is low, the best strategy for the incumbent is to allow entry; although blocking is still a possibility, it has become too costly.

Critique

Although it has proved immensely useful, the Bain-Sylos postulate is seriously flawed as it goes against the basic rationality of economic agents; if there is entry of a challenger, there is a change in the market structure, in the conditions of competition and therefore a new pricing behavior is called for. Stating that such a reaction won't take place is all about *reputation*.[4@]

When trying to assess the plausibility of this hypothesis, one is lead to wonder how costly it may be to build this predatory reputation and also whether there wouldn't be an alternative and more profitable strategy. Dixit (1979) is among the first to provide a game theoretical analysis of entry where these questions start to be answered. His model presented in §10.2.3, uses game theory to assess the cost and benefits of erecting anti-competitive barriers in a framework where firms are unable to make non credible commitment.

There is a wealth of examples showing how quickly firms react to entry. Google provides us with a series of innovation that have spurred rapid reaction. This company unveiled its "desktop search" software on 14/10/04; the same day, AOL announced that

it would soon launch of a similar engine. On 9/12/04, Yahoo announced it would start its own desktop search in january 2005; it was effective on 11/01/05. Finally, MSN, the net division of Microsoft launched its own on 13/12/04. The same story went on with Google's "1GB email account" and with their "browser search bar".

In a slightly different register, Microsoft whose new OS was being scheduled for beta testing in 2006 advanced the launch of a beta version together with the marketing name "Windows Vista" on 22/07/05, only 3 months after Apple's introduction of its "Tiger" OS. Considering the number of years between each of Microsoft OS launches (W95, W98, XP), this is a short span of time which has been interpreted by specialists as an intent by Windows to demonstrate to the general public, that it is too active on the OS technology front.

10.2 Preemption

In this section, we review several preemption strategies that aim at excluding challengers whether they are actual competitors or potential entrants by means of exclusive contracts and/or excessive capacity building.

10.2.1 Exclusive Contracts: Efficiency vs. Exclusion

An exclusive dealing requires a buyer to purchase products or services for a period of time exclusively from one supplier. Prior to the modern era, this practice is understood as a well groomed relationship. Its increasing use can be traced to industrialization when firms start to specialize and cease carrying the entire chain of production down to the final consumer. Prior to the XXth, exclusive dealings are seen through a laissez-faire lens as partial and thus not unreasonable (cf. rule of reason) if founded on risk (supply, price) or protecting investments (cf. §13.3.3). They are routinely upheld in court against legal challenges as soon as there existed a viable alternative (i.e, the market is not monopolized).

During the Progressive Era, the dominant view in the US becomes that a market leader can impose exclusionary contracts to its own benefit and to the detriment of consumers by foreclosing the entry of challengers. The Clayton Act (cf. §8.3) of 1914 address this concern by outlawing tying and exclusive dealing arrangements which *could* lessen competition substantially or *tend* to create a monopoly. Gilbert (2000) notes that, although these arrangements are not qualified as illegal per se, no serious test of infringement is envisioned so that the spectrum of applicability is quite large. Applying the new law, courts have stricken down exclusivity contracts as anti-competitive in cases

involving tying (e.g. §24.1), exclusive dealing (cf. Hollywood movies distribution), and long-term contracts. The most common turn-around to this prohibition for a manufacturer are own stores, franchising and the more recent store-within-a-store concept (cf. Jerath and Zhang (2010)).

Starting in the 1950s, the Chicago school offers a pro-competitive defense of exclusive dealings (cf. Director and Levi (1956) & Bork (1967)) which has proven influential and lead to a greater acceptance of these practices in courts. The case rest on a version of the Coase theorem (cf. §2.4.3): contracts are voluntary, not imposed, and must therefore maximize the combined benefits of the contracting parties. Contractual terms, apart from total price, must therefore be explained as wealth-maximizing, not as the result of relative market power or bargaining power. The first welfare theorem would seem to follow: if a contract maximizes the combined benefits of signatories, it must be efficient. If true, government intervention that limits the set of feasible contracts cannot improve welfare; on the contrary, it impedes parties from achieving all the efficiencies at their reach.

The validity of this claim is however limited since it may fails when there are non competitive third parties or more generally when we take into account externalities. At the same time, simple changes of seemingly innocuous assumptions can sway the conclusion one way or another so that, ultimately, one must work each case very carefully in order to make a call as to whether exclusivity contracts are anticompetitive (exclusionary). Rasmusen et al. (1991) show that uncoordinated buyers may be "tricked" into accepting contracts that, as a group, they would refuse.[5@] Simpson and Wickelgren (2007) then show that the ability to renegotiate contracts brings back efficiency so that exclusive dealing cease to be anti-competitive. Next, if buyers are in fact retailers competing intensively, then exclusive dealing is a vehicle to extract surplus from final consumers for the the joint manufacturer-retailer association; this is inefficient, thus anti-competitive. Abito and Wright (2008) further show that two-part pricing and/or discriminatory pricing and/or sequential contracts (divide and conquer), make exclusion cheaper so that the manufacturer effectively exclude entry, regardless of the degree of downstream competition and any cost advantage of the entrant.

Modern cases of exclusivity deals suspicious of dampening competition are the management of rights for music, football or movies, fees by VISA, licensing schemes by Coca-Cola in Europe, Intel, Yahoo, Apple, AT&T or Google.

10.2.2 Contracts as Barriers to Entry

In a Take-or-Pay Contract, the buyer agrees to purchase from the seller a fixed quantity of a product for a given price over a certain period of time. Irrespective of the quantity which is finally needed and transferred, the buyer is bound by its commitment and is required to pay for the whole volume of sales at the contractual terms agreed upon. In such a contract, there is a transfer price for exchanged units and a (generally lower) breach price for non-exchanged units (that were nevertheless agreed upon). Hence, we are faced with an option contract whereby an option fee is paid up-front for the right to buy at a discounted price in the future.

Such clauses are often seen as anti-competitive and may be prohibited but it is only recently that a formal proof of this assertion was provided. We shall see how exclusivity clauses modify competition, and therefore, the entry decision of potential competitors.

Aghion and Bolton (1987) explore preemption within a sequential model where a challenger has the option to enter a market initially controlled by an incumbent monopoly. In case of entry, a duopoly competition takes place between the incumbent and the challenger. We impose the following rationality condition that is absent from the contestability theory or the Bain-Sylos postulate: neither the challenger nor the incumbent can credibly pretend to behave in a given manner during the duopoly competition (cf. §6.1.7). Both anticipate that during the duopoly competition, both will take whatever action is optimal at that moment. In other words, you can tie your hands for the future by investing heavily into a specialized equipment but you cannot credibly pretend that you will wait tomorrow to make such a costly investment.

Competition without Contracts

An incumbent (monopoly) seller I can offer a contract to a buyer B at date 1 to impeach entry of a competitor E at date 2. Actual exchanges occurs at a later date 3. For analytical tractability, we define a contract as (p_c, d_c) where p_c is the agreed price and d_c the *damage*[6@] paid by B to I if B breaches the contract i.e., buy from E instead of I. In case of entry, I and E compete in prices at date 3. We shall see that contracts constitute a barrier to entry if post-entry profits for the incumbent are lower than the pre-entry profits.

It is public knowledge that the incumbent's cost is $c_I = \frac{1}{2}$, while the entrant's cost c_E is private information; thus, a contract cannot be contingent on c_E. Nevertheless, it is known that c_E can take any value between 0 and 1. The buyer values the good at 1 so that trade is always optimal. The default payoff for E if she does not enter the market is set to zero.

To show that a contract can be useful for the incumbent we first analyze the no-contract situation. The game tree is shown on Figure 10.1 below where N stands for nature who draws the entrant's cost c_E at random in $[0;1]$ (uniform distribution).

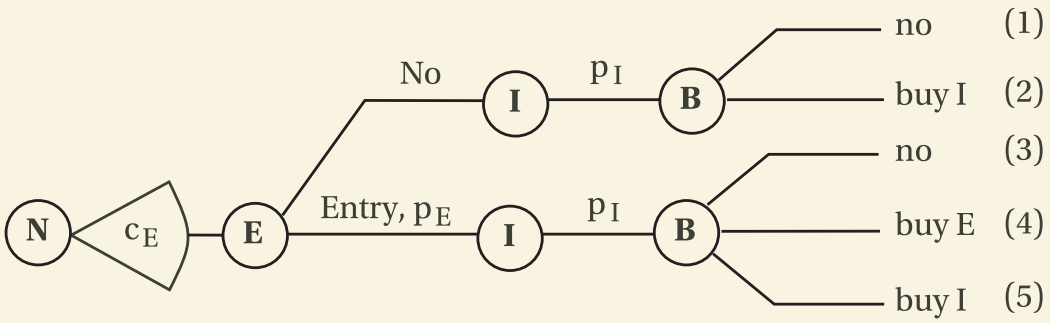


Figure 10.1: Entry Sequence

We analyze this game by backward induction. The buyer chooses to buy if $p_I \leq 1$, his willingness to pay for the good; this gives us the rule to choose among branches (1) and (2) on Figure 10.1. The optimal strategy for the incumbent is thus to propose $p_I = 1$ so that the no-entry profits are $\pi_I^{out} = 1 - c_I = \frac{1}{2}, \pi_B^{out} = \pi_E^{out} = 0$.

If E enters then both firms compete in Bertrand fashion to win the buyer's purchase. The winner is the firm with the lowest marginal cost and the price is the highest marginal cost $\max\{c_I, c_E\}$. If the entrant is the loser (case $c_E \geq c_I$), she earns zero profit and therefore chooses optimally not to enter. There is entry only if the entrant has a better technology i.e., $c_E \leq c_I$, which leads to profits $\pi_E^{in} = c_I - c_E \geq 0$, $\pi_I^{in} = 0$ and $\pi_B^{in} = 1 - c_I$. The probability of entry is thus $\phi = Prob(c_E \leq c_I) = \frac{1}{2}$. We deduce the expected (before Nature chooses c_E) profits of the game without contracting:

$$\begin{aligned}
\pi_E &= (1-\phi)\pi_E^{out} + \int_0^{1/2} \left(\tfrac{1}{2} - c_E\right) \mathrm{d}c_E = \tfrac{1}{8} \\
\pi_B &= (1-\phi)\pi_B^{out} + \int_0^{1/2} \left(1 - \tfrac{1}{2}\right) \mathrm{d}c_E = \tfrac{1}{4} \\
\pi_I &= (1-\phi)\pi_I^{out} + \phi\pi_I^{in} = \tfrac{1}{4}
\end{aligned}$$

Competition with Contracts

When a contract (p_c, d_c) between the incumbent and the buyer is legal, the game tree is modified as shown on Figure 10.2. The incumbent price p_c is now set before the hypothetical entry of the competitor; thus if the contract is accepted by the buyer, the incumbent gains commitment power: he commits not to enter into the price war that was taking place in the absence of a contract. Since this effect is potentially bad for its profit, the incumbent uses the second instrument, the damage d_c, to compensate himself.

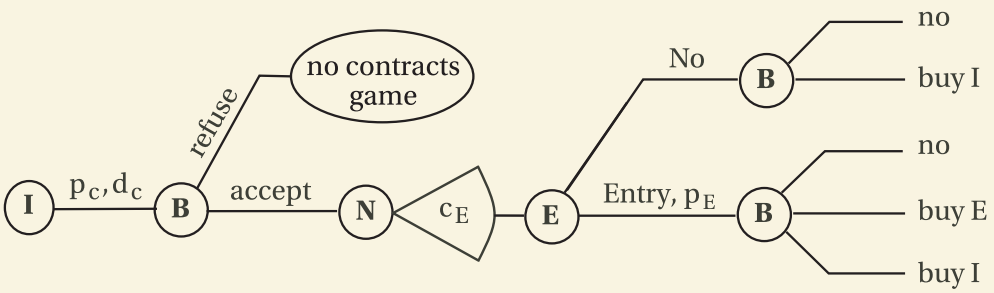


Figure 10.2: Contract as a Barrier to Entry

From the analysis of the no-contract game we know that the buyer will accept the contract only if he expects a profit of $\frac{1}{4}$ (or more). Now observe that after accepting (p_c, d_c), the buyer either gets $\pi_B^{out} = 1 - p_c$ if there is no entry because the price is fixed or $\pi_B^{in} \geq 1 - p_c$ if there is entry because at worst he sticks to the contract. We thus conclude that $p_c \leq \frac{3}{4}$ must hold for the contract to be accepted.

Since B must pay d_c to I if he buys from E, he will do so if and only if $p_E \leq p_c - d_c$. Given this behavior, the optimal price set by E is $p_c - d_c$ but it has to be larger than the marginal cost to make entry profitable, thus the probability of entry is exactly $p_c - d_c$. The main difference with the no-contract case is that the entry decision depends only on the terms of the contract and not anymore on c_I, the technology of the incumbent.

It remains now for the incumbent to design an optimal contract; when the contract (p_c, d_c) is accepted the incumbent's profit is

$$\pi_I(p_c, d_c) = (p_c - d_c)d_c + (1 - p_c + d_c)(p_c - c_I) \tag{10.1}$$

The two parts in (10.1) correspond to entry and no-entry of the competitor. As $\frac{\partial \pi_I}{\partial p_c} = 2(\frac{3}{4} - p_c + d_c)$ and $p_c \leq \frac{3}{4}$, we observe that necessarily $\frac{\partial \pi_I}{\partial p_c} \geq 0$; this means that the optimal price is precisely the limit $\frac{3}{4}$. The intuition is that given the commitment power of setting the price long in advance, the incumbent would like to raise the price a lot. Yet this would make the buyer refuse the contract in the first place, thus the participation constraint $p_c \leq \frac{3}{4}$ will be binding. The incumbent's profit now reads

$$\pi_I(d_c) = \left(\tfrac{3}{4} - d_c\right) d_c + \left(\tfrac{1}{4} + d_c\right) \tfrac{1}{4} \tag{10.2}$$

The optimal damage, maximizing (10.2), solves the FOC $\frac{3}{4} - 2d_c + \frac{1}{4} = 0 \Leftrightarrow d_c = \frac{1}{2}$; it yields a profit of $\frac{5}{16}$ greater than $\frac{1}{4}$, the profit in the absence of contracts. This ultimately proves the usefulness of a contract to limit entry and raise profits.

Observe that, upon signing the equilibrium contract, entry takes place when $c_e < p_c - d_c = \frac{1}{4}$ which is inefficiently low. For that reason, it may be socially desirable to

prohibit exclusivity clauses to guarantee a maximal level of competition. Notice finally by examining (10.2) with $d_c = \frac{1}{2}$ that the incumbent earns more when entry occurs ($\frac{1}{2}$) than when it does not ($\frac{1}{4}$). This implies that the incumbent will never wish to renegotiate the contract making this strategic barrier all the more robust.

> An incumbent seller facing a threat of entry into his market will sign long-term contracts that sometime prevent the entry of a lower cost producer. Although contract are freely entered by the two parties, welfare is reduced on expectation because better technologies are frequently barred from replacing obsolete ones. Long-term contracts between buyers and sellers can therefore be deemed anti-competitive.

This conclusion can be summarized with the help of Figure 10.3 where the distribution of surpluses makes clear how the incumbent is able to absorb part of the welfare addition bring about by the cost innovation of the entrant.

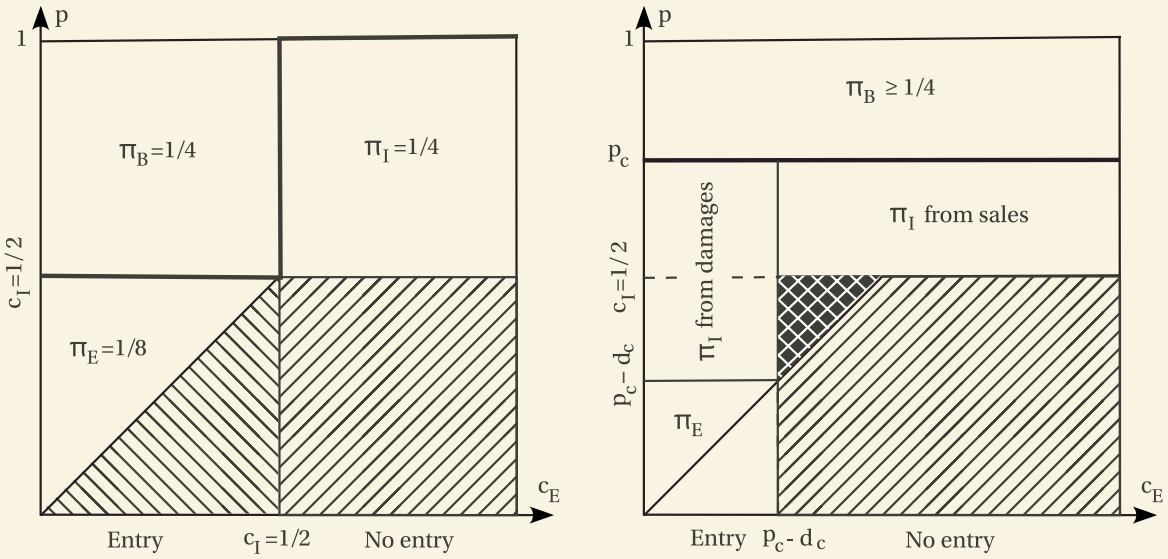


Figure 10.3: Contract as a Barrier to Entry

Exclusion as a Surplus Extraction Vehicle

Jing and Winter (2010) synthesize the whole debate around the potential anti-competitive nature of exclusive dealing as follows: the signatories may end up better off even though their deal is inefficient (welfare-reducing) if they succeed to increase their share by a large margin (of the welfare). They consider an incumbent manufacturer selling to non competing retailers (e.g., local monopolies). To maximize their joint profit, the manufacturer and a retailer minimize their overall cost. When the new technology has a cost

below the exercise price x, challengers enter and Bertrand competition among them force their price towards their cost c. The expected cost is thus $C(x) = \int_0^x c\,\mathrm{d}H + (1-H(x))c_I$. The optimal exercise price solves $C'(x) = 0$ and is $x = c_I$, so that entry is efficient and not blockaded.

Now, if there is a sole challenger (as in Aghion and Bolton (1987)), he will monopolize manufacturing upon entry i.e., he will always sell at x whenever $c < x$. This makes the cost of manufacturing dearer and leads to an inefficient optimal x because the pair incumbent-retailer will want to reduce outsourcing wrt. the previous case. Formally, the expected cost $\tilde{C}(x) = xH(x) + (1-H(x))c_I$ satisfies $\tilde{C}(c_I) = H(c_I) > C'(c_I) = 0$ so that the optimal choice is smaller and exclusion ensues.

Naked exclusion à la Rasmusen et al. (1991) takes into account the minimum scale of operation needed to support entry of a challenger when the downstream market is populated by many firms or consumers. The incumbent can then exploit buyers' lack of coordination to block entry by signing exclusivity contracts with enough downstream buyers to shrink the remaining market. The underlying phenomena is that every buyer who signs an exclusive contract imposes a negative externality on other buyers. When many have already signed or are about to do so, any free buyer realizes that his choice will not affect the entry decision of a challenger and can thus be bribed at a very low cost by the incumbent. At the limit, the incumbent brings all buyers on board. When buyers are able to coordinate, exclusion is profitable only if the amount earned from all buyers thanks to monopoly pricing, exceeds the minimum amount that must be paid to signatories to exclude entry.

Analytically, if m retailers out of n are signed with the incumbent, the zero profit condition among competing challengers implies $0 = mx + (n-m)p - nc - F$ i.e., $p_f = \frac{F+nc-mx}{n-m}$. Indeed, the m captive buyers are lured with a price of x (undercutting the incumbent) while free buyers enjoy the price war between the two entrants. If this price is above the WTP of buyers v (i.e., $c > c_2$), entry is impossible since the incumbent cannot be profitably undercut. On the contrary, if this price is below the incumbent's cost c_I (i.e., $c < c_1$), he can't impede entry so that we have 3 regimes according to whether the new technology cost c stands. In the intermediate case, entry is blocked by lowering the free buyer's price down to p_f. The joint surplus of the signatories is $m(v - \tilde{C}(x)) + (n-m)\pi_f$ where π_f is the profit made over free buyers and $\tilde{C}(x)$ the expected manufacturing price. Observe now that $\pi_f = v(1-H(c_2)) + \int_{c_1}^{c_2} p_f\,\mathrm{d}H$ and it is a matter of algebra to check that this is decreasing with the exercise price x.[7@] Lastly, since $\tilde{C} = xH(c_1) + c_I(1-H(c_1))$ is increasing with x,[8@] the joint surplus rises when pushing the exercise price below the incumbent's cost which is inefficient, thus anti-competitive.

10.2.3 Preemption with Capacity Building

Notable cases of exclusion by means of building an excessive capacity are Alcoa and Dupont as recanted by Ghemawat (1984) (sec. II).

Dixit (1980) is one of the early applications of game theory to industrial organization showing that preemptive (excessive) capacity building can be a deterrence strategy successful to block entry. The idea here is quite simple: if you build a large and costly capacity you credibly convince potential entrants that competition in the future will be fierce; indeed, now that you have a large capacity, there is nothing that can impeach you from flooding the market if you want to. Reasoning backward, the challenger understands that upon entry the market will be flooded, the price will be so low that he will not cover his fixed out of his producer's surplus, thus he rationally abstain from entering.

In this section, we present the model that capture this idea and proceed to demonstrate its correctness. The incumbent firm I can enlarge its production capacity k_I before the potential entrant E decides to enter or not the market (at a fixed cost F). If there is entry, firms compete on quantities à la Cournot for the demand $D(p) = 1 - p$. There is one (best) technology available to both firms; the cost of a unit of capacity is δ and once a capacity k is installed, it allows to produce at marginal cost $c(q) = \begin{cases} c & \text{if } q \leq k \\ c+\theta & \text{if } q > k \end{cases}$ where $\theta > \delta$ indicates that investing ex-ante into capacity reduces the ex-post marginal cost.

Best replies in quantities†

We show that producing beyond capacity does not appear in equilibrium. Indeed the profit of firm i (letting j be the other firm) is $\underline{\Pi}_i(q_i, q_j) = q_i(1 - q_i - q_j) - cq_i$ if $q_i \leq k_i$, $\overline{\Pi}_i(q_i, q_j) = q_i(1-q_i-q_j) - cq_i - \theta(q_i - k_i)$ otherwise. The unconditional argument maximizer of $\underline{\Pi}_i$ is $\frac{1-q_j-c}{2}$ (the plain line on Figure 10.4 below) while that of $\overline{\Pi}_i$ is $\frac{1-q_j-c-\theta}{2}$ (the dashed line). The best reply to q_j, drawn in bold face on Figure 10.4, is $\frac{1-q_j-c-\theta}{2}$ over $[0; k_i]$, then k_i and then $\frac{1-q_j-c}{2}$ over $[k_i; 1]$.

It is worthwhile for firm i to incur high extra production cost (pay θ) to cover the market if q_j is very low because the marginal benefit is greater than $c + \theta$. At the other extreme, if q_j is very large, the marginal benefit for i is very low and it is optimal to retreat to a low production, thus his capacity will not be completely used. For intermediate q_j the optimal response of firm i is to produce up to capacity, no more, no less because the price is between c and $c + \theta$. Analytically the best reply is

$$BR_i(q_j) = \begin{cases} \frac{1-q_j-c-\theta}{2} & \text{if } q_j < \max\{0, 1-c-\theta-2k_i\} \quad i) \\ k_i & \text{otherwise} \quad ii) \\ \frac{1-q_j-c}{2} & \text{if } q_j > \max\{0, 1-c-2k_i\} \quad iii) \end{cases}$$

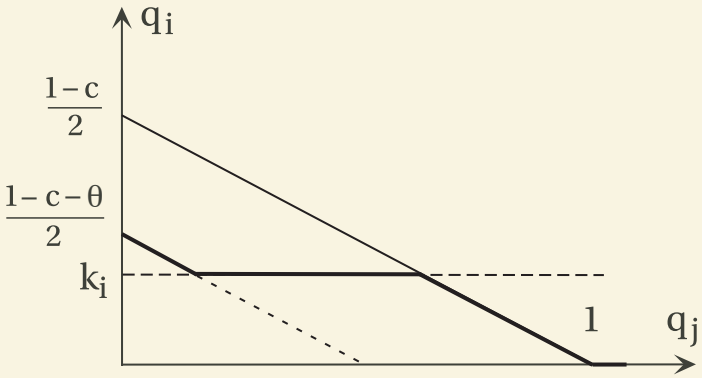


Figure 10.4: Quantity Best Reply

Cournot Equilibrium and Capacity Reactions

As can be seen from Figure 10.4 the best reply curves can intersect at most in 3 ways, over the first segment (i), the second (ii) or the third (iii).

i) The solution of $\left\{q_i = \frac{1-q_j-c-\theta}{2}, q_j = \frac{1-q_i-c-\theta}{2}\right\}$ is $q_i^* = q_j^* = \frac{1-c-\theta}{3}$. It is an equilibrium only if $q_j^* < 1-c-\theta-2k_i$ for $i=1,2$ thus if $k_i \leq \frac{1-c-\theta}{3}$ for $i=1,2$.

ii) In this middle case, we have $q_i^* = k_i, q_j^* = k_j$; it holds when both k_i and k_j are between $\frac{1-c-\theta}{3}$ and $\frac{1-c}{3}$.

iii) The solution of $\left\{q_i = \frac{1-q_j-c}{2}, q_j = \frac{1-q_i-c}{2}\right\}$ is $q_i^* = q_j^* = \frac{1-c}{3}$ together with the necessary condition $k_i \geq \frac{1-c}{3}$ for $i=1,2$.

Case (i) is valid if capacities are small in which case the profit of firm i is $\Pi_i = cte - \theta\left(\frac{1-c-\theta}{3} - k_i\right) - \delta k_i$. As $\theta > \delta$, Π_i is increasing with k_i thus the best reply of firm i is to increase its capacity. Case (iii) is the opposite, it is valid for large capacities and we have $\Pi_i = cte - \delta k_i$, thus the best reply of firm i is to decrease its capacity.

These two possibilities indicate that in a perfect equilibrium, intermediate capacities are chosen i.e., case (ii) applies. The profit of firm i is $\Pi_i(k_i,k_j) = k_i(1-k_i-k_j-c-\delta)$, hence the best reply to k_j is $k_i = \frac{1-\delta-c-k_j}{2}$. In the next part we will use this optimal behavior.

Blocking vs. Allowing

Since the incumbent plays first with k_1, the challenger optimally enters with $\hat{k}_2(k_1) = \frac{1-\delta-c-k_1}{2}$. Its resulting profit is thus $\hat{\Pi}_2(k_1) \equiv \frac{1}{4}(1-\delta-c-k_1)^2 - F$. The incumbent firm can make entry non profitable by choosing $k_1 \geq \bar{k}_1(F) \equiv 1-\delta-c-2\sqrt{F}$, but this need not be the optimal policy especially when $\bar{k}_1$ is large because building capacity is costly. We analyze this choice as a function of F, the fixed cost of entry.

If F is very large, then $\bar{k}_1(F) = 0$ so that the incumbent is an unconstrained monopoly. The optimal capacity is $k_1^* \equiv \frac{1-c-\delta}{2}$ because, although it is paid in two stages, the total marginal cost is $c+\delta$. It yields the monopoly profit $\bar{\Pi}_1 \equiv \frac{1}{4}(1-c-\delta)^2$. This situation is possible if $\bar{k}_1(F) \leq k_1^* \Leftrightarrow F \geq \bar{F}_3 \equiv \frac{(1-c-\delta)^2}{16}$.

If on the other hand F is small, then $\bar{k}_1(F)$ is large which means that the incumbent has to build a wasteful large capacity to block entry. Then, having already paid $\delta \bar{k}_1(F)$, the relevant marginal cost is c so that the optimal sales are $\frac{1-c}{2}$ yielding the total profit $\underline{\Pi}_1(F) \equiv \frac{1}{4}(1-c)^2 - \delta \bar{k}_1(F)$. This solution is feasible only if $\bar{k}_1(F) \geq \frac{1-c}{2} \Leftrightarrow F \leq \bar{F}_2 \equiv \frac{(1-c-2\delta)^2}{16}$. If $\bar{F}_2 < F < \bar{F}_3$, the optimal sales are constrained by the capacity and the total profit is $2F\bar{k}_1(F)$.

With this preliminary information we can proceed to analyze the optimal behavior of the incumbent. Given that blocking entry is costly when $F < \bar{F}_2$, it may be optimal to accommodate entry. In that case, the price resulting from the Cournot equilibrium is $p = 1 - k_1 - \hat{k}_2(k_1)$. The profit is thus $\Pi_1(k_1) = k_1\left(1 - k_1 - \hat{k}_2(k_1) - c\right) - \delta k_1 = k_1 \frac{1-\delta-c-k_1}{2}$ which is 50% of the monopoly profit. The optimal capacity is thus the monopoly one, k_1^*, leading to an optimal reply $k_2^* = k_1^*/2$ and equilibrium profits $\bar{\Pi}_1/2$. It will be optimal to allow entry if $\bar{\Pi}_1/2 > \underline{\Pi}_1(F) \Leftrightarrow F \leq \bar{F}_1$ for another threshold $\bar{F}_1$ that we assume smaller than $\bar{F}_2$ (δ small enough).

In conclusion, the optimal behavior of the incumbent depends on the fixed cost of entry:

- if $F \leq \bar{F}_1$, build the monopoly quantity k_1^*, allow entry and sell k_1^*.
- if $\bar{F}_1 < F \leq \bar{F}_2$, choose an excessive capacity $\bar{k}_1(F) > k_1^*$, block entry and sell $\frac{1-c}{2} < \bar{k}_1(F)$ (some capacity remains idle, there is a waste).
- if $\bar{F}_2 < F \leq \bar{F}_3$, choose an excessive capacity $\bar{k}_1(F) > k_1^*$, block entry and sell up to capacity.
- if $F > \bar{F}_3$, choose k_1^*, block entry at no cost and sell the monopoly quantity.

To conclude on capacity as a barrier to entry, if the fixed cost of entry is

very large: the incumbent is a standard monopoly and behaves as such.

large: the incumbent builds an excessive capacity and uses all of it to block entry.

intermediate: the incumbent builds an excessive capacity and uses part of it to block entry.

small: the incumbent builds a capacity equal to the monopoly sales, allows entry and earns half of the monopoly profit.

10.3 Foreclosure

The economic definition of foreclosure puts the emphasis on the denial of proper access to an *essential facility*, often called a bottleneck. According to the EC, the latter is an infrastructure (e.g., national electricity power grid or internal code of an operating system) which is necessary for reaching customers and/or enabling competitors to carry on their business. It is deemed essential if its duplication is impossible or extremely difficult due to technological, physical, geographical, legal or economic constraints. Denying access to an essential facility may be considered an abuse of a dominant position by the EC, in particular where it prevents competition in a downstream market.

The question that immediately comes to mind is, why would the owner of an essential facility refrain from selling the use of his facility to anyone ready to pay for it? After all, he has a monopoly for that very valuable service. To understand the rationality of foreclosure we can draw an analogy by seeing a bottleneck owner as a durable-good monopolist (cf. §4.3.5): once he has contracted with a downstream firm for access to his essential facility, he has an incentive to provide access to other firms as well. This is an inter-temporal commitment problem where Coase (1972)'s theorem tells us that the monopolist does not obtain the monopoly profit because he creates his own competition: by selling more of the durable good at some date, he depreciates the value of units sold at earlier dates; the prospect of further sales in turn makes early buyers wary of expropriation and makes them reluctant to purchase initially. This is why it is better to limit access in the first place.

10.3.1 Theory

We follow a simple model from Tirole (1988) where an upstream firm U produces an input at marginal cost c; she monopolistically supplies downstream firms A and B. The latter produce homogeneous goods using the input on a one-for-one basis and at zero marginal cost. The WTP for the final product is $P(q)$. The industry's monopoly profit $q(P(q)-c)$ corresponds to the vertical integration of downstream firms with the upstream one. This profit is maximum for an output q^M, price $p^M = P(q^M)$ and yields the maximum profit π^M.

The timing among independent firms is as follows:

1. U offers firms A and B tariffs $T_A(.)$ and $T_B(.)$.
2. Each firm orders a quantity of input. Orders are observable.
3. Firms A and B compete in prices with capacities q_A and q_B because the transformation of the input is time consuming.

We saw in §5.2 that Bertrand competition with capacity constraints leads both firms to name the short term competitive market price $P(q_A + q_B)$; hence the second stage is a Cournot competition. Denote $BR(q)$ the quantity best reply of a firm when expecting the other to order q (cf. §5.1 on symmetric Cournot competition). To analyze the first stage, we need to distinguish whether the tariff offered to one downstream firm is observed by the other or not.

If secret contracts are not available (for either legal or technical reasons) then U can get the entire monopoly profit by offering each downstream firm to buy $\frac{q^M}{2}$ for a total price $t^M = \frac{1}{2} q^M p^M$. Both firm accept, get their zero reservation value, the monopoly output is produced and sold. In this world, there is no rationale for foreclosure since the upstream monopolist need not exclude any of the competitors.

The previous offer ceases to be credible if contracts are secret or can be privately renegotiated. Indeed, U could offer A a new contract (q_A, t_A) that maximizes their joint profit, call it $\Pi_A(q_A)$, conditional on B having naively accepted to buy $\frac{q^M}{2}$. To see this observe that

$$\Pi_A(q_A) = \tfrac{1}{2} p^M q^M + q_A \left(P\left(\tfrac{1}{2} q^M + q_A\right) - c \right)$$

is independent of the transfer t_A and is maximum for $\hat{q}_A = BR(\frac{q^M}{2}) > \frac{q^M}{2}$; this optimal deviation is the Cournot best reply that an integrated U-A firm would choose when facing an integrated U-B firm selling $\frac{q^M}{2}$. Anticipating this future flooding of the consumer market, B would refuse the initial offer (which was leaving him zero rent).

A symmetric offer (q^*, t^*) by U will form an equilibrium only if downstream firms do not fear a profit damaging secret recontracting. Since the best secret deal involves the quantity $BR(q^*)$, a firm will be welcomed it whenever it is lesser than the original q^* because this can only increase the final price. The equilibrium constraint is thus $BR(q^*) \leq q^* \Leftrightarrow q^* \geq q^c$, the Cournot individual production. The meaning is clear: it is only by agreeing publicly to sell much that no downstream firm will fear a secret renegotiation aiming at flooding the market. It remains to tailor t^* to leave no rent to the downstream firms.

> If secret contracts are feasible, the upstream firm cannot refrain from double-crossing each of her customers with the other, thereby generating a flooding of the market that ends up lowering her profits to the Cournot level. This is why contracting with a single downstream firm is optimal; foreclosure is then at work since other downstream candidates will face a rebuttal.

10.3.2 Foreclosure strategies

We now understand why foreclosure can improve the monopoly's situation in the presence of secret contracts. By signing an exclusivity contract with one downstream firm the monopoly commits not to sell to the other downstream firms and restores her profits to the monopoly level (no secret contracts case). The available strategies to achieve this objective are:

- U can enter into an exclusive agreement with a single downstream firm.[9@]
- U may integrate with one of the downstream firms. Indeed, by supplying q^m to its downstream subsidiary, the integrated firm earns the monopoly profit and could only loose from supplying another downstream firm.
- U can offer a protection clause to one client to solves the commitment problem. Since the production of a firm is observable and the technology is deterministic, his purchase of input is also observable, hence U may penalize herself with a large side payment to D_i if she were to sell more input to D_j.
- U offers a resale price maintenance (RPM) (minimum price for the downstream market) together with a return option for unsold input units (cf. §9.2.2).

To repeat, even though the upstream firm is in a monopoly position, her inability to credibly commit gives room for opportunistic behavior and prevents her from achieving the monopoly outcome. Foreclosure is a way to restore profits but it clearly lowers welfare since the final production is bound to decrease. If there are more downstream firms, say n, the commitment problem is worse since the no-foreclosure equilibrium price tends to the marginal cost c as n becomes large.[10@] Thus, the more competitive the downstream industry, the more likely that foreclosure by the bottleneck owner takes place.

Also, it is important to note that foreclosure is more likely to occur the higher the bottleneck appears in the production chain. If, contrary to the case presented above, the bottleneck owner is a downstream monopoly (in direct contact with customers) and buys inputs from competitive upstream firms then she can internalize the negative externality between providers and is thus induced to maintain monopoly prices. The reason is that she can at the same time extract all producer surplus from upstream firms and charge the monopoly price to final consumers. In the former case, the connection between final consumers and the monopoly was indirect; any attempt to extract their surplus was doomed to fail because the intermediaries, the downstream firms, were opportunistically trying to grab a part of it.

Seen from a welfare point of view, the better case is when the bottleneck is high in the production chain. Indeed, if foreclosure can be avoided then lower prices and greater production take place. This has lead antitrust authorities to adopt the "common

carrier" policy[11@] of forcing the bottleneck to operate upstream. For example it is better to have a GSM network owner like VODAFONE in Spain establish a telephone company and buy access to fixed lines (local loop) from the national monopoly TELEFONICA than letting the national monopoly market GSM telephony and buy the network service from VODAFONE. The important task is then to make sure that the historical monopoly does not foreclose the use of its fixed lines network. It is socially desirable to ensure that the most competitive segment of the market has access to final consumers.

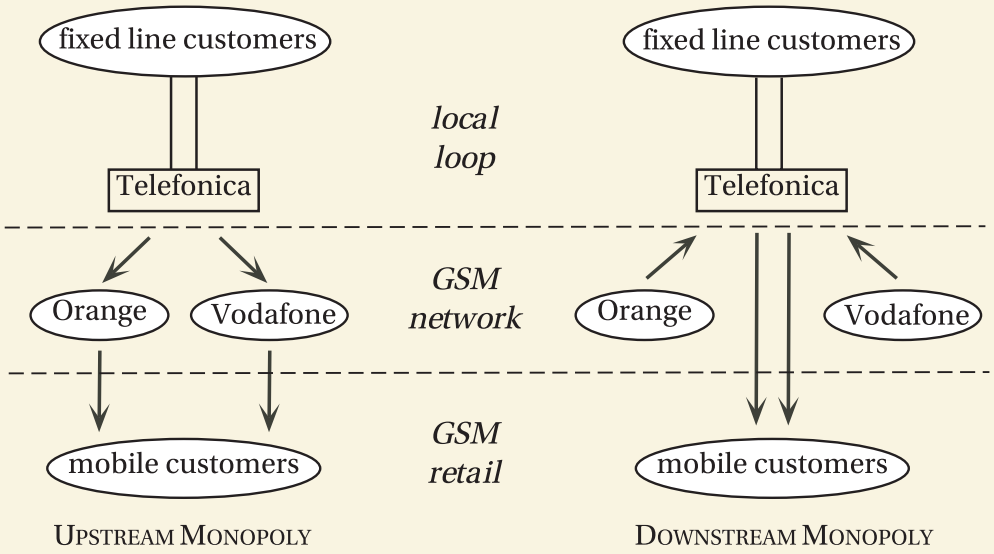


Figure 10.5: Common Carrier Policy

Since in practice foreclosure is often assimilated to the discrimination of downstream firms, it has been proposed to make discrimination illegal. This is a very short sighted response because it gives the monopoly the commitment it most desires. Indeed, if the monopoly is forced to offer the same contract to downstream firms A and B, there is no way for him to make a secret and *different* contract with A because B could easily prove that the input was provided under different conditions. Hence, the upstream monopoly earns the maximum benefit. True, foreclosure has been eliminated but a more inefficient outcome has replaced it !

10.3.3 Cases and Remedies

Some historical cases will help the reader to remind that foreclosure can be found in many places and settings:

- The first case dates from 1912 and involves a key bridge across the Mississippi River in Saint Louis; it was owned by a set of railroads companies who excluded

nonmember competitors from using it. The Supreme Court ruled that this practice of tying was a violation of the Sherman (antitrust) Act (cf. Reiffen and Kleit (1990)). The ECJ followed the same doctrine for the first time in 1978 against the US banana company United Brand for its distribution behavior.

- Associated Press, a newspapers cooperative was condemned in 1945 for blocking membership by competing newspapers.
- In 1984, the US Civil Aeronautics Board forced the major airlines who had created a computer reservation system to give a fair and transparent access to smaller airlines. In 1988, the EC fined the Belgian airline Sabena for a similar access (cf. Fisher (1999)).
- In 1989, the UK Monopoly and Mergers Commission considered that "pubs" were an essential facility to the access the consumer market and forced the majors brewers to divest their chains of pubs (cf. Waterson (2010)).
- In 1991, the ECJ held in the Port of Genoa case that the harbor is an essential facility and that its use should not be reserved to the undertaking managing it.
- For years Intel refused to share technical details of its microprocessor chips with companies that build computers unless those companies agreed to turn over their own technologies in exchange. The US Federal Trade Commission was seeking mandatory non-discriminatory licensing of the data on reasonable terms. A settlement of an antitrust suit was reached in 1999 one day before the start of formal hearings.

Rey and Tirole (2007) identify 6 set of policies used by competition law practitioners to remedy foreclosure:

1. *Structural policies* such as divestitures (e.g., UK brewers analyzed by Slade (2004)), line of business restrictions or common ownership by users (e.g., Associated press) are often considered in last resort, as they may involve substantial transaction costs of disentangling activities. Yet, policy makers may come to the conclusions that it is hard to design proper rules of access for the integrated bottleneck, and that other methods of foreclosure can be prevented under vertical separation.
2. *Access price control*: antitrust authorities link the price of access with some measure of its cost. The ECJ first made such a comparison in the "United Brand" case. As is well known, the measurement of marginal cost is a difficult empirical matter, while the allocation of common costs among product lines has weak theoretical underpinnings. Excluded competitors may be required to prove the overpricing they suffered.

3. *Access quantity control* whereby a significant share of the bottleneck is to be allocated to new entrants as in the "Eurotunnel" case.
4. *Price linkages* is to use prices for retail goods or other accesses as benchmarks for the access price. Examples are the Efficient Component Pricing Rule (ECPR), the same pricing rule for all buyers, a single per-unit price or reciprocity rules.
5. *Common Carrier* policies turn the vertical structure of the industry upside down thereby allowing all actors to access the consumer market (after paying an access price to the bottleneck).
6. *Disclosure requirements* for supply contracts are intended to promote transparency and a greater downstream competition.

The "Common Carrier" policy has been extended into *vertical unbundling* whereby a vertically integrated utility (in a monopoly position) is ordered to divest the network segment (e.g., route, track or canal) which is a natural monopoly and allow competition over the service or superstructure segment (e.g., truck, train or boat).[12@]

10.4 Predation

In this section, we formalize the strategies that firms can use to signal their strength or weakness and how they exercise predation to build a reputation that can be useful in the future. The first model explains how an entrant may signal its willingness to avoid cutthroat competition while the second shows how an incumbent can signal in advance its toughness to avoid later entry. The war of attrition treated in §7.4.2 is an alternative vision of economic struggle where the issue is market exit rather than entry.

10.4.1 Introduction

Definitions

Predatory pricing is a deliberate strategy of setting prices below production costs. This is prohibited by EU competition law because this behavior has no other economic rationale than to eliminate competitors, since it would otherwise be more rational not to produce and sell a product that cannot be priced above average variable cost.[13@] From a theoretical point of view predatory pricing amounts to build a reputation.
Signaling (cf. §21.1.3) by an incumbent is to convey information that discourages unprofitable entry or survival of competitors. The hallmark is credible communication, through the inferences that potential entrants will make from observing costly actions either prior to entry in the case of limit pricing or afterwards in the case of limited entry.

Examples

NutraSweet, a division of US firm Monsanto marketed the low-calorie sweetener Aspartame whose patent for Europe, Canada, and Japan expired in 1987. A price war then occurred when a new company started to make and sell the paten-free Aspartame in Europe. Within two years the price was down by two thirds. The EC later condemned NutraSweet for *predatory pricing* and applied an anti-dumping duty on the imports of NutraSweet in the community.

English firm Laker Airways started as a charter company in the late 1960s and applied in 1973 for a permit to fly regularly between the UK and the US which was granted in 1977. The product called Skytrain in a direct reference to train travel offered flights without reservation and with minimum on-board service at the discount price of one third of the regular fare of big airlines. By 1980, Skytrain's market share was 14% but Laker went bankrupted in 1982 after competitors met his price. The US liquidator then sued several big airlines in US courts for *predatory pricing* arguing their prices were below marginal cost and worse, for pressuring the plane maker not to reschedule the debt payments due by Laker (in order to force immediate bankruptcy). This case created a row between the UK and US government which President Reagan ended by removing the case from the US court. The british judicial case was settled out of court in favor of Laker.

A very frequent case of predation is *cross-subsidization*. Consider firm A which is earning handsome benefits in a protected market with not much future and firm B currently losing money in a very competitive but promising market. If A and B merge, the cash-flow from A's activity today can be used to sustain an aggressive behavior in B's market aiming at evicting competitors so that tomorrow the AB company will enjoy a large market power in B's market and be able to rip large profits.

An example of *Signaling* could be Nestle's intensive advertising for Nescafe; this behavior may be interpreted as a signal for toughness since there are only fringe competitors in most European countries. Likewise, one can interpret as signaling the large amount of advertising by firms with large market shares and a very high goodwill such as Coca-Cola, Vodafone or Nike.

10.4.2 Limited Entry

Gelman and Salop (1983) introduce a neat idea that has proved quite successful in business schools: you can use a larger competitor's strength against him. More precisely limited entry is to build a capacity k small enough to convince the incumbent that it is not worthwhile for him to block entry, through a price war because he, too, would forgo

profits. Let us study the sequential game of entry to ascertain whether this makes sense.

The entrant offers a price p_e and the incumbent respond with p_i. Goods are homogeneous, the demand is $D = 1 - p$ and marginal cost is zero for both firms. The incumbent can undercut the entrant with a price just below p_e, denoted $p_i = p_e^-$, and drive him out of the market for the next period (there is always a fixed cost to enter). However the payoff with this aggressive strategy is $(1 - p_e)p_e$ which may be quite low. Thus, it may be better to choose $p_i > p_e$ and anticipate a residual demand $1 - k - p_i$ since the entrant has a small capacity ($k < 1 - p_e$) and will thus not be able to serve all his demand. The profit in that case is $(1 - k - p_i)p_i$ so that the optimal price is easily computed as $p_i^* = \frac{1-k}{2}$; it yields profits of $\frac{(1-k)^2}{4}$.

The entrant will play on this possibility by choosing k and p_e small enough to convince the incumbent to let him in. The entrant can thus maximize its profit kp_e under the constraints $p_i^* > p_e$ (undercut the incumbent) and $(1 - p_e)p_e \leq \frac{(1-k)^2}{4}$ (leave the incumbent happy). We obtain two conditions on the entrant's capacity that must be satisfied simultaneously i.e.,

$$k \leq \min\left\{1 - 2p_e, 1 - 2\sqrt{(1 - p_e)p_e}\right\} \Leftrightarrow k \leq \min\left\{\tfrac{1}{2}, 1 - 2\sqrt{(1 - p_e)p_e}\right\}$$

Since the entrant's profit is increasing with capacity, he will choose a value that saturates the constraint i.e., $k = 1 - 2\sqrt{(1 - p_e)p_e}$. The profit of the entrant is thus $p_e\left(1 - 2\sqrt{(1 - p_e)p_e}\right)$ and is maximum for $p_e^* = \frac{1}{2} - \frac{1}{2\sqrt{2}} \simeq 0.15$, leading to $k^* = 1 - \frac{1}{\sqrt{2}} \simeq 0.29$; the optimal response of the incumbent is $p_i^* = \frac{1-k^*}{2} = \frac{1}{2\sqrt{2}} \simeq 0.35$ generating sales of 0.36. In percentage of the monopoly profit ($\frac{1}{4}$), the profits achieved under limited entry are $\pi_e \simeq 17\%$ and $\pi_i \simeq 50\%$.

> The entrant builds a small capacity, offers a very low price and earns a third of the incumbent's profits; he uses the strength of the incumbent like a Judo contestant uses the strength of his adversary.

10.4.3 Limit Pricing †

The Bain-Sylos postulate presented in §10.1 lead to the belief by many academics and practitioners that an aggressive pricing policy was predatory because it deliberately barred the entry of a market to a potential entrant. Since this reasoning is based on a hardly satisfied postulate, its truthfulness lies on shaky grounds. In a relatively recent work, Milgrom and Roberts (1982) use the asymmetry of information regarding the cost structure of the incumbent firm to explain why can aggressive pricing be deemed predatory. The basic idea is that a low cost firm can initially price very aggressively to signal its efficient technology to a potential entrant very much like advertising can signal the

high quality of a product to consumers (cf. §11.5.2). Both works are applications of the signaling theory developed in §21.1.3.

In the absence of signaling through limit pricing, the challenger would enter the market because, *on average*, he has a better technology. The incumbent will thus take advantage of an initial period of trade where he has a monopoly situation to signal through its price that he is a strong incumbent in order to convince the entrant not to enter.

To develop this idea assume Cournot competition if the challenger enters. Let c_i equal 0 or $2c_e$ with equal probability. Given the demand $D(p) = 1-p$, a myopic incumbent would produce $\frac{1-c_i}{2}$ in the first period and earns $\Pi^m(c_i) = \frac{(1-c_i)^2}{4}$. If this behavior credibly reveals c_i, then the second period after entry is a game of complete information where firms play a Cournot game whose equilibrium is $q_e = \frac{1-2c_e+c_i}{3}$ and $q_i = \frac{1-2c_i+c_e}{3}$; profits are

$$\Pi_e(c_i) = \frac{(1-2c_e+c_i)(1+c_e+c_i)}{9} \quad \text{and} \quad \Pi_i(c_i) = \frac{(1-2c_i+c_e)(1+c_e+c_i)}{9}.$$

Now, assuming $\Pi_e(0) < F < \Pi_e(2c_e)$ where F is the fixed entry cost, the challenger would enter against a high cost incumbent but not against a low cost one. Since $\Pi_e(c_i) < \Pi^m(c_i)$ for all c_i, no kind of incumbent likes entry, thus the high cost incumbent may try to imitate the low cost one by producing $q_1(0)$ instead of $q_1(2c_e)$. If the challenger believes that $q_1(0)$ signals a low cost incumbent and does not enter then the incumbent earns

$$(1 - q_1(0) - 2c_e)q_1(0) + \Pi^m(2c_e)$$

instead of

$$\Pi^m(2c_e) + \Pi_i(2c_e)$$

if he allows entry. We see that there is a cost to block entry because flooding the market with the large quantity $q_1(0)$ is costly for the high cost incumbent; but there is also a benefit of blocking entry to avoid the profit damaging second period competition. Blocking is therefore optimal if

$$(1 - q_1(0) - 2c_e)q_1(0) > \Pi^m(2c_e)$$

i.e., if $q_1(0)$ is smaller than some threshold $\hat{q}_1$.

Up to now, we have been studying some level 1 of anticipation by the high cost incumbent. Let us pass to level 2: the entrant anticipating this imitation strategy will keep thinking that there is a 50% probability that the incumbent is weak because he believes that both kind of incumbents are producing the same quantity $q_1(0)$ in period one. If the fixed cost F is lesser than $\frac{\Pi_e(0)+\Pi_e(2c_e)}{2}$, then the challenger will enter in the second period.

Passing at level 3 of mental induction, the strong (low cost) incumbent would like to avoid the entry of a challenger that is not able to distinguish the weak from the strong

incumbent. Hence, the strong incumbent must do something that the weak would not imitate: produce so much in the first period (flood the market) that the losses for an imitative weak incumbent would be so large that even being a second period monopoly would not compensate i.e., producing $q_1 > \hat{q}_1$. This is the essence of limit pricing. Precise computations (left as an exercise to the reader) yield

$$\hat{q}_1(c_e) \equiv \frac{1}{2} - c_e + \frac{\sqrt{72c_e^2 - 36c_e + 5}}{6}.$$

This function is convex (U-shape) with a minimum for $c_e \simeq 0,35$. The profit for the low cost incumbent, after having successfully signaled its strength is $(1 - \hat{q}_1(c_e))\hat{q}_1(c_e) + \Pi^m(0)$.

Overall, limit pricing will be profitable if this profit is greater than what he would get letting the challenger enter the market i.e., $\Pi^m(0) + \Pi_i(0)$. The condition is thus $(1 - \hat{q}_1(c_e))\hat{q}_1(c_e) > \Pi_i(0)$ which proves to be equivalent to $c_e < \bar{c}_e \equiv \frac{17+3\sqrt{17}}{68} \simeq 0.43$ (another exercise!).

To conclude, limit pricing can be observed for a large range of entrant cost, even a very small one, a case where the challenger and both types of incumbents have very similar marginal cost.

10.5 Industrial Policy and Regulatory Barriers

Rodrik (2004) recalls that industrial economic activity is ripe with market-failures such as informational asymmetries, externalities in human capital or demand, increasing returns to scale (lumpy investments) and more generally coordination failures. Policies coordinating and stimulating specific economic activities or promoting structural change are thus a legitimate area of government intervention called industrial policy; it is implemented with import-substitution, export facilitation, promotion of foreign investment, free-trade zones and targeted subsidies.[14@]

In the case of industry, there are large spillovers (network externality) because once there is proof by one successful entrepreneur that the deed can be done, countless imitators flock-in and turn the nascent industry into a competitive one where no-one, even the first-comer, earns an exceptionally high return on investment.[15@] Oddly enough, it is in sectors with the highest entry barriers (e.g., government regulation) that innovation springs up since, only there are entrepreneurs better protected (this even-though the initial cost is higher, due to the barriers). This may explain the regulatory barriers upon which we report after.

The first-best policy to foster self-discovery would be to subsidize investments in new,

non-traditional industries but this is largely a theoretical desire, impossible to implement because monitoring is too costly. The situation is somewhat analogous with respect to technological externalities that flow from R&D in which case the first-best is an R&D subsidy but the real (second-best) policy is patent protection. Note also that an optimal subsidy policy requires equating the social marginal cost of investment to the *expected* return of projects in new areas. Hence the successes make up for the failures if, and this is the crucial point, these are phased out in time. A good industrial policy is thus carrot-and-stick; it needs to encourage investments in non-traditional areas, but also weed out projects and investments that fail. A government that makes no mistakes is simply not trying hard enough (think of new drug designs).

Regarding the coordination of investment activities displaying scale economies (so-called big push), the problem is social learning i.e., discovering where the information and coordination externalities lie and therefore what the objectives of industrial policy ought to be and how it is to be targeted. The implementation then require bureaucrats to interact deeply with industrialists at the risk of ending up in bed with them. There is thus a tension between closeness and independence.

The State, motivated by security and quality concerns (cf. §16.2 & §17.1) and also in response to pressure from industries (cf. §16.2.2 on special interest groups) tends to limit entry in many activities. As shown by Conway et al. (2005), there still exists a number of regulatory impediments to product market competition in areas where technology and market conditions make competition viable.

- *State control* (SC) is conducive of inefficiency in the absence of market failure because competition is likely to be distorted when one player, the government is both judge and party.
- *Barriers to entrepreneurship* (BE) are all these administrative opaque procedures and burdens necessary to create a firm on top of legal barriers to competition.
- *Barriers to commerce* (BC) are the tariffs or ownership limits and discriminations applied onto foreigners.

Table 10.1 shows these 3 indexes as well as their average for a variety of OECD countries in 2003. Overall, regulatory impediments to product market competition have declined in the OECD area in recent years (cf. last column).

Country	SC	BE	BC	mean	Δ 98-03
Poland	36	22	24	*27*	−12
Italy	31	14	11	*18*	−9
France	26	15	9	*17*	−7
Spain	27	15	6	*16*	−7
Germany	22	15	6	*14*	−5
Sweden	19	10	7	*12*	−5
Netherlands	19	16	6	*13*	−4
United Kingdom	17	7	3	*9*	−2
Turkey	28	24	16	*22*	−9
Mexico	19	22	23	*21*	−2
Korea	16	16	12	*15*	−9
Japan	15	14	9	*12*	−6
Canada	16	7	11	*11*	−7
United States	11	12	7	*10*	−2
Australia	5	11	8	*8*	−4

Table 10.1: Product Market Regulation in the OECD

Part E

Differentiation and Innovation

Chapter 11

Differentiation and Competition

In the early XXth century, the discrepancies between empirical observations and the predictions of the Marshallian theory of perfect competition were ascribed to exogenous frictions until Sraffa (1926) came to argue instead that these frictions were endogenously created by firms. Translated in today's economic language, he stated that firms try to differentiate themselves from rivals to break the homogeneity of goods that produces the deleterious effect of perfect competition. In his own words: *The causes of the preference shown by any group of buyers for a particular firm are of the most diverse nature, and may range from long custom, personal acquaintance, confidence in the quality of the product, proximity, knowledge of particular requirements and the possibility of obtaining credit, to the reputation of a trade-mark, or sign, or a name with high traditions, ... have for their principal purpose that of distinguishing it from the products of other firms.*

Inspired by these insights, the founders of the theory of imperfect competition Hotelling (1929) and Robinson (1933) recognize that the own price elasticity of demand is negative but not infinite as would predict perfect competition. Thus, when a firm increases her price, she starts losing customers, but neither none nor all of them. Secondly, the perfect competition framework is mute about the destination of the customers lost by the firm; will they refrain from consuming or buy from a rival? The models of imperfect competition we shall develop hereafter seek answers to these questions.

11.1 Horizontal Differentiation

A classification of differentiated products will be of great help to address the questions asked in the introduction. When some consumers prefer a brand to another equally priced while others hold the reverse preference, products are said to be *horizontally* differentiated (e.g., two identical cars except for color). If, on the contrary, everybody agrees on which is the best of two equally priced products, then they are said to be *vertically* differentiated (e.g., a TV for sale in two shops offering warranties of different duration).

According to the type of differentiation, the consequences on competition differ greatly. We treat in turn horizontal and vertical differentiation.

In this section, we first take a look at the general principles of horizontal differentiation before treating in detail the selection of locations in Hotelling (1929)'s classical model whose study was started in §5.2.2. Then we study how strongly firms are willing to differentiation i.e., obtain a cost advantage or a demand one.

11.1.1 Differentiation Principles

As will be illustrated afterwards, the literature has tended to argue in favor of the *maximum differentiation principle*, according to which firms maximally differentiate in order to relax price competition. It should be noted that the result does not depend on the uniform distribution of consumers along the market segment. In the more realistic case where some areas are densely populated, firms will try to locate themselves at the epicenter of these clusters, as local monopolies. Train or tube stations provide an illustration; they are indeed surrounded by numerous shops, but in general one of each kind, as theory suggests.

In §11.2.1, we extend Hotelling's model to capture more issues and explain why his intuition was almost right in the following sense:

> Firms differentiate their products and services along their most importance characteristic but otherwise choose the same attributes.

This is entirely consistent with the casual observation that within the shopping area of a large city, each street seems to be exclusively devoted to one particular good i.e., there is minimum differentiation. Yet, location is a secondary characteristic when compared to brand or design or style that define the product and shopping experience offered to clients. In that primary dimension, firms seek maximal differentiation.

To understand intuitively how such an equilibrium occurs let us take a dynamic point of view. The reason a particular street ends up specialized in clothes, restaurants or theaters has much to do with the activity of the first shop who had an important commercial success in the street. Since the historical shop drains many consumers, new firms settle nearby to attract the attention of these potential customers and steal a bit of the incumbent's business. Then, the word of mouth spreads the information that this particular street hosts several similar shops and this increases the number of shoppers because consumers are certain to find what there are looking for thanks to this extended supply (their opportunity cost of passing by is reduced). The virtuous circle is engaged since this additional demand will motivate more firms to enter driven by their desire to grab

a share of the cake. At some point, there is more supply than demand and to avoid a destructive Bertrand competition, the newly entered firms have to differentiate themselves through other dimensions like quality or specialization into a segment of consumers, a food style, a serving style, etc... One case where this clustering of similar shops is almost impossible (or leads to frequent failures) is pure distribution. There are no two adjacent supermarkets in the world because they all sell the same brands and can only compete on prices "à la Bertrand".[1@] We can conclude that

> Minimum differentiation e.g., geographical agglomeration, is driven by the expansion of demand more than by the desire of firms to steal the business of their competitors.

Another instance where location is crucial is when prices are regulated (often to protect small businesses from undercutting strategies by large firms). Competition then takes place only on the location in order to attract the maximal number of potential customers. In the case of two firms, they both end up in the middle of the market. The same phenomenon occurs with free Hertzian TV where channels offer almost identical program all day long. Likewise, political parties who cannot compete on price terms make electoral promises leaning toward the center in an intent to grab the other side's traditional clientele (at least in bi-polar countries).

11.1.2 Competition for Location

In §5.2.2, we started to study Hotelling (1929)'s classical model of horizontal differentiation based on geography. Two shops selling the same commodity are located at the extremes of a street. We showed their ability to sustain prices higher than their marginal cost if consumers have to bear a transportation cost to go from their home to the shops.

We now tackle the intensity of differentiation i.e., the issue of location: is it better to move away from one's competitor to relax competition (and charge higher prices) or come nearby to capture his clientele? In this process, each firm anticipates the effect her positioning will have on the setting of prices afterwards i.e., our equilibrium concept is the subgame perfect equilibrium (cf. §2.4.2).

Price competition

Given the prices quoted in equilibrium, the demands addressed to the firms are given by the plain lines of Figure 11.1. Hence, by reducing the address b to b', firm B would gain market shares (the utility of B's customers is now given by the dashed line) since $\tilde{x}$ shifts leftward to $\tilde{x}'$. This observation lead Hotelling (1929) to formulate the *principle*

of minimum differentiation according to which firms competing in prices and product attributes tend to supply identical products (choose the same attributes).[2@]

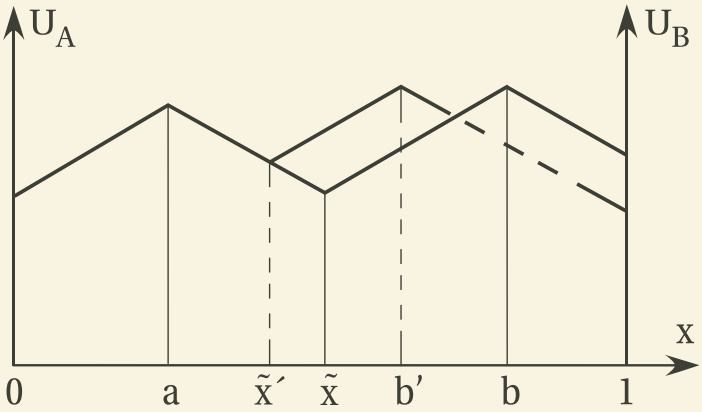


Figure 11.1: The Hotelling Model of Location

We modify the model of price competition developed in §5.2.2 by adding an initial stage where firms choose their locations a and b inside the street (with the convention that $a \leq b$). We also use a *quadratic* transportation cost to facilitate the analysis. Given firms locations a and b (cf. example on Figure 11.1), the utility of a consumer located at x from buying is

$$u_A(x) = \bar{p} - p_A - t\,(x-a)^2 \text{ and } u_B(x) = \bar{p} - p_B - t\,(x-b)^2 \tag{11.1}$$

hence the indifferent consumer is located at the address solving $u_A = u_B$ i.e.,

$$\tilde{x} \equiv \frac{p_B - p_A}{2t(b-a)} + \frac{a+b}{2} \tag{11.2}$$

Since $\pi_A = (p_A - c)\tilde{x}$, we deduce firm A's best reply as

$$p_A = \tfrac{1}{2}(c + p_B) + t(b-a)\tfrac{a+b}{2} \tag{11.3}$$

and from $\pi_B = (p_B - c)(1-\tilde{x})$, firm B's best reply as

$$p_B = \tfrac{1}{2}(c + p_A) + t(b-a)\left(1 - \tfrac{a+b}{2}\right) \tag{11.4}$$

The equilibrium of the pricing game solving system (11.2-11.3) is[3@]

$$\begin{cases} p_A^* = c + \frac{t}{3}(2 + a + b)(b-a) \\ p_B^* = c + \frac{t}{3}(4 - a - b)(b-a) \end{cases} \tag{11.5}$$

Differentiation

We can now study the incentives for firms to differentiate horizontally. The first stage profits of firm A is now a sole function of locations $\Pi_A(a,b) \equiv (p_A^* - c)\tilde{x}(p_A^*, p_B^*)$, thus

$$\frac{d\Pi_A}{da} = \frac{\partial \Pi_A}{\partial p_A}\frac{\partial p_A^*}{\partial a} + (p_A^* - c)\left(\frac{\partial \tilde{x}}{\partial a} + \frac{\partial \tilde{x}}{\partial p_B}\frac{\partial p_B^*}{\partial a}\right) \tag{11.6}$$

Since p_A^* is a best-reply, $\frac{\partial \Pi_A}{\partial p_A} = 0$ (this is the envelope theorem), so that,

$$\frac{d\Pi_A}{da} \propto \frac{\partial \tilde{x}}{\partial a} + \frac{\partial \tilde{x}}{\partial p_B}\frac{\partial p_B^*}{\partial a} \tag{11.7}$$

As we already saw in §6 on strategic moves, the overall effect of differentiation is the sum of a direct demand effect (first term in (11.7)) and an indirect strategic effect (second term) through the change in pricing policy of the opponent. A positive (resp. negative) value in (11.7) would motivate firm 1 to reduce (resp. increase) differentiation. The demand effect is positive since (11.2) yields $\frac{\partial \tilde{x}}{\partial a} = \frac{1}{2} + \frac{p_B - p_A}{2t(b-a)^2} > 0$; thus the firm ought to move to the city center (minimize differentiation) in order to gain market shares. Yet, such a move is met by an increased toughness of firm B in the price competition as $\frac{\partial p_B^*}{\partial a} = \frac{2t(b-2)}{3} < 0$ which in turn lowers profits since $\frac{\partial \tilde{x}}{\partial p_B} = \frac{1}{2t(b-a)} > 0$.

We have here a typical phenomena of economic reasoning where simple intuitive arguments lead us to opposite conclusions. The only way to disentangle the issue is to specify a simple enough model that allow each effect to be computed in order to determine which one dominates the other. In the Hotelling model, we can plug the equilibrium prices given by system (11.5) into the profit functions π_A and π_B to derive the reduced form profits as

$$\begin{aligned} \Pi_A(a,b) &= \tfrac{t}{18}(2+a+b)^2(b-a) \\ \Pi_B(a,b) &= \tfrac{t}{18}(4-a-b)^2(b-a) \end{aligned} \tag{11.8}$$

It is a simple exercise to check that $\frac{d\Pi_A}{da} \propto (2+a+b)(a+b-2) < 0$ and likewise $\frac{d\Pi_B}{db} > 0$ i.e., the strategic effect dominates the demand one. These inequalities mean that, meanwhile $a > 0$, firm A prefers to move its location leftward and symmetrically, meanwhile $b < 1$, firm B prefers to move its location rightward.

> The equilibrium locations choices are simply the boundaries of the city so that the "principle of maximum differentiation" holds.

If new firms can enter sequentially and incumbents can adjust their locations then in equilibrium firms are always located at the same distance one from another. If on top, a firm can open several shops, it will never sit a new shop next to an old one because this

would cannibalize existing customers. On the contrary, a firm sits new shops between his competitors'.

Multi-unit Purchase

Up to now, consumers were only willing to buy one unit of the commodity which fits well the case of durable goods. To treat classical retail competition where people buy many units of many different goods, it is enough to assume that consumers buy varying amounts of a composite good (e.g., that used to construct the CPI). The choice among two possible shopping locations is thus driven by the comparison of final utility levels. In equation (11.1), it is enough to replace the surplus $\bar{p} - p_i$ by the utility level $u(p_i)$ derived from buying an optimal quantity at shop i given the (composite good) price p_i in force there. This way, market shares continue to be determined by the difference in transportation cost of the indifferent consumer $\tilde{x}$. Firms' profits are then the product of sales by the unitary profit margin which is the same for all consumers since we are implicitly assuming away wealth effect (the transportation cost does not influence the amount spend at the shop).

11.1.3 Oligopoly & the Circular City

Although appealing, the Hotelling linear city model of competition does not readily extend beyond the duopoly. Salop (1979) introduces the unit length circular city where firms are located equidistantly on a turnpike.[4@] As an example, observe that many western cities are surrounded by a highway ring (turnpike) next to which sit malls and supermarkets at more or less the same distance. If consumers are uniformly distributed along the ring (and cannot travel inside the disc to reach another point of the ring) then a monopolist will locate anywhere on the circle. In real conditions, such an indifference does not occur because there is always an accumulation point (or road knot) that is nearer to a majority of consumers. Locating at this point reduces overall transportation costs to a minimum which enables the monopoly to charge a maximum price for its services. Under the maximum differentiation principle, an entrant would choose the opposite point of the incumbent monopoly. More generally, upon entry or exit, firms readjust their location (characteristic) to remain equidistant one from another.

When n shops are active on the market, they locate at distance $\delta = \frac{1}{n}$ from each other on the ring.[5@] The westward shop A has a northern neighbor N and southern neighbor S; there are thus a northern and southern indifferent consumer, located at distance $\tilde{x}_N = \frac{\delta t + p_N - p_A}{2\delta t}$ and $\tilde{x}_S = \frac{\delta t + p_S - p_A}{2\delta t}$ from A (this is equation (5.15) in Hotelling's model where the transportation cost per km t is multiplied by the distance δ).

If people are uniformly distributed across the ring, a fraction $\frac{1}{n}$ live between A and each of his two adjacent neighbors; he thus grabs $\frac{1}{n}\tilde{x}_S$ from the south side and $\frac{1}{n}\tilde{x}_N$ from the north side i.e,

$$D_A = \frac{1}{n}(\tilde{x}_S + \tilde{x}_N) = \frac{\delta t + \bar{p}_B - p_A}{n\delta t} = \frac{1}{n} + \frac{1}{t}(\bar{p}_B - p_A) \tag{11.9}$$

where $\bar{p}_B \equiv \frac{p_N + p_S}{2}$ is the mean price of A's direct neighbors. Firm A's demand is thus the fair share of the market, $\frac{1}{n}$, and a bonus proportional to the difference between his price and his direct opponent's average (cf. eq. 5.15).

Assuming that all firms share the same zero cost technology, individual profit is $\pi_A = p_A D_A \propto p_A\left(\frac{t}{n} + \bar{p}_B - p_A\right)$, the best reply is easily characterized as

$$p_A = \frac{1}{2}\left(\frac{t}{n} + \bar{p}_B\right) \tag{11.10}$$

and the symmetric equilibrium is $p^* = \frac{t}{n}$. Market size m is measured by the population density which was taken to be unitary. It is easy to check from the formulas that accounting for it amounts to replace the transportation cost t by $\frac{t}{m}$. We can thus confirm the intuition according to which

> More competitors increases competition in the sense that the unit margin shrinks. A lower density of population or a greater transportation cost enable firms to exercise greater market power.

Tackling the oligopoly case also raises the question of entry; in §6.1.4, we show that open access to the market leads to excessive entry which would be reduced if the market was cartelized.

11.1.4 Urban Economics

Land Rent Distribution

If Hotelling (1929) can be deemed the father of urban economics,[6@] geographer von Thünen (1826) deserves the grandfather title by introducing the spatial dimension to economics, in connection with the classical theory of land rent. This author considers a self sufficient "Isolated State", free from external influences, with one centrally located city. Land is flat while the soil quality and climate are homogeneous. Farmers transport their own goods to the city market across land and act to maximize profits.

In equilibrium, we should observe four rings of agricultural activity surrounding the city. Dairying and intensive farming occur in the ring closest to the city because vegetables, fruit and milk must get to market quickly (otherwise they spoil). Timber for building and firewood are produced in the second ring because it is very heavy and costly to

transport. The third ring consists of extensive fields crops because they last longer than dairy products and are much lighter to transport than wood, thus they can be located further away from the city. Ranching (cattle) is located furthest away because animals are self-transporting. Beyond the fourth ring lies the unoccupied wilderness, which is too great a distance from the central city for any type of agricultural product to be economically viable. As one gets closer to the city market, the price of land increases and so does the productivity of the farming products. This results occurs because farmers balance the cost of transportation and the cost of land to decide which activity to perform at any location. The most cost-effective product for the city market outbids other options to rent the piece of land.

To understand this result we can imagine that, in line with historical property distribution, the land is owned by landlords; they allocate land to the farmers who bid the highest rent at each distance x from the center. For each of the farming categories $i = 1$ to 4, above let p_i be the equilibrium price in the city market,[7@] let f_i be the marginal productivity of the activity[8@] per acre and t_i the transportation cost per unit of distance of one unit of product i. A farmer growing type i at distance x from the city will earn $\pi_i(x) = f_i(p_i - t_i x)$ per acre. If we let $\underline{\pi}$ denote the subsistence level of a farmer, the maximum bid anyone can make for cultivating an acre of land at x with type-i crop is $\pi_i(x) - \underline{\pi}$. Assuming, in accordance with historical reality, that there are many more farmers than landlords, a Bertrand competition occurs among farmers to get the land; hence, in equilibrium, they bid their willingness to pay which is $\pi_i(x) - \underline{\pi}$.[9@] As seen on Figure 11.2, the upper envelope of the bid-curves determine the equilibrium crops at each distance x from the city.

Dairy products appear first because they have a very high transportation cost (steep curve) since they are perishable; although their unit price is not so high they have a very high productivity per acre since breeding requires fewer land than anything else. The next curve, that of timber, is characterized by high transportation cost and good productivity (trees grow towards the sky). Outer rings tend to be larger and larger as they correspond to activities requiring a lot of space. Cultivation ends at some distance x_4 when the lowest economic value of land equals the subsistence level of a farmer R. Observe with the dashed bidding curve that crop 5 is not cultivated because the price of its output is too low to make it worthwhile anywhere.[10@]

City Housing Distribution

The previous framework is easily extended to modern cities to analyze the distribution of income and square space , as shown by Brueckner (1987).

Households working at the city business district (CBD) in the center, consume hous-

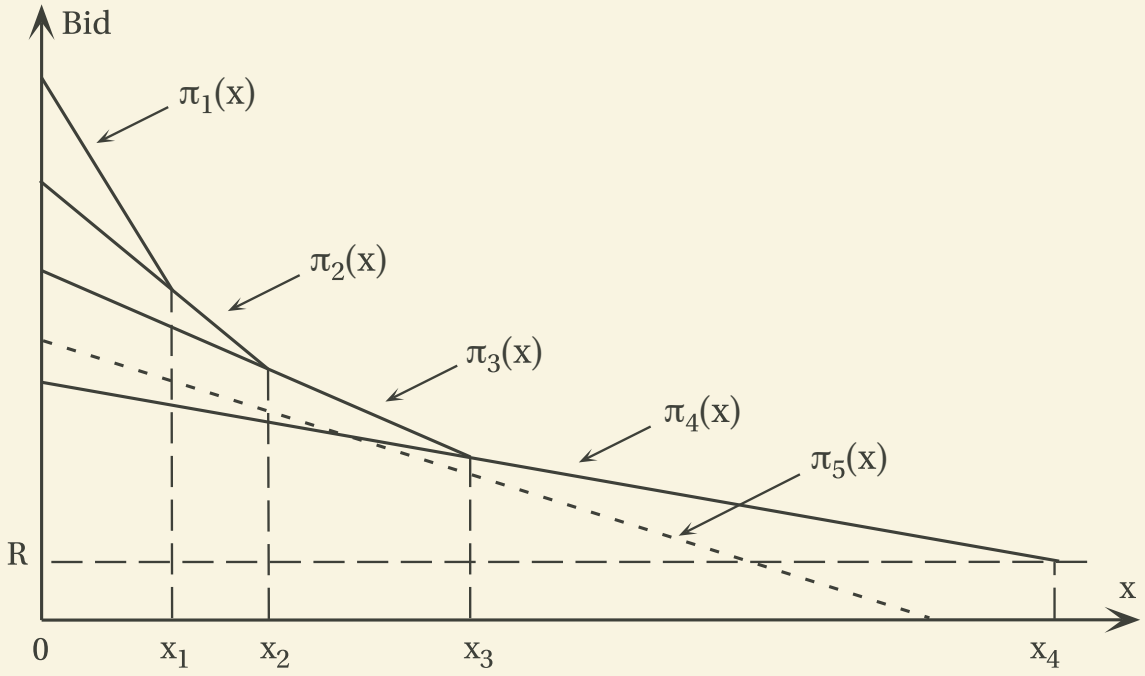


Figure 11.2: The von Thünen Spatial Equilibrium

ing q and a composite good z; given the income w and the price p_x of land at x, the budget constraint is $z + p_x q \leq w - tx$ where t is the transportation cost measuring the opportunity cost of time lost in transportation. Individual demand is a function of distance x and housing price p. The large number of households is then distributed across locations so as to equate their final utility level.[11@] This additional equation allows to work out housing price p_x and consumption q_x parametrized by distance to the CBD. Looking back at the budget constraint, we have $p_x = \frac{w-tx-z}{q_x}$ which can also be interpreted as the WTP or bid of the household to live at distance x from the center.

To learn about the shape of this bidding curve, suppose a person moves away from the CBD; she has now higher transportation cost, thus less disposable income and demands less of everything including housing, should prices be kept constant. She would be worse than before and try to move back. To retain her at this new location, her landlord has to reduce the price p_x.[12@] Thus the bidding curve is decreasing as in the farming story. Also, dwellings away from the CBD are larger since housing is cheaper wrt. other goods, so that people consume more of it.[13@]

If we further assume that housing is a normal good, its demand increases with income which means that poorer households have steeper bidding curve.[14@] Re-interpreting Figure 11.2, crop labels $i = 1,..,5$ become wealth classes, so that in equilibrium, the poorer households outbid the richer ones to live nearby the center, although the surface they buy is smaller. Rich households occupy greater surface but at greater distance from the

CBD.

This model fits quite well with US reality; to account adequately for the rather inverse location structure of European Cities, one has to account the higher opportunity cost of time of richer households who value the accessibility to the central business district (CBD) for work as well as for cultural amenities i.e., t_x increases with income x. As shown by Brueckner et al. (1999), if the ratio of opportunity cost over housing $\frac{t}{d_x}$ rises with income (amenities really matter), then the rich tend to live at central locations. A classic treatment of urban economics is Fujita and Thisse (2002).

11.2 Location and Variety

In this section, we show how to relate product location and product variety as instances of horizontal differentiation. In the last part we recall the agricultural origin of the whole process of geographical differentiation.

11.2.1 Variety and Opportunity Cost

Lancaster (1966) formulates an innovative interpretation of horizontal differentiation without relation to the original geographical considerations of Hotelling: firms develop brands or labels for their products to distinguish them from other brands sold in the same market segment i.e., to reduce the extent of competition. In other words, firms engage in *interbrand* competition like Danone and Nestlé do over yogurts using advertising and innovations on taste and recipe.

Formally, firms differentiate their products and services by selecting differing attributes or characteristics because they appeal differently to the *subjective* tastes of consumers. In Lancaster (1966)'s approach presented on Figure 11.3, each consumer is characterized by an ideal (subjective) combination of attributes; he thus bears an opportunity cost from being forced to buy one of the few brands available on the market (e.g., limited choice of colors, shapes or motors for cars). The result of this form of competition is market segmentation, a segment being the set of consumers who are likely to buy a particular variant. On the cost side, it must be noticed that economies of scope such as having a single marketing department for all varieties of a given product line, are crucial to explain the apparition of varieties.

One could draw the conclusion that if variants are cheap to design, each firm would introduce many of them to occupy all market niches (being nearby clusters of consumers on Figure 11.3). But as variants appear there is less and less differentiation in the whole market since every consumer is now nearby to several available models; thus

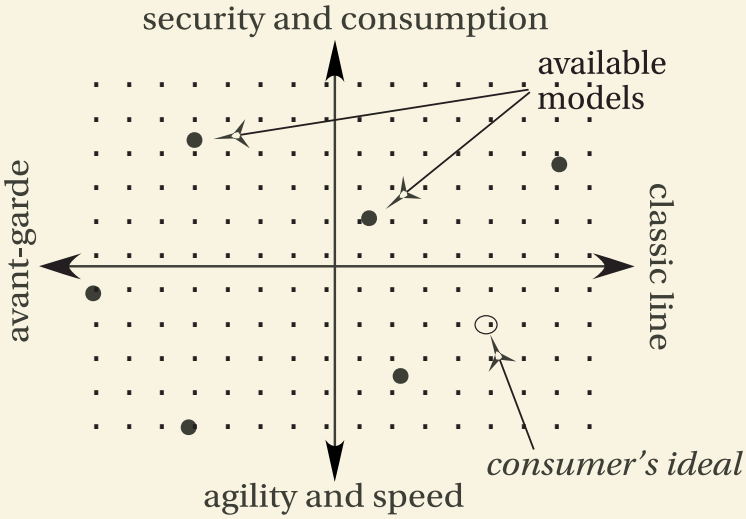


Figure 11.3: The Space of Products Characteristics

there is a tougher competition, not only among firms but also among models of the same company.[15@] One could also wonder whether firms will try to locate their models away or nearby their competitors. A concrete example taken from Fraiman et al. (2003) is presented on Figure 11.4 for clothing were the brands of the Spanish Inditex company are in italics while those of its main contender Cortefiel are underlined (other brands appear in regular type). A total of eight characteristics are plotted over the two panels.

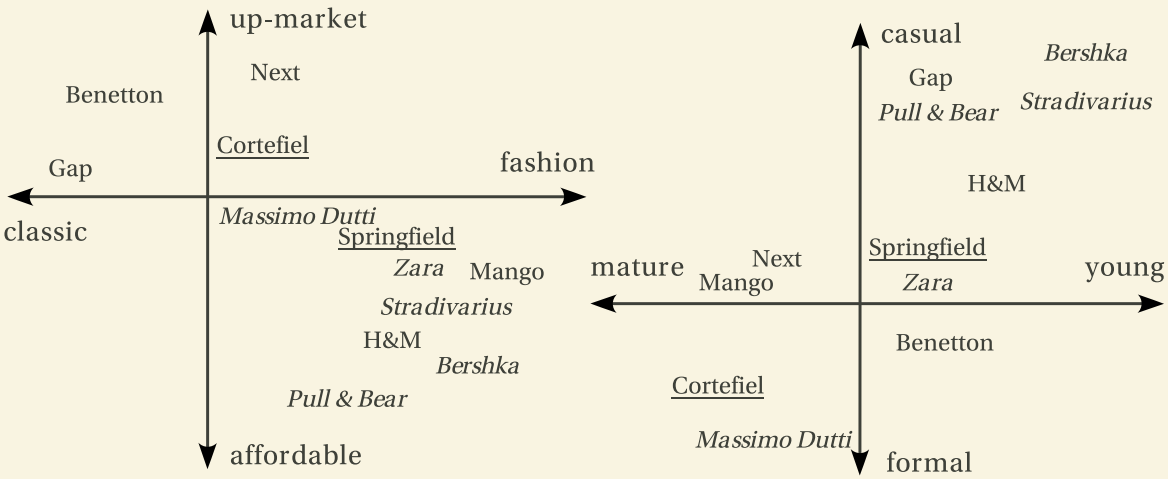


Figure 11.4: Products Positioning for Apparel

Another example of development of products along two important characteristics is given on Figure 11.5 for the aircraft industry. We also notice from historical events that competitors, especially Airbus, tend to introduce close substitutes to existing models except for "jumbo jets" where the Boeing 747 stayed 35 years unchallenged before the

launch of the Airbus A380 in 2004 (cf. Esty and Ghemawat (2002)).

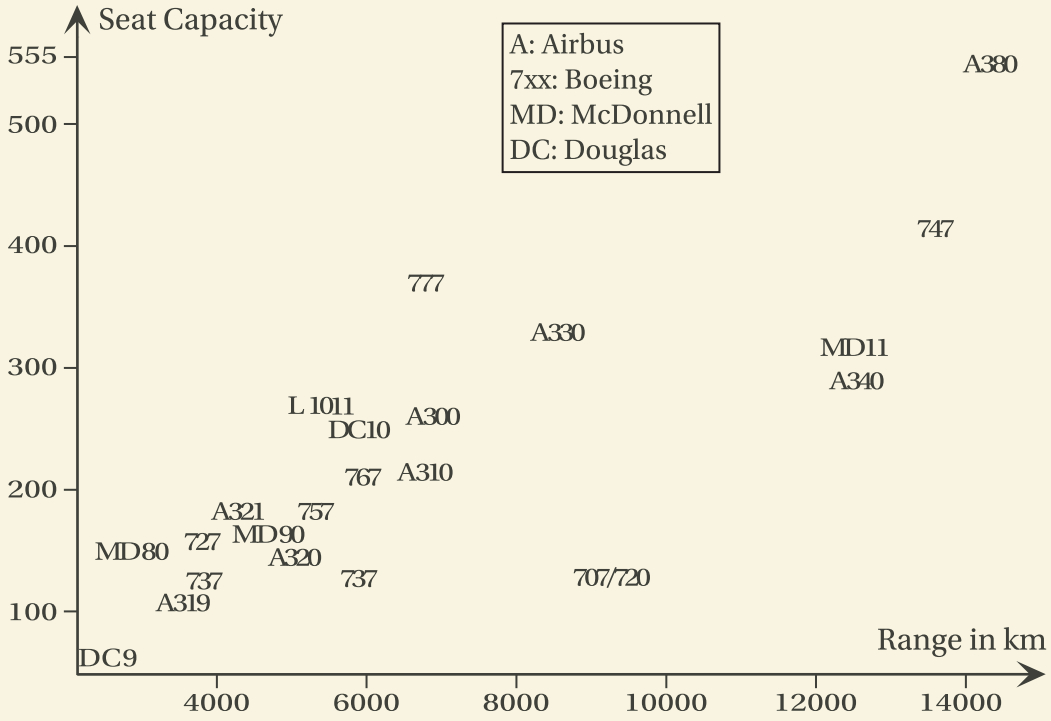


Figure 11.5: Aircraft Characteristics along Size and Range

11.2.2 Multi-dimentional Differentiation

To be able to answer the question of where is it best to locate in the characteristics space, we need a simplification. As we said above, the distance between one's ideal combination of characteristics and one of the available model is an opportunity cost. Suppose that the *design* of a car can be ranked according to technical criteria from the definite classic (0) to the latest avant-garde (1). Similarly there is a $[0;1]$ scale for *performance* axis ranging from thrill (0) to security (1). Each consumer values at price t_i the distance between the two extremes for each characteristic i. The overall opportunity cost from buying the pair of characteristics $z = (z_1, z_2)$ instead of his ideal bundle x is then the linear combination[16@] $t_1(x_1 - z_1)^2 + t_2(x_2 - z_2)^2$. To ease our study of multi-dimentional differentiation, the term $t_i(x_i - z_i)^2$ is replaced by $(x_i^{'} - z_i^{'})^2$ with $x_i^{'} \in [0; \sqrt{t_i}]$. Since attribute #1 is the strong one, the space of characteristics is a stretched rectangle as display on Figure 11.6.

The first result is a reduction of complexity. From the point of view of firms, several attributes uniformly distributed are akin to one synthetic characteristic with a bell-shaped distribution. Consider the left panel of Figure 11.6 where firms' locations are shown by black dots. We can draw a new axis associated with a synthetic attribute z whereby firms

are located at z_A and z_B; we let $\bar{z} \equiv \frac{z_A+z_B}{2}$ denote the locations' mid-point. The uniform consumer distribution over the rectangle gives rise to a bell-shaped density function, it is given at every z by the height of the segment orthogonal to the axis joining the top of the box to the bottom (dashed segment). We thus start with $f(0) = 0$, increase until z_0, remain constant up to z_1 and decrease until $f(1) = 0$. The associated distribution function is thus increasing and S-shaped.

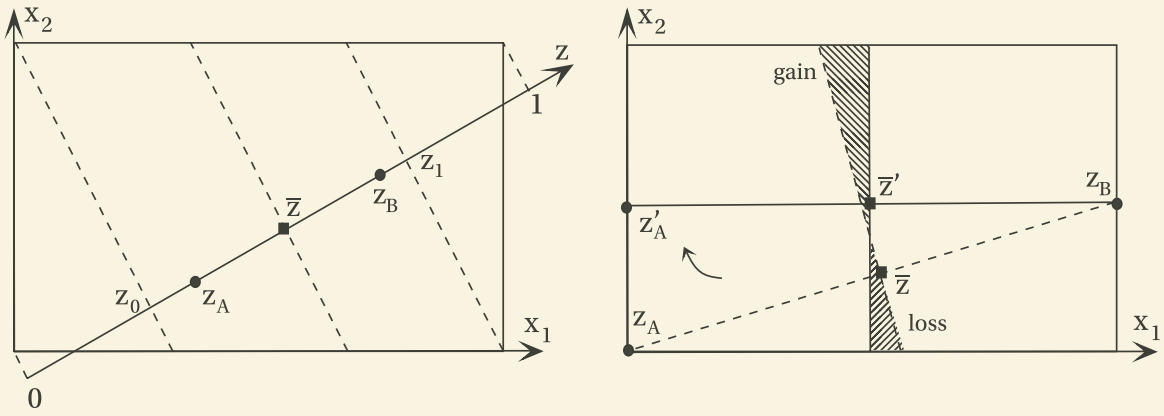


Figure 11.6: Two dimensional attribute space

Let us now study the price equilibrium and the first stage of differentiation. With our quadratic opportunity cost, an indifferent customer is such that $p_B - p_A = \sum_{i\leq 2}(a_i + b_i - 2x_i)(b_i - a_i)$. The location of these individuals is a segment orthogonal to the z axis. It passes through $\bar{z}$ when prices are equal and moves in parallel fashion for unequal prices. We see on Figure 11.6 that if the differentiation choices satisfy $\frac{b_2-a_2}{b_1-a_1} < \frac{t_1}{t_2}$ then the frontier between the two market shares hits the top and bottom of the rectangle; this is the likely situation since attribute #1 is the strong one. As shown on the right panel, a reduction of differentiation by firm A regarding the weak attribute (at constant prices) moves the indifference line (independently of prices) in a manner that makes her more competitive towards a large chunk of clients and less competitive towards a smaller bunch of clients. This result means that firms will tend to imitation over the weak attribute and differentiate in the dominant attribute to relax price competition.

More precisely, Irmen and Thisse (1998) show that demand is $D_A = \frac{p_B-p_A+\gamma}{2t_1(b_1-a_1)}$ as in (11.2) with $\gamma \equiv t_1(b_1-a_1)(b_1+a_1)+t_2(b_2-a_2)(b_2+a_2-1)$. The price equilibrium is as in (11.5): $p_A^* = \frac{2t_1(b_1-a_1)+\gamma}{3} = \frac{t_1\Psi_1+t_2\Psi_2}{3}$ where $\Psi_1 \equiv (b_1-a_1)(2+b_1+a_1)$ and $\Psi_2 \equiv (b_2-a_2)(b_2+a_2-1)$. As in (11.8), profit is $\pi_A^* = \frac{(t_1\Psi_1+t_2\Psi_2)^2}{18t_1(b_1-a_1)}$, thus the optimal differentiation for the weak attribute a_2 maximizes Ψ_2 i.e., $a_2^* = \frac{1}{2}$. We thus see that differentiation over the weak attribute amounts to reduce the sustainable price ($\Psi_2 < 0$) because consumers are more difficult to capture which makes competition more intense. Since firm B will do likewise and choose

$b_2^* = \frac{1}{2}$, we have $\Psi_2 = 0$ so that the profit formula reduces exactly to (11.5) for which we already know that a_1 should be minimized.

11.2.3 Monopolistic Competition

Chamberlin (1933) outlines a theory of indirect oligopoly competition where a firm's demand is more elastic than the entire market demand. The idea is that brands are differentiated enough for own-price elasticity to be finite, unlike in the perfect competition paradigm. Decades later Spence (1976b) and Dixit and Stiglitz (1977) formalize the idea with a simple yet powerful specification of preferences. Instead of relying on complex Hotelling models where heterogeneous consumers pick a brand among many, homogeneous consumers display a taste for variety and thus consume a bit of all brands. The main thrust of this approach is to yield for each firm a demand with constant price elasticity.

Competition

Given a presumably large number of versions of a product (indexed by j), consumers care for the composite good $y = \left(\sum_j x_j^\theta\right)^{1/\theta}$ where $0 < \theta < 1$ is a substitutability parameter.[17@] The representative agent maximizes $u(y, x_0)$ under the budget constraint $x_0 + \sum_j p_j x_j \leq I + \sum_j \pi_j$ where the numéraire x_0 stands for an aggregate of all other goods, I is non capital income and π_j is the profit of firm j. It is fairly easy to prove that the consumer solves his optimization problem in two steps. He first arbitrates between x_0 and y taking the "composite" price to be $\rho \equiv \left(\sum_j p_j^{-1/\beta}\right)^{-\beta}$ with $\beta \equiv \frac{1-\theta}{\theta}$. This yields a "composite" demand $D(\rho)$. Then, he optimizes among available versions with $x_j = D(\rho)\left(\frac{\rho}{p_j}\right)^{\frac{1}{1-\theta}}$. If there is a large number of available versions, then the composite price ρ is unaffected by p_j and this demand displays constant elasticity $\frac{-1}{1-\theta}$. It is also the cross elasticity of substitution between any two versions since $\frac{x_j}{x_i} = \left(\frac{p_i}{p_j}\right)^{\frac{1}{1-\theta}}$.

Each version of the product is produced with the same technology involving a fixed cost F and constant marginal cost c. A key implicit assumption is that when two (or more) firms produce the same version of the product, they engage into Bertrand competition because the resulting product is homogeneous for consumers. Marginal cost pricing then ensues and since fixed cost are present, only one firm can survive (cf. §5.2.1); this enables us to identify from now on a version with a brand. The constant price elasticity of d_j leads to the optimal price $p_j = p^* \equiv (1+\beta)c$ (cf. Lerner rule (3.4)). Symmetry then yields $\rho^* = p^* n^{-\beta}$. Firms enter the market while it is profitable i.e., $0 \leq \pi = \beta c q - F \Rightarrow q \geq \hat{q} \equiv \frac{F}{\beta c}$. Given that the overall expense on the good is $\sum_j q_j p_j = \rho D(\rho)$ and that sales are identical among

brands, we have $q^* = \frac{\rho D(\rho)}{np^*}$, thus the limiting number of firms n^* is the integer part of the solution to $\hat{q} = q^* \Leftrightarrow Fn^{1+\beta} = \beta cD\left((1+\beta)cn^{-\beta}\right)$. If the utility function is Cobb-Douglas with $u(y,x) = y^\alpha x^{1-\alpha}$, then $D(\rho) = \frac{\alpha I}{\rho}$ at the zero profit equilibrium, so that $n^* = \frac{\alpha I(1-\theta)}{F}$.

Efficiency

The efficiency of monopolistic competition can now be assessed along the second and first best lines. The constrained optimum maximizes utility under a no-loss restriction for each active firm. This amounts to minimize the composite price ρ. By convexity, the solution is symmetric with $\rho = pn^{-\beta}$. By duality, the solution is also profit maximization under a composite price constraint $\rho \leq \bar{\rho}$. On an isocurve $\rho = \text{cte}$, we have $\dot{p} \equiv \frac{\partial p}{\partial n} = \frac{p\beta}{n}$ and since profit is $\pi = (p-c)\frac{\rho D(\rho)}{np} - F$, the sign of $\frac{\partial \pi}{\partial n}$ is that of $\dot{p}np - (p-c)(p+n\dot{p}) \propto (1+\beta)c - p$ hence profit reaches a maximum at $p = (1+\beta)c$. Since the price at the free entry equilibrium and at the second best optimum are identical, so are the number of active firms (given that none makes a profit). Hence, we may say that monopolistic competition is constrained efficient.

The unconstrained optimum involves pricing at marginal cost and compensating firms with lump sum transfers to cover their fixed cost. By convexity, the optimum is symmetric, thus $q_i = \hat{q}, \hat{y} = \hat{q}n^{1+\beta}$ and $\hat{\rho} = cn^{-\beta}$ so that utility is $u\left(1 - n(F+cq), qn^{1+\beta}\right)$ since total cost of production have to be borne. The FOC over n yields $\frac{F+cq}{(1+\beta)qn^\beta} = \frac{u_2}{u_1}$ but since this ratio is equal to ρ at the consumer optimum, the solution is again $\hat{q}$, the equilibrium individual sales. From $\hat{p} = c < p^*$, we deduce $\hat{\rho} < \rho^*$ for the same number of firms n^*, hence total composite demand is $\hat{y} = D(\hat{\rho}) > D(\rho^*) = y^*$ which means that more firms must be active (as each sells the same). We then see from the formula for the composite price that the difference $\hat{\rho} - \rho^*$ is increased. The first-best thus involves the same individual firm size but a greater variety (more firms) because the greater sales allowed by marginal cost pricing allow to amortize more fixed cost at the economy level.

It is interesting to contrast this result with that obtained in the Hotelling model of §11.1.3 where full market coverage lead to opposite conclusion. The key difference is that total demand is not fixed but expands with variety (number of firms) hence allows to generate more consumer surplus that can then finance the entry of more firms.

11.2.4 International Trade

The modern theory of international trade has borrowed the models of industrial organization to overcome the limits of the perfect competition paradigm to study commerce among nations. Indeed, in the ubiquitous presence of fixed cost (or scale economies), perfect competition leads to marginal cost pricing and overall losses whereas monopoly

guarantees survival but generates a welfare loss and is a poor match to the reality of markets.

Trade

Centuries ago, Mercantilists saw international trade as a zero sum game i.e., a situation where the gain of one must be the loss of the other. Smith (1776) bailed them out with the absolute advantage whereby a country develops an efficient technology to produce a good at a lower cost than others. Later on, Ricardo (1817) claimed that a least developed country was not doomed to autarchy as it could specialize in the sector where it had a comparative advantage, although in the absolute, it would remain an inefficient world competitor.[18@] Trade was then mostly inter-industry and lead to a high degree of specialization (e.g., Portugal sells wine to England who sells wool to Italy who sells clothes to Portugal).

With the advent of the industrial revolution and large scale production, advanced countries (and regions within countries) started to trade the same goods albeit in versions differentiated by quality or specification. This is now the major component of trade. Krugman (1979) provides the first formal explanation to this phenomenon with a simple application of the monopolistic competition model.[19@]

Intra-Industry Trade

To address the effect of trade on workers' wage, labour is assumed to be the sole production factor with endogenous wage w and exogenous endowment L. We may thus introduce parameters γ and f to write $I = wL$, $c = \gamma w$ and $F = fw$, leading to the optimal relative price $\frac{p}{w} = (1+\beta)\gamma$ and limit number of firms $n = \frac{L(1-\theta)}{f}$ in the Cobb-Douglas specification.

The two countries, Home and Abroad, are identical except for labour endowments L and $\hat{L}$. They have no comparative advantages that would generate inter-industry trade yet intra-industry trade will take place. Upon eliminating barriers to trade, the total number of brands accessible to consumers rises to $n+\hat{n}$ which is exactly that corresponding to the grand economy uniting all the resources of Home and Abroad. Wages can be normalized to unity in each country ($w = \hat{w} = 1$) since only the relative price $\frac{p}{w}$ matters. Now, we observe that workers have no reason to emigrate, so that production continues as before in both countries. Lastly, the equilibrium remains unchanged because $(1+\beta)\gamma$ remains the optimal relative price for all firms whatever their location. The reason why intra-industry trade takes place is now clearcut: since consumers buy all available brands, a proportion $\frac{\hat{L}}{L+\hat{L}}$ of Home's income is spend over foreign brands (and vice versa

for Abroad). The greater number of consumed brands also translate into a greater welfare.[20@] Trade volume is $\frac{L\hat{L}}{L+\hat{L}}$, thus maximum when countries are of equal size.

Inter-Industry Trade

In order to study both inter and intra-industry trade, Krugman (1981) extends the basic model as follows. The economy now involves two sectors (e.g., industry and agriculture) with the same production technologies (same γ) but differing labour endowments $(\lambda, 1-\lambda)$. The previous results imply that $\frac{p_i}{w_i} = (1+\beta)\gamma$ for all sectors. Since firms earns zero profits, total wages in sector i, $w_i L_i$ are equal to their share of national revenue which, for simplicity, we take to be one half (i.e., we set $\alpha = 1/2$ in the Cobb-Douglas utility function). The number of firms in sector i then satisfies $\frac{n_i}{L_i} = \frac{1-\theta}{2f}$.

Consider then two otherwise identical countries with mirror endowments $(\lambda, 1-\lambda)$ and $(1-\lambda, \lambda)$. By construction, the grand (free trade) economy has the same number of firms in each sector, thus wages are equalized across sectors i.e., the scarce factor in each country loses from opening up frontiers. As before, consumers in both countries consume all available varieties, local and foreign, hence there is trade. Home, being the producer of λ% of the good #1 varieties, exports this good for a value of λ% of Abroad's expenditure on the good. A likewise consideration holds for good #2. Now it remains to observe that Abroad spends equally on both goods to deduce that Home export are $(\lambda + 1 - \lambda)\frac{Y^*}{2}$, a constant proportion of GDP. Thus trade volume is constant whatever the degree of specialization λ (insofar as Abroad is inversely specialized). Obviously, for low λ, countries are increasingly specialized and make mostly inter-industry trade.

To appreciate the overall effect of free trade upon worker's utility, we look at the equilibrium utility of a Home worker earning w. Two effects are expected. Firstly, there is a distribution effect due to factor price equalization, it favors the abundant factor in each country but hurts the scarce one (who enjoys a rent under autarchy). Next, the trade union aggrandizes the economy thus increases available varieties to the benefit of all. The commercial union will be welcomed if Home workers in the scarce sector gain from it, which we now study.

Since the underlying preferences leads to equal expenditure on both sectors ($\alpha = \frac{1}{2}$), the equilibrium demand for sector i is the fraction $\frac{1}{2n_i p_i}$ of the consumer's income w. His indirect utility is then

$$v(w) = \sum_{i \leq 2} \log\left(n_i \left(\frac{w}{2n_i p_i}\right)^{\theta} \right)^{1/\theta} = \sum_{i \leq 2} \log \frac{w}{p_i} + \beta \log n_i + cte$$

with $\beta \equiv \frac{1-\theta}{\theta}$. W.lo.g. we set $\lambda < \frac{1}{2}$ i.e., the scarce sector is #1. The utility variation for a

sector #1 worker upon opening frontiers to trade is then

$$\begin{aligned}\Delta v_1 &= v(w_1^*) - v(w_1) = \log\left(\frac{w_1^*}{p_1^*}/\frac{w_1}{p_1}\right) + \log\left(\frac{w_1^*}{p_2^*}/\frac{w_1}{p_2}\right) + \beta\left(\log\frac{n_1^*}{n_1} + \log\frac{n_2^*}{n_2}\right)\\ &= 0 + \log\left(\frac{\lambda}{1-\lambda}\right) - \beta\log(\lambda(1-\lambda))\\ &= (1-\beta)\log\lambda - (1+\beta)\log(1-\lambda) \qquad (11.11)\end{aligned}$$

because $w_1^* = w_2^*$ implies $\frac{w_1^*}{p_2^*}/\frac{w_1}{p_2} = \left(\frac{w_2^*}{p_2^*}/\frac{w_1}{p_1}\right)\frac{p_2}{p_1} = \frac{(1+\beta)\gamma}{(1+\beta)\gamma}\frac{p_2}{p_1} = \frac{w_2}{w_1} = \frac{L_1}{L_2} = \frac{\lambda}{1-\lambda}$ (applying $\frac{p_i}{w_i} = (1+\beta)\gamma$ repetitively). Likewise $\Delta v_2 = \log\left(\frac{1-\lambda}{\lambda}\right) - \beta\log(\lambda(1-\lambda)) > 0$ as $\lambda < \frac{1}{2}$. In both sector, the worker suffers from specialization (first term) but gains from the larger market. For highly differentiated versions of basic products i.e., when $\theta \le \frac{1}{2}$ (or $\beta > 1$), we have $\Delta v_1 > 0$. For $\theta > \frac{1}{2}$, (11.11) has a zero $\bar{\lambda}(\beta) \equiv \frac{1}{1+z}$ with $z = e^{\frac{1-\beta}{1+\beta}}$ such that $\Delta v_1 > 0 \Leftrightarrow \lambda > \bar{\lambda}(\beta)$.

> The model thus confirms that intra-industry trade among similar countries is beneficial for workers of all sectors whereas inter-industry trade, typical of highly asymmetric partners puts under duress the workers enjoying the highest wages under autarky.

11.3 Vertical Differentiation: Quality

In this section we study how quality, aka vertical differentiation, affects price competition among firms. Quality can stand for features of a product such as air conditioning in a car but it also applies to services. As a matter of example, consider two supermarkets; in the first one, employees are well trained, well dressed and ready to help, shelves are nicely decorated. The other supermarket has a reduced number of employees and crude shelve presentation but most probably, this negative aspect will be compensated by lower prices. This does not prevent some consumers to persist in shopping at the dearest shop, simply because they pay more attention to the quality of the service, which compensates them for the higher prices.

11.3.1 Quality and Market Power

Since consumers have heterogeneous tastes and incomes, the willingness to pay for quality differs widely; it therefore makes sense to develop a range of products (based on the same elementary commodity) and use differential pricing to maximize profits in this market. The example presented in Table 11.1 involves a brand of Spanish wine whose price per bottle, quoted in December 2005, varies according to the nurturing (time spend in oak barrels):

Nurturing	cosecha	crianza	reserva	gran reserva
Vintage	2003	2000	1998	1995
Price (€)	1,65	2,65	4,20	8,10

Table 11.1: Wine Qualities

The theoretical relationships between quality and market power are those obtained in §3.3 summarized in Table 3.1. In the rest of the section, we shall concentrate on competition among firms over a market such as wine and assume that each offers a single good which may nevertheless be interpreted as a family of versions. For instance all Spanish wine makers offer the nurturings reported in Table 11.1. Yet, they differentiate themselves along quality characteristics such as grape origin, use of additives, use of mechanical processing or the reputation of their enologist. In the end, these characteristics produce important price differentials.

11.3.2 Price Competition

We study Mussa and Rosen (1978)'s model of quality differentiation (cf. §3.3.1) as simplified by Tirole (1988) to explain how price differentials justify quality wedges. To simplify, each consumers is ready to buy one unit of a durable good. The brand i good has an objective quality index $s_i \in [0,1]$ like those build-up by consumer associations or the specialized press for cars, domestic appliances and other durable goods.[21@] Consumers are characterized by a "taste for quality" x uniformly distributed in $[0;1]$; their willingness to pay for good i takes into account the objective quality in a multiplicative way and is simply xs_i so that the utility of buying good i at price p_i is the surplus $u_i(x) = xs_i - p_i$; refraining from consuming yields a utility normalized to 0. Two firms A and B propose brands of (potentially) differing qualities and then compete in prices. As explained in §2.4.2, this sequential game is analyzed by backward induction: given the qualities initially chosen s_A and s_B, the price equilibrium is identified, then we go to the first stage to solve the quality equilibrium.

At the price competition stage, one quality being higher than the other, we relabel firms as h (high) and l (low); we have $s_h \geq s_l$. The utility functions $u_h(x)$ and $u_l(x)$ are drawn on Figure 11.7. The potential sales of firm i are $1 - x_i$ where x_i is such that $u_i(x_i) = 0$. As can be seen on the drawing, the quality differential advantages firm h. The indifferent consumer is $\tilde{x}$ such that $u_h(x) = u_l(x)$ thus $\tilde{x} = \frac{p_h - p_l}{s_h - s_l}$.

The effective demands addressed to firms are $D_h = max\{0, 1 - \tilde{x}\}$ and $D_l = max\{0, \tilde{x} - x_l\}$. Notice that if $p_l > p_h \frac{s_l}{s_h}$, then $\tilde{x} < x_l$ so that the low quality firm is excluded from the market ($D_l = 0$) in which case the high quality firm becomes a monopoly. The demand

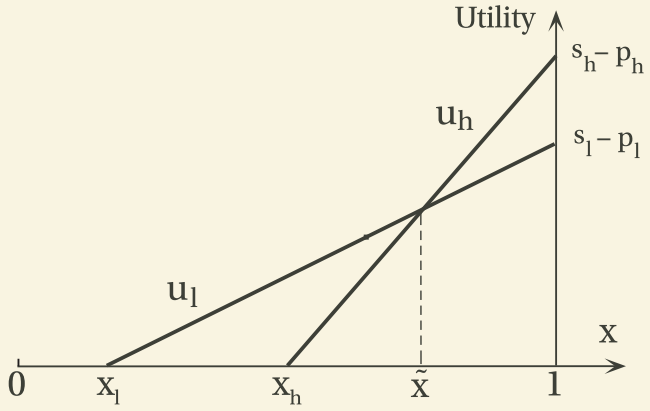


Figure 11.7: Utilities under Differing Qualities

resulting from consumers' choices, given prices, are therefore

$$D_l(p_l, p_h) = \begin{cases} 1 - \frac{p_l}{s_l} & \text{if } p_l < p_h - s_h + s_l \\ \frac{p_h s_l - p_l s_h}{s_l(s_h - s_l)} & \text{otherwise} \\ 0 & \text{if } p_l > p_h \frac{s_l}{s_h} \end{cases} \tag{11.12}$$

and

$$D_h(p_l, p_h) = \begin{cases} 1 - \frac{p_h}{s_h} & \text{if } p_h \leq \frac{s_h}{s_l} p_l \\ 1 - \frac{p_h - p_l}{s_h - s_l} & \text{otherwise} \\ 0 & \text{if } p_h > p_l + s_h - s_l \end{cases} \tag{11.13}$$

Whenever $p_l \geq p_h \frac{s_l}{s_h}$, firm l has zero demand and zero profit; she thus has an incentive to reduce her price in order to grab a positive market share. Hence, only $D_h = 1 - \tilde{x}$ and $D_l = \tilde{x} - x_l$ are relevant for the equilibrium analysis.

The payoff functions in that case are

$$\pi_h(p_h, p_l) = p_h \left(1 - \frac{p_h - p_l}{s_h - s_l}\right) \tag{11.14}$$

and

$$\pi_l(p_h, p_l) = p_l \frac{p_h s_l - p_l s_h}{s_l(s_h - s_l)} \tag{11.15}$$

The best reply derived from the first order conditions $\frac{\partial \pi_h}{\partial p_h} = 0$ and $\frac{\partial \pi_l}{\partial p_l} = 0$ are

$$\psi^h(p^l) = \frac{s_h - s_l + p_l}{2} \quad \text{and} \quad \psi^l(p^h) = p_h \frac{s_l}{2 s_h} \tag{11.16}$$

The price equilibrium solves simultaneously $\psi^h(p_h) = p_h$ and $\psi^l(p_h) = p_l$; we obtain

$$p_l^* = \frac{s_l(s_h - s_l)}{4s_h - s_l} \quad \text{and} \quad p_h^* = \frac{2s_h}{s_l} p_l^*. \tag{11.17}$$

Observe that the nearer the qualities, the lower the prices. Hence, choosing a quality far away form's one rival increases the equilibrium prices and the resulting profits. We may conclude that

> Vertical differentiation enables firms to soften price competition. However the high quality firm is a clear winner as he gets to set a price more than twice that of his competitor.

11.3.3 Quality competition

This first stage where firms choose their qualities is analyzed taking into account that in the second stage of price competition, firms play the Nash equilibrium just characterized; this is to say we impose subgame perfection.

Whatever the example we may think of, quality is costly to achieve, thus a good reason to lower one's quality is save on costs. To study with precision the other reasons that may lead a firm to lower her quality we will assume away the first reason i.e., suppose that quality is costless.

The demands addressed to the firms at the equilibrium prices can be computed using (11.17) inside (11.12) and (11.13); we obtain $D_l^* = \frac{s_h}{4s_h - s_l}$ and $D_h^* = \frac{2s_h}{4s_h - s_l}$. The corresponding profits are thus

$$\Pi_h(s_h, s_l) \equiv p_h^* D_h^* = \frac{4s_h^2(s_h - s_l)}{(4s_h - s_l)^2} \tag{11.18}$$

and

$$\Pi_l(s_h, s_l) \equiv p_l^* D_l^* = \frac{s_l s_h(s_h - s_l)}{(4s_h - s_l)^2}. \tag{11.19}$$

It then remains to consider the first stage of the game where qualities are chosen. The profit of firm $i = A, B$ has 3 parts depending on whether products are differentiated or not. If identical qualities have been chosen then the two products are homogeneous and the price competition is of the Bertrand type leading to zero profit since in the Bertrand equilibrium, firms quote their marginal cost (which is zero here). Otherwise, the previous analysis applies taking into account the fact that i's product might be the high

or low quality one. We can therefore write[22@]

$$\Pi_i(s_i, s_j) = \begin{cases} \Pi_l(s_j, s_i) & \text{if } s_i < s_j \\ 0 & \text{if } s_i = s_j \\ \Pi_h(s_i, s_j) & \text{if } s_i > s_j \end{cases} \tag{11.20}$$

The reader will easily check that $\frac{\partial \Pi_h(s_h, s_l)}{\partial s_h} > 0$ which means that the quality leader would like to propose the highest possible quality $s = 1$, because it boosts her sales and price. On the other hand, one can check that $\frac{\partial \Pi_l(s_h, s_l)}{\partial s_l} = 0$ for $s_l = \frac{4}{7} s_h$; this is the optimal degree of differentiation for the low quality firm. Lastly, no one wants to match the competitor's quality because it would trigger the damaging Bertrand competition. The equilibrium of the first stage is thus $(s_1, s_2) = (\frac{4}{7}, 1)$ or $(s_1, s_2) = (1, \frac{4}{7})$. The high quality brand is not known but this is not crucial given that A and B are names of firms without history. In equilibrium, the high quality firm quotes a price $\frac{7}{2}$ times larger than the low quality one and sales twice as much, thereby achieving a seven fold profit (check as an exercise that $p_l = \frac{1}{14}$ and $D_l = \frac{7}{24}$). We may conclude this analysis by stating

> Firms are likely to relax price competition (Bertrand paradox) through vertical product differentiation.

11.4 Drivers of Differentiation †

11.4.1 Differential Pricing

Business Stealing and Versioning

The "punch line" of this chapter is that differentiation is a *business stealing* vehicle i.e., it is a resolute strategy for firms to make distinctive and more attractive products in order to capture greater market shares. At the same time, differentiation is also intuitively understood as a channel for applying *differential pricing* in which case, we speak of *versioning* (cf. §4.3.4). The idea is to partition a market into segments of homogeneous customers, design an appealing version of the basic product for each segment and price it at a greater margin. For instance,

- Perfumes come in two gender-oriented fragrances sharing the same base.
- Computers and softwares are differentiated across their destined use as in professional vs. home.
- Cars often come with a hatchback, sedan, station-wagon and coupé version.

- Trains or planes offer classes with increasing levels of quality of service (both on board and on ground) to reach segments of users with increasing willingness to pay.
- Book are sold in hardcover and paperback editions, through readers' clubs and as surplus books in discount shops.

The two drivers of differentiation, business stealing and differential pricing, have been considered from two radically different perspectives. In chapter 4, firms design complex price discriminating strategies but are insulated from competition as we endow them with an exogenous degree of market power. In the present chapter, firms directly compete but are forced to use simple unit price strategies.

Differentiation and Discrimination

We follow Armstrong (2007) and bring together differentiation and discrimination in a very simple framework (cf. also §5.3.4). Recalling that a monopoly is always willing to use price discrimination, we are, in fact, testing the robustness of this conclusion to the introduction of direct competition among firms. In the wider interpretation of horizontal differentiation on characteristics, the transportation cost represents how much an individual dislikes buying his less preferred brand; we may thus speak of the *choosiness*. Our findings summarize as follows

> The ability to differentiate prices based on WTP, choosiness and heterogeneity is respectively useless, useful and harmful.

Information on WTP Recall that in the standard Hotelling model (without production cost), consumers display horizontal heterogeneity with a *taste* parameter $x \in [0;1]$ while firms are horizontally differentiated with $x_A = 0$ and $x_B = 1$. We can then add vertical heterogeneity with choosiness or transportation cost $t \in [t_0; t_1]$. The WTP v for the object may be heterogenous but is nevertheless assumed large enough to generate a purchase. We assume that taste, choosiness and WTP parameters are statistically independent.

If a monopolist can observe individual WTPs, we already know that he will sell to each person at her WTP minus the transport cost. The outcome is efficient and maximizes profit. If now, duopolists observe individual WTPs, they fail to gain anything because each buyer takes his decision on the basis of the price difference. The demand received by a firm is thus independent of the WTP of purchasers and so is the equilibrium. Observe also that since the information is not used, the same outcome arise in equilibrium whether one, both or none of the firm knows the WTPs. Hence this piece of

information has no value except to a monopolist. This result extends to situations where consumers buy multiple units and multiple products meanwhile the location parameter remains unknown to the firm for this is exactly why competition makes this information irrelevant.

Information on Choosiness When firms can discriminate over choosiness, they repeat the same basic interaction in all segments. They fight more over people who see their products as close substitutes which is detrimental to profits. At the same time, firms relax competition over people who are addicted to a specification which is beneficial for profits. Overall, firms gain more than they lose when compared to non discriminatory pricing as we now proceed to show.

Firms offer prices p_A and p_B to consumers of segment t. The indifferent customer is $\hat{x}_t = \frac{1}{2} - \frac{p_A - p_B}{2t}$ and the equilibrium prices are $p_A = p_B = t$. Overall, the average price is $\bar{t} = \mathbb{E}[t]$. If firms cannot discriminate, they post a single price valid for all segments. The demand addressed to A becomes the expected indifferent customer $\mathbb{E}[\hat{x}_t] = \frac{1}{2} - \frac{p_A - p_B}{2\hat{t}}$ where $\hat{t} \equiv 1/\mathbb{E}[\frac{1}{t}]$ is the harmonic mean of choosiness. The equilibrium prices are $p_A = p_B = \hat{t} < \bar{t}$ by Jensen inequality (cf. footnote 26.3). This means that discrimination over choosiness raises prices and profits. Picky people (high t) end up paying more under discrimination and make the extra profit. In the absence of information, firms fail to evaluate precisely how more or less choosy people value the price difference so that a discount has a greater business stealing effect, thereby intensifying competition.

Information on Taste The taste parameter x of consumer is the original source of differentiation among firms. When they can discriminate over that characteristic, a price war ensues to the benefit of consumers and to the detriment of firms. Indeed, each firm is now able to extract large chunks of surplus from people close to their offering (in the characteristics space). But this means that the strong market is now specific to each firm, unlike the case of choosiness discrimination. So, when a firm starts price discriminating and lowers the price in her weak market, she is frontally attacking the other side's strong market. The lack of agreement on what is the local monopoly for each firm leads to a prisoner's dilemma as we now proceed to show.

If firms can discriminate on taste only, they offer prices p_A^x and p_B^x. When choosiness t is common, firms compete à la Bertrand for each consumer. A consumer with type $x < \frac{1}{2}$ shops at A if

$$p_A^x + tx < p_B^x + t(1-x) \Leftrightarrow p_A^x < p_B^x + t(1-2x)$$

This means that B is undercut without recourse by the price $t(1-2x)$ which is thus the equilibrium price. Likewise, B takes the other side of the market. Since every consumer

pays less than the price in the absence of information which is here $\hat{t} = \bar{t} = t$, price discrimination is beneficial for consumers and detrimental to firms (who lose half of their profits as individual profit is $\int_0^{1/2} t(1-2x)\,\mathrm{d}x = \frac{t}{4}$).

To complete the analysis, it is useful to look at the case where choosiness is not common to all but, for instance, uniformly distributed over $[t_0; t_1]$. Consider the segment $x < \frac{1}{2}$ and discriminating prices p_A and p_B. By pricing at $t_0(1-2x)$, firm A can exclude B and make on average $\frac{t_0}{4}$. If this is a dominant strategy, it will also be for B for $x > \frac{1}{2}$, so that total profit for each firm is $\frac{t_0}{2}$, lesser than under no information. It remains to study the strategy by which firm A sells dear to choosy people leaving nearby. The price p_A attracts types such that $t > z \equiv \frac{p_A - p_B}{1-2x}$ hence $d_A(p_A, p_B) = \frac{t_1 - z}{t_1 - t_0}$ and by symmetry $d_B(p_A, p_B) = \frac{z - t_0}{t_1 - t_0}$. The standard Hotelling competition leads to best replies

$$p_A = \frac{p_B + t_1(1-2x)}{2} \quad \text{and} \quad p_B = \frac{p_A - t_0(1-2x)}{2}$$

The equilibrium solving this system is

$$p_A^* = \frac{(2t_1 - t_0)(1-2x)}{3} \quad \text{and} \quad p_B^* = \frac{(t_1 - 2t_0)(1-2x)}{3}$$

and leads to profits

$$\pi_A = \frac{(1-2x)(2t_1 - t_0)^2}{9(t_1 - t_0)} \quad \text{and} \quad \pi_B = \frac{(1-2x)(t_1 - 2t_0)^2}{9(t_1 - t_0)}$$

By symmetry for the case $x > \frac{1}{2}$, each firm earns $\mathbb{E}[\pi_A + \pi_B] = \frac{5t_1^2 + 5t_0^2 - 8t_1 t_0}{18(t_1 - t_0)}$. Now, this competition (and its equilibrium) is valid if the proposed equilibrium price is greater than the undercutting strategy which reads $\frac{2t_1 - t_0}{3} > t_0$ i.e., if t_0 is small compared to t_1. In the limiting case where t_0 vanishes, the equilibrium profit tends to $\frac{5t_1}{18} = \frac{5\bar{t}}{9} > \frac{\hat{t}}{2}$, the payoff under uniform pricing. Hence, firms take advantage of price discrimination over taste (x) if choosiness (t) displays a large heterogeneity because segmenting clients by taste amounts to engage in Bertrand competition.

A price discriminating monopolist is never worse off. Likewise, an oligopolistic firm is always better off if it can price discriminate, for given prices offered by its rivals. However, once account is taken of how rivals will react, firms find themselves trapped in a prisoner's dilemma and lose out (cf. §2.4.1). In relation to this theoretical finding, many suspect that, at the end of the XIXth century, US railways companies welcomed the tariff regulation prohibiting price discrimination precisely because this practice was destroying their profits (cf. §9.2.3).

11.4.2 Cost Edge

Tyagi (2007) shows in a price competition framework that similar firms would like to differentiate more. However, if one firm has a strong cost advantage, it may prefer to reduce differentiation in order to take full advantage of its lower cost.

We use a simplified version of the price system (5.18) describing price competition in a differentiated environment with $q_i = 1 - p_i + d(p_j - p_i)$. We set $c_1 = 0$ and $c_2 = c$ so as to give firm 1 a cost edge. Profits are $\pi_1 = q_1 p_1$ and $\pi_2 = q_2(p_2 - c)$. The FOCs of profit maximization are

$$\begin{array}{l} 0 = 1 - 2(1+d)p_1 + dp_2 \\ 0 = 1 + (1+d)c - 2(1+d)p_2 + dp_1 \end{array} \Rightarrow \begin{array}{l} p_1 = \frac{1+dp_2}{2(1+d)} \\ p_2 = \frac{1+(1+d)c+dp_1}{2(1+d)} \end{array} \Rightarrow \begin{array}{l} p_1^* = \frac{2+d(cd+c+3)}{(d+2)(2+3d)} \\ p_2^* = p_1^* + \frac{c(1+d)}{2+3d} \end{array}$$

The relevant range for the cost disavantage is $p_2^* > c \Leftrightarrow c < \bar{c} \equiv \frac{2+3d}{4d+d^2+2}$. The equilibrium profits being $\pi_i = (1+d)\left(p_i^* - c_i\right)^2$ (cf. solution of system (5.18)), we can now inquire whether the two firms like or not differentiation. We find out

$$\begin{array}{rcl} \frac{\partial \pi_2}{\partial d} & \propto & \left(d^2+4d+2\right)c - 2 - 3d = \left(d^2+4d+2\right)(c-\bar{c}) < 0 \\ \frac{\partial \pi_1}{\partial d} & \propto & (1+d)\left(3d^3+18d^2+20d+8\right)c - d\,(2+3d)^2 < 0 \Leftrightarrow c < \underline{c} \end{array}$$

where $\underline{c} \equiv \frac{d(2+3d)^2}{(1+d)(3d^3+18d^2+20d+8)} < \bar{c}$ since $\bar{c} - \underline{c} \propto 24d + 24d^2 + 7d^3 + 8 > 0$.

We conclude that the weak firm always prefers more differentiation (lower substituability parameter d) in order to relax the price competition where she stands at a disavantage. The strong firm has no such clear cut preference; she displays the same preference if the cost advantage is small but otherwise she prefers less differentiation in order to take full advantage of her large cost edge.

11.5 Advertising

The OECD estimates that member countries spend more than 2% of their GDP in advertising. Table 11.2 present estimates in 2003 bn$ from consulting firm Zenith Optimedia regarding advertising expenditure in major media such as newspapers, magazines, television, radio, cinema, outdoor, internet. The recent survey on advertising by The Economist is an excellent update on the topic. At the individual level, The 2002 report on Leading National Advertisers in the US by Advertising Age magazine informs us that many companies among which General Motors (cars), Procter and Gamble (detergents), Ford (cars), Pepsi (beverage), Pfizer (pharmacy), AOL (media) spend more than 2 bn$ in

advertising on the US market alone in 2001.

Region	2003	*share*	2005	*share*	2007	share
North America	158	*46%*	175	*45.0%*	193	*44.5%*
Europe	89	*26%*	99	*25.4%*	108	*25.0%*
Asia, Pacific	70	*20%*	80	*20.6%*	91	*21.0%*
Latin America	14	*4%*	17	*4.4%*	19	*4.4%*
Africa, M. East	14	*4%*	18	*4.7%*	22	*5.1%*
World	346		388		434	

Table 11.2: Advertising Expenditure

These stylized facts make advertising a relevant topic of study. We rely on the extensive review of Bagwell (2007).

11.5.1 Opposite Views

The two basic functions of advertising are to inform and to convince. It has obviously no place in a perfectly competitive economy since consumers are perfectly informed and firms can sell as much as they want at the market price. Advertising is thus a feature of an imperfectly competitive economy.

Persuasive View

Advertising clearly speaks to our subjectivity in trying to build loyalty and make us feel secure; for instance, shoe manufacturers like Nike or Adidas advertise their footwear with famous champions to enable us, buyers, to identify with our idols.

The negative view, dominant in the first half of the 20th century, holds that advertising alters consumers' tastes to create a non genuine brand loyalty. The resulting product differentiation makes the demand for the advertised product more inelastic and thus enable the firm to sustain higher prices. Since advertising is of a fixed cost nature it participate to erect a barrier to entry. Indeed, advertising by established firms creates a reputation for their brands and new entrants can succeed only by developing their own reputation through even more advertising (they must induce consumers to switch from an established and familiar product to a new and unknown one).

It has been long recognized that advertising enables to steal business from competitors so that it any campaign will be counteracted by more advertising but since the market remains the same, at least if we consider a loose definition of the market, resources spend on advertising are wasted since fail to generate any form of wealth for society.

The three arguments of fooling consumers, erecting barrier and resource waste add up to produce the general conclusion that advertising is anti-competitive.

Informative View

According to this positive view, advertising signals the existence of a product and some of its competitive attributes like price, retail location or quality attributes. Consumers therefore save on the search costs of learning price, quality and the very existence of products. When the information revealed is trusted (e.g., local products advertised in local newspapers) the degree of differentiation due to misinformation of consumers is reduced hence advertising is pro-competitive.

Secondly, this socially useful information, is made available at a lower cost when produced by firms for the simple reason that the mass media permits the diffusion of a single message to large audiences at a reasonable price. In the 1960s, Chicago economists like Stigler and Becker went further in arguing that advertising was the endogenous market response to the market failure generated by the imperfect consumer information. They reason that the advertiser's demand curve becomes more elastic (price reductions become more effective business stealing instruments) i.e., promotes competition among established firms. As well, advertising facilitates entry as it provides a means through which a new entrant can publicize its existence, prices and products. Their conclusion is that advertising can be pro-competitive.

Complementary View

This last and most recent view tries to implant advertising inside the consumer preferences to avoid the ad-hoc formalizations of the old views. Advertising is then seen as a complementary good to the advertised product. What the consumer values is the good $z = g(a)q$ where q is the product quantity and a the amount of advertising. The positive function g captures the character of advertising, either useful ($g' > 0$) or annoying ($g' < 0$). For example, consumers may value "social prestige", and the consumption of a product may generate greater prestige when the product is advertised. An important implication is that standard methods may be used to investigate whether advertising is supplied to a socially optimal degree, even if it conveys no information.

It is often infeasible to separately and directly sell advertising to consumers. Instead, advertisements is given away like for direct mail ads or sold jointly with the other products like for a newspaper. The former case may be understood as a situation in which advertising is a good (or at least not a bad) that is given away, the quantity of advertising is determined by the producers, and each consumer simply accepts (consumes) all of the

advertising that is received. The latter case corresponds to a situation in which each consumer determines his consumption quantity of the joint good, given the price of the joint good. As advertising is complementary, it may be sold at a subsidized implicit price. Indeed, if advertising is a bad (e.g., TV ads may lower utility), then its implicit price is negative (advertisers include free and enjoyable programs to compensate the viewer for watching the ads).

Historical notes

Early on, Marshall (1890) saw a constructive role for informative advertising but also noted that it could be socially wasteful when involving repetitive messages whose purpose is to steal the customers of competing firms (the resources spent by firms fail to generate anything useful for the economy).

Robinson (1933), among others at the time, pursues more systematically the integration of advertising in her theory of monopolistic competition. Even though she does not model consumer behavior, her reasoning leads to the conclusion that advertising is both persuasive and informative (cf. next paragraph) but is overall anti-competitive.

Kaldor (1950) adds scale economies as an indirect positive effect of advertising. The advent of large-scale advertising enables manufacturers to establish brand names and leap over wholesalers to establish a direct connection with consumers. The manufacturing sector then becomes more concentrated, and additional scale economies associated with mass-production techniques are realized. The resulting structure is thus characterized by low production costs and high selling costs (marketing and advertising). This theory has some resonance in today's business which we proceed to discuss using examples from two sectors.

Illustrative Cases

Food It is well know that products displayed on front shelves sell much more. Large retailer chains like Carrefour or Ahold exploit this multiplier effect to charge brand name producers an extra fee to get these precious locations in their stores. This behavior has been deemed anti-competitive in several European countries but it is quite difficult to fight. The only way to regain bargaining power with respect to retailers is to be more famous than they are in the mind of consumers. An illustration of this is the very large amount of advertising on European TV channels made by Danone[23@] and Nestlé[24@] ; by spending large resources in advertising both try to reach the degree of visibility of the world leader in image, Coca-Cola. The idea is quite simple: a supermarket cannot default on Coke without losing instantaneously most of its clients.

Hotel Since the advent of the internet and communication technologies (ICT), most leisure travels are sold through the 4 Computer Reservation Systems (CRS) Sabre, Amadeus Galileo and Worldspan that gather offers from dozens of suppliers and dispatch them to thousands of travel agencies. Nowadays, with the internet, we are millions to access their databases through the sites of Expedia or Travelocity. Hotel companies with brand names like Hilton or Sheraton fear to be diluted in the catalogue of these powerful aggregators; in that case they would have little bargaining power to negotiate the fee they pay to get a good position in the window that the consumer will browse to pick its travel. They have thus started intensive campaigns of "branding" to increase the fidelity of clients and have them connect directly to the hotel internet site instead of looking for a rebate on some wholesaler site. Likewise, airlines companies in the US have joined effort and created Orbitz, an internet site selling their own seats at better price and conditions than competing CRSs to avoid being squeezed.

Empirical findings

It has been shown empirically that brand advertising has a significant effect on the brand's current and future sales (goodwill effect) but sales appear more responsive to price and quality.[25@] With respect to competition, market share are negatively related to rival advertising and an increase in own brand advertising appears to induce rivals to respond with more advertising. Overall, the total demand seems responsive to advertising.

Using data on the Coca-Cola and Pepsi-Cola markets over the 1968-86 period, Gasmi et al. (1992) show that Coca-Cola was a Stackelberg leader in price and advertising until 1976. Afterward, competition took the form of duopoly conduct characterized by collusion in advertising and possibly price. In this context, the empirical estimates suggest that advertising in the cola market is largely combative.

Kadiyali (1996) analyzes the U.S. photographic film industry. In the 1970s, Kodak had a virtual monopoly but accommodated entry by Fuji in 1980. The parameter estimates for the pre-entry period (before 1980) indicate that Kodak maintained its monopoly position by using limit pricing and high advertising. Estimates for the post-entry period suggest that Kodak was compelled to accommodate Fuji who enjoyed demand and cost advantages, that Kodak and Fuji then colluded in price and advertising and finally that advertising expanded market size and constituted a public good across firms.

11.5.2 Theories

Monopoly

The standard model of monopoly advertising by Dorfman and Steiner (1954) can be presented as follows: consumer demand depends not only negatively on price p but also positively on advertising expenditures a i.e., $q = D(a,p)$ with $D_a > 0$. We dot inquire here into the origin of this effect.

The monopoly profit is $\Pi = pD(a,p) - C(D(a,p)) - a$ and the FOCs are the traditional Lerner(cf. eq. 3.4) equation

$$\frac{p - C_m}{p} = \frac{1}{\epsilon_p} \tag{11.21}$$

where ϵ_p is the price elasticity of demand (cf. eq. (2.14)) and the novel equation

$$(p - C_m)D_a = 1 \quad \Leftrightarrow \quad \frac{a}{pD} = \frac{\epsilon_a}{\epsilon_p} \tag{11.22}$$

where ϵ_a is the advertising elasticity of demand. The RHS of (11.22) states that a profit-maximizing monopolist spends on advertising a proportion of revenue determined by a simple elasticity ratio.

Application to the persuasive view

Consumers, willing to buy one unit of the good (e.g., a car), are differentiated by their basic willingness to pay v. Advertising then acts as a multiplier yielding social prestige upon owners; the net WTP is thus $vg(a) - p$ where $g(0) = 1$ and $g' > 0$. Assuming a uniform distribution of consumers of total mass one, the demand addressed to the firm is $D(a,p) = 1 - \frac{p}{g(a)}$ since the indifferent consumer has WTP $\frac{p}{g(a)}$. We have $\frac{\partial D}{\partial p} = \frac{-1}{g(a)}$.

As the persuasive view suggests, when advertising is increased, the demand function becomes more inelastic since $\epsilon_p = \frac{p}{g(a)-p} \Rightarrow \frac{\partial \epsilon_p}{\partial a} < 0$. The persuasive view also holds that the profit-maximizing price p_a^M, solving (11.21), rises when the level of advertising is raised. To see if this occurs in the present setting, we can study the FOC of price optimality $(p - C_m)\frac{\partial D}{\partial p} + D = 0$ which simplifies into $2p_a^M = g(a) + C_m$ for our particular demand. It is then clear that an increase in advertising a, increases the optimal price p_a^M through the effect on social prestige ($g' > 0$) but nothing can be said outright for the effect of technology.[26@]

If there are diseconomies of scale in production ($C_m \nearrow$), an increase in advertising increases sales and marginal cost so that the price is increased to preserve the margin as measured by the Lerner index (cf. §11.21); the claim is thus correct. If, on the contrary, the technology shows strong economies of scale, then the now negative scale effect can

outbalance the direct advertising effect so that greater advertising leads the monopoly to lower his price. To conclude, the model is useful to estimate when the persuasive view is correct or not relative to the effect of advertising on price.

Application to the informative view

If advertising only informs consumers of the existence of a product, the demand addressed to the monopolist is the product of the individual demand $d(p)$ by the informed consumer base $g(a)$ where we can assume $g' > 0 > g''$ (decreasing returns to scale of advertising). In such a case, the price elasticity of aggregate demand is equal to the price elasticity of individual demand (the advertising effect $g(a)$ cancels out in computing ϵ_p).

Upon examining the Lerner equation (11.21) for the optimal price conditional on advertising, we see that the impact of advertising is entirely dictated by the volume or scale effect (the quantity in C_m) i.e., the direct effect present in the previous application is absent. The effect of advertising is thus straightforward: if there are economies (resp. diseconomies) of scale in production then more advertising lowers (resp. raises) the optimal price but in contrast with the previous application, if the marginal cost is constant, then advertising has no effect on the monopoly price.

Dixit and Norman (1978) provide a foundation for the normative theory of persuasive advertising. If the consumer welfare is measured relative to a fixed standard, then a monopolist provides price-increasing advertising (as in the first application) to an extent that is socially excessive. Proponents of the informative and complementary views, however, argue that the fixed-standard approach ignores consumer-welfare gains from advertising that are associated with information and social prestige. Under their alternative approach, a monopolist provides price-maintaining and price-decreasing advertising to an extent that is socially insufficient. In conclusion, the judgment we can hold on advertising depends on moral values and on whether "social prestige" should enter welfare calculations.

Duopoly

In oligopoly markets, advertising is an important instrument of competition. The advertising of one firm may steal the business of other firms and thus lowers their profits. This *business-stealing* externality raises the possibility that advertising may be excessive because once a firms starts spending on advertising, a challenger may be forced to spend twice as much to retain or regain his customers. Given the intensification of competition, it is unclear whether advertising will be socially excessive or not. This issue is the earliest example of economic conflict assessed analytically in the literature (cf. §7).

To inquire this quandary we use Grossman and Shapiro (1984)'s version of the Hotelling model with extreme locations. The transportation cost is t, the marginal cost of production is c and the advertising technology works as follows: the cost to reach (or inform) a proportion x of the street (potential clients) is $A(x) = ax^2/2$. Let us interpret the parameters. The total transportation cost for consumers to learn one price by themselves (searching) is $\int_0^1 tx\,\mathrm{d}x = \frac{1}{2}t$. The cost for a firm to inform everybody of its price is $a/2$, thus the cost for suppliers to inform buyers is a. As we shall see, the equilibrium will be highly dependent on which of $\frac{1}{2}t$ or a is the greatest i.e., who owns the most efficient information transmission technology.

Firms choose prices p_1 and p_2 and advertising levels x_1 and x_2. Given the potential demand x_1 addressed to firm 1, $x_1(1-x_2)$ consumers are captive as they ignore p_2, they buy if $p_1 < S$, the reservation price. A proportion x_1x_2 on the other hand is fully informed and solve the classical Hotelling trade-off. Demand is thus $D_i = x_i\left(1 - x_j + x_j\frac{p_j-p_i+t}{2t}\right)$ for $i = 1,2$. Profit is then $\pi_i = (p_i - c)D_i - A(x_i)$. The elasticity of demand at $p_1 = p_2 = p$, $x_1 = x_2 = x$ is $\frac{xp}{(2-x)t}$ and is thus increasing with advertising. The FOC of optimal pricing leads to

$$p_i = \frac{p_j + t + c}{2} + \frac{1 - x_j}{x_j}t \tag{11.23}$$

where the second term is the mark-up enabled by the lack of information of some consumers. The FOC of optimal advertising leads to

$$x_i = \frac{p_i - c}{a}\left(1 - x_j + x_j\frac{p_j - p_i + t}{2t}\right) \tag{11.24}$$

Looking for a symmetric equilibrium (p, x), (11.23) simplifies into $p = t+c+2t\frac{1-x}{x}$ while (11.24) yields $2x = \frac{p-c}{a}(2-x)$. Using the former, the latter further simplifies into

$$\frac{2x}{2-x} = \frac{p-c}{a} = \frac{t}{a}\left(1 + 2\frac{1-x}{x}\right) \quad \Leftrightarrow \quad x^* = \frac{2}{1+\sqrt{2a/t}} < 1$$

if $t < 2a$. The equilibrium price is then $p^* = c + t + 2t\frac{1-x^*}{x^*} = c + \sqrt{2at} > c + t$ leading to profit $\Pi = 2a\left(1+\sqrt{2a/t}\right)^{-2}$. We observe from these formulas that if $t \geq 2a$ then competitive forces push firms to inform the whole street ($x^* = 1$) and price the good at $p^* = c + t$, the complete information Hotelling price. This occurs because the information technology of producers is much more efficient than the searching technology of consumers.

In this model, the price is weakly larger than under full information and is increasing in t but at a slower rate (over the range $t < 2a$). The reason is that more differentiation makes advertising more valuable for firms thus increases the amount of information revealed to consumers in equilibrium; competition is increased thanks to the possibility

of advertising. Oddly, the equilibrium profit is increasing with a, the cost of revealing information; although the direct effect of a is negative, its strategic effect is positive and larger. Given our previous analysis this is not so much of a surprise because a larger a basically means that firms find it more costly to reveal information, hence they reduce their advertising which enables to sustain larger prices. Therefore, a little tax on advertising (disguised as a moral requirement) may raise profits for firms and the tax collection for the State so that we, consumers, loose on both sides !!

Signaling Quality

We argue here following Kihlstrom and Riordan (1984) that advertising can be a signal of quality in an oligopoly setting if a high quality producer is patient enough. Since the exposition bears considerable similarity with §21.1.3 on "education as a signal of ability", the model is only sketched.

Whenever there is no cheap and reliable agency that can emit a certificate of quality, firms are not able to inform correctly their potential customers about the quality of their products. For such *experience goods*, the opinion of consumers is either confirmed or revised upward or downward according to what they bought and what they anticipated. The following "hit-and-run" strategy immediately comes to mind to take advantage of this delay in the revelation of true quality: package a cheap and low quality product as if it was top-notch, sell it at a high price for one period and then lower the price since nobody will get fool now that the word has spread regarding what was the real quality inside the gleaming package.

If we let π_i^j denotes the profit during one period made by selling quality i when consumer think it is quality j then we obviously have the ranking $\pi_L^H > \pi_H^H$ since cost are lower for a lower true quality but also $\pi_L^H > \pi_L^L$ since consumers are ready to pay more for (what they think is) a higher quality. The present value of aggregated profits over many periods for the "hit-and-run" strategy is $\pi_L^H + \frac{1}{r}\pi_L^L$ where r is the interest rate and is obviously greater than the profit $\pi_L^L + \frac{1}{r}\pi_L^L$ gained by a truthful maker of a low quality product.

To force out the deceptive strategy of "hit-and-run", a high quality producer can simply spend an amount of advertising a so large that a low quality producer would not want to imitate this behavior for the sake of being see at "HQ". The condition is

$$a > \pi_L^H - \pi_L^L \tag{11.25}$$

Yet the high quality producer should take care of not spending too much in advertising because he still the option of switching to low quality himself at no cost. This won't

occur if

$$\pi_H^H - a + \frac{1}{r}\pi_H^H \geq \pi_L^L + \frac{1}{r}\pi_L^L \quad \Leftrightarrow \quad a < \frac{1+r}{r}\left(\pi_H^H - \pi_L^L\right) \tag{11.26}$$

A first necessary condition for advertising to be a signal of quality is $\pi_H^H > \pi_L^L$ i.e., high quality products must generate more profits than low quality ones in a perfectly informed market. As we saw in §11.3, the condition is satisfied. Comparing (11.25) and (11.26), we derive a second necessary condition for advertising to be a signal of quality: $r < \frac{\pi_H^H - \pi_L^L}{\pi_L^H - \pi_H^H}$ i.e., the producer is patient and cares mostly for future profits.

> A high quality producer can signal his characteristic in a discriminatory fashion (avoid mimicking by low quality imitators) only if his product has a long life cycle.

Chapter 12

Research and Development

In a market economy, the ultimate ruler is the consumer. Generally, he/she agrees to pay more for quality and although higher quality products tend to be more costly to produce, they generate higher profits as we showed in §11.3. It is thus rational for a firm to invest financial and human resources in order to become a quality leader. This constitutes an example of sunk cost because these resources cannot be reoriented without suffering a minimal loss (opportunity cost). The subjective traits of "quality" have been studied with advertising in the previous chapter. In the present one, we deal with the objective side.

Innovation according to Schumpeter (1942) includes the entire process of inventing, demonstrating, bringing to the market and improving a novel product. It involves two complementary activities, *invention* and *melioration*, respectively carried by small and large firms. Big firms specialize on improving of existing products (e.g., new features, increased reliability, enhanced user friendliness, new uses) whereas small firms specialize on discovering new products. Because few inventions have market value and even fewer are successfully brought to the market, entrants suffer a rate of failure (exit). Incumbents, on the other hand, follow a routinized and conservative approach to achieve strategic objectives such as meeting the R&D effort of competitors. Their bureaucratic control of R&D ensures modest, predictable and incremental changes whereas inventions have the opposite characteristics. We shall not delve further into the intricacies of invention and melioration.

In this chapter, innovation or *Research and Development* (R&D) stands for all activities aimed at creating and improving products and services or at reducing the production and marketing cost; for instance, "Smart" is an innovative 2-doors car created by Daimler-Chrysler for city use whereas the "just-in-time" inventory management is a cost reducing innovation (cf. OECD for updates on Innovation).

We shall first discuss the value of innovation for society and how it is protected. We then study how and why firms spend in R&D spending in a variety of market structures. We assess optimality of using welfare as the measure of social desirability. An important

theoretical development is the chaining of R&D and standard market competition. We also show how R&D can become a strategic device for governments. The last section presents a wealth of cases of violations of Intellectual Property Rights (IPRs).

12.1 Social and Legal Matters

12.1.1 Social Value

Health related research, both public and private is obviously beneficial to the human kind, directly in terms of better quality of life and longevity but also indirectly by permitting a longer and more productive working life. Innovations like the Instant Noodles or Tetra Pak are desirable for society because they increase consumer satisfaction (and welfare). Technological advances such as radio, disc, movie, TV, CD, DVD, PC and the internet have change the way we entertain. Other innovations of a more technological kind reduce production costs:

- The *natural gas combined cycle* is an electricity turbine that doubles the energetic efficiency or halves the cost of producing electricity with respect to single fuel turbines.
- The invention of *dynamite* by Alfred Nobel enabled many public works by drastically reducing the cost of drilling mountains.
- The distillation of kerosene from petroleum and later of gasoline made air and road transport possible.[1@]

A few great inventions like penicillin came by surprise but most are the outcome of a long process starting with investment in human and physical capital (e.g., brains and laboratories); in other words they are very costly. DiMasi et al. (2003) in their study of R&D by the US pharmaceutical industry show that less than 1% of the molecules examined in the pre-clinical period make it into human testing and of these, only 20 % pass the development process and get approved by health authorities. The average R&D cost for drugs marketed in the late 1990s is $800 millions.[2@]

Since the limited human and physical resources of the economy have many useful alternative uses, we are warranted to study whether the levels of R&D expanded by private firms are efficient i.e., whether this activity should be supported or discouraged. One very important argument in favor of public subsidies for R&D other than health related ones is the positive externality generated by R&D upon growth in industrialized countries as attested by empirical investigation.

12.1.2 Legal Background

R&D is so costly that firms often need to cooperate to finance projects; it is also frequent to observe public agencies organize and finance networks of applied research among national firms to reach specific industrial goals. At the European level, the EC dedicates large amounts to bring forth international teams gathering researchers from the public and private sectors (cf. European Research Area). The case of fundamental research is not treated here; since this activity has no direct market value and is a public good, no private firm can find it profitable to finance it, so that financing rests entirely on the public power.

Now, the idea of cooperation being akin to collusion and anti-competitive behavior, we might worry that R&D generates more costs in terms of increased market power than benefits in terms of innovation. As we saw in §9, collusion may result in a lessening of competition and cause negative market effects with respect to prices, output, innovation or the variety and quality of products. On the other hand, R&D cooperation can be a means to share risk, save costs, pool know-how and launch innovation faster. In particular, for small and medium-sized enterprises, cooperation can be of great help to adapt themselves to the changing market place. d' Aspremont and Jacquemin (1988) study the descriptive as well as normative implications of collaboration in R&D in the presence of spill-overs: the fact that the individual research effort of one firm is useful to both firms because of a network effect (they belong to the same industry and both interact with public research). These authors globally find out that R&D is insufficient with respect to the efficient level which would fully internalize the network externality. If the spill-over is large enough then allowing cooperation in R&D but not in production increases total R&D and total production towards the efficient levels. A full normative assessment is more difficult because more cooperation (either none, only in R&D or at both stages) generates more R&D but tend to decrease output.

These arguments together with the positive externality on growth alluded before have been recognized by the EU in its competition law. The *Horizontal Agreement* is an arrangement between actual or potential competitors, operating at the same level of the production or distribution chain. The agreement can cover R&D, production, purchasing or commercialization.[3@] A *Block Exemption* is an authorization by the EC for certain types of agreements which are exempted from the prohibition on restrictive agreements (cf. §8.2.4). Block exemptions exist, for agreements of vertical integration, R&D, specialization, technology transfer and car distribution. Consortia between shipping companies for the joint operation of liner transport services benefit of a block exemption and are thus free to restrict competition within the common market and to affect trade between member states.

12.1.3 The Patent System

Copycats and Free riders

Consider the example of continuous innovation displayed on Figure 12.1: the first item is a traditionnal *glass* bottle, the second is a *plastic* bottle introduced in the 1960s, the thrid is a *svelte* plastic bottle designed to facilitate handling (1980s), the fourth is an *easy to dispose* and environmentally friendly model (1990s) and finally we have the *body shaped* model, easy to carry everywhere thanks to the incorporated handle. Each of these innovations is readily reproducible by a copycat as it is based on a simple idea that improves the service delivered to the consumer. For more complex discoveries, reverse engineering is necessary in order to discover how to copy the innovation (cf. DVD history 12.3.4).[4@]

Figure 12.1: Easily Imitable Innovations

Free riding occurs when an entrepreneur builds a large showroom and train her staff to demonstrate consumer appliances. Too often, visitors flock into the shop to test the products from world known brands but then go to buy at a nearby shop whose business strategy is to do without such costly niceties.

It is crucial to understand that, in the absence of protection for easily duplicable innovations or ideas, Bertrand competition "kills" innovation because the invention, like a public good, suffers from *free riding*. Indeed, when one firm innovates, either through a cost reduction or through a new product, other firms promptly copy the innovation so that the innovator enjoys monopoly profits for a short period only. Indeed, they soon compete on equal foot in a Bertrand fashion and are lead to price nearby their marginal cost so that their economic profit is almost nil. Recall that the economic marginal cost remunerates the capital employed in production but not the initial investment in R&D

because it is a sunk cost. Hence, the rational decision of any firm present in that market could well be to wait for someone to sunk the money necessary to come by with the innovation and then free ride on the innovator as quickly as possible. In equilibrium, investments into R&D are delayed and shrunk, the global pace and intensity of innovation is thus severely hampered.

The iPod success story is a reminder that such a fate is avoidable. The complementarity between the iPod hardware and the iTunes software has given Apple an edge that even experienced and wealthy competitors like Microsoft have fail to catch up with (cf. also §24).[5@]

Legal Protection

In such circumstances, there seems to be a market failure that keeps R&D from reaching an efficient level in the industry because innovation cannot be adequately rewarded. The *patent* system (cf. §12.3.1 for a precise definition) is a simple and quite efficient answer to this problem. As soon as a firm discovers a technical device, a molecule for a drug or a software code, it *rushes* to the patent office to apply for patent protection. If the application is accepted, the patent holder is bestowed a legal monopoly over the next 20 years[6@] for the use of the patented item; after that period the product has generally become obsolete.[7@] This patent race displays the "winner gets all" paradox because the first to obtain the innovation earns monopoly rents net of her R&D investment while the other firms who participated in the race earn negative economic profits. In this race, there is no simple (pure strategy) equilibrium because if one firms invests then everyone else should either invest nothing (not participate) or invest more in order to be the first to discover the innovation (on expectation obviously).

The long-run dynamic view of the R&D process is that challengers who are barred from imitation by the patent protection will try to develop an even better innovation. Hence, if she wants to maintain her rents, the patent holder must retain technological leadership which means that she has to keep investing and innovating. This idea of a contestable market somehow justifies the temporary monopoly granted by the patent protection because the continuous stream of innovation greatly increases consumer welfare. The analysis of this dynamic interaction is beyond the purpose of this book; let us nevertheless report the accepted conclusion:[8@]

> When all potential competitors stand on equal foot, aggregate expenditure on R&D is too large for two reinforcing reasons: there are too many firms and each invests too much.

As argued by Posner, shortening the patent term would reduce the rent dissipation

by reducing the revenue from a patent; it would also reduce the transaction costs of licensing, because more inventions would be in the public domain.

12.2 The Pace of Innovation

In this section, we study the incentives of firms to spend on R&D in various market structures: static monopoly, static competition and potential entry. Finally we address the issue of simultaneity in the innovation process.

12.2.1 Innovation in a static market

In this section, we assume that the market structure remains unaffected by the innovation process.

Monopoly

Following Arrow (1962), we analyze the incentives of a monopoly to invest into R&D. Consider a market for an homogeneous product with demand $D(p)$. If, thanks to an innovation, the marginal cost falls from $\bar{c}$ to $\underline{c}$ then the consumer surplus increases by $V^* = \int_{\underline{c}}^{\bar{c}} D(p)\,\mathrm{d}p$ as indicated by the two striped areas at the bottom of Figure 12.2.

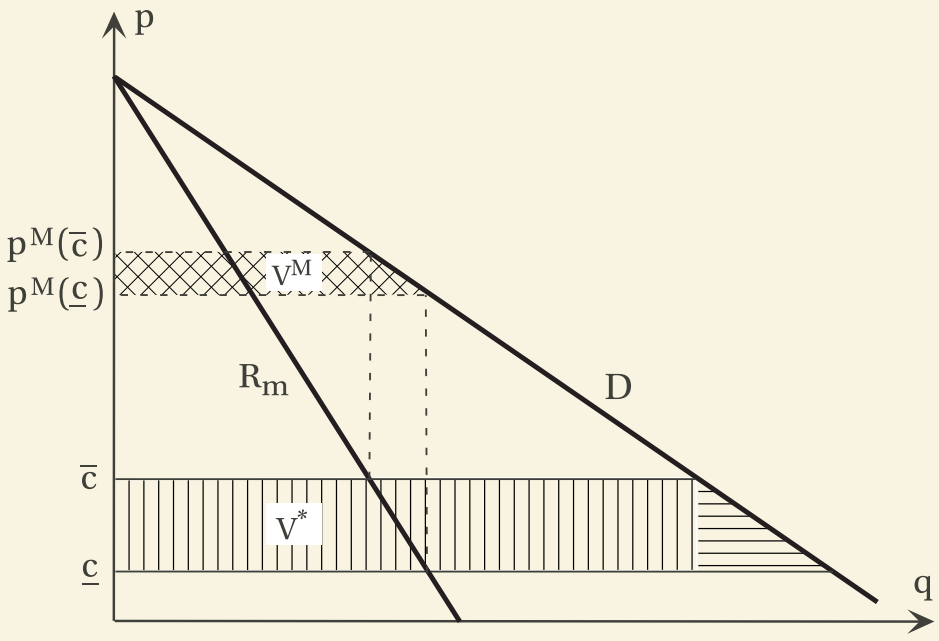


Figure 12.2: Replacement Effect

The value for society of developing the innovation is exactly V^* since in a competitive market the producer surplus is zero (both before and after the innovation development). If, on the other hand, the market is controlled by a monopoly, his profit $\Pi = (p-c)D(p)$ is maximized for p_c^M equating marginal revenue to marginal cost c. When the marginal cost falls from $\bar{c}$ to $\underline{c}$, the profit increases by some amount V^M, the grid area on Figure 12.2 which is clearly smaller than V^*.[9@] The intuition is clear-cut: for a small cost variation, the value of innovation is proportional to sales, whether the efficient ones $D(c)$ or the monopoly ones $D(p_c^M)$; thus the monopoly is less motivated to invest into R&D because his optimal production decision is much smaller.

> A monopoly underinvests into R&D because she values the innovation with respect to her limited sales while society values the innovation with respect to the larger efficient production.

Competition

Arrow (1962) also considers an initially competitive situation; he shows that the incentives to innovate for a competitive incumbent greatly improve upon the previous monopoly case. More precisely, consider an initial situation where Bertrand competition prevails, thus forcing all firms to price at their marginal cost $\bar{c}$ and earn zero (extraordinary) profit (cf. §5.2). When one firm innovates, her marginal cost falls from $\bar{c}$ to $\underline{c}$ and she enjoys profit $V^B \equiv (\bar{c}-\underline{c})D(\bar{c})$ because Bertrand competition among the other firms keeps the price at $\bar{c}$. As can be checked on Figure 12.2, V^B is equal to the vertically striped area which is V^* minus the small horizontally striped triangle.[10@]

> Under Bertrand competition, the value of innovation for a competing firm is lesser than the social value but much greater than the value to a monopoly.

The underlying reason for this discrepancy is the differing opportunity cost of a monopoly and competitive firm. In the competitive market, the innovator jumps from zero economic profits to almost monopoly profits while under monopoly, the monopoly is bound to replace himself, thus the "replacement effect" coined by Arrow (1962).

The previous sharp conclusion is severely weakened if firms compete in quantity. Indeed, for the linear demand $D(p) = a - bp$, we can use the Cournot model with asymmetric cost of §5.1.2 to assess the innovation value. The pre-innovation efficient output is $\bar{Q}^* = a - b\bar{c}$, thus the value for society is roughly $V^* = (\bar{c}-\underline{c})\bar{Q}^*$ (and half of this for a monopolist as she sells one half). In the symmetric pre-innovation duopoly, each firm earns $\frac{2}{9} \simeq 22\%$ of the welfare. The innovation bestows its holder a cost advantage which

turns into a higher profit because she sells more and enjoys lower cost. The innovation value, measured by the profit difference, simplifies into $V^C = \frac{4}{9}(a - b\underline{c})(\bar{c} - \underline{c})$.

Imitation

Protecting one's innovation from imitation is of the essence for a firm. Indeed, if all firms can adopt it in a regime of Bertrand competition, profits remain zero before and after which ultimately means that the innovation value if zero. Under the softer Cournot regime, the industry takes advantage of the innovation to sell more and earn more, thus the innovation value to a firm is her share of industry profit. In a symmetric duopoly, a firm earns $\frac{2}{9} \simeq 22\%$ of the welfare so that the individual value $V^C_{im} = \frac{2}{9}V^*$ is 22% of the social one, not a large figure but still more than zero. Using the innovation value when the implicit patent considered above, we can compute the WTP for the patent $V^C - V^C_{im} \simeq V^C_{im}$ i.e., about 22% of the social value of the innovation.

A corollary of these results is that the first welfare theorem does not apply wrt. innovation. Even when we start from the ideal situation of perfect competition (Bertrand), the incentives to innovate are lower than would be socially optimal. The reason is that innovation being costly, it must be protected by a patent to avoid imitation (and its deleterious effect). But, once the patent is in force, the competitive market gives way to a less efficient monopolistic competition. Indeed, either the innovation decreases the production cost of the patent holder and he can translate it into market power (our model) or the innovation allows a greater differentiation (either in quality or characteristics). In both cases, the competitiveness in the downward market is reduced.

Search

Another dimension through which a competitive market structure fosters more innovation than a monopolized one is the trial and error process aka. the search for profitable innovations. Indeed, in a competitive environment, external industry wide chocks are filtered out when comparing profitability with a challenger. An oligopolistic firm thus gets better signals regarding whether an innovation is working or not. The monopoly, on the other hand, lacks this benchmarking or yardstick ability. Since a firm active in a competitive environment is better able to figure out the usefulness of a new strategy or technology, she is more willing to try them. We thus observe more and faster innovation in competitive industries. As a side result, we obtain an additional reason why a competitive or oligopolistic industry is more efficient than a monopoly (beyond the deadweight loss seen in §3.2).

12.2.2 Patent Licensing

Our plan in this section is first to motivate licensing, then examine basic strategies such as royalties and license fees, compare them, combine them, look at the outsider/insider distinction, compare different kinds of royalties and finally deal with the inherent commitment problem of the patentee.

Introduction

In an era of global competition, firms cannot rely solely on their own R&D, they must plan the external sourcing of technology. At the same time, licensing their own findings becomes an additional source of revenue (cf. Zuniga and Guellec (2009)). We shall mostly deal with an independent laboratory that comes up with a cost reducing innovation, successfully patents it and designs an optimal licensing policy towards the members of the industry where the innovation can be used. An example would be the Dolby sound system for movie theaters. The basic strategies at the disposal of the innovator are selling licenses for a fixed fee, auction a limited number of licenses, set a per-unit or ad-valorem royalty or combine these basic strategies. The case where the innovator is an industry incumbent, aka an insider, is analyzed later on.

The ideal strategy for the patentee would be to sell one license, have the licensee expel the rest of the industry thanks to its drastic efficiency improvement, let him set the monopoly price and earn the monopoly profit (associated with the new technology) and then siphon this payoff by way of the upfront license fee. In any model of oligopolistic competition for the end product, this happens only if the new monopoly price is below the old marginal cost of the industry i.e., when the innovation is *drastic*.

Aside from this rather unfrequent case, the innovator is bound to earn less from her discovery. When considering basic strategies as exclusive alternatives Kamien and Tauman (2002) show that, in the absence of uncertainty, the auction prevails over fees and royalties. The empirical prevalence of royalties can then be explained by issues such as asymmetric information, uncertain quality of innovation, product differentiation, moral hazard, risk aversion, leadership structure or strategic delegation.

Royalties

The royalty system is the most popular mean of licensing an innovation. Yet, from a theoretical point of view, it is a tax that inefficiently reduces the total output sold by firms, thus reduces the industry profit and ultimately the payment the innovator can extract from the industry. This claim, as well as the comparison with fees and auction, is developed within a model of oligopolistic Cournot competition.

We consider a fixed market structure with n firms sharing the same technology characterized by marginal cost c. The innovation allows to reduce cost by δ. When the innovation is licensed through a (per-unit) royalty r, a licensee achieves a per-unit saving $\delta - r$. Thus, he is willing to license if and only if $r \leq \delta$ as he can still compete with the old technology. Under such a condition, all firms license and the new Cournot equilibrium is $q_r = \frac{a-c+\delta-r}{n+1}$. The patentee thus earns $\pi_0 = nrq_r$, a quadratic expression in r whose unconstrained maximizer (cf. §1.4.3) is easily found to be $r^* = \frac{a-c+\delta}{2}$. If $\delta \geq \bar{\delta} \equiv a - c$, we have a *drastic* innovation and $r^* \leq \delta$ i.e., the ideal royalty can be used. For non drastic innovations (the usual case), the optimal royalty is the maximum one, the cost saving δ. Notice that the price and consumers surplus remains the same. Welfare improves by the total cost saving but firms fails to take advantage of it since it all goes to the patentee who earns $\pi_0^R = \frac{n}{n+1}\delta(a-c)$.

Per-Capita vs. Ad-Valorem As recalled by Vishwasrao (2007) royalties are more often than not based on the revenue generated by the product incorporating the innovation, as opposed to being based on the number of units sold. Boccard (2010c) show that for both an outsider and an insider, ad-valorem royalties are preferred because they make licensing revenues increasing with the price which reduces the degree of competition in the market. This allows an outsider to extract more profit and an insider to strategically reduce her aggressiveness so as to sustain a higher price, a lower output and ultimately higher profits.[11@]

Fees & Auction

Under fee or auctions, any number m of firms can end up owning a license. Using eq. (5.14), the equilibrium with coexistence of licensed (L) and non licensed (N) firms is $q_{N,m} = \frac{a-c-m\delta}{n+1}$ for a non licensee and $q_{L,m} = q_{N,m} + \delta$ for a licensee. Observe that for $m \geq \bar{m} \equiv \frac{a-c}{\delta}$, non licensed firms are expelled from the market as their output is driven to zero.[12@] Let us study first the fee and auction regime under non exclusion i.e., over the domain $m \leq \bar{m}$.

Under a license fee regime, a potential acquirer must compare her equilibrium payoff in two situations: either she is one of m licensees and earn $q_{L,m}^2$ or she stays out and earns $q_{N,m-1}^2$ given that only $m-1$ firms have the license. The WTP for a license, conditional on m being bought in equilibrium, is thus

$$v_m = q_{L,m}^2 - q_{N,m-1}^2 \propto (a-c+(n+1-m)\delta)^2 - (a-c-(m-1)\delta)^2 = 2\delta^2 n\left(\tfrac{2\bar{m}+n+2}{2} - m\right) \quad (12.1)$$

Upon varying v_m the patentee can attract exactly m licensees and earn $\pi_0 \propto m\left(\frac{2\bar{m}+n+2}{2} - m\right)$,

thus the optimal number is $m^F = \frac{2\bar{m}+n+2}{4}$ (insofar as it is between n and $\bar{m}$ for otherwise the optimal value is whichever bound is hit first (cf. §1.4.3).).[13@]

When the patentee auctions m licenses, a potential bidder reasons that if he stays out, there will still be m licensees[14@] and he will earn less than when there are $m-1$ licensees as above (i.e., $q_{N,m-1} > q_{N,m}$). His WTP for the license is thus the greater $q^2_{L,m} - q^2_{N,m} \propto \bar{m} + \frac{n+1}{2} - m$ and the optimal number of licenses is $m^A = \frac{2\bar{m}+n+1}{4} \leq m^F$ (under the same proviso as above). Regarding profits, we have $\pi_0^F = m^F\left(q^2_{L,m} - q^2_{N,m-1}\right) < m^F\left(q^2_{L,m} - q^2_{N,m}\right)$ because $q_{N,m-1} > q_{N,m}$. Now, the latter expression is less than π_0^A by the very fact that m^A is an optimal choice. We have thus shown that auctioning dominates fees.

When the patentee decides to exclude non licensed firms from the market by setting $m \geq \bar{m}$, competition only takes place between advanced firms (with cost $c-\delta$), hence the more numerous the new industry is, the lower the industry profits, and the lower the patentee's revenue. It is thus never optimal to sell more than $\bar{m}$ licenses. The truly optimal number of licenses is thus m^A if $m^A < \bar{m} \Leftrightarrow n < 2\bar{m} - 1$ and $\bar{m}$ otherwise.

Lastly, observe that selling $\bar{m}$ licenses (with a fee or an auction), we obtain $q_N = 0$, $q_L = \delta$, and $Q = \bar{m}\delta = a - c$, thus $p = c$, as if there was perfect competition with the old technology. The patentee thus earns $\pi_0 = \bar{m}q_L^2 = \delta(a-c) > \pi_0^R$. This particular example shows why an ex-ante method is superior: it allows to reduce the degree of competitiveness of the market, thus sustain a higher profit that is siphoned through fees or auction. This completes the proof of our claim that $\pi_0^A > \pi_0^F > \pi_0^R$.

Optimal Combination

Once we allow the combination of basic licensing strategies, the optimal policy is to cartelize the industry using royalties with the entire industry as a price raising instrument in order to diffuse competition and allow large profits that are siphoned back through license fees.

Giebe and Wolfstetter (2008) study the optimal combination of royalties and auction for an outside laboratory. The optimal policy is to set high royalty to all so as to sustain a high equilibrium price and near cartel profit and then siphon part of this with the auctioning of licenses. A strategy for the patentee is thus to auction m licenses and set a royalty rate r_L for licensees and a larger rate r_N for non-licensees.

As we already saw for the pure royalty case, $\delta \geq r_N$ must hold in order to be effective; next we must have $r_N \geq r_L$ for otherwise no one would bid for a license.[15@] Letting $R_i = r_i q_i$ be the royalties paid by a type-i firm, operating profit is $\bar{\pi}_i \equiv \pi_i + R_i = (p - c + \delta)q_i$. The patentee profit is

$$\pi_0 \quad = \quad m\left(r_L q_L + \pi_L - \pi_N\right) + (n-m)r_N q_N = m(\bar{\pi}_L - \bar{\pi}_N) + n r_N q_N \tag{12.2}$$

$$= \quad m\bar{\pi}_L + (n-m)\bar{\pi}_N - n\pi_N = \bar{\pi}_0 - n\pi_N \tag{12.3}$$

where $\bar{\pi}_0$ is the industry operating profit, gross of any payment to the patentee. Notice that the formula is correct even when the patentee excludes some firms from the market. At this point, we notice that $\bar{\pi}_L - \bar{\pi}_N$ in (12.2) is exactly (12.1) which means that the patentee will make more profit than by auctioning licenses alone thanks to the royalties paid by non licensees. Royalties paid by licensed firms are only useful to tune the end market cost conditions suitably.

Meanwhile $p < c$ in equilibrium, non licensed firms are crowded out and the equilibrium is symmetric among licensees. By (5.12), raising r_L reduces total output but (unexpectedly) raises $\bar{\pi}_0$ because output is closer, though still larger, to the monopoly one when the innovation is not drastic. When r_L reaches δ, all firms are paying the royalty δ, thus have the same "old" marginal cost c i.e., $p \geq c$ must hold in equilibrium meaning that no firm is being excluded under the optimal policy.

Having proved that all firms are active, we see that raising r_N towards its maximum δ increases $\bar{c}$, thus simultaneously increases industry operating profit $\bar{\pi}_0$ (cf. argument above) while decreasing obsolete profit π_N by (5.14). Since this change unequivocally increases π_0, we must have $r_N^* = \delta$. The patentee profit thus simplifies into $\pi_0 = Q(a - c + \delta - Q) - nq_N^2$ with $Q \equiv mq_L + (n-m)q_N$ being market output. Working out (5.14) with $c_N = c$ and $c_L = c - \delta + r_L$, we obtain $q_N = \frac{a-c-z}{n+1}$ and $Q = \frac{n}{n+1}(a-c) + \frac{z}{n+1}$ where $z \equiv m(\delta - r_L)$ so that π_0 is a function of z only i.e., the optimal scheme is indeterminate.[16@] The FOC for optimality is

$$\begin{aligned} 0 = \frac{\partial \pi_0}{\partial z} \Leftrightarrow 2(Q^M - Q)\frac{\partial Q}{\partial z} = 2nq_N \frac{\partial q_N}{\partial z} \Leftrightarrow \frac{2(Q^M - Q)}{n+1} = -\frac{2nq_N}{n+1} \\ \Rightarrow \quad Q^M = Q - nq_N = \frac{n}{n+1}(a - c + z/n - a + c + z) = z \end{aligned} \tag{12.4}$$

Since the patentee sets $r_L < \delta$, we have $\bar{c} < c$ i.e., the industry average cost falls; this means a greater output and a greater welfare (as payments to the patentee do not matter). It might even be the case that the royalty for a licensee is negative i.e., firms are subsidized to use the innovation in the end market. Indeed $r_L < 0 \Leftrightarrow (n-1)\delta < Q^M \Leftrightarrow \delta < \frac{a-c}{2n-3}$ i.e., the innovation is small.

Insider Licensing

We now study the optimal licensing policy of a patentee that is also an incumbent of the industry where the innovation is used. An example would be Sony with the Blu-Ray technology. Kamien and Tauman (2002) show that for a large oligopoly or equivalently

for a large innovation, there is a reversal in the sense that royalties now dominate auctioning.

The revenue for an incumbent of auctioning licenses does not change wrt. the outsider case when n is large. Indeed, the optimal number of licenses is $m^A = \bar{m}$ which amounts to exclude non licensees from the end market. The price paid for the license is thus the operating profit of an incumbent firm (since her outside option is now zero). This means that one earns likewise as an insider getting a free license or as an outsider auctioning the license. From, $q_N = 0$, we deduce $q_L = q_N + \delta = \delta$, thus the auctioning profit is $\pi_I^A = \bar{m} q_L^2$.

The royalty scheme must be studied anew when the patentee is also an incumbent. We already know that a higher rate r increases her cost advantage over the rest of the industry, thus raises her output q_L while lowering that of challengers q_N; this means that her total profit $\pi_I^R = q_L^2 + (n-1) r q_N$ unequivocally rises, thereby making $r = \delta$ optimal. Using (5.14), we compute $q_L = \frac{a-c+n\delta}{n+1}$ and $q_N = q_L - \delta = \frac{a-c-\delta}{n+1}$, thus the incumbent innovator earns $\pi_I^R = \frac{\delta^2(\bar{m}^2+\bar{m}(n^2+2n-1)+1)}{(n+1)^2} > \pi_I^A \Leftrightarrow (\bar{m}-1)^2 > 0$. This proves the claim that royalties dominate the auction for n or δ large.

Regarding incentives to invest into R&D, an outsider having a zero opportunity cost has always greater incentives than an insider who already earn a Cournot profit.

Commitment

The licensing of the patented technology is prone to the hold-up problem (cf. §14.2). Having granted n licenses, the patent holder is always tempted to sell further licenses (at a lower fee) which thereby depreciates the value of the existing n licenses since the original licensees now face more competition than expected.[17@]

Such an expropriation is ex-post profitable for the licensor, but reduces his ex-ante profit. Indeed, if many downstream firms hold a license, intense competition destroys their profits, thus lowers their ability to pay the license fee in the first place. Therefore, a patent holder would like to promise that he will emit a limited number of licenses.[18@]

A similar point can be made for the franchising of a brand. Franchisees are unlikely to pay much to franchisors if they do not have the guarantee that competitors will not set shop at their doorsteps. The phenomenon at play here is *lack of commitment* similar to that of the durable good monopolist who cannot refrain from lowering its price as time passes (cf. §4.3.5). A solution to excessive licensing might be to integrate vertically with a firm competing in the product market.[19@] Let us compute the optimal strategy between licensing to all competitors and integrating with one of them in a simple model.

Imagine that an independent laboratory has develop a method to reduce the marginal cost of production from c to 0. If the patent holder integrates with one downstream firm, she achieves a cost advantage; in that case, she will not be tempted anymore to sell

additional licenses to firms that compete with her. If on the other hand, she does not integrate and starts to sell patents then, as we argued previously, she will sell licenses to all n competing firms present in the market. Thus, Cournot competition takes place with different cost structures.

We consider the simples demand is $D = 1 - p$. In the "all licensing" case, the Cournot equilibrium profit is $\pi^c = \frac{1}{(n+1)^2}$ (cf. §5.1.3). The patent holder can therefore asks a fee $F = \pi^c$ and nets a total profit $\pi^{lic} \equiv \frac{n}{(n+1)^2}$ (this is a two part tariff rent extraction). The patentee would be better off issuing a single license if she could commit to it. If not, then she shall license to all n competitors in the downstream market. In the "exclusive licensing" case, the patent holder integrates with one firm and thereby achieve the necessary commitment. From that moment on, it is rational for her to refuse to license to others because this would lower her future profits in the downstream market. Competition thus involves an asymmetry of costs i.e., a unique leader and many followers. Applying the general result of eq. (5.14), we have $\bar{q} = \frac{1+(n-1)c}{n+1}$ for the licensee and $q = \frac{1-2c}{n+1}$ for the other firms, thus the aggregate quantity is $Q = \bar{q} + (n-1)q = \frac{n-(n-1)c}{n+1}$. Observing that the equilibrium price is $\bar{q}$, the profit of the leader is simply $\pi^{int} \equiv \bar{q}^2 = \frac{(1+(n-1)c)^2}{(n+1)^2}$. The comparison we were looking for is $\pi^{lic} > \pi^{int} \Leftrightarrow n < \bar{n} \equiv \left(\frac{1-c}{c}\right)^2$. To conclude:

> Licensing dominates integration only when there is much market power but it is less likely to happen if the innovation is drastic (large cost reduction).

12.2.3 Innovation Race

Arrow (1962)'s results are criticizable because they fail to account for the Schumpeterian process of creative-destruction whereby old and inefficient incumbent monopolies are displaced by young and innovative entrants.

Entry Threat

Once we consider the threat of entry, the story change dramatically. The simple analysis of Gilbert and Newbery (1982) shows that the willingness to spend on R&D of a monopoly is greater because he stands to lose much; a strategic analysis even reinforces this conclusion with hints towards an anticompetitive behavior of preemption.

Consider an incumbent monopoly (he) and a challenger (she) ready to enter the market. If the monopoly innovates first, he will patent the innovation and most likely bar the challenger from entering. Indeed, the latter has a high cost technology, thus upon entering she will either earn zero profits if there is Bertrand competition or a low profit if

there is Cournot competition. If her sunk cost of entry is greater than this level, entry is barred. The monopoly therefore enjoys the large profit $\Pi_{\underline{c}}^M$ after having innovated first.

The interesting case is when the challenger is first to innovates; she will then enter because she has a cost advantage. If there is Cournot competition, profits are respectively Π_i^C and Π_e^C for the incumbent and the entrant with $\Pi_i^C < \Pi_e^C$. We can now assess the value of being first to innovate (and patent). This value is $V^i \equiv \Pi_{\underline{c}}^M - \Pi_i^C$ for the monopoly and $V^e \equiv \Pi_e^C$ for the challenger. We saw in our analysis of oligopoly that producer surplus (aka industry profits) decreases with the number of active firms i.e., $\Pi_e^C + \Pi_i^C \leq \Pi_{\underline{c}}^M$. We can thus conclude[20@] that

> The value of innovation is larger for the monopoly than for the challenger once we account for entry in the innovation process.

How does the greater willingness to spend on R&D of the incumbent relative to the challenger impinges on their respective behaviors? It is clear that the more you spend on R&D, the quicker you will develop the innovation, yet the connection need not be linear because luck has a role to play. It is nevertheless useful to start with a deterministic relationship to immediately see that the incumbent will preempt the innovation. Indeed, the challenger won't spend more than V^e on it,[21@] while the incumbent's own limit is the greater V^i, thus he can surely beat her by spending slightly more and this leave him with a net profit of $V^i - V^e = \Pi_{\underline{c}}^M - \Pi_i^C - \Pi_e^C$ which is the loss in producer surplus generated by entry.[22@]

Policy Implications

From an efficiency point of view, the innovation will take place at the right time (as if a competitive challenger would innovate) but since the monopoly will have succeeded to block entry, there is a welfare loss due to his pricing conduct.

The policy interpretation of this simple result is that the monopoly may try to erect a strategic barrier to entry by *over-investing* into R&D to secure the best technology and accumulate patents; this attitude is popularly known as building a "thicket" or "bucket" of patents. The broader reading of this preemptive behavior reads as follows: if entry occurs, the large technological advantage of the incumbent will force the entrant to differentiate either to a low quality or to compete with a higher marginal cost. As we previously saw in the corresponding models, the entrant's economic profit is low in both cases. This means she won't be able to recover her sunk cost of entry (her own R&D investment), thus she will rationally abstain from entering, thereby leaving the incumbent free to price at monopoly levels and enjoy monopoly rents.

Observe also that developing a new technology (not a cost reduction as in our model) or acquiring a patent does not necessarily mean that the monopoly will bring it to the market; in other words, a patent might be "sleeping". The innovation comes to the market with certainty only when entry actually takes place for otherwise it might be cheaper or more profitable to continue exploiting the good-old fashioned technology. The strategic barrier to entry we have just commented here is fully legal although it is inefficient since resources are wasted (from a social point of view) to protect a monopoly position.

12.2.4 Market Rivalry

In this section, we show how investments into a better technology (lower cost or better product) impinge on the market competition and how the intensity of the later feeds back the incentives to invest in the first place.

Why innovate?

We explain in §3.3.1 on quality that a consumer's WTP for an item depends on its perceived quality, whether a good or a service. It is furthermore shown that product development such as market research, innovative R&D, or advertising is akin to a reduction of the production cost (of the item under consideration). This cost reduction can proceed through many avenues such as improving management or customer care or simply updating the production technology, either by purchasing an advanced (patented) technology or through in-house R&D. In any case, the cost reduction activity displays decreasing returns to scale.[23@]

We thus use "R&D investment" as a generic name for all the possible strategies that firms may use to reduce their cost or improve their product quality. Our aim is to study the determination of the optimal level of R&D and embed this issue in models of market competition to understand how the intensity and mode of rivalry affect the choices of individual firms.

The *direct* effect on profits of a marginal cost saving is proportional to current output i.e., from $\pi = (p - c)q$ we deduce $\Delta\pi = q\Delta c$. This technological melioration also makes the firm look tough (hurts other firms), whatever the competition mode. The cost saving can also display a *strategic* effect if there exists a relationship between the equilibrium quantity q^* and the marginal cost c. This is not the case under monopoly, perfect competition or price-setting regulation because there is no strategic interaction between firms and thus no indirect effect of a cost reduction. When the innovator competes in an oligopoly framework, there is a *strategic* effect because competitors respond to the firm's change of technology. As shown in §6.2.4 on business strategies, quantities are strategic substi-

tutes while prices are strategic complements and since the cost saving makes the firm tougher, the strategic effect is positive in the Cournot case but negative in the Bertrand one.[24@] We show below that this dichotomy remains true though weakened. Cournot competition reinforces the natural willingness to spend on R&D by 33% in duopoly and up to 100% in large oligopolies while Bertrand competition reduces it by 33% in duopoly but no more than 15% in large oligopolies.

The one technical assumption common to the various model exposed below regards R&D and is borrowed from Brander and Spencer (1983). We assume existence of a patent free technology whose constant marginal cost is c. To reduce her marginal cost by some amount x, a firm invests into R&D an amount $\psi(x) = \lambda x^2$ for some parameter λ that might be firm dependent. This way, the cost reduction activity displays DRS.

Quantity competition

Given R&D effort x_i, marginal cost is $c_i = c - x_i$ for $i = 1,2$. As we saw in §5.1.2 on Cournot duopoly competition, the profit of firm i is $\pi_i = \frac{1}{b}q_i^2$ where the equilibrium quantity is (5.10)

$$q_i = \frac{1}{3}(a - 2bc_i + bc_j) = \frac{1}{3}\left(a - bc + 2bx_i - bx_j\right) \tag{12.5}$$

The marginal benefit of innovation is thus

$$\frac{\partial \pi_i}{\partial x_i} = \frac{2}{b}q_i\frac{\partial q_i}{\partial x_i} = \frac{4}{3}q_i = \frac{4}{9}(a - bc + 2bx_i - bx_j) \tag{12.6}$$

which is 33% greater than the direct effect q_i as we claimed above. For an oligopoly with n firms, it is enough to observe that the $\frac{2}{3}$ coefficient in the duopoly equation (5.10) is replaced by $\frac{n}{n+1}$ of (5.14) in the oligopoly case. Thus coefficient $\frac{4}{3}$ in equation (12.6) is replaced by $\frac{2n}{n+1}$. In a large oligopoly, the ratio tends to 2 meaning that firms have an indirect investment effect equal to 100% of the direct one (cf. §26.1.5 of the appendix for the complete analysis).

To simplify algebra in the duopoly case, we introduce parameter $\sigma_i \equiv \frac{9}{2b}\lambda_i - 2$ in the R&D cost function; this way $\psi_i(x_i) = \frac{2}{9}(\sigma_i + 2)bx_i^2$ and the marginal expense of reducing the production cost is thus $\frac{\partial \psi_i}{\partial x_i} = \frac{4}{9}(\sigma_i + 2)bx_i$.

We can now characterize the optimal R&D investment by solving the FOC

$$\frac{\partial \pi_i}{\partial x_i} = \frac{\partial \psi_i}{\partial x_i} \Leftrightarrow a - bc + 2bx_i - bx_j = (\sigma_i + 2)bx_i \tag{12.7}$$

The best reply of firm i against the choice x_j of her rival is $x_i = \frac{a - bc - bx_j}{b\sigma_i}$. This best reply makes sense only if $\sigma_i > 0$ i.e., λ_i is large enough. We observe that the strategic substi-

tutability of quantities under Cournot competition translates into the R&D stage: the more my opponent invest to reduce his cost, the less I will invest myself (cf. §6.2.4) . Given the symmetric formula for firm j, the equilibrium of the R&D game can be computed as

$$\hat{x}_i = \frac{a-bc}{b}\frac{\sigma_j - 1}{\sigma_i \sigma_j - 1} \tag{12.8}$$

This result makes sense only if the σ parameters are both greater or both lesser than unity (equivalently that the λ's are large) i.e., cost reduction are either hard or easy to achieve for both firms, which we shall assume. We notice that equilibrium R&D increases with market size (a) but decrease with the elasticity of demand (b).

For identical R&D technologies ($\sigma_i = \sigma_j = \sigma$), we obtain $\hat{x} = \frac{a-bc}{b(\sigma+1)}$, thus $\hat{q} = \frac{a-bc}{3}\frac{\sigma+2}{\sigma+1} > q^C = \frac{a-bc}{3}$, the output under no innovation at all. We compute and $\hat{\pi} = \frac{1}{b}\hat{q}^2 - \frac{2}{9}(\sigma+2)b\hat{x}^2 = \frac{(a-bc)^2}{9b}\frac{(\sigma+2)\sigma}{(\sigma+1)^2} = \pi^C \frac{(\sigma+1)^2-1}{(\sigma+1)^2} < \pi^C$.

> The ability to innovate is taken on by firms who thus reduce their cost and produce more than previously. This is welfare increasing on counts of allocation (greater consumption) and efficiency (lower unit cost for the industry). Firms however are trapped in a prisoner's dilemma since payoffs are greater for both when innovation is not an option.

The problem faced by firms is that each has an incentive to be the first to innovate. Yet they end up investing too much i.e., more than the return they get from it.

Price competition

To assess Bertrand competition, we use the Hotelling model of duopoly price competition to avoid the difficulties associated with the Bertrand paradox. As shown in footnote 26.3, when marginal cost differ, the equilibrium price charged by a firm is $p_A^* = t + \frac{2c_A + c_B}{3} = t + c - \frac{2x_A + x_B}{3}$ with sales $q_A = \frac{1}{2} + \frac{x_A - x_B}{6t}$ and profit $\pi_A = 2tq_A^2$. The total effect of a cost reduction is then

$$\frac{d\pi_A}{dx_A} = 4tq_A\frac{dq_A}{dx_A} = 4tq_A\frac{1}{6t} = \frac{2}{3}q_A \tag{12.9}$$

which represents a 33% decrease over the direct one. The extension of the Hotelling model to the circular city oligopoly (cf. §11.1.3) allows to perform a similar analysis. In §26.1.5 of the appendix, weshow that investment incentives are reduced by 20% for n=3 and to no more than 15% in the large oligopoly limit (cf. for the complete analysis).

We can now solve for the equilibrium of the R&D investment. Assuming that the same cost of investment $\psi(x) = \lambda x^2$ for both firms, the best reply solves $\frac{2}{3}q_A = 2\lambda x_A$. In a symmetrical equilibrium $x_A = x_B = x$ must hold, thus $q_A = q_B = \frac{1}{2}$ from which we deduce

$x^* = \frac{1}{6\lambda}$ and the final price $p = t + c - x^*$. Profit is then $\pi = t/2$. In this simple model, profits are not impacted by the R&D efforts but since consumer enjoy a lower price, their surplus increases and so does welfare.

Timing of Innovation

A simple setting enables to draw interesting conclusions with respect to the timing of the decision to perform R&D. Imagine that each firm can engage into R&D at a fixed cost F; the probability of making a discovery is α and the market value of the discovery is V.

The expected payoff for a monopoly is $\pi_1 = \alpha V - F$ and it is rational to undertake R&D if $F \leq \alpha V$. Now, if there are two firms that can possibly engage into R&D the technological uncertainty is supplemented by a market uncertainty since the other firm may also discover the innovation in which case the market prize (cake) V would have to be shared equally.[25@] The expected payoff for one firm is thus $\pi_2 = \alpha(1-\alpha)V + \frac{\alpha^2}{2}V - F$ so that R&D is undertaken by both only if $F \leq \left(1 - \frac{\alpha}{2}\right)\alpha V$. For an intermediate R&D cost, one firm can successfully engage into R&D but not the other one. The innovation game between the two firms is one of coordination with a unique equilibrium up to the identity of the firm who does R&D.

From a welfare point of view, we may consider maximizing the sum of profits. In that case we have to compare $W_1 = \alpha V - F$ the welfare if a single firm does R&D (we neglect consumer surplus) and $W_2 = 2\pi_2$ the welfare when both firms do R&D. The cut-off is $W_2 \leq W_1 \Leftrightarrow F \leq (1-\alpha)\alpha V$. We may thus conclude that individual decisions are efficient except for $(1-\alpha)\alpha V \leq F \leq \left(1 - \frac{\alpha}{2}\right)\alpha V$ i.e., the low cost of R&D push both firms into R&D while it would be optimal to have a single one doing it.

12.2.5 Public Subsidies †

As we already mentioned public subsidies for private research are frequent and involve large amounts. The US, supporting Boeing (B), and the EC, supporting Airbus (A), came to an agreement in 1992 according to which governments can back up loans up to one third of the development cost of a new aircraft. In October 2002, the US government accused the EU member countries to violate this agreement by lending to Airbus on better terms than the company could get from a commercial bank. The EC responded that Boeing cross subsidizes civil airplane development with military contracts. This bitter dispute after reaching a mutual menace of WTO litigation was settled late 2004 by an agreement to dismantle the existing subsidies in the near future.

The underlying reality of hidden subsidies can be understood using the model of Brander and Spencer (1983) who show quite clearly the distortion of competition that this

subsidizing behavior can create. As we saw in §12.2.4 on innovation and market rivalry, firms competing in Cournot fashion have incentives to invest into R&D and reach an equilibrium level of R&D (cf. eq. (12.8)) characterized by own technology parameters σ_i and σ_j. The equilibrium profits net of R&D cost can be simplified into

$$\hat{\Pi}_i = \Pi_i(\hat{x}_i, \hat{x}_j) = \frac{(a-bc)^2}{9b} \frac{(\sigma_i+2)\sigma_i \left(\sigma_j - 1\right)^2}{\left(\sigma_i\sigma_j - 1\right)^2} \tag{12.10}$$

The symmetric situation where $\sigma_j = \sigma_i = \sigma$ is the most natural to start with; it occurs when firm finance R&D by leveraging funds on the world market at the same conditions. In that case, the common equilibrium level of R&D is $\hat{x} = \frac{a-bc}{b(\sigma+1)}$.

It can be checked that the equilibrium level of R&D is greater than the level minimizing cost because improving one's technology forces the competitor to reduce production and therefore adds an additional benefit. In that sense, the authors speak of over-investment in R&D. Yet, the truly efficient level of R&D has to maximize welfare net of the R&D investment. More precisely, welfare here is that generated in a competitive market where the good is sold at marginal cost; since marginal cost is constant, profits are nil thus the welfare reduces to consumer surplus computed at the marginal cost. Using the formula seen in §2.2.2, the consumer surplus at price c_i is $W_D(c) = \frac{(a-bc_i)^2}{2b}$. Hence, the welfare net of R&D cost is

$$W(x) = \frac{1}{2b}(a - bc + bx)^2 - \frac{2}{9}(\sigma+2)bx^2;$$

hence the optimal value is $x^* = \frac{9(a-bc)}{b(4\sigma-1)} > 2\hat{x}$ (check this using cross product), the amount expanded by the industry in the Nash equilibrium. We can conclude that the competition leads firms to invest sub-optimally into R&D.

We now introduce governmental subsidies by assuming that each firm comes from a different country that is solely interested by the profits its national champion can gather on the world market and bring back home. To subsidize R&D, a government can offer a tax rebate that is to say, profits invested in R&D become deductible from income tax. In our model, this policy amounts to reduce the cost factor down to $\sigma_i < \sigma$. Now observe that $\hat{\Pi}_i \propto \lambda \equiv \frac{(\sigma_i+2)\sigma_i}{\left(\sigma_i\sigma_j-1\right)^2}$ (cf. eq. (12.10)), so that $\frac{\partial\hat{\Pi}_i}{\partial\sigma_i} \propto \frac{\partial\lambda}{\partial\sigma_i} = -2\frac{\sigma_i+\sigma_i\sigma_j+1}{\left(\sigma_i\sigma_j-1\right)^3} < 0$. This derivative means that government i has an incentive to subsidize the R&D of its champion by decreasing the σ_i factor; in response to this fiscal incentive, firm i will increase its R&D $\hat{x}_i$ (check in eq. (12.8)), become a tougher opponent in the world market and earn more profit.

The optimal support σ_i^* is difficult to characterize because we only known the benefits of subsidization while we ignore the cost of the public funds that are lost due to the

subsidy (cf. §17.1.2). Nevertheless, government j, being himself rational, will react by subsidizing its own champion so that the overall levels of R&D will increase.

Consider two firms pondering an investment into R&D aiming at reducing their production cost and become better contestants in the ensuing market competition. The first result of their model is that market competition "à la Cournot" leads firms to invest sub-optimally into R&D.

Governmental subsidies are then introduced by assuming that each firm comes from a different country that is solely interested by the profits its national champion can gather on the world market and bring back home. To subsidize R&D, each government can free from income tax, profits invested back in R&D. The second result is that each government has an incentive to subsidize his champion although the optimal level is difficult to characterize because we only known the benefits of subsidization while we ignore the cost of the public funds that are lost due to the subsidy. Unless this governmental race overshoots the efficient level alluded before, the final amount of R&D is still insufficiently low which is in accordance with the findings of most empirical studies. Our last conclusion will therefore be that we should not bother for the EU-US dispute since these hidden subsidies bring us better planes and improve our travel experience.

12.3 Intellectual Property Rights

Technological innovation in the form of a less costly production technology or a new product, is at the root of all material progress in human societies. It is the random outcome of a very costly investment into R&D. When successful, the innovation bestows an advantage in quality or cost upon its creator who becomes a better competitor, frequently a monopoly; the innovator is thus able to obtain extraordinary profits that enable him to recoup his initial R&D investment. As we saw in the chapter on R&D, innovations that are easily copied must be protected by legal means to maintain the incentive to invest into R&D.

This protection issue was long recognized by governments of industrialized countries. To encourage innovation they issue patents that protect creators over the domestic market and enable them to earn monopoly rents for some period in order to recoup the cost of their R&D investment. The remedy has proved successful since some industries like pharmacy do not hesitate in spending as much as a quarter of there income into R&D.

More broadly we are interested in this chapter by Intellectual Property Rights (IPRs) such as patents and trademarks; they give their developer certain exclusive rights over the exploitation of their work and Know-How.[26@] An example will help to grasp the very idea of intellectual property. To access the Nokia discussion forum on mobile phones, one

must become a member and accept the terms among which the “License to Nokia for Any Submitted Content” stating (in a single sentence):

By submitting any information or materials such as feedback, data, text, software, music, sound, photographs, graphics, video, messages, answers, questions, comments, suggestions, scores, hints, strategies, concepts, designs, ideas, plans, orders, requests or the like, or any other material (Content) to Nokia, you license and grant Nokia and its affiliates and sub-licensees a non-exclusive, royalty-free and free of charge, perpetual, worldwide, irrevocable, and fully sub-licensable right to use, reproduce, modify, adapt, communicate to the public, make available, publish, translate, copy, modify, adapt, create derivative works of, distribute, and display such Content or any concept described in it throughout the world in any media, product and/service, including, without limitation wireless devices, mobile phones and any related products, services and accessories, advertising, marketing and promotional materials, and digital reproductions, without compensation, restrictions on use, acknowledgement of source, accountability or liability, and with waiver of all moral rights and rights of attribution, integrity and identity.

12.3.1 Patent

A patent is a set of exclusive rights granted by a government to an inventor for a man-made, technical creation with the purpose of encouraging the development of new inventions and their disclosure.[27@] Patents for new drugs are probably the most famous ones although it is not always the molecule that is patented as in the case of *aspirin* where the patent was awarded to Bayer for the industrial process of synthesizing the drug.

The first[28@] patent law is the “Inventor Bylaws” awarded by the Venetian Republic in 1474. England introduced its “Monopoly law” in 1624 and other industrialized countries followed in the next century. Today, a patent confers for 20 years[29@] the right to exclude others from making, using, selling, offering for sale, or importing the patented invention. As an alternative to legally monopolizing the market, the patentee can license the patent to other firms; in any case, the investment made during the development of the invention can be recouped. Three conditions must be fullfilled to obtain a patent.

- *Novelty*: nobody else must have invented it before; the patent office checks this by performing a search of “prior art".
- *Non-obviousness*: the invention must not be an obvious answer to a problem for someone who is skilled in that technological field.
- *Utility*: the invention must be useful, have a practical application.

After the expiration of the 20 years monopoly period, anyone is free to practice the invention and since disclosure is mandatory to obtain the patent, this is very easy to do. Thus, the patent system can be seen from two perspectives, either as a way to reward individualistic innovators or a way to promote the diffusion of innovative knowledge in the whole society. The patent system is nevertheless the object of much debate. Regarding

the protection provided by a patent, Arundel (2001) shows empirically that secrecy and lead-time (being first to market a new idea) are viewed by inventors, especially small firms, as a better protection from imitation.

The information technologies are generating a problem around novelty and scope of patents. Owners of some old patents claim they cover some of the most widely adopted practices of the internet like hyperlinking,[30@] graphical formats,[31@] or software code as in the most influential case where SCO Group, inheritor of some elements of the Unix operating system, sued IBM in 2003 claiming that IBM's Linux misappropriated the Unix code belonging to SCO.[32@] More recently, the US patent office has awarded patents for the translation of old business methods to the internet like shopping with a virtual cart,[33@] buying with a virtual credit card,[34@] or recommending items.[35@]

Jaffe and Lerner (2004) argue from their econometric analysis that the intensification of lawsuits is due to regulatory reforms from the 1980s. They converted the US patent system from a stimulator of innovation to a creator of litigation and uncertainty that threatens the innovation process itself. Indeed, firms fear to introduce products using well known devices or computer code because someone has succeed to patent it and is threatening to sue those who would not ask for a licence, which is likely to be expensive. As a result, development and production costs are increased. In a reaction to the SCO-IBM case, large software makers like Microsoft, Sun, Kodak or Novell recently announced they will cover the cost of litigation whenever one of their customers is sued because some products he bought contain copyright, patent, trademark or trade secret infringement.

The EU is better protected against these misuses because the European Patent Office issues a public notice of information before granting the patent; this way, other parties can offer information relevant to the determination of novelty. The proposed new directive on Patents seek to avoid any drift towards patents for business methods or computer programs which do not provide any technical contribution to the state of the art. The US system lacks this feature and is furthermore financed by the fees of applicants so that frivolous patents are frequently issued in a clear case of regulatory capture. The US also use a "first to invent" rule contrary to the "first to file" rule used in the rest of the world; this generates a lot of legal wrangling about priority.

12.3.2 Copyright

Copyright is the right to copy and publish (in a broad sense) a particular literary or artistic original work. It also covers copying in electronic form, the making of translated versions, the creation of a television program based on the work, writing computer pro-

grams, and putting the work on the Internet. The copyright protects only the expression of an idea, so it is legal to take the plot from somebody else's mystery novel, and write your own mystery novel based on that plot. Most countries, including the EU and US, have adhered to the Berne Convention[36@] granting automatic copyright protection for the artist's lifetime plus 70 years upon creation of the work. In some countries, registration with a Copyright Office has some benefits like being able to sue, or to get bigger damages if you win.

Copyright laws let authors decide how, where and when to exploit the works they created. Usually, works are sold in the form of copies on tangible media, such as books or compact discs. The author of a work can request a fee for every copy made of his work. He can also restrict the number of copies being made, for example to keep the work exclusive.

Copyright protection is much longer than the patent one and generally involves a smaller if not a zero investment; it is supposed to be an economic incentive for the *next* producer of art, not a guarantee for the *established* one. A prime element of explanation for this difference with patents is the greater moral value put upon intellectual creations in comparison with technical ones, the second one is the necessity to avoid the congestion that would result for the diffusion of the popular works if they were free of royalties.[37@]

The copyright industries involve press, publishing, music, broadcasting, movies, theater, advertising and computer software; they contributed 5.3 % of the 15 EU GDP and 3.1 % of total EU employment in 2000. Similar figures obtain for the US economy. More information is available on the EC internal market website.

12.3.3 Trademarks

A trademark is an exclusive right giving its holder the right to exclude (or stop) others from using the mark. It is mainly aimed at preventing unfair competition; it helps to identify good and services and to create a good image (in legal terms goodwill). Many counterfeit copies of trademarked goods enter the market in the hopes of confusing potential buyers. Because such counterfeit copies are typically of inferior quality with respect to the original product, they can seriously harm the value of the trademark and the image created. This explains why many trademark holders are very active in prosecuting copy cats and destroying counterfeit goods.

When a trademarked product lawfully enters the market (e.g. because the trademark holder manufactured it and sold it in a store), the buyer may want to resell it, either on the same market, or at some entirely different market in a different country or region. As selling a trademarked product is use of the trademark in commerce, the trademark

holder could block this resale by claiming trademark infringement. As this is not always fair, various countries have developed an "exhaustion" or "first sale" doctrine regulating when a trademark holder can and cannot act against a reseller of 'his' products.

The universality principle states that if the trademarked product legally entered the market somewhere, reselling the product anywhere in the world is not infringement. However, it ignores countries legal differences and was never applied for historical reasons (protectionism). The converse approach, known as the territoriality principle states that a trademark is legal only in the country where it is registered.

In the USA, the *first sale* doctrine states that the trademark holder has full control until the first authorized sale within the USA. He cannot act against domestic resellers of products he put on the market himself or if the product was manufactured by a third party with his permission (a licensee). To sell Levi's jeans in New York, one can buy them anywhere in the US but not in Brazil unless the Brazilian trademark "Levi's" is held by the same company as the US trademark. Still the US trademark law can be used to bar this parallel import on the ground that the foreign product would be considered different from the domestic product by the American consumer (different quality).

In the European Union, the Trademark directive supports a similar idea called the exhaustion principle: when trademarked products are put on the market in a member country with the trademark owner's consent, he cannot later oppose the import or re-import of these products into another country in the EU. Furthermore the European directive do not permit member countries to have more liberal exhaustion laws, for example by allowing worldwide exhaustion.

The Trademark directive has been refined by the ECJ in the Silhouette judgement of 1998. Contrary to the US view, to be able to sell products bought outside the EEA into the EEA, the importer must have the unequivocal consent of the trademark owner. So, even when a trademark holder sells a product in Brazil without any restrictions, he can later block parallel import of the product at the European borders. But if he put them on the market in Belgium, he can't block parallel import to the Netherlands, nor a re-import from the Netherlands back to Belgium.

12.3.4 IPR Cases

Trade Secrets Litigation is an alternative way to seek profits since a court victory can entail a considerable market advantage. It is customary to see firms mutually threatening themselves with legal actions.

The *Barbie* vs. *Bratz* series of lawsuits from 2004 to 2010 is a case in point. Under most legal systems, all creation made by an employee within a firm belongs to that firm.

Hence, when a person changes employer within the same industry, there are frequently lawsuits initiated by the former employer against the new employer relative to the illegal transferring of intellectual property by the defector.

Mattel, the world's largest toy maker, has seen sales of its famous Barbie dolls suffer from competition by the edgier, urban-styled Bratz line which MGA launched in 2001.[38@] In April 2004, Mattel sued the doll designer for secretly working with MGA while still employed by them (back in October 2000); the latter countersued contending that his employment agreements with Mattel were invalid and the Bratz remained just an idea until after he left Mattel. In April 2005, MGA increased the stakes and sued Mattel, claiming its My Scene toy line was unfairly similar to Bratz and that Mattel was using its clout with retailers to stifle competition. In November 2006, Mattel retaliated by accusing MGA of copyright infringement, misappropriation of trade secrets and racketeering, demanding in short full ownership of the Bratz doll. In October 2007, MGA hired a prestigious (and expensive) legal counsel. In July 2008, the designer settled with Mattel for an undisclosed amount. A week later, a federal jury ruled against MGA and awarded .1bn$ in damages to Mattel (out of 2bn$ requested).[39@] The crux of matter then was whether MGA owed Mattel profits derived from the first original dolls or from all the subsequent dolls and related products. In December, the judge ordered MGA to recall all Bratz dolls from retailers, to destroy "specialized plates, molds and matrices" used to make the dolls and to transfer all trademark rights in the Bratz name to Mattel. MGA appealed the ruling and in the meantime introduced an alternative line of dolls. In January 2009, the judge granted a stay of execution allowing MGA to operate during that year. In February, MGA changed stance and tried to settle the case with Mattel by selling them the Bratz line for less than .5bn$, but to no avail. In April, the judge ordered to hand over the control of the company to a temporary receiver only to overturn the decision in May while fastening the transfer of the Bratz ownership to Mattel. MGA appealed and in December, the court suspended the transfer order (i.e., MGA could keep selling its Bratz dolls) and finally reversed it in July 2010 stating: "It is not equitable to transfer this billion-dollar brand — the value of which is overwhelmingly the result of MGA's legitimate efforts — because it may have started with two misappropriated names". The entire case will probably be have to be retried. In this story, the money spend into building offensive and counter-offensive lawsuits neither improves the quality of products nor positively stimulates demand. The two firms are simply engaged into a war of attrition with a view to either exclude the challenger from the market or get a bigger share of the cake (cf. §7.4.2).

DMCA The Digital Millennium Copyright Act passed in 1998 in the US aims at updating intellectual property rules for digital content. This extreme law has been used by private parties to exert control over their markets in a way never intended by its creators.

In 2003, a Kentucky court granted a preliminary injunction to Lexmark International against a company that makes generic replacement cartridges for Lexmark printers. The court found that a chip in Lexmark cartridges that identify the refills as "official" could be protected under the DMCA, and thus, cannot be cloned. This is an obvious intent by a company to limit the access of its clients to less expensive ink cartridges (cf. cases of illegal tying 24.1).

Publishing results that show the weaknesses in the security of a software package can violate the DMCA. The most famous case is that of russian programmer Sklyarov who was arrested in California and jailed for 3 weeks in July 2001 after giving a speech about his company's software that could bypass protections on Adobe Systems' eBooks. Prosecutors argued the russian software violated the DMCA, which outlaws offering software that can circumvent copyright protections. The company faced charges related to directly designing and marketing software that could be used to crack eBook copyright protections, plus an additional charge related to conspiring to do so. The defense said the software was designed to allow people to make backup copies of eBooks they already own or transfer the material to a different computer. In december 2002, the jury acquitted Sklyarov's company of all charges.

Trademark Levi-Strauss has been confirmed into its right to be the only seller in Europe of Levis jeans trough its networks of exclusive Levis shops. The British supermarket chain Tesco was condemned by the european court of justice (ECJ) for buying genuine Levis jeans in the US and selling them in its stores.[40@]

One may point that the prohibition of parallel imports is not in accordance with the raison d'être of trademark law, which is to assure the originality of goods and to protect consumers from piracy and fraud. Furthermore, the current EU legislation favors anticompetitive behavior by providing Trademark holders with the possibility to segment global markets and apply price discrimination. In the face of the argument that international exhaustion would reduce the economic value of trademark rights and damage innovation, one should note that all the cases judged in this respect by the ECJ pertain to high end products whose reputation was long established so that investments were covered a long time ago (e.g., Levi Strauss, Davidoff, Sebago, Christian Dior).

A few years ago, Microsoft filled a trademark lawsuit against Lindows, a company offering a low-cost Linux-based operating system that was compatible with popular Mi-

crosoft file formats. Microsoft claimed that "windows" was clearly a trademark while Lindows argued that “windows” is a generic term for a certain type of software interface that predates the Microsoft product. MS successfully barred Lindows in a number of countries and finally came to an agreement with its potential infringer in 2004.

For the record, Apple unsuccessfully sued Microsoft in 1992 for copying the look and feel of its Macintosh desktop software. In its successful defense, Microsoft argued that windows, icons and menus were developed in 1970s at Xerox corporation!

Copying of Books and Writings Scientific publishers MIT Press, Elsevier and John Wiley have opened a lawsuit against a shop selling photocopies of their copyright protected books to students at the University of Los Angeles. The case was settled with the shopkeeper agreeing to pay damages and honor all of the plaintiffs’ copyrights in the future.

A musician was once sued by the company holding his own copyrights for writing a new song too similar to an older song written by himself twenty years before (cf. Columbia Law School list of recording artists copyright infringements).

In still another case where real names do not matter, company A purchased a *single* subscription of a Market Analysis newsletter from company B but used to disseminate it to all its employees by photocopying, faxing and even posting an electronic version on its intranet. The content producer was awarded $20 million in damages. Likewise, the American Geophysical Union sued the Oil company Texaco in 1992 for illegal photocopying; the case was settled before the start of the trial with the defendant paying a large (undisclosed) amount to the right holder (more examples).

DVD Protection The media format known as Digital Versatile Disc (DVD) is the successor of the Compact Disc (CD). Initially, Sony and Philips advocated a disc with a single side of engraved data while a group of seven (mostly japanese) firms advocated for a double-sided disc. Recalling the bitter competition in the 1970s between the home video recording format VHS and Beta (cf. §24.3.5) , the two sides reached consensus on hardware and software specifications and created the DVD Forum in 1995 to license the new standard.

It has been customary for hollywood studios to stagger movie releases across continents, first the US, then Europe and Japan a few months later and finally the rest of the world some months later; it was important for studios that DVD would hit markets in the same staggered way. Hollywood therefore required and obtained from the DVD Forum that each DVD player would be given a code for the region in which it would be sold.[41@] This meant that a disc bought in one country would not play on a player bought

in another country. Such a feature could be viewed as an illegal vertical restraint of trade enabling unwarranted price discrimination, but since most DVD-players are now shipped with a hack removing the region coding, the issue has become irrelevant.

DVD Copying A highly mediatized case of copyright infringement related to the DVD technology is DeCSS. The Content Scrambling System (CSS) is an encryption and authentication scheme intended to prevent the perfect digital copy of DVD content (colloquially known as "ripping"). Although there is no charge to obtain a CSS license, it is a lengthy process. In 1999, only computers operating under the Windows or Mac systems could offer a DVD-player thus, in order to play a DVD on a computer equipped with the Linux operating system, a group of computer addicts "reverse engineered" the CSS code to produce a code breaker named DeCSS which was immediatly posted on hundreds of websites.

The Motion Picture Association of America (MPAA) quickly reacted in court by invoking infringement of the DMCA but has failed to defend effectively its monopoly of DVD decryption. In 2003, Jon Johansen who posted DeCSS on the internet was acquitted by an Oslo court of the charges of theft. He admitted copying only legally purchased DVDs using the program, and the court ruled that he was entitled to do this.

More generally, the decompilation of computer programs to achieve inter-operability has been ruled lawful in partial contradiction with the DMCA provision that forbids decompilation of programs protecting digital versions of copyrighted works; examples of such protection algorithm are present inside the WMA format from Microsoft and the AAC format from Apple to sell music digitally.

DivX & MP3 As a standalone application DeCSS is useless for piracy but combined with a "DivX" encoder it enables to burn on a single CD a good quality copy of a DVD movie. This brings us to the second infringement, that of the patents for compressing music and video. The original DivX software published in 1999 combined the reverse engineering of Microsoft's proprietary compression algorithms for sound and video. To avoid legal problems, developers quickly turned to the MPEG-4 ISO specifications (cf. §24 on standards) in order to develop trouble-free video compression algorithms.

As for audio, the most popular format over the internet is the MP3. This lossy compression algorithm was developed in the late 1980s by the German Fraunhofer Institute using funding from the EC and became the audio part of the MPEG-1 ISO standard (cf. §24 on standards). The idea is to use models of human auditive perception to eliminate much of the redundant information stored in a CD so as to obtain the same listening experience but using twelve to twenty time less space (this technological constraint was

quite relevant back in the 1980s). [42@]

The reason why MP3 has become the digital audio standard is probably due to the initial lax enforcement policy of the original patent holders, Thomson Multimedia and Fraunhofer IIS. From 1992, date of the ISO release, until 1998, software developers have been able to explore the code to develop players (to listen MP3s on the computer) and encoders (to rip CDs into MP3s). By contrast, several other digital audio formats developed by companies like Lucent, Yamaha, and Microsoft, have kept their formats proprietary and limited how outside developers can employ their technology. Releasing the technology in the open has probably helped this standard to become dominant thanks to the user-friendly softwares that have been developed around it. Starting late 1998, the patent holders have been seeking licensing from commercial developers.

Internet File Sharing Many recent cases of IPR infringements derive from the aforementioned "ripping" and encoding technologies which permit the sharing of digital versions of songs and movies on peer-to-peer (P2P) distribution networks. With the ever higher penetration of high speed internet connection in households (e.g., cable, ADSL), the sharing of movies in digital format has increased steadily.

Starting in December 1999, the Recording Industry Association of America (RIAA) battled for 18 months the Napster network which permitted the free sharing of copyrighted music in the MP3 format. The courts ruled in this case that the software developer was directly involved in the copyright infringement because its server was performing the searching and indexing of available files on the user's computers. Developers responded with decentralized Peer-to-peer networks. In response, the RIAA and the MPAA started to sue individual users for *direct* infringement of the DMCA. A second strategy adopted by record labels and movie studios has been to sue developers for *indirect* infringement of the DMCA. Both the lower and the appeal courts ruled that peer-to-peer software developers were not liable for any copyright infringement committed by people using their products, as long as they had no direct ability to stop the acts. The case has gone since to the Supreme Court who ruled in June 2005 that file-sharing software companies could be held legally responsible for copyright infringement on their networks. Since then, many countries have enacted laws going in the same direction.

Defenders of these P2P networks are quick to point at the similarity with the case of VCR (cf. §24.3.5): Sony, a manufacturer of VCRs, was accused in 1976 by content maker Universal Studios of permitting the copying of copyrighted material.[43@] The final judgement of the US supreme court in 1984 confirmed that "the non commercial home recording of material broadcast over the public airwaves was a fair use of copyrighted works and did not constitute copyright infringement". The underlying argument is that

a distributor cannot be held liable for users' infringement so long as the tool (VCR or software) is capable of substantial noninfringing uses (cf. review). Several famous professors of economics denounce the diversion of this judgment into the simplistic conclusion that one should never use indirect liability for products merely capable of substantial noninfringing use (fair use). They argue on the contrary that indirect liability is the optimal mechanism when direct deterrence is ineffective because of the high costs associated with identifying and pursuing individual violators. As in the case of online auctioning of prohibited items, the auctioneer is not held responsible as long as he acts expeditiously to remove infringing content as soon as he becomes aware of its existence. On the other side, Intel, the giant semiconductor manufacturer claims that expanding the scope of indirect liability would chill innovation and stifle the development of new products, including some designed to enhance lawful access to copyrighted works!

Part F

Integration: Limits to the Boundaries of the Firm

Chapter 13

Firms vs. Markets

This Part focuses on firms and their relations to markets; it also contains, for historical reasons, the theory of contracts that is of broader scope.

Industrial Organization traditionally limited its scope to the interaction *among* firms in the marketplace. The interactions taking place *inside* firms were investigated by academics from management, sociology and organization theory. Yet, as shown on Table 13.1, a large share of value creation (GDP) in modern economies is generated within firms, not on markets (cf. also Lafontaine and Slade (2007)).[1@] For manufacturing, the ratio is one third, for services, it is twice that and on average, over one half. These stylized facts prove that firms are a worthwhile object of study for economists.

US	1998	2003	2008
Agriculture	40	43	43
Mining	55	57	52
Utilities	61	58	58
Construction	47	50	50
Durable goods	36	36	36
Perishable goods	33	33	28
Information	52	52	51
Wholesale trade	72	70	65
Retail trade	71	71	67
Transportation	52	52	48
Finance	55	56	55
Real estate	71	70	69
Business services	64	67	66
Leisure	56	57	56
Education,Health	61	61	61
Government	*65*	*62*	*60*
Total	**55**	**56**	**54**

France	1999	2004	2009
Agriculture	49	46	38
Food, Tobacco	24	25	21
Energy	33	29	25
Construction	44	43	44
Consumer goods	33	30	28
Automotive	19	18	16
Equipment goods	29	28	26
Intermediate goods	33	30	29
Trade	55	53	51
Transport	48	47	50
Finance	51	49	46
Real estate	78	80	81
Business Services	57	55	53
Personal Services	54	53	53
Education, health, social	79	78	77
Government	71	70	71
Total	**51**	**50**	**50**

Table 13.1: Value Added as % of Production

The first section centers on the firm and builds on the fact that productive activities are regulated only to a limited extent by prices and markets and much more by authority. We investigate the explicit and implicit incentives schemes at work inside firms to promote hard work and cooperation. The second section then contemplates various theories regarding the formation and evolution of the firm, including the "black box" one dominant in this book according to which the firm is a single rational minded economic agent called the "entrepreneur" interacting with other entrepreneurs on markets.

13.1 Inside the Firm

In this section, we briefly recall the legal nature of the firm and identify the basic incentive problem residing inside the modern corporation, the fact that executives and workers' objectives need not be aligned with those of the owners. We then show how explicit contracts like wage schemes enable to realign incentives. When opportunism, free riding or moral hazard cannot be eliminated using enforceable contracts, some implicit contractual arrangements are nevertheless available. Holmstrom and Tirole (1989) and Simon (1991) cite *authority*, *rewards* and *identification* (loyalty), *external control* and *bureaucracy*.

13.1.1 What is a Firm ?

Legal Forms

A small business is called a (sole) *proprietorship* if owned and managed by a single individual while it is called a *partnership* if several people unite to own and manage the firm. Setting up a partnership is the most sensible and economical way to start a business; yet the owners are personally liable without limits for their firm's debts or commitments.

When owners start to recruit managers to run the business on their behalf, they often turn their firm into a corporation, a legal entity able to borrow or lend money, sue or be sued and pay taxes.[2@] At the creation of a corporation, its shares are often held by the company's creators and a few backers (venture capitalists or family). If the firm grows successfully, it will need additional capital. This leads the owners to list their shares in a stock exchange where new shares are easily issued, quoted and traded. The corporation becomes *public* (not to be confused with public ownership by the state) and owned by dispersed stockholders (or shareholders) who range from the small individual investor to the gigantic pension fund. However, partnerships remain prominent in the areas of law, accounting, medicine, investment banking, architecture, advertising, and consulting

because it is difficult for clients to assess the quality or ability of the people serving them (cf. §21.1.2).

The stockholders own the corporation but do not manage it; their active participation is limited, in European law, to elect[3@] an administrative board (directors) who then appoints top management and make sure that managers act in the shareholders' best interests (a not so clear concept to be elucidated later). This separation of ownership and management contributes to stabilize corporations both ways: if there's a change of managers (they quit or are dismissed), the corporation can continue its business and develop according to the desires of the board; likewise if a stockholder sell his shares to a new investor this will not disrupt the operations of the corporation.

The legal definition of the corporation also encompasses limited liability of the owners which means that shareholders will lose at most the price they paid for their shares. The other side of the coin is that shareholders are only entitled to the firm's residual cash flow i.e., what remains after employees, suppliers, lenders and the government have been paid. The corporation has also some disadvantages when compared to a partnership because there are bureaucratic costs linked to the necessary communication in the authority chain illustrated on Figure 13.1. There are also conflicts of interest in this chain which generate inefficiency costs. Formal contracts being unavailable inside the firm, some alternative methods are designed to cope with opportunism, free riding, moral hazard and hold-up; they are presented hereafter.

Figure 13.1: Communication and Authority Ladder

13.1.2 Explicit Incentives: Compensation

A craftsman has always the right incentives towards hard work because he is the residual claimant of his efforts. Motivating his apprentice toward the maximization of the shop's profit is also quite elementary using a bonus for good performance or the promise to inherit the workshop later on.[4@] Transposing this result to the more general and ubiquitous modern corporation where employees almost invariably work in teams is more difficult. We comment here some issues that are developed analytically in §20.2 of the chapter on incentives.

Team vs Individual Compensation

Paying an employee according to a performance measure reflecting his individual contribution is the cheapest way to provide him with incentives towards effort. Also, if the observed signal is precise, the risk exposure of the agent is low so that incentives are cheaply provided (cf. formal model of §20.2). Whenever agents work in teams, the performance measure necessarily includes contributions from other people which mechanically increase the risk exposure of everyone; incentives thus become more expensive to provide so that team compensation would look inadequate.

However, joint accountability enriches the performance measures available to the manager, which helps to mitigate multi-task problems. To understand this issue, recall that the technological advantage afforded by specialization also means that agents perform their specialized activity in different places or over different products. The exception is line assembly where the worker applies his specialized effort repetitively and can thus be paid at a piece rate to guarantee optimal effort. A scientist or manager, however, will typically work over at least two projects simultaneously and the productivities of these tasks may not be perfectly reflected in the observed performance measures.

Put differently, the profit of a product line is the sum of profit attributable to, say, R&D and marketing; yet revenues, an observable measure, may magnify the marketing activity with respect to the R&D one. This means that if an employee works only in that line of business and his wage is linked to revenues and tuned so as to give him adequate incentive to expand effort in R&D, he will mechanically expand too much effort in marketing. Now, if the agent performs the same task, say marketing, in different product lines, his wage can depend on the two revenues and can be tuned so as to give him the right incentives to perform the efficient time allocation among the two tasks. However, he will be exposed to a greater risk since his wage will become more variable.

In §20.2.4, we develop Corts (2007)'s model exposing this arbitrage between risk and multi-tasking. We conclude that when the degree of risk aversion is low, or when signals of effort are informative or when the multi-task problem is serious, team work dominates individual accountability.

Relative vs. Joint Performance Evaluation

Having shown that team work can be the optimal organization inside the firm, we can delve further into this idea. We seek to uncover optimal performance measures and to assess whether an employee's compensation should be tied to the performance of his peers.

One option is to link positively the wage to the group's performance with *joint* perfor-

mance evaluation (JPE) i.e., set a bonus linked to the division's profit, the output rate of the assembly line or the group-based lending adopted in the Grameen bank (microcredit). The other option, *relative* performance evaluation (RPE), takes an opposite point of view and recommends to penalize the employee (relative to others) if the entire group fare well. This is a form of contest as studied in §7. For instance CEOs are more likely to be dismissed from their jobs after bad industry and bad market performance indicating that their peers performed better The "Tour de France" cycling competition displays both kind of evaluations. Every day, the first runner passing the finish line gets a time reward, the second a smaller one and so on; this is a case of RPE. On the "team" day, the time of each runner is the time of the latest from the team to pass the finish line; this is a case of extreme JPE where teammates take great care of the weakest among them.

The economic theory of incentives (aka. moral hazard) largely focuses on short-term relationships between employees and employers; it advocates the use of incentive schemes that are sensitive to individual performance measures and induce competition among workers for instance via tournaments for access to a higher position (cf. §13.1.3). Moral hazard in teams extends the previous focus to the employee's peer group. If the performance measure of workers has a common noise component, then RPE can be attractive since it insulates the workers from the risk of common shocks and thus generates a stronger incentive than other schemes. This allows the employer to provide incentives at low cost. Yet RPE, by entailing extreme competition among team members, fails to foster cooperation. In environments where workers interact closely with each other it leads to an inefficient peer pressure: the group forces a "rate buster" to reduce his output so as not to make other team members look bad. In many activities team work has been implemented to foster cooperation among employees via frequent and long-term interaction, via empowerment (decentralized authority) and via peer monitoring and no outside supervision. Incentives to effort are given through a JPE wage scheme.

The model of interaction between employers and employees developed in §7.4.3 comes to the following conclusion.

> For short term relationships, it is optimal to use RPE i.e., to pay someone when his signal is good but when the partner's signal is bad. The key to this result is that a good signal is more informative of effort if the partner's signal is bad. Thus, paying a bonus solely in that asymmetrical situation generates an inexpensive motivation to work.
> Long term relationships call on to use the opposite JPE scheme because it is important to avoid retaliation by workers who could be penalized by the RPE scheme. To guarantee effort during the whole duration of the relationship, a mutual bonus is optimal.

13.1.3 Implicit Incentives: Motivation

Motivation refers to all the incentive schemes that go beyond short run piece rate; it includes authority, rewards, identification with the firm, external controls and specialization.

Authority

Two features of employment relationships are key to understand its widespread use. Take the example of a shirt manufacturer and its dressmaker. Firstly, there is uncertainty for the employer: shall he produce white or black shirts? Fashion will tell! Secondly, the employee is likely to be obedient because he is indifferent about the color of the fabric to be sewed. Those elements makes the employment contract very attractive for both parties because it saves on *transaction costs*, there is no need to negotiate a contract for the furniture of white shirts and another one for black shirts or similarly there is no need to go back to the local trading post each time the manufacturer needs to buy a new bundle of shirt in the latest fashionable color.

But the authority does not only issue commandments; more often it asks a result to be produced ("repair this sewing machine"), or a principle to be applied ("sew body before sleeves"), or goal to be achieved ("sew as cheaply as possible consistent with quality"). There is delegation not only of the task but also of the method used to perform the task. Doing the job well is not so much responding to commands but rather taking initiative to advance organizational objectives. To be sure, obeying rules literally is a favorite method of work slowdown, an alternative to strike used by all sorts of controllers (customs, fiscal).

The employer's ideal command would be "always decide in such a way as to maximize company profit" but checking its obedience is almost impossible without losing the benefit of delegation. Even if the employees were (loyal) robots, complete discretion would lead to chaos and would surrender the large efficiencies attainable by specialization in decision-making work. We need to delegate within guidelines, which creates the problem of monitoring the observance of guidelines without recentralization what has just been delegated. Indeed, to ascertain and judge the decision criteria of the employee, the employer would have to review the whole decision process thus completely losing the benefit of delegation. The solution is then to motivate the employee in order to be able to trust him.

Rewards

Employees may be motivated to accept authority and contribute to the organization's goal by giving them monetary rewards, promotions, or perquisites. The "workaholic" behavior of many managers is well explained by their career concerns and the implied lifetime income stream (cf. §20.2.5). However reward schemes work only if contributions are measurable; it is then clear why wages are linked to sales, piecework or the firm's profit. Yet if a key characteristic like quality of the product is not observed, the incentive scheme may cause quantities to grow at the cost of lowered quality. Likewise safety rules may be ignored or problems passed to other divisions within the firm.

It seems obvious that rewards systems should be aligned with desired behavior but management theorists tell us that this is not common practice; instead there is an overemphasis on "objective" criterions, quantifiable standards, highly visible behaviors and also hypocrisy by the rewarder claiming that the target was not met although it was. Some examples, inspired by Kerr (1975), are:

- Economy and prudence should rule governmental budgeting but the biggest budgets are allotted to the biggest spenders.
- Voters punish candidates who frankly discuss where money will come from; thus they vote for acceptable but void goals.
- Physicians tend to over medicate patients (label a well person sick), because the reverse error (labeling a sick person well) has far severe consequences although both kind of errors should be minimized.
- Academics ought to manage the dual responsibilities of teaching and researching but since in most countries they are rewarded for one task only, they tend to neglect the other one.
- In sports, it is team performance that matters but individuals who obtain rewards such as fame and endorsement contracts.

An enduring system of motivation is found in yardstick competition and tournaments whereby a group of employees are benchmarked one against the others. Promotion is an extreme case wherein a large prize is awarded to the best performer. It has proved to be an efficient system of motivation and to make the prize of promotion even more attractive to the many contenders, many firms pay their top executive more than her marginal productivity. The threat of dismissal can also discipline workers in a firm but only if they face a possible unemployment period or a lower wage in the new job. Hence firms pay experienced people much more than youngsters to increase the difference in expected salary (efficiency wage theory).

Since rewards tend to turn employees into competitors, the cooperative attitude so important for the firm's success might be lost; it would therefore be dangerous to rely solely on rewards as a source of motivation.

Identification

We do not observe much free riding or shirking inside firms; instead many employees display pride in work and loyalty to their organization. *Identification* with the structure is a widespread behavior that induces employees to accept organizational goals and authority as bases for their actions.

Natural selection can explain why such a behavior is common. Darwinian models and experiments support neither the idea that nice guys can make it to the next generation nor for the idea that people only pursue selfish personal economic goals. Instead, the observation that humans depend on the surrounding society for nutrition, shelter, safety and apprenticeship has lead theorists to predicts the appearance of *docility* (being tractable, manageable, and teachable) as a key to building fitness and improving survival. Darwinian fitness calls for a substantial responsiveness to social influence: motivation to learn or imitate, willingness to obey or conform; hence Simon concludes that docility is not altruism but enlightened selfishness.[5@] This topic is also discussed in §24.5.

Docility is used to inculcate individuals with organizational pride and loyalty. Identification with the "we", which may be a family, a company, a nation, or a sports team, allows individuals to experience satisfaction from successes of the unit they belong to. Thus, organizational identification becomes a motivation for employees to work actively for organizational goals. Furthermore our bounded rationality acts as a reinforcing feedback; indeed we cannot grasp the complex reality in its entirety and therefore focus our attention on specific aspects among which the goals of our organization.

External Control

As stated above, the well functioning of a necessarily decentralized organization rests on the active participation of subordinates to make constructive propositions and take value enhancing decisions. When instead, they interfere to promote their own well being or their career, they generate so-called influence costs.[6@] Beyond the arguments of internal discipline seen above, several external forces make managers "behave" , several of which appear as models in §23.3.

- Labor market discipline states that misbehavior almost always end-up noticed, thus made public which ruins the manager's reputation and his ability to find another

job.

- For investments, managers may have a shorter or longer horizon (reputation) than the firm; they may end up taking sub optimal decisions. Stock-options are then a good proxy of an option on the firm's value which is also a good proxy of an option on the manager's reputation.
- Product market discipline argues that whenever there is competition it is possible to judge a manager's performance against that of competitors, hence one can write incentive contracts. This comparability also enhances the reputation concern of the manager (cf. §3.2.2).
- Capital market discipline tells us that incumbent managers must do their best to maximize the firm's value because otherwise they will be ousted by a hostile take-overs and replaced by better able and willing managers.

Bureaucracy

Weber (1922) characterizes bureaucracy as the centralization of administration into a corps of well trained professionals relying on the division of labour within the organization, a pyramidal authority structure and impersonal rules that regulate the relations between organizational members.

The key of a bureaucracy is its system of control based on rational rules meant to design and regulate the whole organization on the basis of technical knowledge in order to achieve maximum efficiency. The specialization of tasks is achieved by an extensive definition of the duties and responsibilities of each individual. Hierarchical relations are impersonal and authority is legitimized by the efficiency of the administrative rules, unlike feudal order where relationships are personal and based on the sacredness of tradition. Hence loyalty is oriented towards an office and not towards who holds it. The bureaucrat is not selected on basis of family or political loyalty but on formal qualifications related to his specialized duty;[7@] also the position cannot be sold or pass to an heir.[8@] Entering the bureaucratic organization is for a "life's work" i.e., to make a career. Remuneration ignores short-term productivity by relying on a fixed salary but stimulates long-term productivity through promotion for achievement; it also fosters fidelity by accounting for seniority.

The popular view of bureaucracy is often that of "red tape" and inefficiency. To explain this perversion we have to notice that if bureaucracy implies reliability and predictability it also lacks flexibility and tend to turn means into ends. Indeed, the emphasis on conformity and rigid rules induces the bureaucrat to internalize them; the formal aspect of bureaucracy becomes more important than the substantive one, the efficient working

of the organization.

On the theory front, Niskanen (1968) shows that a bureau tends to overproduce wrt. the efficient situation where marginal cost is equated to social WTP (cf. §17.2.4). Since public services have near natural monopoly attributes (cf. proof in §17.2), delegation to the private sector (or privatization) would create the de-facto monopoly whose tendency is opposite i.e., lower output so as to charge a higher price. This typical "Chicago school" author, is nevertheless aware that privatization is no solution unless an effective regulation forces the awardee to behave in the people's interest. The question then is whether a (of necessity public) control agency is better at regulating a bureau or a concessionaire (cf. §16.4.1).

More recently, Acemoglu et al. (2008) explain why pensions, health care and education are managed by the State. The social aim of schooling is to increase human capital by adequate teaching but privatizing schools introduces yardstick competition. First, institutions wastefully advertise to attract customers. Secondly, they are judged on characteristics observable in the short-run meanwhile the social objective is a long run one. Hence, schools substitute socially valuable investment by visibility enhancing ones. On the human capital side, teachers and school managers train children like monkeys to succeed at state exams. Law enforcement displays the same problem with over-powered incentives under yardstick competition. Regarding pensions, a private fund manager will distort the composition to display short-run profitability when the long-run is what matters. Regarding health, private hospitals tend to substitute medical equipment by amenities to attract more clientele instead of focusing on curing them. Even within state agencies, as soon as bureaucrats are rated per output, they start to go after easy preys (may be innocent people) instead of socially harmful ones. For instance, IRS agents pursue divorcees who failed to pay the pension to their ex-wife instead of trying to catch big-time defrauders. All of these cases point toward low powered incentives characteristic of the bureaucracy in order to sustain a good result (though clearly second-best).

13.1.4 Internal Organization of the Firm

Functions vs. Products

During the first half of the XX^{th} century, firms took advantage of scale economies by organizing themselves around *functions* or *processes* i.e., they had a different *departments* for marketing, manufacturing, purchases, finance, engineering, R&D, legal matters, human resources and so on. Chandler (1992) identifies market leadership and the building of a competitive advantage as the result of a three-pronged strategy: investment in large-scale production to lower unit cost, investment in marketing, distribution, and purchas-

ing networks and finally, recruitment and organization of professional managers. Thereafter, many firms became conglomerates producing a range of vastly different products. Under such an organization, most decisions are taken at the company's headquarters, ensuring better coordination of activities across departments but also leading to congestion at the top. This system emphasizes the importance of *inputs* and takes advantage of the specialization of labor; at the same time the identification of contributions to overall profits is made difficult which can result in demotivation, lack of incentives but above all confuse the allocation of cash-flow towards the best opportunities.

The opposite organizational form focuses on the *outputs* or *products* i.e., emphasize consumer vs. business clients, US vs. European geography, Civil vs. Military use, agricultural vs. industrial specificities. This classification leads the firm to group complementary tasks into self-contained *divisions*.[9@] Thanks to this decentralization, the better informed local managers are taking better decision in the day-to-day operations but also wrt. long-term strategy. Under this latter organization, the contributions of divisions are easier to identify thanks to analytical accounting and therefore easier to reward. Being quasi-autonomous profit centers, divisions compete for the firm's cash-flow in what Williamson (1975) called a "miniature capital market" with potential gains in efficiency over external markets deriving from manager's superior knowledge of the firm's opportunities compared to that of investors or bankers.

The tension between the two organization forms can be illustrated as follows: when the VW automaker bought Seat and Skoda in the 1990s in addition to its original brands Volkswagen and Audi, it intended to run the 4 firms as independent competitors each with a distinct image. The group then quickly realized that important scale economies could be achieved by designing platforms upon which models of the 4 brands could be build; roughly speaking some elements of the cost of designing a new car would cut by a factor 4. That was a move back towards a functional organization. This integration process was successful and heightened at about 60% of the components of a car. The downside was too much similarity of the models that came to compete one against the other instead of competing against outside brands. The current strategy of the VW group seems to be a strengthening of the brand image of each subsidiary, thereby reducing significantly the commonalities inside the group (a move back to the divisional organization).[10@]

Authority vs. Information

Another fundamental characteristic of a firm is its distribution of authority in relation to the information structure, that is to say, who is to take the coordination decisions when the activities of the firm are hit by a shock ? Maskin et al. (2000) and Qian et al. (2006)

study this trade-off in a simple model where the product is successfully completed (sold) only if all complementary parts fit together (e.g., car, computer, congress organization, software or an audit). Although operations follow a plan, there are always unexpected (exogenous) contingencies that create either a mismatching called an *attribute shock* or a line failure called a *capacity shock*. Examples of the former kind would be when a piece does not fit or a new kind of piece is needed while examples of the latter kind would be when a piece does not come in time or the product does not sell anymore. These shocks call for adjustments and require coordination given the complexity of the whole process.

It is at this point that the organization of authority and information circulation can make a difference. There are only two authority layers, one CEO and two division managers. We then assume that division managers have perfect information with respect to shocks hitting their division but none with regard to other divisions while the CEO receives a more or less distorted information of all divisions. The authority structures are thus :

- Decentralization : the division managers take coordination decisions.
- Centralization : the CEO takes all decisions in which case the firm shape is irrelevant.

The concepts of internal organization can be represented schematically. Imagine an entertainment firm selling two items, videos and music with two processes, production and sales. The two organizations are represented on Figure 13.2.

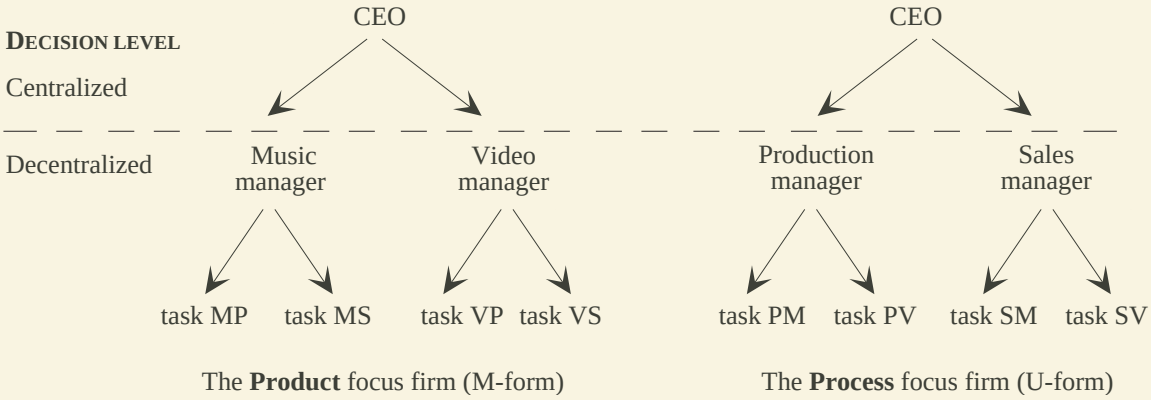


Figure 13.2: Two Organization Schemes

In the simplified scheme of Figure 13.2, the decentralized product focused organization (M-form) is more efficient if there are only attribute shocks (mismatching) because in the decentralized U-form firm, no adjustment can be carried out for the failure to communicate and because the centralized form suffers from an imperfect communication. If

there are only capacity shocks then the decentralized U-form firm dominates the M-form one as soon as there are some storage costs. Otherwise the M-form organization can anticipate the effect of any shock with the following policy "transfer 50% of the expected need of the other division if i am not hit by a shock myself". When no shock occurs transfers cancel out and if only one division suffers a shock it is insured by the automatic transfer of the other so that production are equalized in the two divisions. When both type of shocks can occur centralization is likely to become the best organization form if the communication channels are sufficiently efficient.

Summarizing, a *process* focused firm copes well with capacity shocks because she is good at managing substitution or reallocation of resources from a product line to another when needed. A *product* focused firm conversely copes well with attribute shocks since all processes are affected by the shock and the close integration of those processes within the division allows their managers to coordinate on an effective reaction. To conclude, the internal organization of a firm will evolve in reaction to the evolution of consumer tastes, technological changes such as the information processing enabled by the use of computers and networks or the better specialization of tasks enabled by robotization.

13.2 Contract Theory

In this sub-section, we explain why real life contracts are so incomplete. The reasons we advance borrow from the transaction cost and property rights theories of the firm studied in the next section. An primer on the topic is Salanié (2005), while Bolton and Dewatripont (2005) is a more exhaustive reference.

13.2.1 Incomplete Contracts

Contracts are useful to make commitments, to delegate task, to allocate decision and control rights; their ultimate goal is to allow for more efficient exchanges (in a wide sense). As the examples displayed below demonstrate, contracts are often incomplete in the sense that much is left to be decided later on.

Market Whether at the city market or at a shop, the buyer picks an item, pays the listed price (or bargain a little over it) and walks away with the item. The exchange contract is complete in the sense that nothing we may think of is capable of derailing that simple interaction.

Haircut The posted price is agreed by the customer and the beautician, but the former gets to specify "how she wants it to look," and the latter gets to decide on the instruments and how they are used.

Restaurant For the listed price on the menu, the contract gives you the right to receive one serving of the dish and eat it at your table. It is tacitly understood that you can decide how fast you eat, how loud you talk, and how big a mess you leave behind, while the restaurant can decide on how and from which ingredients the dish is prepared, how soon it is served, when you get your drinks, which napkins you are given, who sits at the other tables, and so on.

University The contract explicitly gives the buyer the right to attend classes and use libraries for an academic year; he has also discretion regarding timing and selection of the classes to attend. In turn, the university can decide which courses are offered, when and where they are offered, and who teaches them. It also decides on the contents of the libraries, the temperature of the pool, and so on.

Custom-built house The contract includes a large set of architectural drawings specifying particular production methods and materials. However, a tremendous number of decisions are still left out. The customer can make decisions on colors, fixtures, and a number of other decorative items. Conversely, the builder can decide on the brands or suppliers of several materials, as well as the exact location of many nails, joints, and boards.

Let us then try to classify what should be put into a contract and why less is in fact used. Firstly, it is difficult or costly to fully specify the contracting parties' obligations because

- Parties must speak the same (business) language.
- Writing a contract is time consuming.
- Future contingencies are difficult to describe, thus to incorporate.
- Contingencies must be *observable* by both parties for otherwise they'll disagree regarding what happened.
- Contingencies must be *verifiable* by an external enforcer such as a judge or a private arbitrator for otherwise they can't be enforced.
- Wealth constraints and the protection afforded to parties by *limited liability* put limits on the use of money to elicit the parties' willingness to pay for specific services or decisions.
- Collusion: complex contracts use parties to check on each other's truth-telling and obedience; there is then an incentive for parties to side-step the contract and collude against its enforcer to avoid penalties.
- Enforcement being provided by human beings, the latter may be lazy or fail to understand the contract or let their own preferences (or legal precedents and prin-

ciples) bias their decisions or even collude with the parties against the spirit of the contract.

- Ex-ante asymmetric information among the signatories forbids them to include everything that might be relevant in the contract since the better informed party prefers to maintain his informational advantage.
- Renegotiation: cf below

For these reasons, contracts are incomplete and only specify general rules of behavior or decisions and control rights such as who is to decide what to do if some specific event occurs or who is to decide on how to use an asset in the future.[11@] As noted by Tirole (1999), one must understand that simple institutions such as authority, property rights, and patents are popular because they have good robustness and learning properties i.e., they are not too far of the mark when the parties make mistakes in their view of the world or become less rational and they are universal so that one can learn how to use them in a context and apply them in a different environment.[12@]

The major problem generated by incompleteness is *renegotiation*. As events unfold and new information becomes available, decisions must be taken to achieve the objectives of the relationship; in other words, the parties need to constantly update and adapt their plans. There is little doubt that the parties will be able to identify the efficient course of action, that which maximizes the value of their relationship. By acting efficiently instead of following a default behavior (that thought to be correct ex-ante when the contract was designed), they generate a quasi-rent (cf. §2.4.3 and Fig. 2.9). The problematic issue is its division among the parties.

In the TCE vision, parties cannot help but haggle over the quasi-rent, each trying to get his way. In the end some of that wealth is wasted i.e., there is a deadweight-loss. Integration of the parties under a common hood can then alleviate this problem by replacing negotiation with authority, assuming that the intensity of haggling is reduced when parties are hierarchically related.

In the PRT vision, the Coase theorem holds (cf. §2.4.3) which means that some middle ground is immediately found and there is no deadweight-loss at the outset of the bargaining over the quasi-rent. The problem then is "hold-up", the fact that if one party has to share the fruits of her efforts with the other, then she has less motives to invest into meliorating the relationship (cf. §14.1.3 and §14.2).

13.2.2 Simple Contracts

Real life contracts are remarkably simple in that they frequently omit potentially useful and feasible provisions; they willingly ignore verifiable information or avoid to specify

specific behaviors. In particular, a lot of decisions made during the life of a contract are absent from it and instead tacitly delegated to one of the parties. For instance, employment or service contracts essentially focus only schedule, duration and compensation. At most, a vague notion of the mission to accomplish is stated. Thereafter, the principal orders the agents a task for which the latter has considerable latitude to execute (cf. also the examples above).

Wernerfelt (2007) explains that attempts to economize on bargaining costs imply that two parties may write a contract which is incomplete in the sense that each party tacitly cedes some decision rights to the other. Recall indeed that if future decisions are jointly taken, parties will reach an efficient decision but will also have to bargain over the division of the quasi-rent which is wasteful (under the TCE vision).[13@] On the other hand, if one party is delegated the decision right then no haggling takes place but she will fall short of efficiency for two reasons. Firstly, she will lack the input of the other party and secondly, she will sway the decision towards her personal best. Beyond the obvious role of the bargaining cost which calls for delegation, a party is more likely to cede the decision-right over an attribute if she does not care much for it, has less information about it or has little difference of opinion about it. Furthermore, inso far as the decision taken is not irreversible, the excluded party can call for a renegotiation when she feels that the candidate decision is too far from her ideal. This threat will force the decision maker to behave in the first place and thus yield a better outcome, closer to the jointly optimal one but without any actual costly negotiation.

Kessler and Leider (2010) offer a behavioral explanation[14@] to contractual simplicity based on *norms*: a simple "handshake" over a norm of good behavior ("do the right thing") is enough to motivate the parties to abide by the spirit of the contract and provide a "consummate performance" when the most that an enforceable contract could achieve would be a "perfunctory performance".[15@] The underlying driver is the fact that people suffer for violating the norm they choose to follow. Thus, negotiating a high norm results in parties take high intensity actions close to the ideal level required for the success of their relationship.[16@] In games with strategic complementarity, the norm sensitivity must be high to compensate the natural proclivity to undercut the partner's action. It is also shown empirically that establishing a high norm is an effective substitute for an enforceable restriction. This is so because an enforceable restriction allows parties to justify living down to the letter of the contract, rather than fulfilling its spirit. Parties thus see the inclusion of a minimum requirement as implicit permission to take only that action, a notion known as "crowding out" of intrinsic motivation.[17@] Given this substitution effect, the optimal contract is extreme i.e., either trivial and norm-oriented or meticulous and clause-oriented.[18@]

13.2.3 Flexible Contracts

The alternative, following Hart and Moore (2008), is to consider not simple vs. complex but flexible vs. rigid. By definition, more is known to the parties ex-post than ex-ante, so that more can be contracted upon. Yet, many issues remain unverifiable i.e., although parties observe what's going on there is no way to assert it in front of an expert. Each party is thus left free to perform as she sees fit. The span of moral hazard or free will is even larger if the contract deliberately delegates some contractible actions to a party. To assess how parties perform those out-of-contract actions, we take a behavioralist approach. If there is trust, each party will abide by the spirit of the contract and perform in a consummate manner.[19@] If, however, one party feels "shortchanged" or aggrieved, she will retaliate by sticking to the letter of the contract and giving a perfunctory performance, thereby creating a deadweight-loss (wrt. optimal performance). For instance, a seller can cut quality, delay delivery, exploit ambiguous terms, contest facts, refuse to cooperate or make a last minute change or threaten to do any of these things. They key is that the retaliatory action is cheap for him but expensive for the buyer. Apart from quality, the buyer can behave in a likewise fashion.

How can mistrust appear ? At the ex-ante initial contracting stage, parties trust themselves and see the contract as balanced and just.[20@] Indeed, being negotiated under competitive conditions, the contract constitutes a salient reference point because market forces define what each party brings to the relationship. Once the parties are immersed in their bilateral relationship, social norms of behavior and performance lose some of their grip and the self-serving bias kicks in.[21@] Each now tends to claim for herself any wealth newly created, especially the quasi-rents that appear during the life of the contract. Whatever the sharing rule in use (either exogenously fixed ex-ante or endogenously bargained ex-post), at least one party will feel aggrieved and will generate a deadweight-loss by its revengeful attitude.[22@]

One way to put an end this inefficiency is to reduce quasi-rents by rigidifying the ex-ante contract so as to eliminate most if not all ex-post adaptations which are the source of quasi-rents. The downside is obviously the inefficiency of such a contract unable to adapt to changing circumstances. This trade-off then helps to understand why parties tend to put restrictions (disallow renegotiation) on conflicting variables such as price or compensation but leave more room for more consensual ones such as the nature or characteristics of the good to be traded or the task to be executed (e.g., I care about my salary but not so much whether I teach micro or macro or whether I teach in the morning or the afternoon).

13.3 Theories of the Firm

13.3.1 Organizations vs. Markets

The Puzzle of Organizations and Markets

Before reviewing the most important theories relative to the formation of the firm, Simon (1991)'s tale will help set the analysis framework: A martian approaches the Earth from space, equipped with a telescope that reveals social structures. Firms (and public institutions) reveal themselves, as green areas with faint interior contours marking out divisions and departments. Market transactions show as red lines connecting firms, forming a network in the spaces between them. Within firms the approaching visitor also sees blue lines of authority connecting bosses with workers. Our visitor might see one of the green masses divide, as a firm divested itself of one of its divisions or it might see one green object gobble up another (acquisition or merger).

No matter whether our visitor approaches North America, urban China or Europe, the greater part of the space below him would be within the green areas, for almost all of the inhabitants would be employees, hence inside the firm boundaries. Organizations would be the dominant feature of the landscape; a rough description would speak of large green areas interconnected by red lines instead of a network of red lines connecting green spots. Hence our visitor would be quite surprised to hear the structure called a *market economy* instead of the obviously more appropriate *organizational economy*.[23@]

The constant progress of technology since the start of the industrial revolution has enabled firms to achieve great economies of scale and scope; their size has increased in relation to markets which is why most of the population work as employees or managers but few are entrepreneurs.

Economics of Organization

As rightly argued by Gibbons (2005), theories of the firm ought to be called "theories of the boundary of the firm". In this section, we pit the neoclassical view of a firm as being defined by production constraints against the more recent proposals of institutional economics, namely the *property rights theory* (PRT) and the *transaction cost economics* (TCE). None of these views are incompatible, rather, they differ in scope as one can notice with Table 13.2 borrowed from Williamson (2000). Four levels of organization are distinguished; for each, we indicate the time frame of its evolution, the objectives (not necessarily explicit) it tries to achieve, the nature of outcomes and lastly the theory fitting best the case.

Level	**Foundation**	**Institution**	**Governance**	**Transaction**
Time	century	decade	year	continuous
Objectives	religious beliefs	social rules	private rules	rationality
Output	norms customs	law bureaucracy	contracts organization	price quantity
Theory	sociology	property rights	transaction costs	neoclassical

Table 13.2: Social Organization and Time

The foundation level has no clear origin, it seems to appear spontaneously, being driven mostly by religion; it is always taken as fixed by economists (except historians).[24@] According to North (1990), institutions are the humanly devised constraints that structure human interaction. They are made up of formal constraints such as rules, laws or constitutions and informal constraints such as norms of behavior, conventions and self imposed codes of conduct (descendant from the foundation level). Together they define the rules of the game i.e., the incentive structure of societies.[25@] A classification of *institutions* pertinent for our purpose is:

- Political institutions, including the constitution, the legislative and judicial branches, form the superstructure above the government.
- The government determines economic policies on the short-term and designs economic and political institutions for the long term.
- Economic policies such as wealth redistribution (taxation, subsidization) or public investment are decisions implemented using of economic institutions.
- Economic institutions cover the areas of law regarding contract, commerce, tort, patents, property and finance to name a few.

The fundamental property of institutions is that they outlived the people who are part of them; they bring stability and credibility to the rules they establish for all actors whether the government or firms. Regarding the organization of economic activity, the proponents of the property rights theory (PRT) claim that in the absence of a well designed legal system, chaos dominates. The purpose of defining and enforcing property rights is to eliminate warlike competition for the control of economic resources and replace it with peaceful–non destructive–market competition. Private property rights are then the (human) right to use an asset, appropriate its returns, change its form, substance, or location. The PRT is nevertheless demanding as it relies on a far reaching and complex law to decide in every possible circumstance who holds the property right and on a powerful State to enforce this right. It therefore assumes no transaction cost.[26@]

It is however clear that technological progress and the advent of specialization have changed the structure of the economy; the relevant variables to be contracted upon are

more and more complex so that property rights become more and more difficult to enforce. As a consequence, contracts are necessarily incomplete and parties are forced to adapt in the face of unforeseen events. In the course of renegotiating their original agreement, parties haggle over the value of what needs to be adjusted. There are cost of transaction because each party is at the same time an opportunist predator and a pray to his partner's opportunism.

Transaction costs economics (TCE) then stresses the need to go beyond the description of economic institutions and focus instead on the actual unfolding of the economic activity. Attention is shifted on the attributes of the transaction and the properties of alternative modes of governance, where the latter is understood as an effort to craft order i.e., to mitigate conflict among parties in order to achieve the mutual gains from exchange. The last level is the neoclassical one dealing with marginal analysis, the fine tuning of prices and explicit incentives, the fact that prices and output continuously respond to changing market conditions. We now proceed to review the three economic theories in the reverse sense.

13.3.2 The Neoclassical Firm

The premise of this theory is the technological frontier relating inputs and output studied in §2.1.2. It is known as the "black box" since its internal operation fall outside the realm of economics. It assumes that both can be bought and sold on competitive (spot) markets so that price taking behavior is always optimal. Given that output is sold at a constant market price, the profit maximizing firm has the adequate incentives to minimize its costs by choosing the best combination of inputs, applying the correct managerial effort and so on.[27@] The result is the cost function with its variable and fixed parts. The optimal production equates price with marginal cost but takes care to cover fixed cost i.e., price has to be greater than average cost.

The average cost curve is often pictured as U-shaped because as soon as the firm operates it has to pay some fixed costs for plants and machines. Over some production range, the marginal cost is constant because duplication of output only requires duplication of inputs in a fixed proportion. Thus the average cost is decreasing. There is however a scale of production beyond which the duplication of inputs like preferred location or managerial talent becomes difficult, then marginal cost rise and so does average cost. This gives a simple explanation to the limited size of a firm.

Large output levels enable firms to specialize employees in specific tasks, to apply more efficient processes and invest in cost reducing techniques. Typically, a plant with several production lines will not stop producing when a failure occurs because the inputs

destined to the failing line can be reassigned to other lines. Likewise the supply of many (more or less independent) markets rather than a unique one guarantees the firm an almost constant demand in the short term (insurance from conjuncture shocks). Those arguments are captured in the neoclassical economic theory by the ideas of scale and scope economies (cf. §2.1.5).

Hart (1995) recalls that this theory has successfully linked production decisions with prices of inputs and outputs, explained the aggregate behavior of an industry or the strategic interactions when output markets are not competitive (oligopoly). Still this theory says nothing about the size of a firm or about its internal organization. True, a manager's talent is limited but hiring a second one should solve any duplication problem related to this specific input. In a nutshell, the neoclassical theory is a theory of plant size not of firm size. Indeed, the legal definition of the firm relates to the ownership of capital, thus among other thing plants. A single tycoon could own all plants on the globe, each one having an ideally chosen size. Alternatively, all plants could be independent firms. In fact any combination is compatible with the neoclassical theory and this is why it fails to explain how firms are born, grow and die.

13.3.3 Transaction Costs Economics

Transaction Modes: Make or Buy

Starting with Coase (1937), economists started to argue that institutions such as firms or public agencies serve the purpose of facilitating exchange and can best be understood as optimal responses to contractual constraints rather than production constraints.[28@] The original quandary regarding the nature of the firm is "make or buy": should an input be bought on the market or made inside the firm? Should a product be sold to a retailer or directly proposed to the end-user ? The options contemplated here are backward and forward integration. If the firm goes this way, it is bound to grow in size since it will need more assets to produce internally.

The major difference between the two modes regards the relationship of the involved parties. *External* procurement through the market involves a business relation between a seller and a buyer who stand on equal foot. The market price paid for the item provides parties with incentives to behave adequately; they end up maximizing the joint value of their relationship. More generally, contractual clauses establishing payments contingent on some specific performances are incentive devices. *Internal* provision involves an authority relation between a boss who orders production and a subordinate who carries on the order.[29@] The integration of a productive asset therefore shifts the terms of the relationship from a price mode to a quantity mode. More generally we can classify the

way a firm buys an input or sells its output:

- *Spot market*: buyers and sellers meet anonymously to exchange standardized products.
- *Bilateral relationship*: parties sign a long term contract specifying in a detailed manner how they make and exchange a product.
- *Integration*: one party of the previous relationship buys the other to switch the relation mode from price to quantity.

The choice of one mode over other alternatives obviously responds to monetary considerations (i.e., cost vs. gain). Whenever, an activity takes place between independent firms, there are so called *transaction costs* while there are *management costs* if the activity takes place within the firm. A simple marginal substitution argument tells us that the limit to the size of a firm is reached when the management cost of marginally extending the firm's activities is equal to the transaction cost.

At this point, it would seem that internal and external provision are fundamentally different. This is not so because all the transaction modes previously described are governed by contracts, the market transaction being the simplest of all; it is only the items contracted upon that differ. Even so, there remain many similarities. For instance, a sales and labour contract both stipulate a quality, a quantity and a price. In the former case, quality refers to the description of the good to be produced while in the former case it refers to the job design. Quantity then can refers to output to be traded (sale) or input to be provided (labour). Lastly, the price can incorporate cash or in-kind transfers or perks.[30@] The difference between the internal and external modes will therefore lie in finer details regarding what is observable and verifiable with sufficient precision to be incorporated into a contract.

Contract theory study how contracts are designed and what purposes they perform. A very active branch is agency theory (cf. Part H) which summarizes the firm as a bilateral relation between a subordinate, the *agent* and a supervisor, the *principal*. Its premises are that effort is unpleasant or costly and difficult to observe. To guarantee that employees or contractors work hard, it is thus necessary to pay them in function of their output. The agency theory has elaborated on this basic story to explain the nature of incentive contracts that are observed within firms, between firms or between firms and governments.

Yet this theory does not put a limit to the number of contract the owner of the firm can sign with its employees or commercial partners. Once again, it fails to pin down the boundary of the firm. Also, the fact that a buyer and a seller are linked by a contract has no implication on their legal relation; they may be independents firms or divisions of a

single firm; hence the agency view is consistent with the existence of a unique huge firm in the world or infinitely many small ones. It tries to explain the working of relations inside firms or among firms but not the fundamental question of this chapter "why do firms exists ?"

Transaction Costs

In an attempt to answer this quandary, Williamson (1975, 1985) has developed the Transaction Costs Economics (TCE) out of two observations pervasive of bilateral contracting. Firstly, it is difficult or costly to fully specify the contracting parties' obligations as we explained in §13.2. Secondly, parties develop "relationship-specific assets" of the following kind:

- Site specificity: parties commit immobile physical assets.
- Physical and Human asset: parties invest in specialized physical assets or human capital.
- Dedicated assets: parties build productive capacity for which there is insufficient demand absent their trading relationship.

In the presence of relationship-specific assets, parties become "locked in" to one another over the course of their relationship because the next best use of their specific assets is low. Indeed, site specificity raises transportation costs, specialized assets have a high switching cost and dedicated assets have a low rate of return. The "lock-in" then generates ex-post "quasi-rents", because the value of trade within the relationship far exceeds the value of outside trading opportunities (cf. §2.4.3 on bargaining and in particular Figure 2.9).

When both contractual incompleteness and assets specificity are present, renegotiation is bound to occur so that the contracting parties face ex-post opportunistic behavior: each may engage in inefficient rent-seeking in an attempt to "hold-up" the other party and obtain a larger share of the quasi-rents (cf. §16.3 for a detailed introduction to rent-seeking). Whenever the two independent firms are merged, it is the common boss of the two divisions who decides how to share the quasi-rent, thereby reducing the quibbling of the formerly independent division managers. Our study of rent-seeking in §7 (p193 and p62) supports this view on two grounds. Firstly, asymmetrically empowered rent-seekers, such as the boss and a division manager, jointly spend less in wasteful rent-seeking activities than if they are symmetrically empowered (bargaining between two bosses). Indeed, since the boss has a comparative advantage, he can fire the manager, it is not worthwhile to seek rent with great effort for the manager, so that the boss can also reduce his own effort. Integration, in this respect, gives more power over assets

to one party and takes away power from the other party. Secondly, the bargaining over quasi-rents by two independent parties often involves asymmetric information regarding rent-seeking abilities. In that case, there is a welfare loss because parties expand effort to seek rent. Integration then eliminates this waste.

Even if the sharing of quasi-rents does not lead the parties into wasteful rent-seeking,[31@] the incentives towards investments into specific assets are diluted because each is a partial claimant to the fruits of his efforts; under-investment then results. A merger then, has the potential to improve upon this situation since the common owner is residual claimant. This point is at the heart of the property right theory (PRT) developed in the next paragraph. Whinston (2003) argues also that the merger improves the coordination of investment and trading but increases bureaucratic costs. A central prediction of the TCE is that greater absolute levels of quasi-rents increase the likelihood of vertical integration (when contracts are incomplete). We study several models of TCE in §14.4.

The wider implication of TCE is that the choice of an optimal governance structure moves from "anonymous market transactions" to "administrative rules" as difficulties appear such as complexity of the product, information asymmetry, moral hazard, hold-up, absence of property rights. Each move in the chain of Table 13.3 entails more security, less incentives and more bureaucratic costs.

spot market → short term contract → long term (incomplete) contracts → integration into private firm → regulated firm → bureaucracy

Table 13.3: From Market to Bureaucracy

Formalization

To understand the impact of "make or buy" " onto the production technology, Nickerson and Vanden Bergh (1999) recur to the dichotomy between *mass* and *job-shop* production. On the one hand, a job-shop utilizes flexible re-deployable tools albeit with a limited output capacity (per unit of capital equipment). Mass production, on the other hand, develops inflexible speedy machines that increase production capacity. If capital expenditures are identical, the job-shop has a smaller production capacity and more generic capital equipment than mass production. Increasing the asset specificity then amounts to switch towards a larger capacity and it is then possible to identify asset specificity with total output of the firm.

Now, the "make or buy" choice boils down to comparing two traditional cost curves. From the discussion on transaction cost, we may assume that the fixed cost of using standardized market exchange is less than the fixed cost of setting up a hierarchical relationship. Conversely, the additional cost of adapting to an increased specificity is

greater when organized in the market compared to when organized via hierarchy. But such a cost variation is that due to a larger capacity requirement which is usually known as the marginal cost. If the two cost curves cross, as on Figure 13.3, the market procurement is optimal for low levels of specificity (or low sales volume) while the hierarchical organization dominates for high levels of specificity (or high sales volume). Notice that the cost curve which is effectively observed has a kink and the property that when hierarchy is in use, marginal cost is low and fixed cost is high while the reverse holds when the market is used. Thus, comparisons of the two modes using descriptive statistics can suffer from sample selection bias . One mode should not be compared to the other, but to a counterfactual scenario for without this correction, it may falsely appear more efficient and cost effective than the other (cf. discussion at the end of §14.4.1). Lastly, for situations where one curve is entirely above the other, there is an unconditional optimal organization of production.

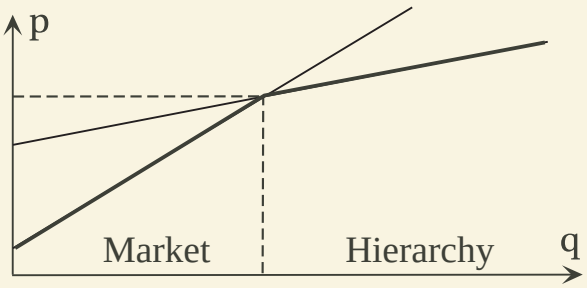


Figure 13.3: Optimal Internal Organization

Empirical Findings

There are few econometric studies on the TCE but they support it fairly well. Masten (1984) studies the procurement decisions of a large aerospace company over 1,887 components. The two variables having a positive effect on the likelihood of integration are the degree of component complexity which measures the difficulty of making complete contracts and the degree to which the component was specific to this firm. He also found that a combination of the two features reinforces the probability that the component would be made "in-house".

Monteverde and Teece (1982) explore the level of internal versus external procurement for 133 components used by GM and Ford in 1976. Unlike Masten (1984) , they do not focus on physical asset specificity, but on the (un-patentable) know-how that is generated during the design development. In their words: "The existence of transaction-specific know-how and skills and the difficulties of skill transfer mean that it will be costly to switch to an alternative supplier. An assembler (GM or Ford) will tend to choose

vertically integrated component production when high switching costs would otherwise lock the assembler into dependence upon a supplier and thereby expose the assembler to opportunistic re-contracting or to the loss of know-how". The key variables are the level of engineering effort to measure know-how acquisition and the specificity of the component (whether usable in other cars). These authors find evidence that increases in both variables raise the likelihood of integration.

Joskow (1985) studies the coal procurement decisions of electric utilities and finds that vertical integration is more likely for "mine-mouth" electric generating plants, those which sit next to the source of their fuel.

Hold-up

The first allusion to the hold-up is Marshall (1890) (V.XI.33) who observes that partial monopolists, such as railway, gas, water and electrical companies, attempt to raise their charges on the consumer who has adapted his plant to make use of their services; he cites a Pittsburgh steel mill who after switching from coal to natural gas is faced with a sudden doubling of the price. Once the mill owner has sunk the cost of setting-up his machines, the cost of making steel includes only labour and energy so that it is still worthwhile to produce although the energy component is dearer; this is so because the capital component has been sunk i.e., does not enter into the equation anymore. This *ex-post opportunism* of the power utility amounts to *hold-up* the profits of the mill and can even reduce them so much that there is not left enough to repay the initial investment (sunk cost).[32@] Absent a binding contract or integration of the two firms, the utility cannot commit not to act opportunistically because this is just being rational. Reasoning backwards, the mill owner will invest less or not at all if he anticipates that this future hold-up will prevent him to recoup today's investment.

Williamson (1971) and Goldberg (1976) reintroduce the topic in modern economics by comparing short term, long term contracts and the potentital benefits of vertical integration. Klein et al. (1978) discussing the historically famous case of GM and Fisher Autobody, already mentioned by Coase (1937): Fisher made car bodies and sold them at a high price to GM who then asked Fisher to build a plant adjacent to the GM plants. Fisher refused and GM eventually bought Fisher out. The economic analysis goes as follows: suppose that after having signed a contract for delivery of car bodies, Fisher builds a plant nearby GM's ones. Once Fisher's plant has been set up, GM can threaten to break the relationship pretending to buy car bodies from another supplier; this would leave Fisher in a very difficult situation because he has specialized its assembly line to fit GM's models and can hardly sell those bodies to another car maker. Thanks to this ex-post opportunism, GM could force a renegotiation and obtain a better deal. The story

goes saying that Fisher fearing not to be able to recoup his investment refused in the first place to make that investment (the plant located in Detroit) which in turn triggered the decision by GM to buy them altogether to make that desired investment. The topic is studied formally in §14.2.

13.3.4 Property Rights Theory

A critic

Alchian and Demsetz (1972) developed the Property Rights approach in economics as a critical development of the TCE which, according to their view, fails to explain the decision to integrate an asset. Grossman and Hart (1986) and Hart and Moore (1990) provide us with the first models of asset integration and argue that although transaction cost arguments explain well the switch from short term anonymous spot market relations to long term ones (repeated bilateral trade), they are silent with respect to the decision to buy the partner. Likewise, the problems of moral hazard or adverse selection identified by the agency theory are not improved by integration; it is difficult to sustain that more relevant information becomes observable after integration or is more easily used by parties to design better contracts. Whatever audit that the new integrated firm can perform to extract information and reduce agency costs could be mandated from outside if firms were not integrated. In other words, opportunism is likely to remain a plague within firms where decentralization and delegation is the rule.

Ownership vs. Control

Legally speaking, the ownership of an asset is associated with ownership of the revenues accruing from that asset. What the property right approach to the firm stresses is that ownership also gives power over the use of the assets. Grossman and Hart (1986) borrow from the theories of agency and transaction costs to build a contract theory oriented on the use of assets:

- A firm is the list of assets it owns.[33@]
- Ownership of an asset confers all rights over its use and the returns it generates.
- The owner can cease a *specific right* like a specific use of the asset.
- Transaction cost are large, contracts are incomplete, hence residual rights matter.
- *Residual rights* give the power to decide on the use of the asset whenever something unexpected arises.

Consider the example of a publisher Alan, a printer Bart and his press, the asset. A specific right is the use by Alan of Bart's press to print copies of a book. The contract among Alan and Bart can be fairly precise about the obligations of Alan regarding quality of the material to be printed or the obligations of Bart regarding quality of the printed material but it cannot possibly specify all details of the use of Bart's press in every conceivable situation, so that when something unexpected occurs it is Bart who decides what to do. For instance, it is not written in which language Alan's material must be submitted but the jurisprudence has stated that Alan has the residual right to choose any roman language while Bart can legally prohibit asian languages (that are far more costly to print). Now, if Alan buys Bart's press, he can decide to print in Chinese since he holds the *residual control rights* over Bart's asset (the press); the added cost will be borne by Alan.

The decision for Alan to buy Bart's asset is really about the size of Alan's business; it is related to long term risk. By having the residual control rights over Bart's press, Alan makes sure that whatever happens in the future, he will be able to have the asset work for him if he needs it (to print rapidly more copies). He will not have to design a complex and costly contract to convince Bart to do what he wants. The potential negative side of integration is the distortion of the operator's incentives: Bart is likely to be less careful with the press once he has sold it to Alan. The property rights theory thus predicts that the owner of an asset should be the agent most able to use it; for instance my house or my car are my property because it maximizes my incentive to take care of them, hence to maximize their value of use.

The Insurance example

Grossman and Hart (1986) illustrates their property rights theory with the following analysis of the insurance industry classified into direct writers who own their clients database and indirect writers who don't. In the former case, the sellers of insurance might well be called employees since they are not able to sell insurance products from a different firm. In the latter case, the sellers are (independent) agents able to sell various brands of the same insurance product. From the point of view of the previous theory, there is a single asset, the client list, and the question simply revolves around its ownership (non-integration has no meaning). As we shall see now, the markets where one form dominates the other corresponds fairly well to the prediction of the theory.

The investments undertaken by an insurer can be training of salesmen, advertising, product development, policyholder services. A retailer, whether employee or broker, invest into building loyalty with its clients; it is a long lasting investment that can produce long lasting premiums. Due to their non-verifiability, these investments cannot be

contracted upon; instead insurers and retailers contract upon sales (initial ones and renewals). In this typical moral hazard situation (cf. §20), retailers are induced to work hard only if companies design a commission scheme that goes beyond the signing of a client, for instance they can integrate renewals bonuses.

If for reasons of cost or competition an insurer decides to give-up a profile of clients over a particular product in a particular region, then the ownership of the client list will make a difference for retailers. A sales employees who has spent a lot of time to sign clients will lose them together with the renewal bonuses; on the other hand, a broker is protected against such a situation since he will be able to costlessly switch its clients to a new company and maintain its level of satisfaction.

Symmetrically when a company develops a new product it must agree to share future profits with independent brokers in order to have access to their clients. Brokers are also able to play one company against the other to extract more surplus from consumers. These opportunistic attitudes affect negatively the investments of insurers; there is a hold-up (cf. §14.2 on the hold-up).

A life-insurance policyholder has less tendency to switch company than a the holder of a car insurance, also his renewal is much less sensitive to the retailer's effort (there's no claim until death occurs); hence the commission scheme can put most weight on the signing a new life policy and still ensure optimal effort from retailers. The retailer is now much less sensitive to the insurer's ex-post behavior and will therefore easily relinquish ownership of the client list. On the other hand, for all products like car, house or fire, where renewal is not guaranteed and depends on the current conditions of the policy as well as the care taken by the retailer, the latter is more likely to hold the client list. Stilzed facts support this view since in the US, 65% of property-casualty insurance is generated by brokers against 12% for life insurance.

Wage bargaining

Unions and firms periodically bargain over wages and employment. The standard negotiation is on "wage rate" only, leaving firms free to choose later on the level of employment. Another popular alternative is the "wage contract" negotiation whereby a minimum amount of labor-days is also agreed.[34@] Grout (1984) observes that in the UK (as opposed to the US), the legal protection bestowed upon unions prevents them from committing to any agreement i.e., they can always go on strike to grab an additional advantage (this was true in the 1980s). Whatever the agreement initially signed, the firm anticipates the future renegotiation towards higher wages. Since technical efficiency in production commands to equate the MRTS (cf. eq. 2.1) to the ratio of wages over the cost of capital, the firm will under-invest. On the contrary, in the US where agreements are

binding (enforced by the law), firms are lead to invest optimally.

The same argument indicates that specific human capital investment by employees (e.g., learning to use an in-house software) is likewise inefficiently low since any worker (or his union) fears the future expropriation by the management of the additional profits his devoted attitude will generate. Generally speaking, building a reputation for sticking to its word permits to reduce the hold-up and thereby improves efficiency within the relationship.

Chapter 14

Vertical Integration

Our starting point in this chapter is the observation that nearly all the goods and services we consume are the outcome of a complex process whereby raw materials are transformed into basic goods or components that are finally assembled into salable items. In parallel to this industrial vertical chain, we also find a chain of support services such as accounting, administration or marketing helping. Vertical integration occurs when two firms active at different but contiguous levels of these vertical chains, decide to merge. The reverse process is called outsourcing or spin-off.[1@]

The object of this chapter is to formalize the arguments developed in the previous one regarding the forces behind the creation, expansion or contraction of firms. In the first section, we present the bilateral monopoly whereby two vertically related firms trade. The application of market power by each of them (aka double marginalization) is conducive of inefficiency which is a first reason to either integrate or sign restrictive agreements. We then turn to the hold-up problem which originates in the specific investments that parties to a trade need to undertake in order to make the most of their relationship. The last two sections analyze the cost and benefits of vertical integration according to the property rights and transaction cost theories.

14.1 Bilateral Monopoly

In this section, we shall see that vertical relationships between independent entities may generate an harmful "double marginalization". We then study under what conditions it is beneficial for the parties to integrate vertically (in order to eliminate this problem). We also look at alternatives to integration and their potentially anti-competitive consequences.

14.1.1 Complementary Monopolies

Cournot (1838)'s treatment of the bilateral monopoly is in fact a study of side-by-side monopolists as would be Intel (firm A) and Microsoft (firm B) when selling respectively, chip and operative system, to competitive computer makers.

Given that each unit of final product incorporates components A and B, the unit cost of a buyer depends on the sum of prices and so does his demand (whatever his market power in his own market). From the point of view of A, B's price depresses the WTP of buyers; it is as if B's price was an added marginal cost for him. Now when firm B exercises market power, she sets her price above her own marginal cost, so that A perceives a demand further depressed. His own profit maximization will proceed as if B's marginal cost was higher than it really is, thus he chooses to reduce sales by naming an even greater price. It is now clear that the two monopolies have entered a spiral of increasing prices that moves the sales away from the joint profit maximizing solution where only true marginal cost would enter calculations. The limit is reached when both prices are best reply to each other; this constitutes a Nash equilibrium.

Let us illustrate this phenomenon analytically. Given that each unit of final product incorporates components A and B, the demand received by A and B is identical and a function $D(p_A+p_B)$ of the total price of the two components. Firm A's profit thus reads $\pi_A = p_A D(p_A+p_B) - C_A(D(p_A+p_B))$. Since $D(p_A+p_B)=q$ can be inverted into $p_A+p_B=P(q)$, we can rewrite profit as $\pi_A = (P(q)-p_B)q - C_A(q)$ so that the FOC for optimal sales is $R_m = C_m^A + p_B$ as if the marginal cost of firm A was inflated by firm B's price.

If firms A and B were integrated or acted collusively, their joint profit would be $\pi = (p_A+p_B)D(p_A+p_B) - C_A(D(p_A+p_B)) - C_B(D(p_A+p_B)) = qP(q) - C(q)$ where $p = p_A+p_B$ and $C(.) = C_A(.) + C_B(.)$. The optimal choice $\bar{q}$ would solve $R_m = C_m = C_m^A + C_m^B$. Letting $\bar{p}_i \equiv C_m^i(\bar{q})$ for $i = A,B$, we observe that the optimal price $\bar{p} \equiv P(\bar{q}) > R_m(\bar{q}) = C_m(\bar{q}) = \bar{p}_A + \bar{p}_B$.

We can now easily show that the sum of uncoordinated prices (side-by-side monopolists) overshoots the joint profit maximizer. If firm B was to price at $\bar{p}_B$ then firm A's optimal sales would be $\bar{q}$ and she would achieve that objective by setting the price $p_A = P(\bar{q}) - \bar{p}_B > \bar{p}_A$. In response to that choice, firm B would solve $R_m = C_m^B + p_A$ which would entails a quantity $q < \bar{q}$ and therefore a price $p_B = P(q) - p_A = P(q) - P(\bar{q}) + \bar{p}_B > \bar{p}_B$. This proves that the pair $(\bar{p}_A, \bar{p}_B)$ is not an equilibrium of the indirect interaction among A and B and that a stable situation involves greater prices for both firms. The Nash equilibrium is the solution of the system in 3 unknowns $R_m(q) = C_m^A(q) + p_B = C_m^B(q) + p_A$, $p_A + p_B = P(q)$.

Consider for example, constant marginal cost c_i for $i = A,B$ and linear demand $D(p) = \alpha - \beta p$; the efficient quantity is $q^* = \alpha - \beta(c_A + c_B)$, the joint monopoly one is $\bar{q} = \frac{1}{2}q^*$ and it is easy to check that the Nash equilibrium is $p_A = \frac{\alpha+\beta(2c_A-c_B)}{3\beta}$, $p_B = \frac{\alpha+\beta(2c_B-c_A)}{3\beta}$, leading

to sales $\hat{q} = \frac{1}{3}q^*$. Uncoordinated firms further restrict trade as compared to the optimal choice of a joint monopoly.

14.1.2 Vertical Monopolies

The situation most interesting for modern economics is the face-to-face bilateral monopoly illustrated below where an upstream potential maker of an item, the seller S exchanges it with a downstream potential user, the buyer B. An example would be Samsung negotiating with Apple the making of flash memory hard drives for the iPod or a manufacturer like Nestlé negotiating with a retailer like Carrefour:

	Manufacturer		Retailer	
Raw	$\longrightarrow$	Intermediate	$\longrightarrow$	Final
input	*Upstream*	good	*Downstream*	good

Pareto Optimality

Let us start with some definitions. If a quantity q is produced and exchanged, the seller incurs production cost $C(q)$ while the buyer derives a net revenue $R(q)$ (treating the item as a free input). Given an agreed total price F (or unit price F/q), profits are $\pi_S = F - C(q)$ for the seller and $\pi_B = R(q) - F$ for the buyer. The joint profit is thus $\pi(q) \equiv \pi_S + \pi_B = R(q) - C(q)$. This expression is maximized by the choice of $\bar{q}$ solving $R_m = C_m$ and yields a maximum profit $\bar{\pi} \equiv \pi(\bar{q})$.

As we showed in §2.4.3 on bargaining, the parties will insist on trading $\bar{q}$ since any other trade q, whatever the unit price it involves, can be improved by this particular one to the benefit of both parties. There is indeed an additional value $\bar{\pi} - \pi_S - \pi_B$ that is created when moving form q to $\bar{q}$; it can be shared among the two parties, making both of them willing to accept the quantity change. If the renegotiation of the initial contract is costless, then this is necessarily the outcome (cf. Coase theorem §8.1.3). What remains unclear is the average or unit price, since the cake can be shared in any way among the parties.[2@] Incorrect solutions to the bilateral monopoly problem will be presented at the end of the section.

Incomplete Contracts

Modern contract theory offers a compelling explanation of why the competitive price $\bar{p} \equiv R_m(\bar{q}) = C_m(\bar{q})$ is most likely to be observed. Consider an initial agreement between S and B to deliver $\hat{q}$ units at unit price $\hat{p}$. The seller can at very low cost renege on his scheduled deliveries by blaming the delay on unexpected problems; he can nevertheless

insist on being paid for the delivered units. Likewise, the buyer can renege on the scheduled quantity by arguing that the undesirable units are unsatisfactory; he thus only pays those he accepted. What this trivial analysis of contract law suggests is that only the actually exchanged units have to be paid since it is easy to demonstrate to a judge that they were delivered and accepted (both signatures appear on the delivery receipt). Forcing both parties to perform with respect to the scheduled quantity is much harder because so many excuses are acceptable at first hand; it would take time and expertise to identify those which are acceptable and those which result from malevolence.

The quantity optimally delivered by the seller is thus the minimum of the agreed one and the competitive one at the specified average price. As illustrated on Figure 14.1, given an agreement over quantity q_0, the seller fulfills only if the average price is greater than p_2, otherwise he delivers less than scheduled. The buyer, on the other hand, fulfills only if the average price is lesser than p_1. Thus, only prices between p_2 and p_1 are conducive of fulfillment of promises by both parties. As the quantity is increased towards the joint profit maximizing level $\bar{q}$, the range of acceptable prices shrinks towards a unique level, the competitive price $\bar{p}$.

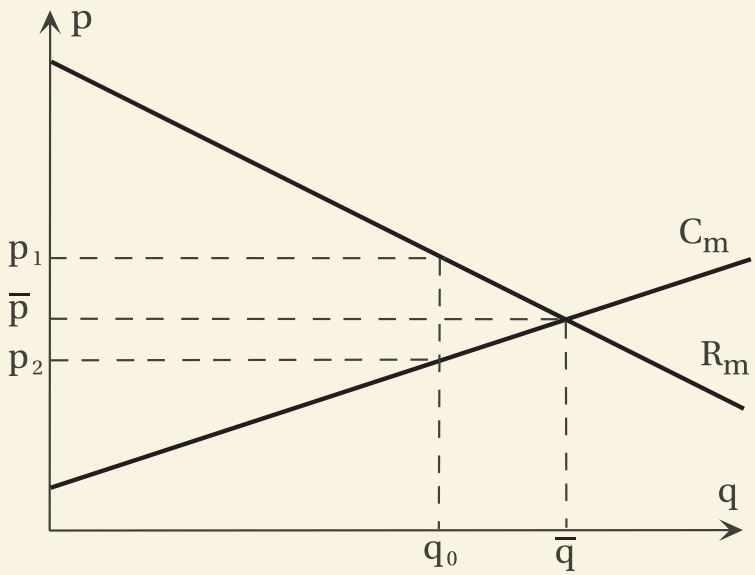


Figure 14.1: Sharing driven by contractual incompleteness

Pure bargaining

In many situations, it is possible to write a binding contract covering not only quantity but also price. Bargaining then takes place ex-ante at the design stage to determine how the value created by an optimal trade is shared among the parties. Let us assume that

if negotiation fails, each party can secure a default payoff π_i for $i = S, B$. The quasi-rent to be shared by coordinating the sale is $\delta \equiv \bar{\pi} - \pi_B - \pi_S$ and this is where bargaining is involved.

The bargaining theories recalled in §2.4.3 converge in the case of certainty towards equal sharing of the quasi-rent. To conclude, the jointly optimal quantity (yet often inefficient from the social point of view) will always be traded.

14.1.3 Double Marginalization

The Issue

Many economists dissented with the above bilateral monopoly analysis; they adhered to a different and sub-optimal solution akin to Stackelberg (1934) leadership: one party, often the seller, names a unit price p while the other party, often the buyer, names a quantity q, trade then proceeds.[3@] If the buyer applies his market power upon the downstream segment of the market, he maximizes $\pi_B = R(q) - pq$ which leads him to demand the quantity solving $R_m(q) = p$. The marginal revenue generated by the item for trade thus appears to be also the willingness to pay of the buyer. Now, the seller who is a first mover, anticipates this demand in order to maximize her own profit; she applies market power over the buyer's WTP. The seller profit being $\pi_S = qR_m(q) - C(q)$, the FOC is $C_m = R_m + q\frac{\partial R_m}{\partial q} < R_m$ since marginal revenue is decreasing; this inequality means that the the seller selects a quantity lesser than the optimal one $\bar{q}$.[4@]

A *double marginalization* is at work here because the buyer's demand is more inelastic than the market demand at the origin of the revenue $R(q)$. Another way to see the sub-optimality at play is to reason that only if the seller forgoes his market power and sets price competitively (i.e., $p = C_m$) will the buyer demand the right quantity. The overall sum of profit thus falls short of its maximum $\bar{\pi}$ when monopolies limit themselves to such simple allocation rules ("you name a price, I name a quantity"). It should therefore be clear that unless some external conditions impeach the parties to either integrate or design binding contracts covering both quantity and price, no sane businessmen would ever continue to trade under such sub-optimal rules. An extremely simple resolution is to allow the first mover to price discriminate the other party by making for instance a price listing or a bundle offer (cf. §4.1.3 on consumer surplus extraction).

Remedies

Double marginalization does not show off if the downstream market is competitive since only the upstream firm has the ability to apply market power. Even in that case, the

seller might fall short of maximum profits if the production of the final good uses several inputs that are, to some extent, substitutable. Consider the example of Tetrapak a leading maker of carton packages for beverages. When Tetrapak, the upstream firm, raises its price to a monopoly level, the downstream retailers like the dairy producer Danone start to substitute the expensive carton for the now relatively cheaper plastic or glass bottles; they optimally adjust their mix of inputs so as to keep their technical rate of substitution equal to the price ratio. As a consequence, the downstream production uses an inefficient combination of inputs and the demand for the upstream monopolist's input falls (Tetrapak sells less carton). The ability to recoup full monopoly profits is lost for the presence of a substitute (plastic) to the upstream input (carton). The only way to restore profits is to integrate downward to control the choices of inputs or use the alternative contract presented hereafter.

The solution uses price discrimination (cf. §4.1.3) with a two-part tariff where the unit price is set equal to the marginal cost; this option incentivizes the retailer to sell the monopoly quantity. It remains then to negotiate the franchise fee so as to share the joint monopoly profit among the upstream and downstream firms (cf. §2.4.3 on bargaining). If there are several downstream firms, the subscription will need to be individualized (cf. §4.2.1 on perfect discrimination). The problem is that the fee usually paid on an annual base tends to be very large putting the retailer at risk of bankruptcy especially when demand can be fluctuating.

To conclude, we might say that even if there is market power at all levels, there are many ways for the upstream firm to restore optimality (NOT efficiency) i.e., maximize the sum of profits for the firms involved in the production of the final good.

Investment Incentives †

We present here a preview of the general result of the next section on relation specific investments and the hold-up problem, namely that firms invest less into their relationship than if they were under common ownership because they anticipate that future gains from trade will be shared according to a flexible rule, sensitive to their behavior.

Imagine that revenue and cost can be enhanced by specialized investments i.e., expenses dedicated at increasing revenue or reducing cost but exclusively for the exchange of the item of interest for the two parties. Let k_B and k_S denote the amount spend by B and S respectively into that activity. We also denote $C_q \equiv \frac{\partial C}{\partial q}$ and $C_k \equiv \frac{\partial C}{\partial k_S}$ and assume $C_{qq} < 0$ i.e., DRS and $C_{qk} < 0$ i.e., investment improves the marginal cost. The maximum joint profit being $\pi(\bar{q}) = R(\bar{q}) - C(\bar{q})$, the marginal value of investment into the production technology for joint profits is $\frac{d\pi}{dk_S} = \frac{\partial \pi}{\partial k_S} + \frac{\partial \bar{q}}{\partial k_S}\frac{\partial \pi}{\partial q} = -C_k$ by definition of $\bar{q}$ which continues to satisfy $R_q = C_q$.

We have previously argued that the only stable exchange price among two independent firms is $\bar{p} = C_q(\bar{q}) = R_q(\bar{q})$. Profits are thus $\pi_S = \bar{q}\bar{p} - C(\bar{q})$ and $\pi_B = R(\bar{q}) - \bar{q}\bar{p}$. Differentiating the FOC of optimality of $\bar{q}$ with respect to k_S, we obtain $\frac{\partial \bar{q}}{\partial k_S} = \frac{C_{qk}}{R_{qq}-C_{qq}} > 0$ which will be used hereafter. The marginal value of investment into the production technology for the seller is $\frac{d\pi_S}{dk_S}$ made of three terms. The first and direct effect is $\frac{\partial \pi_S}{\partial k_S} = -C_k$ as in the joint profit case. The second term is the indirect effect through the price $\frac{\partial \pi_S}{\partial p}\frac{\partial \bar{p}}{\partial k_S}$ while the third term is the indirect effect through the quantity $\frac{\partial \pi_S}{\partial q}\frac{\partial \bar{q}}{\partial k_S}$. If we are able to show that the sum of indirect effects is negative then the seller puts less value on investment than a joint owner. Observe first that $\frac{\partial \pi_S}{\partial p} = \bar{q}$ and $\frac{\partial \pi_S}{\partial q} = \bar{q}\frac{\partial \bar{p}}{\partial q} + \bar{p} - C_q = \bar{q}\frac{\partial \bar{p}}{\partial q}$ since $\bar{p} = C_q(\bar{q})$ by definition of the agreed trade. Then, the indirect effect is

$$\bar{q}\frac{\partial \bar{p}}{\partial k_S} + \bar{q}\frac{\partial \bar{p}}{\partial q}\frac{\partial \bar{q}}{\partial k_S} = \bar{q}C_{qk}\left(1 + \frac{C_{qq}}{R_{qq}-C_{qq}}\right) = \bar{q}\frac{R_{qq}C_{qk}}{R_{qq}-C_{qq}} < 0$$

as soon as revenue displays DRS, a feature of most markets.

A symmetrical inequality obtains when comparing the private marginal value of investment into the revenue technology and the jointly optimal value. This result is a formalization of Williamson (1971)'s under-investment into assets that are relation specific when parties anticipate that future gains from trade (π) will have to be shared according to a flexible rule.

Indeed, if it was possible to specify a fixed total payment F in return for delivering the optimal quantity then each party would face the ideal incentives to invest.[5@] The reason why F is difficult to choose is because uncertainty can provoke wide variations into R or C that could leave one party bankrupt if the payment F had to be fulfilled. It is therefore quite likely that F will be renegotiated, but if this is so, parties will anticipate this fate and modify accordingly their investment choices.

14.1.4 Resale Price Maintenance

We illustrate here how a resale price maintenance (RPM), seen in §9.2.2 on antitrust, can be an alternative to integration in order to overcome the double marginalization problem (cf. §14.1.3). Imagine that a manufacturer sells a product to competing retailers using a two-part tariff (p, F). There are neither production nor distribution costs ($r = c = 0$). Each retailer faces a market demand $D(p) = 1 - p$ divided in two segments, one of size α is protected from the competition of other retailers, for instance because the involved customers have a high opportunity cost of searching for lower prices. The remaining segment of size $1 - \alpha$ is open to price competition among retailers. The dichotomy could alternatively be between households and firms since the latter are more able to dedicate

resources to finding a better deal.

The industry maximizing behavior is to set the monopoly price $p = \frac{1}{2}$ in order to earn $\pi^M = \frac{1}{2}$ on each retail market (and $\frac{1}{2}$ on average). Let us denote w the *wholesale* price set by the manufacturer. In absence of a RPM, the equilibrium retail prices are $\frac{1+w}{2}$ over the monopolized segment and w over the competitive one, leading to individual retailer profits of $\alpha\frac{(1-w)^2}{4}$ which the manufacturer absorbs using a franchise F. Since the sales to each retailer are $\frac{1-w}{2}$ and $1-w$ over the two respective segments, the manufacturer's profit for each market is

$$\pi(w) = \alpha\frac{(1-w)^2}{4} + w\left(\alpha\frac{1-w}{2} + (1-\alpha)(1-w)\right) = (1-w)\left((4-3\alpha)w + \alpha\right)$$

and is maximum for $w_\alpha \equiv 2\frac{1-\alpha}{4-3\alpha} > 0$, so that the final prices in the competitive and monopolized segments are ranked as follows: $w_\alpha < \frac{1}{2} < \frac{1+w_\alpha}{2}$. The absence of RPM and the oligopoly structure in the downstream market therefore leads retailers to adopt a price discrimination that is wasteful for industry profits; although the manufacturer recaptures all profits through franchise fees, he earns only $\pi(w_\alpha) = \frac{1}{2}\frac{(2-\alpha)^2}{4-3\alpha}$, a convex function that reaches π^M at $\alpha = 0$ and 1 only.

To restore full monopoly profits when there is heterogeneity of consumers ($0 < \alpha < 1$), any of the three RPM will do the job:

- set $F = 0, w = \frac{1}{2}$ and *fix* w as a retail price.
- set $F = 0, w = \frac{1}{2}$ and set w as a *maximum* price.
- set $F = \frac{1}{4}, w = 0$ and set a *minimum* price of $\frac{1}{2}$.

The first strategy is tantamount to integration while the last one is the usual explanation for minimum prices: the wholesale price is set at the marginal cost (here 0) to avoid double marginalization and a minimum price guarantees that retailers will not loose their margin in price wars. The second strategy is a little known rationale for maximum retail prices: the wholesale price is set at the monopoly level corresponding to the competitive segment while the maximum price prevents retailers from overexploiting captive customers (double marginalization again).

In other words, if there are n classes of customers upon which price discrimination can be applied, the usual double marginalization problem is multiplied by n, so that the manufacturer needs $n-1$ new instruments to guide the pricing behavior of retailers towards industry maximum and rip the whole profit using the franchise fee (with adequate sliding). If, like in many markets, there are roughly two segments of captive and competitive (bargain seeker) customers, then a single RPM in each class does the job. The econometric study of Shepard (1993) on gasoline price at stations integrated or not with refiners confirms the theoretical findings of the model.

Observe that in this example, the average price without RPM is $\alpha\frac{1+w_\alpha}{2}+(1-\alpha)w_\alpha = \frac{1}{2}\frac{4-\alpha(2+\alpha)}{4-3\alpha} > \frac{1}{2}$, the average price with RPM; thus *RPM is welfare improving* since for homogeneous goods, total quantity is a faithful indicator of market welfare. This need not be true in more general models. Still, the case against RPM cannot be concluded in a clearcut manner from the state of actual research.

14.2 Specific Investment (aka Hold-up)

In §13.3.3, we commented upon the rational opportunistic behavior of a party to a bilateral trade who tries, by haggling, to grab the greatest possible share of the value to be generated by the exchange. When a party holds-up the fruit of the other party's effort to improve the relationship, the latter rationally refuses to invest efficiently; integration (vertical or horizontal) might then be a solution to force a cooperation at the ex-post phase thanks to the imposition of agreement by fiat (authority).

In this section, we present formally this theory and the contractual solutions that have been proposed as well as their limitations for two families of problems: trading a variable volume or a unique asset. We assume ex-post efficiency i.e., parties have the necessary information to identify an efficient course of action once all the relevant information has been revealed. In other words, they are able to bargain quickly over the rent to be shared without haggling or entering into wasteful rent-seeking behavior. The property rights theory (PRT) taken up in the next section, carries on with this hypothesis and develops models of the limit to the boundaries of the firm based on the resolution of the hold-up. The transaction cost approach (TCE), to be studied in the last section, emphasizes on the contrary the frequent inefficiency of ex-post haggling over quasi-rents and makes it the main reason for integration.

14.2.1 Ex-post Opportunism

Specific Trade

When a firm buys a quantity q of input on the world market at price p, she has efficient incentives to invest into her technology. Letting k denote the monetary investment, her profit is $\pi_B = R(k,q) - pq - k$. Even if the market price is random, the FOC of optimal investment k^* is $\frac{\partial R}{\partial k} = 1$ which is efficient. This is so because the unit price, however shaky it may be, is completely insensitive to the investment effort procured by the firm (recall it is world price determined by the aggregation of literally hundreds of bidders). As we shall see, bilateral trading changes this.

Consider now the case where the specific needs of the firm B force her to buy from a dedicated seller S whose cost is $C(q)$. They negotiate a quantity and a price. The potential value of trade is $R(q) - C(q)$ and will optimally be maximized by the parties if there is symmetric information i.e., each knows the valuation (revenue or cost) of the other party. The efficient (ex-post) quantity is $\bar{q}$ solving $R_m = C_m$; it leads to a surplus or quasi-rent $\delta \equiv R(\bar{q}) - C(\bar{q})$. If the negotiation fails, each party continues with his activity and earns a default profit assumed nil for ease of exposition.[6@] Notice that uncertainty could be present and affect the ex-post valuations R and C.

It remains to negotiate the exchange price and this is where the hold-up arises. We assume w.l.o.g. that the surplus δ is equally shared[7@] i.e., $\bar{p} = \frac{R(k,\bar{q})+C(\bar{q})}{2\bar{q}}$. The buyer's profit is now $\pi_B = \frac{1}{2}(R(k,\bar{q}) - C(\bar{q})) - k$, so that investment incentives have been halved (the marginal gain of an additional investment is only 50% of the wealth so created) A completely symmetric argument holds for the input seller.

The hold-up refers to the fact that the inescapable negotiation over the price characteristic of a bilateral relationship forces the buyer to share the fruits of his investment k with the seller, he thus become a *partial claimant* whereas efficiency requires him to be a *residual claimant*. Whenever some aspect of the investment can be contracted upon, it is in the interest of the parties to do so and specify that the efficient level k^* should be carried on. If this is not possible (or too costly), then integration might solve the investment problem since the common owner will internalize the externality that the seller's opportunism imposes on the buyer.

In §14.1.3, we obtained the same qualitative result when the "competitive" price was the equilibrium of the bargaining; this occurred due to the impossibility of signing a binding sales contract.

Commitment

We now show the fundamental role of commitment in the appearance of the hold-up i.e., the incentives towards investment differ according to whether the trade is negotiated after or before the investment is sunk. Let the maximum ex-post surplus be $\bar{\pi}(k)$. If the trade is agreed before the investment, the buyer's default payoff (in case of disagreement) is zero while if the trade is agreed after the investment, the buyer's default payoff is the resale value ρk of the capital invested. We assume that $\rho < 1$ because the investment was specially geared towards the trade with the seller, thus has less value in alternative uses. The two settings also differ with respect to where the investment cost is imputed into the buyer's payoff; if trade is agreed ex-ante then the investment cost is integral part of the negotiation while it is not if trade is agreed later on. We can now easily derive the buyer's ex-ante payoff with rational anticipation in the two settings and compare them.

variable	ex-ante trade	ex-post trade
ex-post surplus W	$\bar{\pi}(k)-k$	$\bar{\pi}(k)$
default payoff $\underline{\pi}_V, \underline{\pi}_B$	$0,0$	$0,\rho k$
rent $\delta = W-\underline{\pi}_V-\underline{\pi}_B$	$\bar{\pi}(k)-k$	$\bar{\pi}(k)-\rho k$
B's ex-ante cost c_B	0	k
$\pi_B = \frac{1}{2}\delta+\underline{\pi}_B-c_B$	$\frac{1}{2}(\bar{\pi}(k)-k)$	$\frac{1}{2}(\bar{\pi}(k)-k)-\frac{1}{2}(1-\rho)k$

Table 14.1: Hold-up and Commitment

When the agreement to trade is binding, the buyer is a residual claimant of his investment although he nets only one half of the absolute surplus generated by the exchange. When the agreement to trade is not binding (i.e., renegotiated at will), the buyer suffers a hold-up due to the opportunistic behavior of the seller; his optimal investment is thus inefficiently low (cf. §13.3.4).

Contractual Instruments

Chung (1991) shows that a simple contract can resolve the mutual hold-up if two conditions apply: one party can be given full bargaining power ex-post and it is possible to specify a default trade $(\hat{q},\hat{p})$ or at least an outside option to the other party. Regarding the first condition, we argued in §14.1.2 that parties will try to act opportunistically at the delivery–reception stage so that enforcing a specific quantity performance is difficult. To make a party dominant ex-post, it is possible to write in the initial contract that this party, say the buyer, will lead ex-post negotiation and make a "take-it-or-leave-it" trading offer. The seller then is forced to accept or walk away from the relationship.[8@]

The general formulation of the hold-up sees both parties invest into specific value enhancing assets so that revenue is $R(k_B,q)$ while cost is $C(k_S,q)$. The efficient ex-post quantity $\bar{q}$ continues to solve the same FOC but is now contingent on past investments. The trading surplus is $\pi = R(k_B,\bar{q})-C(k_S,\bar{q})$ while the default payoffs are $u_B = R(k_B,\hat{q})-\hat{q}\hat{p}$ and $u_S = \hat{q}\hat{p}-C(k_S,\hat{q})$. The quasi-rent to be divided during the ex-post negotiation is thus $\delta = \pi-u_B-u_S$.

Having all power, the buyer's offer grabs the entire quasi-rent so that his profit is $\pi_B = u_B+\delta-k_B = \pi-k_B-u_S$. By construction, the buyer has the right incentive to invest. Since the seller has no bargaining power, his ex-post gain is u_S and ex-ante, his profit is $\pi_B = \hat{q}\hat{p}-C(k_S,\hat{q})-k_S$. Given the efficient seller investment k_S^* solving $\frac{\partial C(k_S,\bar{q})}{\partial k_S}=1$ (where, recall, $\bar{q}$ is a function of k_S), it is possible to tune the constant $\hat{q}$ so that $\frac{\partial C(k_S,\hat{q})}{\partial k_S}=1$ (thanks to the intermediate value theorem) in order to induce the seller to make the efficient investment.

14.2.2 Buy-Out

We leave aside bilateral trading and concentrate on the exchange of a single asset such as an infrastructure, a software, a work of art. The asset requires specific investments by both sides to capture all its potential value. If the agent's investment could be contracted upon, there would be no hold-up since the principal, as the asset final owner, would be the residual claimant of the value added by her own investment. She would thus instruct the agent to invest efficiently and would have the natural incentive to invest efficiently herself. However, many dimensions of the agent's duty, although observable, are non contractible. This means that investment incentives must be provided for him to act diligently. To alleviate the hold-up problem, it is possible to supplement the initial contract with a clause whereby the principal can force the agent to "buy-out" the asset in order to discipline him. Alternatively, the contract first sells the asset to the agent and specifies a repurchasing price for the principal and a later date at which he may exercise this option or "opt-out".

Demski and Sappington (1991) show that if the parties can stick to their initial agreement i.e., commit not to renegotiate, then efficient investments obtain for both parties. Edlin and Hermalin (2000) however qualify this result because the proposed contract is not robust to renegotiation i.e., the principal can hold-up the agent by threatening to opt-out. Efficiency obtains only if investments are substitutes at the margin.

An animated example

Edlin and Hermalin (2000) illustrate the issue at stake with the Pixar-Disney collaboration for the movie *Toy Story*. Each party brought unique talents: Pixar had the 3-D animation technology and Disney had the distribution and marketing expertise for animated films. Prior to release, Disney could observe the quality of the film Pixar produced, but it is difficult, if not impossible, to describe quality in a contract or demonstrate it in court unambiguously. Giving Pixar appropriate incentives to make a great movie would, then, seem problematic. Disney could commit to buy the movie at a fair price but then Pixar, acting opportunistically, had no incentive to work hard which constitutes the archetypical case of moral hazard. If instead, the price is negotiated after the making of the movie then Pixar is victim of the *hold-up* problem for Disney will opportunistically (but rationally) bargain over the price and capture some of the value that Pixar created; Pixar anticipating this fate will not work so hard to make the perfect movie.

A good compromise is to fix a price initially and give Disney the option to cancel the deal. Buying the film at that price will only be attractive to Disney if Pixar makes a sufficiently good movie. The optimal price should equate the film's final value to Disney

assuming Pixar exerts optimal effort. Absent renegotiation, this contract is efficient, but Disney has an incentive to let its option expire and subsequently renegotiate a lower price for the film threatening to leave Pixar alone to market the movie. Even if the parties make the option non-expiring, renegotiation is still a relevant threat: Disney has an incentive to delay promotion and distribution until it can renegotiate a better deal.

Hence, the option contract provides little protection from the hold-up problem. Yet, there is a second and dynamic dimension to this story: Pixar's effort (making a movie) strengthens his bargaining position, since if bargaining breaks down Pixar ends up owning a great movie and valuable know-how that could interest Disney's competitors like Time-Warner, Dreamworks or Universal.

Facts confirm this theory. In 1991, Disney and Pixar signed an agreement for the development and production of up to 3 movies whereby Disney kept 85% of all profits but financed production, distribution and marketing costs. After the success of *Toy Story* in 1995, Disney and Pixar signed in 1997, a new agreement replacing the former, for 5 movies whereby production costs and profits were equally shared. Later on, after the tremendous success of *Finding Nemo*, Pixar decided it would not renew its association with Disney after delivering the fifth agreed movie. Finally, after several more blockbusters, Pixar had acquired so much goodwill that Disney had to buy it in a remake of the General Motors vs. Fisher Autobody story analyzed by Coase (1937).

Solution with Commitment †

The following simple agency model partly formalizes the Pixar-Disney story by introducing a one-sided hold-up problem and showing how it can be solved if firms can stick to their initial agreement i.e., commit not to renegotiate. In the next section we shall see under what conditions, the unavoidability of renegotiation breaks this positive result.

A principal (she), who owns a transferable asset like a patent, a store, or a movie idea, hires an agent (he) to realize a project by applying labor $L > 0$ upon the asset. Later on, the principal invest $K > 0$, for instance in advertising; both decisions are non contractible[9@] and are therefore chosen according to rationality i.e., each party takes the action maximizing its objective at that moment. The expected profit generated by the asset is $R(K,L)$ with $R_L > 0$ and $R_K > 0$. We denote $\Pi(L) \equiv \max_K R(K,L) - K$, the maximal value of the project given the amount of effort L previously applied; it satisfies $\Pi_L = R_L$ by the envelope theorem. Lastly, the reservation utility of the agent is set w.l.o.g. to zero so that the efficient effort (first-best) is L^* maximizing $\Pi(L) - L$ i.e., solving $R_L = 1$.

The optimal contract maximizes the principal's expected profit subject to participation of the agent. Consider an option contract (p_1, p_2) offered by the principal to the agent. Upon acceptance of the contract, the principal transfers ownership of the asset

to the agent for a price p_1. After the agent has applied his effort, the principal has the option to buy back the improved asset at price p_2. If she declines to exercise her option, the agent retains ownership. After deciding whether to exercise her option or not, the principal chooses her investment K. The final owner of the asset receives the return $R(K,L)$.

Demski and Sappington (1991) propose $p_1^* = \Pi(L^*) - \epsilon - L^*$ and $p_2^* = \Pi(L^* - \epsilon)$ where $\epsilon > 0$ is small. We analyze the game tree by means of backward induction to check whether this option contract implements the optimal actions. Assume the agent has accepted (p_1^*, p_2^*), exerted some effort L and that the principal has exercised her buy-back option. Conditional on L, she will optimally invest and earn $\Pi(L)$. Thus, she should exercise the option only if $\Pi(L) \geq p_2^* \Leftrightarrow L \geq L^* - \epsilon$. Knowing this, the agent can choose $L \geq L^* - \epsilon$ to earn $p_2^* - p_1^* - L$ which is maximum for $L = L^* - \epsilon$; this way he nets $p_2^* - p_1^* - L^* + \epsilon = \epsilon$ (by Taylor's formula using the optimality of L^*). If, on the contrary, the agent chooses $L < L^* - \epsilon$, then the principal does not exercise and invests $K = 0$ so that he will earn $R(0,L) - L - p_1^* = R(0,L) - L - R(K^*, L^*) + \epsilon < \epsilon - L^* < 0$ by the optimality of L^*.

The contract (p_1^*, p_2^*) was obviously designed to make the first choice dominant; this way the agent earns a net utility of $\epsilon > 0$ so that accepting the contract is initially an optimal choice (dominant strategy). Letting ϵ tend to zero the principal can obtain the first-best surplus $\Pi(L^*) - L^*$ if the agent is rich enough to pay that amount at the initial stage.

This wealth requirement is problematic in an incomplete financial market and makes the proposed solution quite infeasible. Indeed, the agent is typically not very rich (think of the young company Pixar back in 1990) while the up-front payment can be very large (think of the profits that a Disney blockbuster can generate), thus the agent must raise the money from the financial markets. Now bankers who are not animated movie specialists have a difficulty to identify whether the whole idea will be profitable; they therefore tend to underestimate the expected return of the asset and either refuse to fund or ask a large risk premium to lend the agent the required amount $\Pi(L^*) - L^*$.

Inefficiency of Renegotiation †

Committing not to renegotiate is problematic as it goes against one of the Pareto principle: rational agents understand the benefits of renegotiating an agreement leading to Pareto dominated payoffs. If there is perfect information and negotiation is costless then the final agreement is bound to be Pareto efficient. Edlin and Hermalin (2000) show that this possibility hinders the previous analysis; if renegotiation cannot be prohibited, then the previous option contract must be amended and can still implement the first best if and only if investments are substitutes at the margin.

If the principal, upon having sign the contract (p_1, p_2), refuses to buy back the asset, the agent is left with a profit of $R(0,L) - p_1 - L$ because the principal will not invest at all. By definition of $\Pi(L)$ we have $\Pi(L) > R(0,L)$ thus the asset value is $R(0,L) - p_1 < \Pi(L)$ for any positive price p_1 which means that there is always scope for a renegotiation of p_2 (whatever its original value) in order to reach the better outcome where the principal buys back the asset and invests efficiently in it, in order to increase the asset value to $\Pi(L)$. To simplify the analysis, we assume equal bargaining power.

The amount over which they bargain is $\delta \equiv \Pi(L) - R(0,L) + p_1^*$ thus the renegotiated price is $p_2' = P(L, p_1) \equiv R(0,L) - p_1 + \frac{1}{2}\delta < \Pi(L)$ (we are in the case of renegotiation thus $R(0,L) - p_1 < \Pi(L)$). We now see clearly that if the principal offers the previously optimal contract (p_1^*, p_2^*), the agent will refuse it because he anticipates that the future exercise price will be $P(L^*, p_1^*) < p_2^* = \Pi(L^*)$ which means he won't cover his own investment L^*.

This problem can be overcome when efforts are substitutable, that is when the agent's effort improves greatly the value of the asset **without** the help of the principal; an example might be the Pixar case since Disney had no role in the making of the movie. If it is indeed true that $R_{KL}|_{(K^*,L^*)} < 0$, we can set $\hat{p}_1 = p_1^*$, $\hat{p}_2 = P(L^*, p_1^*) < \Pi(L^*)$. Observe that $P_L = \frac{R_L(0,L)+\Pi_L}{2} > \Pi_L > 0$ because $\Pi_L = R_L(K(L), L) < R_L(0,L)$ by the envelope theorem and the substitutability of efforts.

Suppose now that the agent expands the effort $L > L^*$; if the principal decides to renegotiate the option price $\hat{p}_2$, she ends up paying $P(L, p_1^*) > P(L^*, p_1^*) = \hat{p}_2$ because $P_L > 0$. Thus, it is better to exercise the option at $\hat{p}_2$ without renegotiating. Symmetrically, it is beneficial to enter the renegotiation process if $L \leq L^*$. Over the domain $L > L^*$, the payoff of the agent is the decreasing function $P(L^*, p_1^*) - L$; thus the optimal effort lies below L^*. Over the domain $L \leq L^*$, the payoff of the agent is $U(L) = P(L, p_1^*) - L$ with $U_L > \Pi_L - 1 > 0$ because $\Pi_L(L^*) = 1$. Overall, L^* is the optimal choice (a corner solution though) and since the initial price is p_1^*, the previous proof applies.

If efforts are complementary, then no option contract can implement the first best effort from the agent because the renegotiated price P increases too slowly with effort ($P_L < 1 \Rightarrow U_L < 0$). Unfortunately this is the most relevant case since the very reason why the principal and the agent decided to form a team was to internalize possible synergies i.e., because they expected their efforts to be complementary.

An example of complementarity vs. substitutability is as follows. Suppose that a US teen goes to the movies either if a friend recommends it or if he is impressed by the marketing or TV advertising. Let $z(L)$ be the probability that a friend does not recommend it ($z_L < 0$) and let $y(K)$ be the probability that he is not impressed by the marketing ($y_K < 0$). Assuming independence, the probability he goes to the movie is $1 - z(L)y(K)$, which has a negative cross partial derivative, implying that efforts K and L are substitutes and that

the first best is implemented by an option contract. On the other hand, a European teen being more solicitous will go to see the movie only if he receives two good signals so that the probability is $(1-z(L))\big(1-y(K)\big)$. Noticing that this expression has a positive cross partial derivative, we deduce the complementary of efforts.

The general conclusion is that the consumer market plays an important role for the optimal allocation of incentives in production: never forget that the customer is king!

14.3 On Property Rights

In this section, we emphasize the role of property rights over assets as an instrument guiding the incentives of the parties to perform relationship-specific investments.[10@] Whereas TCE focuses on the level of quasi-rents as the main reason for integration, PRT focuses on the investments that can increase the quasi-rents.

14.3.1 The Control of Incentives

According to the synthesis offered by Hart (1995), parties engaging in a transaction write contracts that are ex-ante incomplete, but complete them ex-post once uncertainty resolves. The ability to exercise residual control rights improves the ex-post bargaining position of an asset owner and thereby increases her investment incentive as well as the incentives of those who enjoy gains from trade with her. As a consequence, it is optimal to assign asset ownership to those who have the most important relationship-specific investments, or who have indispensable human capital.

Private Benefits

To illustrate how firm boundaries affect decision making, consider two production units or divisions A and B. Each unit $i = \{A, B\}$ generates monetary profits v_i and private (nontransferable) benefits w_i in the form of job satisfaction for those working in the unit. If the units are independent, manager i being the boss of unit i, maximizes $v_i + w_i$ since she owns the profit from unit i and cares about her own private benefits.[11@] In contrast, if units A and B are integrated, then what happens depends on who is the overall boss. If A is the overall boss, she maximizes $v_A + v_B + w_A$, since she owns the profit from both units, and cares about her own private benefit but not B's. A similar formula holds if B is the overall boss. Finally, if a (professional) outsider is brought in to be the boss, she will maximize $v_A + v_B$, since she owns all profits and does not care about the private benefits of A and B.

Given that the social optimum is achieved by maximizing total surplus $v_A + v_B + w_A + w_B$, we can conclude that under independence, bosses maximize the right thing, profits plus private benefits, but are parochial since they fail to take into account their externality on the other unit, while under integration they maximize the wrong thing, their sole profit but have a global scope. This demonstrates that both integration and independence have merits and defects.

Diversion of Benefits

In §13.3.4, we argued that the owner of an asset should be the economic agent able to use in the more efficiently. A simple example will help to understand this claim. Consider a machine like the printing press of our original example that necessitates only managerial effort e, measured by its opportunity cost and generates a deterministic profit $\pi(e)$ (return) without creating any externality. In the presence of moral hazard i.e., when effort cannot be controlled, the agency theory shows that the manager's wage should be tied to profits in order to give him efficient incentives for effort; yet this does not makes him the owner of the asset.[12@] In the present case, the efficient effort e^* solves $\pi'(e) = 1$ (marginal benefit equals marginal cost). The manager will optimally expand e^* only if his wage is $w(e) = \pi(e) - F$ i.e., if he owns the return (which must be clearly identifiable). Yet, there is no need for him to own the machine i.e., have the right to decide on the future use of the machine.

In more realistic situations there is still a difference between ownership of asset and return. Indeed, the manager's effort hardly generate cash per-se; instead it generates an input for another production process. In our example, the printer makes books that must be sold later on by the publisher. The latter is thus able to divert some benefits of the printing activity, say a share λ, in a way that is non contractible. For instance, he uses the marketing activities organized around the publication of the book to promote other works by the author (under contract with him). Hence at most $1 - \lambda$ of the return can be contracted upon between the printer and the publisher. If the printer's wage is $w(e) = (1 - \lambda)\pi(e) - F$ then his effort will be lower than e^* because the publisher "holds-up" a share λ of the profits created by his efforts. The only way to restore efficiency is for the printer to buy the publisher in order to reclaim the missing $\lambda\%$ of the benefits.

14.3.2 Relation Specific Investments

The basic idea

Grossman and Hart (1986)'s model as recast by Gibbons (2005), considers investments

that are observable by the parties but not verifiable by a third party.[13@] Let $i = N, B, S$ denote "Non integration", "Buyer integration" and "Seller integration". After the parties have decided over their ownership structure, they invest in assets that are specific to their future trade. Later on, uncertainty resolves and from this moment on, the relevant characteristics for the trade become contractible (complexity has been reduced). Let us analyze the outcome if there is no renegotiation of the initial contract. If one party has integrated the other, she controls all relevant dimensions of trade and thus chooses the characteristics maximizing her sole profit. Under independence, each party retains control over some relevant dimensions and enters a rent-seeking process where she tunes the characteristics under her control to maximize her profit taking the choice of the other party as fixed; there is a Nash equilibrium. We thus have in each case a default trade q^i that depends on the ex-ante investments and the current conditions.

However, when the trading period arrives, both parties reason that their previous investment together with the ex-post trading conditions make a particular trade q^* efficient i.e., it maximizes $W(q) = \pi_B + \pi_S$ and is different from all the default trades mentioned above. The ex-post quasi-rent is thus $\delta = W(q^*) - W(q^i)$ and being shared equally among parties, the ex-ante profit of the buyer is thus

$$U_B = \pi_B(q^i) + \frac{1}{2}\delta = \frac{1}{2}\left(\pi_B(q^i) + \pi_B(q^*)\right) + cte \tag{14.1}$$

with a symmetric formula for the seller. The buyer's incentives to invest ex-ante are 50% of the marginal return at the efficient trade q^* plus 50% of the marginal return at the inefficient trade q^i. Whatever the ownership structure, it is never the case that a party to this relationship has first-best investment incentives. Grossman and Hart (1986) then observe that if the ex-post objective of each party mostly depend on the characteristics she controls then q^N is nearby q^* so that non integration is almost efficient which makes it the dominant choice. If on the other hand, a party's objective hardly depends on the characteristics she initially controls, then integration under the control of the other party is almost efficient which makes this ownership structure the dominant one.

A refined model

Whinston (2003) develops more formally the argument. Consider a bilateral trade setting involving a buyer B and a seller S. Ex-ante, some specific contractible investments are negotiated between B and S to enhance the value of their relationship but some elements k_b and k_s are non contractible (e.g., human capital) and are therefore chosen independently at a cost $c(k) = \frac{1}{2}k^2$ for each party. At that time, the parties can choose to integrate (i) i.e., the buyer buys the seller's assets[14@] or not (n) i.e., remain independent.

Ex-post, the relevant information about the quantity and quality of the good to be traded is revealed to the parties. We make an assumption that greatly simplifies notation by dropping reference to the ex-post efficient quantity or quality and only keep the ex-ante investments. The buyer's ex-post payoff from trading under design $j = i, n$ is denoted $\pi_b^j = \alpha_b + \beta_b^j k_b + \gamma_b^j k_s$ where the β parameter indicates the own effect of one's investments while γ indicates the cross effect. A symmetric formula holds for the seller.

The key argument to all incomplete contracting theories is that, whatever the ex-ante contractual design, be it integration (i) or independence (n), it becomes inadequate once all useful information has been revealed, thus an ex-post renegotiation can take advantage of unexpected trade opportunities or adjust some dimensions of the trade to take into consideration unexpected developments be they positive or negative. Unlike TCE which emphasizes the cost of such a renegotiation, the PRT assumes away transaction cost. Ex-post, parties maximize their joint ex-post payoff whose value is denoted $\pi(k_b, k_s) \equiv \alpha + \mu_b k_b + \mu_s k_s$. Since the joint value of the trade without investments α is greater once ex-post adjustments are made we have $\alpha > \alpha_b + \alpha_s$. Likewise, the value of each investment is greater if ex-post adjustments are made thus the slopes satisfy $\mu_b > \beta_b^j + \gamma_s^j$ and $\mu_s > \beta_s^j + \gamma_b^j$ for design $j = i, n$. The first-best characterized by contractible investments would maximize the net surplus $W \equiv \pi(k_b, k_s) - \frac{1}{2}k_b^2 - \frac{1}{2}k_s^2$; efficient investments are then μ_b and μ_s. Since the point here is that those human capital investment cannot be contracted upon, both designs will yield sub-optimal investment. The optimal design is then that which is less inefficient.

To simplify further the study, we assume equal bargaining power so that the quasi-rents $\delta^j \equiv \pi - \pi_b^j - \pi_s^j$ are evenly split among the parties for design $j = i, n$. The buyer's ex-ante profit is thus,

$$\begin{aligned} \Pi_b^j &= \pi_b^j + \tfrac{1}{2}\delta^j - \tfrac{1}{2}k_b^2 = \tfrac{1}{2}\left(\pi + \pi_b^j - \pi_s^j - k_b^2\right) \\ &\propto \alpha + \alpha_b - \alpha_s + \left(\mu_b + \beta_b^j - \gamma_s^j\right) k_b - k_b^2 + \left(\mu_s - \beta_s^j + \gamma_b^j\right) k_s \end{aligned}$$

for design $j = i, n$. The optimal investment for the buyer is easily characterized as

$$k_b^j = \tfrac{1}{2}\left(\mu_b + \beta_b^j - \gamma_s^j\right) \text{ for design } j = i, n \tag{14.2}$$

The formula for the seller is entirely symmetrical.

Since there is always ex-post cooperation, the total welfare achieved under design $j = i, n$ is

$$W^j \equiv \pi\left(k_b^j, k_s^j\right) - \tfrac{1}{2}(k_b^j)^2 - \tfrac{1}{2}(k_s^j)^2 = \alpha + \mu_b k_b^j + \mu_s k_s^j - \tfrac{1}{2}(k_b^j)^2 - \tfrac{1}{2}(k_s^j)^2$$

The likelihood of vertical integration is seen though the changes in $\Delta \equiv W^i - W^n$ induced by changes in the parameters. For instance, what is the change due to an increase in the intrisic value of the buyer's human capital (μ_b):

$$\frac{\partial \Delta}{\partial \mu_b} = k_b^i - k_b^n + \left(\mu_b - k_b^i\right)\frac{\partial k_b^i}{\partial \mu_b} - \left(\mu_b - k_b^n\right)\frac{\partial k_b^n}{\partial \mu_b} = \frac{1}{2}\left(k_b^i - k_b^n\right) \tag{14.3}$$

Define the selfish value of investment as the difference between own value β_b^j and the value for the partner γ_s^j. One would intuitively assume that the buyer's investment has more selfish value under integration than under independence. If this is so, then $\beta_b^i - \gamma_s^i > \beta_b^n - \gamma_s^n$ (check with (14.2)) so that the buyer invests more under integration than under independence. We can now interpret (14.3): an increase in the intrisic value of the buyer's human capital μ_b makes integration more frequent. Likewise, if the seller invests more when he remains independent ($k_s^n > k_s^i$), an increase in the intrisic value of his human capital μ_s makes integration less frequent since $\frac{\partial \Delta}{\partial \mu_b} = \frac{1}{2}(k_s^i - k_s^n)$. As the PRT claims, integration should occur if one's investment becomes more crucial to generate value.

State Owned Enterprise

Schmitz (2000) studies the optimal ownership of an asset such as a network used for the provision of a public service. When the State is the owner one speak of public provision while privatization refers to private ownership and a delegation to provide the service. Lastly, joint ownership or Public-Private-Partnership (PPP) refers to the situation where both parties hold veto power over the use of the asset.

Privatization of a public enterprise changes the manager's objective towards profit maximization so that cost cutting takes right of way; this is conducive of more efficiency yet too much cost cutting tends to deteriorate the quality of the service so that a negative externality is imposed. Under public provision, the State can design service innovations and implement them at will. This is conducive of efficiency yet, too much innovation inflate the cost of the operator so that another negative externality is imposed. The following analysis examines how the aforementioned ownership structures mitigate those externalities generated by a double moral hazard problem.

The firm S, seller of the service, invests k_S now to cut his future cost of service or equivalently earn an extra revenue $R_S(k_S)$ though it deteriorates the objective of buyer B, here the government (e.g., consumer surplus) by a damage $d_B(k_S)$. The government spends k_B on innovation to improve her future objective by some extra revenue $R_B(k_B)$ though it augments the agent's future cost by $d_S(k_B)$. The decision to implement an

activity is taken ex-post by the asset owner once he has gathered all the relevant information. The decision maker is aware that his decision might hurt the other party and may be willing to negotiate ex-post an agreement. We assume that ex-post quasi-rents are shared equally. Let $\delta_S = R_S(k_S) - d_B(k_S)$ and $\delta_B = R_B(k_B) - d_S(k_B)$ denote the ex-post value or quasi-rent associated with each activity. Assuming both positive, it is always efficient to carry on each activity so that maximum welfare is $W = \delta_S - k_S + \delta_B - k_B$ and the first best levels thus solve $\delta'_S = 1$ and $\delta'_B = 1$. We shall compare the investment incentives under various ownership structure with the first-best.

Under public provision, G can decide to carry on both activities but would only implement the innovation since the other one hurts him. The stake for negotiation is thus δ_S i.e., she will carry on the cost cutting activity if the agent pays her more than her potential loss $d_B(k_S)$. In terms of investment incentives, G sees only his gain $R_B(k_B)$ and ignores the negative externality $d_S(k_B)$ imposed on the agent, she thus over-invests because she is a parochial residual claimant. Regarding, cost cutting, the agent gets only one half of the quasi-rent δ_S, he thus under-invests because he is a partial claimant of the global welfare. Under private provision, M can decide to carry on both activities but would only implement the cost cutting effort since the other one hurts him. The stake for negotiation is thus δ_B i.e., he will carry on the innovation activity if the principal pays her more than her potential loss $d_S(k_B)$. In terms of investment incentives, M sees only his gain $R_S(k_S)$ and ignores the negative externality $d_B(k_S)$ imposed G, he thus over-invests being a parochial residual claimant. Regarding, innovation, G gets only one half of the quasi-rent δ_B, she thus under-invests because she is a partial claimant of the global welfare. Under joint ownership, nothing is implemented by default since each party would veto the activity hurtful to him. The stake for negotiation is thus $\delta_B + \delta_S$ and being shared equally among the parties, each is a partial claimant of the global welfare, thus under-invests.

Privatization is the optimal ownership scheme if the cost cutting externality is limited and innovation is unimportant. Public ownership is optimal when the innovation externality over the firm is limited and cost cutting is unimportant. If the externality are large, regulation can be optimal.

14.4 On Transaction Cost

As we show in Table 14.2, Coase (1937)'s study of the "make or buy" decision has several interpretations. The terms of the alternative are between a flexible but risky scheme (outsourcing, privatizing) and a safe but inflexible one (in-house, bureaucracy). We shall study two models of this dilemma, one emphasizing the adaption to unexpected events,

the other the limits to contractual design.

contracting mode (m)	internal (i)	external (e)
Production	make (in-house)	buy (out-source)
Hierarchical Organization	centralized	decentralized
Procurement	costs plus	fixed price
State public service	public provision	private delegation

Table 14.2: Contracting modes

14.4.1 Contractual Design

Procurement refers to the situation whereby a *principal* hires an *agent* to perform a task or realize a project.[15@] As mentioned in ch. 7, a contender will make both a productive investment to develop a good project and a power (lobby) investment to gain influence over the jury. If influence and capture play a role, the winner enjoys a privilege because he won only thanks to his personal connections with the jury members, not on account of objective economic qualities.

One way to circumvent this inefficient (and unfair) outcome is to force by decree the standardization of the procured object over a large economic area (e.g., EU, US) to create a market for an homogeneous good or service that can then be served by many firms on price terms only (e.g., in a procurement auction). A liquid market active all year long will emerge and enable entry and thus price adjustment downward to the long term cost of procurement. The drawback of this new regulation is the rigidity of the product description that will never perfectly fit the need of each buyer.

When the project is idiosyncratic and cannot be standardized, the relationship between the principal and the agent becomes one of agency, examples of which include:

- A contractor builds a museum on behalf of a government.
- A computer maker hires an electronic firm to design a new graphical chip.
- A fashion designer hires a cheap manufacturer to produce articles bearing his/her name.
- A high-tech firm hires a low cost manufacturer to assemble the product it designed.

The realization of such projects is always plagued by imponderables and uncertainty. For instance,

- The original architectural design is impossible to carry on as such; a reinforcement of the building's infrastructure is needed.[16@]

- A new regulation prohibiting certain designs or heights for buildings might be passed after planning but before completion.
- Archeological remains might be found when digging for the foundations, forcing a delay and extra cost to preserve them (whenever this is mandatory).
- A complex electronic device is sensitive to the (variable) price of a metal component.

When the agent is paid for the project on the basis of costs plus a bonus (denoted C+), the relationship is close and flexible as if the agent was a division within the principal's firm; this situation is akin to internal provision and characterized by authority of the principal over the agent. If, on the contrary, the parties sign a fixed price contract (denoted FP), their relationship is at arm's length i.e., distant and inflexible ; this is akin to external delegation and relations are characterized by mutual agreement or bargaining.

The "make or buy" trade-off is as follows: FP provides better ex ante incentives than C+ for minimizing production cost, it also avoids the excessive administrative cost typical of an internally run operation. Yet, whenever unforeseen events force parties to change their original project ex-post, C+ displays a greater adaptability than FP which translates into a lower transaction cost of adaption.

Ex-post Uncertainty and Opportunism

We now study formally the development of this contractual relationship following Bajari and Tadelis (2001)'s model. The completed project has a value V but unexpected contingencies arise during the realization forcing the parties to modify the project if they want to finish it.

The cost of adjusting the project is denoted F_m for modes $m = i, e$. When parties are integrated or use a C+ contract, the agent applies the necessary modification and the principal pays the true adjustment cost F_i. Whenever the agent works as an external contractor he is supposed to support all costs. Yet, as the modification is due to events not contemplated by the original contract, the agent can ask for a renegotiation at which point he will act *opportunistically*[17@] and this will generate an inefficiency in the sense that the adjustment cost will increase towards $F_e > F_i$.[18@] The agent is typically less impatient than the principal regarding project completion; he can thus extract an additional payment (holding-up the principal) but in the course of this haggling, time and resources are lost. Alternatively, his external position enables him to inflate the adjustment costs billed to the principal and divert part of this inflation. Once again, there is an inefficiency because some of the bill inflation corresponds to an inefficient use of resources such as the substitution of a competitive subcontractor by a family related one. Another prevalent explanation for the adjustment cost increase is the presence of

asymmetric information among the parties; it is developed at the end of the section. To conclude, the weakness of arms length relationship lies in its inherent rigidity.

Design Complexity

To avoid paying for the modification cost F_m, the principal can design a very precise project that contemplates many future contingencies. The quality of a design is measured by the probability τ that no modification will be needed later on. However, the better the design, the costlier it is; we assume that the fixed cost of a design of quality τ is $d(\tau) = -\alpha \ln(1-\tau)$ where α is an exogenous indicator of the complexity of the project.

Under a FP contract with price p, the agent supports all (scheduled) construction costs and is the residual claimant of any effort or investment to bring them down. Suppose then that the anticipated project cost is c_e. The payoffs are $\pi_e = V - d(\tau) - (1-\tau)F_e - p$ and $u_e = p - c_e$. The principal chooses the price p to meet exactly the agent's participation constraint $u_e \geq 0$, hence her profit is

$$\pi_e = V - d(\tau) - (1-\tau)F_e - c_e \tag{14.4}$$

Under a C+ contract, the principal pays all the expenses so that the agent does not care to contain cost. As shown in ch. 20 on moral hazard, such an attitude leads to an inflated project cost $c_i > c_e$. The payoffs are $\pi_i = V - d(\tau) - (1-\tau)F_i - c_i - p$ and $u_i = p$. Now, the price p of the C+ contract is chosen to meet exactly the agent's participation constraint $u \geq 0$ so that the principal profit is

$$\pi_i = V - d(\tau) - (1-\tau)F_i - c_i \tag{14.5}$$

Comparing (14.4) with (14.5), we see that the upside of delegation (outsourcing) is the lower ex-ante project cost while the downside is the greater ex-post adjustment cost.

For the same quality of design τ, the FP contract dominates the C+ contract if the cost savings $c_i - c_e$ from better work incentives are larger than the expected inflation of adjustment costs $(1-\tau)(F_e - F_i)$; we immediately notice that this condition is true for a well designed project i.e., a large τ. To make a more precise comparison, we must take into account the fact that the optimal design will differ according to which contract governs the relationship. For $m = i, e$, the optimal design solves $d' = F_m$ (cf. eqs. (14.5) and (14.4)) and is $\tau_m = 1 - \frac{\alpha}{F_m}$ (assuming $0 < \alpha < F_i$). We can make two observations; firstly, as $F_e > F_i$, a FP contract leads to choose a better design than a C+ one (for a given complexity) and secondly, the design quality decreases with complexity whatever the contract governing the relationship. Plugging the minimized design cost $d(\tau_m)$ in the principal's payoff, we

obtain $\pi = V - \alpha(1 + \ln(\alpha)) + \alpha \ln(F_m) - c_m$ for $m = i, e$. The comparison of the two contracts becomes[19@]

$$\pi_e > \pi_i \Leftrightarrow c_i - c_e > \alpha \ln(F_e/F_i) \Leftrightarrow \alpha < \alpha^* \equiv \frac{c_i - c_e}{\ln(F_e) - \ln(F_i)} \tag{14.6}$$

The general conclusion can be seen on Figure 14.2: as complexity increases from none to infinite, the optimal contract is initially FP with a decreasing quality design. At the threshold α^*, there is a change of regime toward a C+ contract together with a drop in design quality; further on, design quality continues to decrease. The force of the model is not so much to corroborate these intuitive findings but to relate the cut-off position with the fundamentals of the model in equation (14.6).

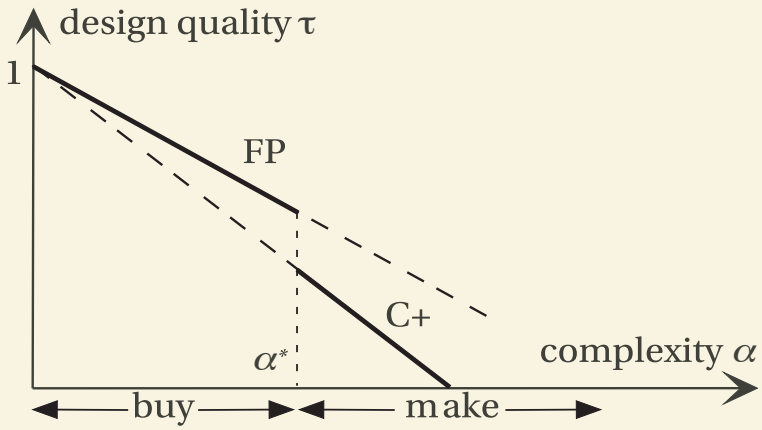


Figure 14.2: The “make or buy” Decision

Discussion

When applied to the “ “make or buy” ” quandary, this result tells us that an easy to define component will be bought on the market, whereas a complex component will be procured internally. For the aerospace industry, Masten (1984) shows that both a higher degree of specialization (specificity) and a higher level of complexity will increase the probability of internal procurement. For the automobile industry, Monteverde and Teece (1982) show that more complexity, identified by more engineering investment, will increase the likelihood of internal procurement.

An obvious corollary noted by Gibbons (2005) is that the firms we observe are less efficient than the markets we observe, even though the firms we observe are more efficient than the markets they replaced. On Figure 14.2, the design quality of observable C+ contracts is worse than that of observable FP contracts although the observable C+ contracts fare better than the (non observed) FP alternatives. Applied to public services, the publicly provided services we use are of worse quality than the delegated ones we

use although better than the privatization alternative (we do not use). As we comment in §2.4.3 (cf. Figure 13.3), the raw comparisons of the two modes can suffer from sample selection bias (cf. discussion at the end of §13.3.3).

Asymmetric information

The previous model entirely rests on the greater inefficiency created by an external procurement when adjusting for unexpected events. We show that asymmetric information is a likely cause of inefficiency.

The fact that only the agent knows without doubt the true adjustment cost F_i creates an asymetry of information that inflates adjustment costs in the external contracting mode. As we explain in §2.4.3, when the (uninformed) principal bargains with the (informed) agent over the compensation for the adjustments to the project, she can either under or over estimate the additional cost and thus offer a compensation lower or greater than the true cost supported by the agent. Alternatively, she rejects an offer that is more or less than the true cost. This means that a renegotiation failure occurs with probability σ, in which case the project is not completed and its entire value is lost. Although we do not know the respective payoffs u^R and π^R of the agent and the principal after a sucessful renegotiation, we still know that their sum is $V - F_i$ since they bargained over a monetray transfer.

The ex-ante payoffs are then $\pi_e = \tau V + (1-\tau)(1-\sigma)\pi^R - p - d(\tau)$ and $u_e = p - c_e + (1-\tau)(1-\sigma)u^R$. The principal chooses the price p to meet exactly the agent's participation constraint $u \geq 0$, hence her profit is $\pi_e = \tau V + (1-\tau)(1-\sigma)(V-F_i) - c_e - d(\tau)$ which is identical to (14.4) whenever $F_e = (1-\sigma)F_i + \sigma V > F_i$.

14.4.2 Limited liability

From an incentives point of view, carrots and sticks are identical as they work through differential payments. They diverge however in the absolute base level and the conditioning event; indeed, the carrot is paid frequently as it rewards each good performance while the beating with the stick is applied infrequently since its role is to deter shirking. It thus seems clear that the stick ought to be the preferred instrument for the principal whenever available. The problem is that for a variety of reasons, sticks are hard to come by with. When an individual has a *limited liability* i.e., scant wealth, imposing a negative payment upon him is akin to slavery; in other words, the only way to lower the utility of a destitute is physical or psychological harm. In market democracies, principles of higher order than economics prohibit exploitation or coercion in labour contracts involving private parties: the worse for an employee is thus dismissal and a mild

shame exposure. This also means that a large and wealthy corporation might engage into harmful activities because although it is liable with almost no limit, its board of directors is not.[20@] It is for that reason that some actions deemed harmful to economic welfare (e.g., fraud) are punished by jail terms to deter the individuals at the command from perpetrating them.

Limited liability, by restricting the use of sticks, gives rise to a new transaction cost insofar as the necessary rise of the base salary makes contracting dearer to the firm. We might however qualify this transaction cost of limited liability as an *exogenous* one since it is somehow imposed on the legal system regulating private contracting by moral considerations. We shall now see that there exists an *endogenous* transaction cost of limited liability resulting from an internal weakness of the legal system which is quite similar to the issue of rent-seeking (cf. §16.3). Whenever a party to a contract, implicit or explicit, faces a lawsuit for wrongful doing and anticipates an expected loss e.g., a fine, a damage payment or a bad reputation, it has an incentive to contest the outcome, with a view to overturn it or to reduce the loss. Whereas "rent-seeking" firms contest a State awarded rent (a carrot), a "penalty-avoiding" firm contests a State punishment (a stick). As in the case of rent-seeking, penalty avoidance can run through legal or illegal means. In the first category, we find lobbying for soft laws and regulations, delaying through appeals or establishing the legally responsible entity in a hard to prosecute overseas territory. In the second category, corruption and outright violence are of the essence.

Glaeser and Shleifer (2003) show how the legal system is optimally changed from a flexible to an inflexible one when limited liability becomes a more severe constraint i.e., when the maximum fine that can be realistically levied upon a firm shrinks (as a proportion of GDP or sales). Consider some activity generating a risk of accident. Safe firms have a constant accident probability p_0. Risky firms display a higher probability p_n but can adopt a safety standard to reduce it by p_l down to $p_n - p_l$ (possibly lesser than p_0). The cost of complying with the safety standard is c. The available legal rules to cope with accidents are:

- *Laissez-faire*: no responsibility whatsoever
- *Regulation*: mandatory safety standard with a fine for discovered violators
- *Negligence*: in case of accident, the fine applies to non safe firms
- *Liability*: in case of accident, the fine applies

The degree of "law and order" in the legal system is measured by the maximum fine F that will not be challenged by capture or corruption. For a risky firm under the liability regime, the expected fine saving of adopting the safety standard is $p_l F$ while under negligence it is the greater $p_n F$ because once the safety standard is installed, the firm

is never fined. Lastly, under regulation, the expected fine saving of complying is $p_r F$ where p_r is the probability of control (audit) by authorities. We assume $p_l < p_n < p_r$. The first inequality means that the safety measure does not eliminate risk while the second means that controlling for the presence of the standard (ex-ante) is comparatively easier than proving responsibility in case of accident (ex-post).

For all legal rules, a firm compares the expected fine saving to the cost c of adopting the safety standard. What matters is the position of the ratio $\frac{c}{F}$ with respect to the aforementioned probabilities. Clearly, when F is small, the government is powerless and cannot promote the adoption of the safety standard. The optimal policy is then laissez-faire as it avoids the administrative cost of setting up a true policy but also the wasteful corruption that such a policy would trigger. As F increases, the safety objective becomes feasible with regulation, then with negligence and lastly for F large with any of the three instruments. There is however a social ranking of the legal modes.

Liability is a first best policy because it only motivates risky firms to adopt the safety standard, so that no safe firm is forced to waste resources in adopting a standard that does not reduce their (already low) riskiness. Negligence is also first best when the maximum fine is intermediate with $\frac{c}{p_n} < F < \frac{c}{p_0} \Leftrightarrow p_n F > c > p_0 F$. This is so because the LHS inequality guarantee adoption by risky firms while the RHS one means that safe firms are not (wastefully) lead to adopt the standard. Lastly, regulation is only second-best as it forces safe firms to adopt the safety standard, which is a social waste of resources. Based on our assessment of what can be done according to the level of the fine and the social benefit of each policy, we conclude that[21@]

> The optimal legal system is
>
> *Laissez-faire* in powerless states ($F < \frac{c}{p_n}$)
> *Regulation* for low level of enforcement ($\frac{c}{p_n} \leq F \leq \frac{c}{p_r}$)
> *Negligence* for intermediate level of enforcement ($\frac{c}{p_r} \leq F \leq \frac{c}{p_l}$)
> *Liability* for high level of enforcement ($F \geq \frac{c}{p_l}$)

Since the cost of adoption c is roughly proportional to firm size q, Glaeser and Shleifer (2003) argue that industrialization amounted to a rising q accompanied by a proportional rise of harm so that damages awarded by justice ought to have grown proportionally; yet the cost of subverting justice remained stable so that the maximum fine F remained the same. This evolution of the underlying parameters then made the change from liability to negligence and lastly to regulation optimal in the US legal system which rationalizes the actual evolution at the turn of the XX^{th} century. At the same time, the US congress

passed laws aiming at increasing F by recalling corrupt judges and by developing bureaucracy (federal commissions) which is a State apparatus more immune to influence. The recess of regulation over the last twenty years can also be interpreted in the light of this model as a decrease of the cost c of meeting the highest safety and quality standards and, as an increase of the maximum fine F due to the consolidation of democratic institutions free from the influence of big trusts (relative to the situation a century ago) i.e., the large fines imposed on corporations by anti-trust authorities or the damages awarded by popular juries are not disputed beyond legal wrangling.

14.4.3 Employment vs. Performance

Using a model of incomplete contracting under moral hazard (cf. ch.20), Levin and Tadelis (2010) show that the employment vs. performance dichotomy regarding labour contracts illustrates the internal vs. external provision ("make or buy") as well as the private vs. public provision.

A principal contracts the services of an agent to provide a service of measurable quality q such as health of patients, educational achievement of students or satisfaction of customers. The agent's activity or effort has two dimensions, one extensive and contractible L akin to time spent on the job while the other one is intensive and non contractible beyond a minimum K_0; it is akin to human capital or organizational design. The technology transforming the agent's inputs into the output is a standard Cobb-Douglas $q = KL$. We shall see later that a quality objective q can be achieved by a simple employment contract stipulating assistance $L = q/K_0$. Writing and enforcing a contractual requirement is always costly but extensive effort L is obviously less difficult to monitor than the quality index q.[22@] We assume that the differential cost between the latter and the former is an increasing convex function $d(q)$ of the final quality q desired by the principal. This is so because greater quality requires more and more difficult and complex to describe sub-objective and tasks.

The agent's cost per unit of time is the increasing convex function $c(K)$. We assume that competition drives the market wage to $c(K_0)$. A contract is $(\bar{w}, \bar{L}, \bar{q})$ where $\bar{L}$ and $\bar{q}$ are minimum requirements and where the salary or fixed fee $\bar{w}$ is tuned to guarantee participation of the agent. We prove first a simplifying principle: it is useless to set both a time and a performance requirement.

For a low quality requirement $\bar{q} \leq K_0\bar{L}$, the optimal investment is the minimal one $K^* = K_0$ since the required output is met with the required input i.e., no extra costly investment is needed. Setting $\bar{q} = 0$ would thus yield the same outcome. For a large quality requirement $\bar{q} > \bar{L}K_0$, we show that the input requirement can be removed (set

$\bar{L} = 0$) by considering where lies the optimal investment K^*. If $K^*\bar{L} \leq \bar{q}$ then the agent voluntarily chooses $L^* \geq \bar{L}$ because both K and L are costly i.e., must maintain $KL = \bar{q}$. Hence, setting $\bar{L} = 0$ has no effect. If $K^*\bar{L} > \bar{q}$ then the agent would like to pick $L^* < \bar{L}$ which means that imposing the input requirement $L \geq \bar{L}$ forces the agent to spend more extensive time in order to attain the output requirement and this is also costly for the principal since $\bar{w}$ must then be adjusted upward.

We have thus shown that setting either $\bar{L}$ or $\bar{q}$ at zero can only improve the outcome for the principal. Observe then that $\bar{L} = 0$ is akin to a performance contract since only the output is paid for while $\bar{q} = 0$ is akin to an employment contract. The cost w^{pe} of achieving one unit of output under a performance contract is the minimum over all pairs (K, L) of the agent's cost $c(K)L$ under the constraint $KL \geq 1$ which necessarily binds at the optimum as we already argued; the agent thus minimizes $\frac{c(K)}{K}$ i.e., choose the optimal investment K^* equalizing average and marginal cost of human capital.[23@] The idea here is that once delegated the service provision, the agent chooses an optimal mix of extensive and intensive efforts in order to produce any additional level of quality. The moral hazard problem is alleviated because the agent is the residual claimant of any cost reduction. The cost of the performance contract is thus $w^{pe} = c'(K^*) = \frac{c(K^*)}{K^*}$. Under an employment contract, the cost of achieving one unit of output is simply $w^{em} = \frac{c(K_0)}{K_0} > w^{pe}$ since K^* minimizes the average cost. The performance contract leaves the agent free to choose both K and L while the employment one somehow fixes K down to its contractible minimum.[24@]

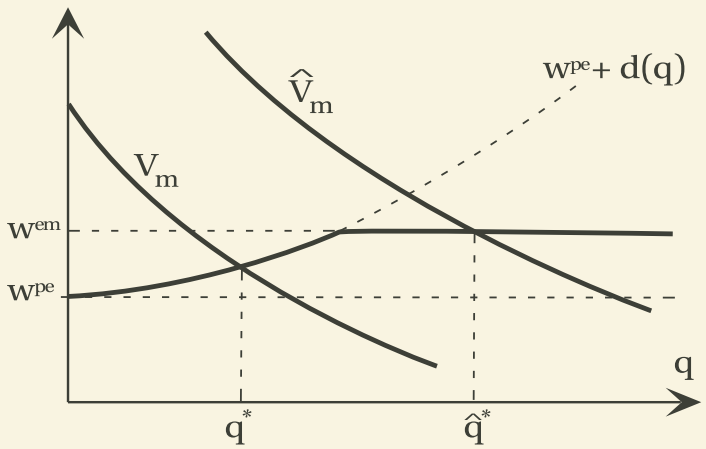


Figure 14.3: Employment vs. Performance

To compute the marginal cost of quality for the principal, we must take into account the cost $d(q)$ of designing a specific objective so that $C_m(q) = \min\{w^{em}, w^{pe} + d'(q)\}$ as shown on Figure 14.3 by the continuous increasing curve. The principal gross payoff is $V(q)$ which can either be a profit or the consumer surplus generated by service of quality q.

We assume decreasing returns to quality i.e., the marginal benefit V_m decreases with quality. Interpreting the intersection of C_m with V_m and an alternative $\hat{V}_m$ where quality is more valued by end-users, we can conclude:

> If the principal cares moderately for quality, a performance contract is optimal while if quality is crucial for the service, employment is the optimal scheme.

Chapter 15

Horizontal Integration

Horizontal Integration deals with mergers and acquisitions (M&A) between firms involved in the same markets. We firstly presents some stylized facts on the huge size of firms in advanced economies and discuss the incentives to grow bigger.

We then inquire about the efficiency effects of mergers and the resulting intromission of public authorities. Recall indeed that when two competitors in a given market merge, the competitiveness of the market is reduced and so is its productive efficiency. From the legal point of view, a merger strengthens or creates a dominant position which may give rise to abuse.[1@]

A related issue is the merger paradox according to which the gain in market power from a merger is not a sufficient reason for making the operation profitable. For instance, the newly formed company might use cross-subsidization or simply its bigger size to unlawfully increase its market power. For these reasons, mergers and acquisition are carefully watched by antitrust authorities which leads to legal disputes centered on what exactly is the relevant market, its participants and the extent of one firm's harmful market power.

The last two sections deal with the practical methods used to define a market and measure market power.

15.1 Merger Activity

15.1.1 Large Firms

We provide here some limited stylized facts, more detailed tables are available.

Forbes magazine compiles information about the 2000 largest firms in the world since 2003.[2@] Table 15.1 displays the distribution among countries and sectors for the period 2004-2010. All figures are % of group total, all tables are sorted by decreasing share of profit. The HHI index of concentration (cf. §15.3) for countries, whether computed on # of

firms or sales, decreases from 2500 to 1500 over the period. Among sectors, concentration is around 600 and slightly increasing.

Loc	Firms	Sales	Profits
US	32.5	34.7	33.6
UK	5.9	6.9	7.0
JP	14.8	14.1	6.4
FR	3.3	6.6	4.9
DE	3.0	6.4	3.9
CN	2.8	2.1	3.7
CA	2.9	2.1	2.9
RU	1.0	1.0	2.8
CH	2.0	2.3	2.8
ES	1.6	1.9	2.6

Industry	Firms	Sales	Profits
Oil & Gas	5.5	12.2	16.8
Banking	15.8	10.3	16.3
Pharma & Biotech	2.1	2.8	5.8
Food & Tobacco	5.3	6.7	5.6
Utilities	5.9	5.0	5.5
Materials	5.7	3.8	5.3
Insurance	5.4	7.3	4.9
Div. Finance	8.1	4.6	4.7
Telecom	3.4	4.2	4.3
Conglomerates	2.1	2.9	3.5

Table 15.1: Distribution of Large Firms

Table 15.2 presents the winners and losers among countries and sectors during this period with the mean annual percentage change of the number of firms present in the Forbes 2000 list.[3@]

Win	%
CN	30.4
LU	15.9
RU	14.4
BR	11.1
CL	10.1
IN	9.5
IL	7.0
AT	6.6
HK	5.5
BM	5.1

Lose	%
UK	6.2
US	4.9
NL	4.8
JP	2.4
DE	2.0
SE	1.6
IT	1.4
FI	1.3
CA	1.2
AE	0.0

Win	%
Capital Goods	6.1
Oil & Gas	4.9
Materials	4.2
Chemicals	2.9
Defense & Space	2.5
Div. Finance	1.7
Telecom	1.0
Trade	0.4
Conglomerates	0.4
Food & Tobacco	0.3

Lose	%
Hotels & Leisure	9.0
Health Care	4.7
Automotive	4.3
Retailing	3.9
Business Services	3.2
Media	2.6
Semiconductors	2.4
Technology	1.6
Software	1.4
Consumer	1.4

Table 15.2: Winners and Losers: Countries & Sectors

Table 15.3 shows the average year-to-year rotation over the period and the proportion of firms who have stayed in the entire period in a given tier. The first two columns of Table 15.4 presents, among firms always present in the list, those who have climbed and tumbled most the ranking over the period.[4@] For instance ATT climbed from the 400's to the top-ten while Citigroup fell from the top spot to the 400's. The impact of the financial crisis is most apparent for losers while telecommunication and energy firms are those most profiting. Regarding entry and exit, the last columns show the over and under achiever. Most exits are due to M&A such as SBC buying ATT in 2005, then rebranding as ATT and buying Bellsouth in 2006. Bank One was bought by JP Morgan in 2004.

In a different tack, Table 15.5 selects, over the period, the largest gains and losses and then sums these among firms in order to display the over and underachievers (in mean

bn$ per year). Vodafone is noticeable for appearing both as a large loss maker and a high climber of the ranking. Although it became profitable lately, it has always retained a large market capitalization (investors trusted it would become profitable). Table 15.6 displays which firms can be deemed the "big players".

Top	Rotation	Permanence
500	75	21
1000	52	38
1500	30	51
2000	12	58

Table 15.3: Movement within the list

Climbing
ATT @
Suez @
Telefónica @
Apple @
Vodafone @
Xstrata @
France Tel @
Vale @
América Móvil @
Generali @

Tumbling
Citigroup @
American Int @
Fannie Mae @
Freddie Mac @
ING @
RBS @
UBS @
Toyota @
Daimler @
Altria @

Entry
ICBC @
China Cons Bk @
Bof China @
EDF @
Inbev @
Rosneft @
Itaúsa @
Bof Comu @
Kraft Foods @
Saudi Basic Ind @

Exit
SBC @
HBOS @
Wachovia @
Merrill Lynch @
ABN Amro @
Bk One @
Washington Mut @
Bellsouth @
Fleetboston Fin @
Lehman Bro @

Table 15.4: Moving on the ladder

Gain	bn$
ExxonMobil @	32.4
Shell @	19.9
BP @	18.4
General Elec @	17.1
Chevron @	14.0
Total @	13.4
Microsoft @	13.4
Gazprom @	12.1
Petrochina @	12.0
WalMart @	11.7

Loss	bn$
Fannie Mae @	-18.9
American Int @	-15.7
Vodafone @	-13.6
General Motors @	-11.7
Freddie Mac @	-10.7
RBS @	-5.9
Fortis @	-5.6
Nextel @	-5.0
Viacom @	-4.6
Ual @	-4.3

Table 15.5: Average maximum Gains and Losses

Table 15.8 displays employment over the last five years using the Fortune Global 500 list of public companies.[5@] Table 15.9 pictures the returns per employee[6@] which are overtly dominated by mineral resources companies with exceptions from GoldmanSachs, Google, Microsoft or Apple as can be seen from Table 15.9 which averages performance over 2008-2010: left panel for large employers and right panel for mid-sized ones.

Revenue	bn$
WalMart @	342
ExxonMobil @	316
Shell @	302
BP @	273
Toyota @	252
Chevron @	179
Total @	167
Daimler @	159
General Elec @	159
Ford Motor @	159

Asset	bn$
RBS @	1996
Bnp Paribas @	1919
Barclays @	1848
HSBC @	1736
Citigroup @	1730
Deutsche Bk @	1572
UBS @	1495
ING @	1481
Bof America @	1479
JP Morgan @	1464

Stock	bn$
ExxonMobil @	366
General Elec @	273
Microsoft @	252
Petrochina @	219
WalMart @	207
Shell @	189
BP @	172
Johnson&Johnson @	171
Procter&Gamble @	170
Pfizer @	164

Table 15.6: 2004-2010 Mean Revenue, Assets and Stock value

Firm	Rank
General Elec @	1
ExxonMobil @	2
HSBC @	3
Shell @	4
BP @	5
Bof America @	6
JP Morgan @	7
WalMart @	8
Berkshire @	9
Total @	10

Firm	Rank
Bnp Paribas @	11
Chevron @	12
Santander @	13
IBM @	14
Verizon Comu @	15
Wells Fargo @	16
Samsung @	17
Barclays @	18
Eni @	19
Procter&Gamble @	20

Firm	Rank
NTT @	21
Pfizer @	22
Petrochina @	23
Nestlé @	24
Goldman Sachs @	25
Axa @	26
Siemens @	27
Microsoft @	28
Toyota @	29
Deutsche Bk @	30

Table 15.7: Average Rankings

Apart from the Chinese newcomers, most firms in the previous tables are US or European. The rest of Asia also host very large undertakings but they are organized in a different manner. Japan is the paradigmatic example. When the country opened up to the western world during the XIXth century, family-controlled vertical conglomerates(cf. zaibatsu) came to dominate the economy. During the post-1945 economic recovery, six major corporate groups (cf. Keiretsu) emerged: Mitsui, Mitsubishi and Sumitomo, Fuyo, Sanwa and Dai-ichi Kangyo. Each is led by a bank providing finance and a general trading company (cf. sogo shosha) coordinating business deals. Being organized as a trust, a keiretsu is immune to hostile takeover. Although their weight in the economy is decreasing, JFTC (2001) reports that they still occupy 13% of the total capital of the Japanese companies and make 11% of total sales. Other important (and newer) Keiretsus are vertically integrated. In the Automobile industry we find Toyota, Nissan, Honda, Daihatsu, Isuzu while in the electronics industry we find Hitachi, Toshiba, Sanyo, Matsushita and Sony.

Firms	Emp.
Wal Mart Sto @	1991
State Grid @	1381
China Nat Petro @	1313
US Postal @	764
Sinopec @	664
Carrefour @	472
Deutsche Post @	463
Agri Bk of China @	452
Hon Hai @	448
China Tel @	439

Firms	Emp.
Siemens @	430
Gazprom @	421
United Parcel @	419
McDonald's @	417
Compass @	391
Hitachi @	378
ICBC @	374
IBM @	374
Tesco @	354
Target @	352

Table 15.8: Top Employers (in thousand)

Large Firms	π/Emp.
Petronas @	376
Exxon Mobil @	358
Goldman Sachs @	292
BHP Billiton @	277
Chevron @	266
Shell @	231
BP @	222
Statoil @	210
Petrobrás @	210
Microsoft @	180

Small Firms	π/Emp.
EnCana @	923
Occidental Petro @	587
Petro Canada @	417
CNP Assurances @	346
GasTerra @	283
PTT @	247
Formosa @	231
Murphy Oil @	199
SABIC @	196
Legal & General @	192

Table 15.9: Top Profits in 1000$ per Employee

15.1.2 Reasons to merge

As we shall see later on with the merger paradox, mergers are hardly motivated by the desire to increase market share or market power when competition is driven by production capacities. In chapter 24 on standards and components, we explore some of the technological motives for mergers; a small list is:

- Fixed cost elimination
- Economies of scale and scope (synergies, downsizing the workforce)
- Network effects (complementarity of products or imposing a standard)

Holmstrom and Tirole (1989) explore an altogether different driver for mergers which is based on managerial incentives (cf.§23.3) and the desire to expand the size of the business they manage. Among the reasons why managers tend to oversize the firm they run, we find

- The technology may be such that the effort of the manager and the size of his staff are complementary, hence staff inflation frees the manager from stress and hard-work.

- Staff size may influence the market perception of the manager's ability. A powerful king is always followed by many courtesans, thus the display of courtesans might signal power.
- A larger staff makes it more costly to fire the manager since they all have to go with him (being inefficiently numerous).
- Increasing the firm's size tend to relax competitive pressure and it is often a strategic move used to prevent entry (cf. Part 10).
- Whenever managerial compensation is linked to the firm's market share there is an incentive to grow bigger as shown in §6.4.
- A bigger staff contains more high ranking level jobs which motivates subordinates (middle managers) longing for promotions. Inversely, rewards through promotion rather than year-to-year bonuses force the firm to grow in order to supply the new positions that such promotion-based reward systems require.

To grow bigger a firm has two basic options, internal and external growth. The former means expanding sales using its current assets (capital and employees) to conquer new markets or increase market shares in core markets. The latter method consists in acquiring or merging with another company, a way chosen by a majority of the companies listed in the above tables.

The creation of large unified commercial areas in Europe or North-America has triggered a wave of mergers and acquisitions over the last 20 years. Harford (2005) examines the causes and timing of industry-level merger waves; his statistical analysis reject a standard explanation, that mergers are triggered by (too) high stock market valuations. Rather, he concludes that industry merger waves occur in response to economic, regulatory or technological shocks that require large scale reallocation of assets. However, a sufficient liquidity in capital markets is necessary for the wave to actually take place, in order that transaction cost do not hinder the fundamental motives of mergers. This condition causes industry merger waves to cluster in time even if industry shocks do not. The following alphabetical listing of cases and their origin is Harford (2005)'s table 2.

Aircraft (1999): Big, older fleets require increased maintenance, repair and overhaul. Increasingly outsourced from carriers, who want "one-stop shops".

Banking (1985): Deregulation allows interstate banking, particularly in California.

Banking (1996): Deregulation and Information Technology.

Business Services (1986): Partially IT-driven mergers as IT becomes important.

Business Services (1998): Fragmented, smaller players combine, share cost structures, offer more complete line of services to customers; industry grows as outsourcing takes off.

Business Supplies (1997): Paper and pulp industry consolidates from fragmented price takers to gain market power and avoid costly duplication of capital intensive production facilities.

Candy & Soda (1992): Snapple and other non-carbonated beverages make strides, leading to activity to beat or buy them.

Chemicals (1995): Large cash flows, over capacity in production, need to consolidate research.

Communication (1987): Break-up of AT&T in 84 was followed by entry into long distance, investment in fiber optic capacity, etc.

Communication (1997): Telecommunications Act in 1996, consolidation, technological changes.

Computers (1998): Internet

Consumer Goods (1986): Mature market and the need to offer full line leads to consolidation.

Electrical Equipment (1986): Several companies seek growth through acquisition to compete better with industry leaders Westinghouse and General Electric.

Electronic Equipment (1999): OEM's growth leads to demand for electronic equipment manufacturers to shift from small regional players to larger global players capable of infrastructure, IT, etc. to grow with their customers.

Entertainment (1987): Deregulation allows firms to own many stations.

Entertainment (1998): Studios seek diversified production sources and strong libraries. Telecom act of 1996 relaxes media ownership limits.

Food Products (1999): Retail consolidation pushes distribution consolidation and or sale of distributors to bigger retailers who want to buy rather than build distribution channels.

Healthcare (1996): Service providers consolidate to have bargaining power with HMOs.

Insurance (1998): Bigger is safer, leading to consolidation, especially in reinsurers.

Machinery (1996): Large manufacturers decreased number of suppliers they were willing to deal with in bid to improve efficiency. This forced consolidation in a number of capital goods industries; many smaller players were bought in "roll-up" deals.

Measuring and Control Equip (1998): Depression in semi-conductor industry (big customer).

Medical Equipment (1998): Two motives: first, acquisitions in core areas to grow, then acquisitions outside core areas to offer broad products to increasingly consolidated customer base (hospitals).

Personal Services (1996): Consolidation in legal and funeral services industries.

Petroleum and Natural Gas (1997): Increasing prices, record drilling, increasing costs lead drive to increase size to be more efficient.

Pharmaceutical Products (1998): Mid-sized companies merge to garner size necessary to fund increasingly large costs of development.
Restaurants, Hotels, Motels (1985): Saturation and similarity, trends toward take-out, competition from supermarket delis.
Restaurants, Hotels, Motels (1996): Operators such as Starwood have buying sprees. Others buy properties to gain sufficient bulk to compete in corporate account business market.
Retail (1986): Shift to specialty stores as aging department stores consolidated; value of land and buildings in revitalized urban centers.
Retail (1996): Strong growth and impact of internet Shipbuilding,
Railroad Equip (1998): Shrinking defense budgets finally forced the issue of overcapacity in the industry.
Steel Works (1997): Collapse in demand from Asia leads to falling prices forcing consolidation.
Transportation (1986): Mostly still working-out issues following deregulation.
Transportation (1997): End of Interstate Commerce Commission, overcapacity in shipping, open-skies agreements, railroad consolidation started with a few big mergers and then forced responses to balance.
Utilities (1997): Deregulation in some markets plus elimination of a law prohibiting mergers between non-contiguous providers.
Wholesale (1996): Simultaneous consolidation in several wholesale sectors as growth slows and firms move to add breadth, take advantage of new IT ability, grow by acquisition.

Table 15.10: Merger Waves in the US

15.1.3 Merger Scrutiny

The report EC (2001) on merger and acquisitions (M&A) of 2001 ascertains between 9000 and 17000 yearly operations involving a European firm over a ten years period. The UK alone accounts for one third of all activity followed by Germany (16%), France (13%), the Netherlands (6%) and Italy (6%). The figures for the US are slightly superior and are ahead the European trend by 2 years as can be seen on Figure 15.1.

Each year, more than 15000 EU firms exchange over 5% of their titles with other firms (for an amount of at least 1 M€). There is a merger or an acquisition whenever a majority of ownership titles is involved. According to EU law, a merger has an effect on trade between member states if the aggregate worldwide turnover of all the undertakings concerned is more than 2.5 bn€; in that case the operation must be notified to the

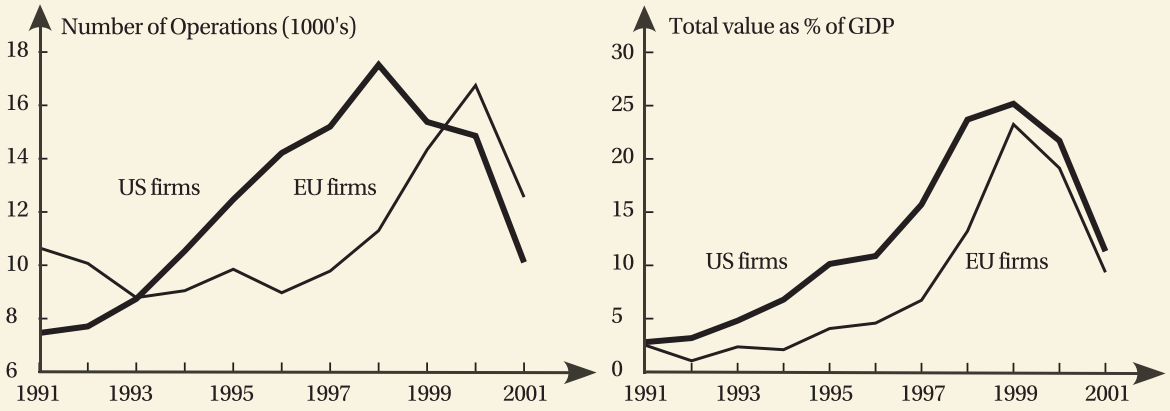


Figure 15.1: M&A in Europe and the USA

EC. Figure 15.2 presents the trend in merger notifications to the EC since the inception of the Merger Regulation in 1989 (cf. latest data). Over a total of 3368 notifications i.e., 190 per year, 3% are withdrawn, 90% are declared out of scope or compatible, 3% lead merging firms to apply remedies and finally in 19 cases (a mere 0.6%) was the proposal prohibited.

When competitiveness is endangered by a proposed merger or acquisition, the EC asks the parties to apply remedies for instance a commitment to exit from a joint-venture, granting access to an infrastructure or technology, terminating an exclusive agreement or transferring a market position (divesture). The latter instrument, by far the most frequent, intends to find a suitable purchaser who would be able to use the transferred asset to exercise a sufficient competitive force on the merging parties. For instance, to accept the Total-Fina acquisition of Elf, the EC required the new company to sell hundreds of highway gas stations in France and Spain. Similarly, the merger of two large tissue and diaper manufacturers, Kimberley-Clark and Scott Paper in 1996 was accepted by US authorities upon their agreement to divest many assets including the second company's brand name "Scotties". A recent ex-post review of the EC regarding these remedies shows a frequent inadequate scope of the divested business e.g., the omission of key assets that were necessary for the viability and competitiveness of the divested business.

The scale of notifications of mergers and acquisitions to the FTC in the US is about tenfold that of the EU in part because the threshold for compulsory notification is a smaller 50m$. Bergman et al. (2009) compares the EU and US merger policies. Focusing on dominant-firm mergers, they find that EU is tougher than the US on average and on mergers resulting in low market shares. Also, US policy is more affected than EU policy by a range of market considerations.

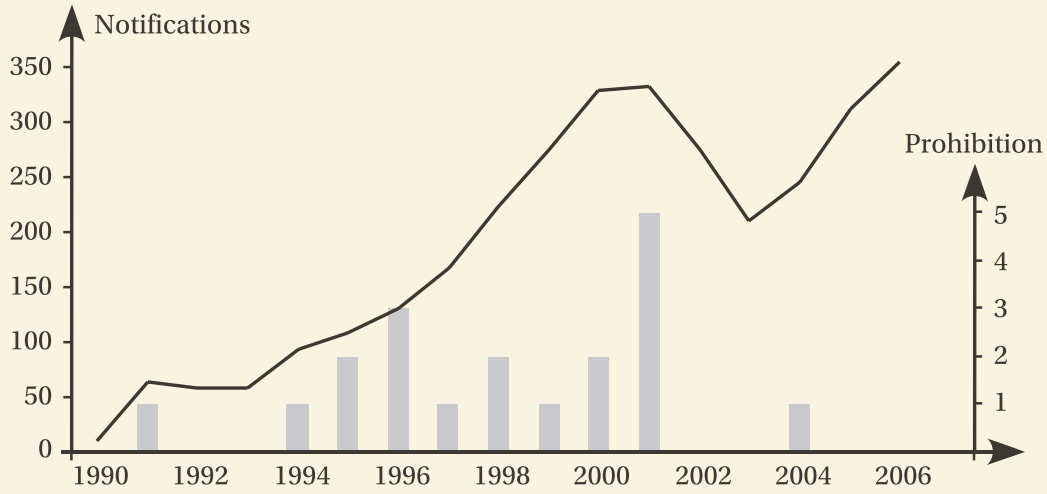


Figure 15.2: European Merger Notifications

15.2 Merger Paradox

Absent scale economies or issues of cost reductions, a merger always seems profitable given that the new firm has more market power; indeed, it can internalize the previously destructive behavior of its former members against each other and thereby achieve a higher total profit than by acting independently. This intuition is actually true but only temporally because it needs to be driven to its logical limit: if the merging firms change their behavior then the non merging ones will also react to this change. The question is then, is this good or bad for the insiders? The answer depends on whether competition is driven by prices or quantities.

15.2.1 Cournot Competition

A merger between firms acting in an homogeneous goods market is viable only if their *absolute* share of industry profits is bigger after the merger. Salant et al. (1983) show that this is very unlikely to happen in a Cournot setting, thereby contradicting the intuition presented above. The explanation of this paradox has to do with the chain of reactions created by the merger.

In a Cournot setting where firms compete in quantities, merging firms (insiders) immediately reduce their production to internalize their new interdependency; this obviously increases their total profit. Yet, since quantities are strategic substitutes, the non merging firms (outsiders) react by expanding their output which hurts the merger's profit although he reacts with a second reduction of output. That will again trigger an expansion from the outsiders and a further loss of profit for the merger. This process

of action-reaction will continue until a new post-merger equilibrium is found; it surprisingly happens that the merger's profit is less than the pre-merger profit of the insiders.

This can be checked within the simple model of Cournot competition seen in §5.1.3. Consider n plants with the same constant returns to scale (CRS) technology that are initially owned by n different firms competing in quantities; in equilibrium, the individual quantity produced by each firm is[7@] $q_n = \frac{a-bc}{n+1}$ and profit is $\pi_n = \frac{1}{b}q_n^2$. Whenever a merger occurs, the CRS hypothesis makes the new owner indifferent on how to distribute production among plants (cf. §2.1.3 for the general multi-plant treatment). We therefore assume a centralized behavior: the merging firms act as a single firm which means that market concentration is reduced.

In this context, a merger among $k+1$ firms leaves only $n-k$ independent firms in the new market configuration. The merger is viable only if the profit of the conglomerate π_{n-k} is larger than the sum of profits of the merging entities $(k+1)\pi_n$. This condition can also be written from the point of view of an acquirer: it is worthwhile to buy k firms only if the profit increase $\pi_{n-k} - \pi_n$ due to reduced market competitiveness is greater than the acquisition cost. The latter is $k\pi_n$ given that each original owner must renounce to π_n. In our example, the ratio $\frac{(a-bc)^2}{b}$ appearing in profits can be eliminated so that the feasibility condition reduces to

$$\frac{1}{(n-k+1)^2} \geq \frac{k+1}{(n+1)^2} \Leftrightarrow k \geq \frac{2n+1-\sqrt{4n+5}}{2} \Leftrightarrow \lambda \equiv \frac{k+1}{n} \geq h(n) \equiv 1 + \frac{3-\sqrt{4n+5}}{2n}$$

It is easy to check that the proportion of firms participating in the merger $h(n)$ reaches a minimum of $\frac{4}{5}$ for $n = 5$, hence

> *Merger paradox*: In a market for homogeneous good where firms compete in quantities, no merger with less than 80% of the initial members of the industry will ever take place.

Practically, this means that the only case where two firms find it profitable to merge is when they already control the whole market (duopoly). Reversing the formal result we have just obtained enables to rationalize a frequently observed conduct: a firm buys another one but leaves it operating as an independent subsidiary although their brands are competitors. Indeed, we have just seen that a full merger with integration of sales forces reduces the overall profit, thus it is better to keep the sales teams as independent competitors and streamline the production side (costs). Such a decentralized behavior is achievable, say, by instructing or motivating the managers of the subsidiaries to operate so as to maximize their individual firm profits.[8@]

In the automotive industry we find the case of PSA with the brands Peugeot and Citroen who sell very similar cars,[9@] Fiat with the quality differentiated brands Fiat,

Lancia and Alfa-Romeo or VW AG with the horizontally and vertically differentiated brands Volkswagen, Audi, Skoda and Seat. This behavior from the parent companies can be profitable: although their subsidiaries compete against themselves they also compete against rivals so that the conglomerate gains market shares. Notice that when VW bought the Spanish firm Seat and the Slovak firm Skoda to market them in the entire European market, it increased overall European competition, thus reduced the cake, but succeeded to increase its own share in absolute value.[10@]

15.2.2 Bertrand Competition

For markets where firms compete in price over horizontally differentiated goods, Davidson and Deneckere (1985) confirm that mergers increase market power under the condition that merging firms are not able to alter product's characteristics. In that situation, any merger is beneficial for all firms, merging and non merging ones alike; also, the bigger the merger, the larger the benefit.

To prove this result, we use the demand $D(p_i, p_j) = a - bp_i - dp_j$ seen for duopoly in equation (5.18) of §5.2.3 and extend it to oligopoly replacing the competitor's price p_j by the average price of all other firms $\bar{p}_{-i} = \frac{1}{n-1}\sum_{j\neq i} p_j$. The profit of a single firm i is $\pi_i = (p_i - c)D(p_i, \bar{p}_{-i})$ and the profit of a k firms merger (involving the first k labels) is $\Pi^k = \sum_{j\leq k}\pi_j$. The FOC for the price of an insider in the post-merger equilibrium can be decomposed into the direct effect of her price on her own profit plus the indirect effect of her price on the profit of the other merging firms; formally

$$\begin{aligned} 0 &= \frac{\partial \Pi^k}{\partial p_i} = \frac{\partial \pi_i}{\partial p_i} + \sum_{j\neq i}\frac{\partial \pi_j}{\partial p_i} \\ &= a + bc - 2bp_i + d\bar{p}_{-i} + (p_i - c)d\lambda_k \end{aligned} \tag{15.1}$$

where $\lambda_k \equiv \frac{k-1}{n-1}$. Notice that for $k = 1$, we have the FOC of an insider in the pre-merger equilibrium or the FOC of an outsider in both situations.

Since the indirect effect is positive for an insider within a merger ($\lambda_k > 0$), it motivates the insider to raise his price beyond what was optimal when he was an independent firm ($\lambda_k = 0$). We deduce that any merger increases the best reply of any insider. Thus, starting from the pre-merger equilibrium p^*, all insiders will increase their price so that all outsiders, in reaction, will increase their own (but less). Then, insiders react to the outsiders move and increase again their price generating a lesser raise by the outsiders. This process eventually converges to a post-merger equilibrium where all insiders play r and all outsiders play s satisfying $r > s > p^*$.

We illustrate this process on Figure 15.3. We use equation (15.1) with $k = 1$ for an

outsider j and the fact that $\bar{p}_{-j} = \frac{kr+(n-k-1)s}{n-1} = \lambda_{k+1} r + (1-\lambda_{k+1}) s$ to deduce

$$\begin{aligned} 2bs &= a + cb + d\left(\lambda_{k+1} r + (1-\lambda_{k+1}) s\right) \\ \Leftrightarrow s &= \Phi(r) \equiv \frac{a + cb + d\lambda_{k+1} r}{2b - d(1-\lambda_{k+1})} \end{aligned} \tag{15.2}$$

Next, we use (15.1) with $k > 1$ for an insider i and the fact that $\bar{p}_{-i} = \lambda_k r + (1-\lambda_k) p$ to deduce

$$\begin{aligned} (2b - \lambda_k d) r &= a + (b - \lambda_k d)\, c + d\left(\lambda_k r + (1-\lambda_k) s\right) \\ \Leftrightarrow r &= \Psi^k(s) \equiv \frac{a + (b-\lambda_k d)\, c + (1-\lambda_k) d s}{2b - 2\lambda_k d} \end{aligned} \tag{15.3}$$

The pre-merger equilibrium is $p^* = \frac{a+bc}{2b-d}$ at the intersection of Φ and Ψ^1 while the post-merger one is at the intersection of Φ and Ψ^k and has the announced characteristics since $k > 1$ implies $\lambda_k > 0$ and $\Psi^k > \Psi^1$.

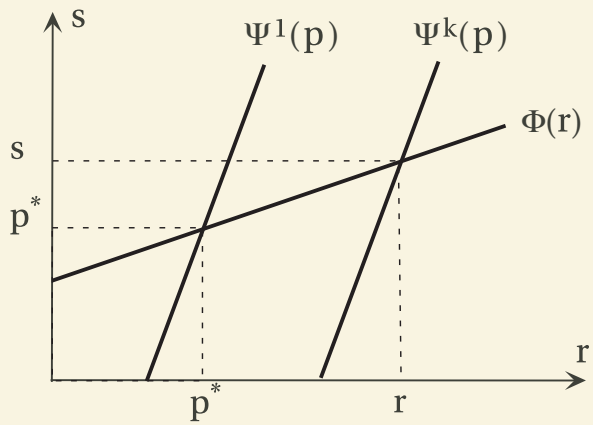


Figure 15.3: Merger Profitability under Price Competition

To understand this result, we first have to recall that in a framework of Bertrand competition, a price increase by any firm is beneficial to all others. This, in turn, means that when some firms merge into a trust, each has a greater incentive to raise her price because her interest is now the trust's profit instead of her own. Indeed, a price increment has now two effects. Firstly, it participates directly to the firm's profit, thus to the trust benefits. Secondly, it raises the profits of the other trust members. What we have shown is that after a merger, all trust members increase their price so that all outsiders, in reaction, will increase their own (but less). Then, trust members will again react to the outsiders move and increase again their price generating a lesser raise by the outsiders. This process eventually converges to a post-merger equilibrium where all trust members propose a price greater than that of outsiders which is itself greater than the

pre-merger common equilibrium price.

15.2.3 Efficiencies and Welfare

Discussion

Two doctrines of economic competition rival for leadership. The Structure Conduct Performance (SCP) paradigm focuses on allocative or *static* efficiency. It studies how firms compete simultaneously in the current market; its policy aim is to create workable competition in today's market. Building on Schumpeter (1942)'s "creative destruction", the Chicago critic holds a *dynamic* efficiency view; it studies how firms compete sequentially for the market and emphasizes innovation. A temporary monopoly that enables the innovator to recoup its investment is thus seen as a necessary evil on the path towards the higher goal of long term progress.

According to tenants of the SCP paradigm, mergers and acquisitions increase firms' market power which leads to higher prices and hurts consumers. On that ground, significant mergers ought to be opposed. For long, antitrust practitioners were in agreement with this view and sought to protect consumers from monopolization. The Chicago school responds from two angles. The "efficiency defense" started by Williamson (1968) broadens the SCP's vision in asserting that mergers can contribute positively to welfare by bringing in efficiencies. This author also argues that the adequate criteria for antitrust authorities is (total) welfare, not consumer surplus. More radically, Demsetz (1973) contends that the SCP confuses correlation and causation. True, there is a positive correlation between market concentration and industry profitability but which one drives the other cannot be identified with a comparative statics exercise, the very tool used by the SCP paradigm. The Chicago schools then endorses a reverse causal chain: the more efficient (innovative) firms capture greater market shares, earn more and tend to buy out the less efficient firms (or drive them out of the market). The policy implication with respect to mergers is also reversed: allowing innovative firms to acquire obsolete ones promotes efficiency and ultimately welfare.

Market Power and Cost Asymmetry

Cournot (1838) introduces the standard model of quantity competition in his chapter 7 and observes that a firm with low marginal cost produces more than higher-cost firms, and that some of the latter might be forced to exit. He also notices that a given total industry output would be produced at higher cost by competing asymmetric producers than if a monopolist made their production decisions (because their marginal costs are

not equal at equilibrium).

Williamson (1968) exploits this intuition and shows that if a merger generates a synergy (i.e., marginal cost reduction) then the welfare loss due to the price increase may be compensated by the cost saving. Although the original model is crude, the intuition is so strong that it remains robust to generalization (cf. next §).

Cowling and Waterson (1976) show in a Cournotian model that the average profit-revenue ratio is equal to the concentration-elasticity ratio. Although no causality can be deduced from this formula, it has been adopted as a foundation for the "market power" rationale. Clarke and Davies (1982) further show that concentration increases with the variance of firms' marginal costs; this proves that concentration is greater when some firms have a cost advantage, a result that lands support to the dynamic efficiency rationale. In the same vein, Salant and Shaffer (1999) show that if the average marginal cost in the industry is constant, then so are the aggregate output and the consumer surplus. When a shock makes such an industry more cost-asymmetric, concentration, aggregate profit and welfare all increase together.[11@] This finding provides a rationale for government support of "national champions" at the expense of other domestic firms with the same initial technology.

Let us develop formally those properties of the Cournot equilibrium under asymmetric cost. There are n active firms with constant marginal cost c_i for $i \leq n$. We denote $Q \equiv \sum_{i\leq n} q_i$ the aggregate output, $\epsilon \equiv \frac{-P}{QP'}$ the elasticity of demand, $H \equiv \sum_{i\leq n} \left(\frac{q_i}{Q}\right)^2$ the HHI concentration index (cf. §15.3), $\pi_i = q_i(p - c_i)$ the individual profit, $\Pi \equiv \sum_{i\leq n} \pi_i$ the aggregate profit and $\bar{c} \equiv \frac{1}{n}\sum_{i\leq n} c_i$ the average marginal cost.

The FOC of profit maximization is

$$P(Q) + q_i P'(Q) = c_i \Rightarrow p - c_i = \frac{p q_i}{\epsilon Q} \tag{15.4}$$

thus

$$\pi_i = q_i(p - c_i) = \frac{pQ}{\epsilon}\left(\frac{q_i}{Q}\right)^2 \Rightarrow \Pi = \frac{pQ}{\epsilon} H \tag{15.5}$$

which is the Cowling and Waterson (1976) formula. FOC (15.4) also reads $\frac{q_i}{Q} = \epsilon \frac{p-c_i}{p}$, thus

$$H = \epsilon^2 \sum_{i\leq n} \left(\frac{p - c_i}{p}\right)^2 \Rightarrow \frac{p^2 H}{\epsilon^2} = \sum_{i\leq n} \left(p - \bar{c} + \bar{c} - c_i\right)^2 = n\left(p - \bar{c}\right)^2 + n\sigma_c^2 \tag{15.6}$$

where σ_c^2 is the variance of the sample of marginal costs. Observe now that summing

(15.4), we obtain $n(p-\bar{c})=\frac{p}{\epsilon}$. Plugging in (15.6), we get $\frac{p^2H}{\epsilon^2}=\frac{p^2}{n\epsilon^2}+n\sigma_c^2$, thus

$$H=\frac{1}{n}+\frac{n\epsilon^2\sigma_c^2}{p^2}=\frac{1}{n}+\frac{(1-n\epsilon)^2}{n}v_c^2 \tag{15.7}$$

where v_c is the coefficient of variation. This is the Clarke and Davies (1982) formula. Salant and Shaffer (1999) observe that summing the left version of (15.4), one gets $P(Q)+\frac{1}{n}QP'(Q)=\bar{c}$ i.e., aggregate output Q and consumer surplus S depend on $\bar{c}$ only. Now, by combining (15.5) and (15.7), we get $\frac{n\epsilon}{pQ}\Pi=1+(1-n\epsilon)^2v_c$. When $\bar{c}$ is constant, so are Q,p,ϵ, thus welfare $W=\Pi+S$ increases with dispersion of technologies. The limit is reached when a maximum number of firms are driven out of the market (or at least lose their economic rent) while the remaining ones achieve zero marginal cost, a result in the line of Cournot's observation regarding industry cost.

Merger Synergies

Let us start with a typology of merger efficiencies. One speaks of *rationalization* when production is reshuffled among plants to lower variable cost or when knowledge diffuses to bring all plants towards the production frontier.[12@] *Scale economies* occur when the merged firm reaches the minimum efficient scale or saves on fixed cost duplication (but remain on the same production frontier). Lastly, there is a *synergy* when the merged firm succeeds to innovate i.e., do something impossible to achieve unilaterally e.g. cut slack in the organization, combine complementary hard-to-trade assets or achieve faster a R&D discovery.

Williamson (1968)'s argument relies on synergies brought about by the merger. Consider a Bertrand duopoly among two identical firms; the price equilibrium is at their common marginal cost $p^*=c$. The efficiency occurs as follows: by merging, the firms are able to improve their technology and reduce their marginal cost by δ. Being a monopoly, the new firm sets $p^M=\frac{1+c-\delta}{2}$ and earns $\pi^M=\left(\frac{1-c+\delta}{2}\right)^2$ whereas the change in consumer surplus is

$$\Delta W_D=-\frac{1}{2}\left(p^M-p^*\right)\left(q^*+q^M\right)=-\frac{1}{8}\left(3-3c+\delta\right)\left(1-c-\delta\right)$$

As $\Delta W=\Delta W_D+\pi^M$, welfare improves if

$$\pi^M>-\Delta W_D\Leftrightarrow(1-c)^2<3\delta\left(2(1-c)+\delta\right)\Leftrightarrow\frac{1-c}{\delta}<3+2\sqrt{3}\Leftrightarrow\frac{\delta}{1-c}>0.15$$

i.e.,the cost reduction is greater than 15% of the original net WTP.

Farrell and Shapiro (1990) show that one cannot preclude concentration, price **and** welfare from growing together i.e., the profit increase generated by efficiencies may offset

the "dead weight" loss. Yet, they show that without synergies or scale economies, the market price is bound to rise so that consumers suffer from a merger.

The first result is rather simple to demonstrate. The profit of firm i is $\pi_i = pq_i - C_i(q_i)$ while the consumer surplus $W_D(q) = \int_0^q P(x)\,\mathrm{d}x - qP(q)$ (cf. eq. (2.19), thus welfare reads

$$W = W_D(q) + \sum_{i\geq 1}\pi_i = \int_0^q P(x)\,\mathrm{d}x - \sum_{i\geq 1} C_i(q_i)$$

and the welfare change due to a change in output decisions is

$$dW = P(q)dq - \sum_{i\geq 1} C_{m,i}dq_i = \sum_{i\geq 1}(P(q) - C_{m,i})dq_i = -P'(q)\sum_{i\geq 1} q_i dq_i$$

by (5.11). Since $H = \sum_{i\geq 1}\left(\frac{q_i}{q}\right)^2$, we have

$$\textstyle\sum_{i\geq 1} q_i dq_i = \frac{1}{2}d\left[\sum_{i\geq 1} q_i^2\right] = \frac{1}{2}d\left[q^2 H\right] = qHdq + \frac{1}{2}q^2 dH$$

so that $dW = -Hq^2P'(q)\left(\frac{dq}{q} + \frac{1}{2}\frac{dH}{H}\right)$ and we can have the unusual situation where the market output shrinks ($dq < 0$), concentration increases ($dH > 0$), yet welfare still rises ($dW > 0$) because a large chunk of production has been moved from inefficient to efficient firms.

The second result is that absent synergies, total output contracts after a merger.[13@] We show it when marginal cost are constant. Letting w.l.o.g. $c_1 \leq c_2 \leq .. \leq c_n$, we saw with eq. (5.12) that total output is $q = \frac{n}{n+1}(a - bc)$ where $c \equiv \frac{1}{n}\sum_{j\geq 1} c_j$ is the average industry cost.[14@] Upon merging firm i with any better one, the new owner will use exclusively the more efficient technology i.e., shuts down the obsolete plant. We thus have index i disappear from the list of active firms, hence the total output becomes $\widehat{q} = \frac{n-1}{n}(a - bc_{-i})$ where $c_{-i} \equiv \frac{1}{n-1}\sum_{j\neq i} c_j$. We have

$$\begin{aligned} n(n+1)\left(q - \widehat{q}\right) &= a - b\left(n^2 c - (n^2-1)c_{-i}\right) = a - b\left(n\textstyle\sum_{j\geq 1} c_j - (n+1)\sum_{j\neq i} c_j\right) \\ &= a - b\left(nc_i - \textstyle\sum_{j\neq i} c_j\right) = (n+1)q_i \quad \text{by (5.14)} \end{aligned}$$

thus total output shrinks even if the worst technology is bought by the best one. As claimed, price rises and consumer surplus falls.

Rationalization

As shown in Boccard (2009), a simple rationalization can be welfare improving. Consider the demand $D(p) = 1 - p$, n "cutting edge" firms with marginal cost c_1 and n "obsolete" firms with marginal cost $c_2 = c_1 + \delta$. We let $Q_1^* \equiv a - c_1$ and $Q_2^* \equiv Q_1^* - \delta$ be the efficient

(competitive) market outputs under the two technologies. The *technological gap* is the dimensionless ratio $\gamma \equiv 2\frac{Q_1^*-Q_2^*}{Q_2^*} = \frac{2\delta}{a-c_2}$ measures the strength of the cost advantage.

Let us first characterize the equilibrium in the Cournot game of quantity competition. The FOCs of profit maximization are $2q_1 = a - bc_1 - (n-1)q_1 - nq_2$ and $2q_2 = a - bc_2 - (n-1)q_2 - nq_1$ so that the equilibrium is $q_1 = \frac{a-c_1+\delta n}{2n+1}$ and $q_2 = \frac{a-c_2-\delta n}{2n+1}$ where we notice that $q_1 - q_2 = \delta$. Obsolete firms participate only if $q_2 > 0 \Leftrightarrow a - c_2 > \delta n \Leftrightarrow \gamma < \frac{2}{n}$, a condition which we shall assume to hold.

Upon merging an efficient firm to an obsolete one, we still have n efficient firms but $n-1$ obsolete ones since the new owner will use exclusively the efficient technology i.e., shuts down the obsolete plant. The output changes between the old and new equilibrium computed using (15.2.3) are $\Delta q_2 = \Delta q_1 = \frac{q_2}{2n} > 0$ i.e., the increased concentration benefits all remaining firms. Yet one obsolete firm has been shut down so that the aggregate change is $\Delta Q = (2n-1)\frac{q_2}{2n} - q_2 = -\frac{q_2}{2n}$.

The changes in the various elements constitutive of the welfare are

$$\begin{aligned}
\Delta CS &= \left(Q + \tfrac{1}{2}\Delta Q\right)\Delta Q = -\left(nq_1 + nq_2 - \tfrac{q_2}{4n}\right)\tfrac{q_2}{2n} && (15.8)\\
\Delta \Pi_1 &= n(2q_1 + \Delta q_1)\Delta q_1 = n(2q_1 + \Delta q_1)\tfrac{q_2}{2n} && (15.9)\\
\Delta \Pi_2 &= (n-1)(q_2 + \Delta q_2)^2 - nq_2^2 = (n-1)\left(2q_2 + \Delta q_2\right)\Delta q_2 - q_2^2 \\
&= q_2^2\left((n-1)\left(2 + \tfrac{1}{2n}\right)\tfrac{1}{2n} - 1\right) = -(3n+1)\left(\frac{q_2}{2n}\right)^2 && (15.10)
\end{aligned}$$

Since all variations in (15.8),(15.9),(15.10) are proportional to $\frac{q_2}{n}$,

$$\begin{aligned}
\frac{n}{q_2}\Delta W &= -\left(nq_1 + nq_2 - \tfrac{q_2}{4n}\right) + n\left(2q_1 + \tfrac{q_2}{2n}\right) - (3n+1)\frac{q_2}{2n} \\
&= nq_1 - \left(n + 1 + \tfrac{1}{4n}\right)q_2 = n\delta - \left(1 + \tfrac{1}{4n}\right)q_2 && (15.11)
\end{aligned}$$

Let us interpret (15.11) using the rightmost expression. The first term is the positive welfare balance brought about by efficient firms whereas the negative second term is slightly more than the output of the obsolete mothballed firm. In equilibrium, an efficient firm produces more than an obsolete one; let us call "output gap" the difference $q_1 - q_2 = \delta$. A merger between two asymmetric firms is welfare improving if the combined output gap of all efficient firms $n\delta$ is slightly larger than the output of the retired obsolete firm.

15.3 Measures of Concentration

The exercise of market power is primarily linked to the size of a firm so that indices of concentration within an industry are of crucial importance. The first task is thus to

define precisely the market where concentration is measured.

15.3.1 What is a Market?

To answer this question we need to identify the firms that might sell products relatively similar for consumers, the later being themselves identified as an homogeneous group.

Products

A product or service can be perfectly described by its technological characteristics and the process of its production. Since a perfect substitute is impossible to find one could conclude that there is a unique producer in each market. Yet society does not implode when a product is missing because we always manage to find something similar called a substitute; now if any product has some, even imperfect, substitute then there is a single market in the whole economy!

Those two schizophrenic views can be useful to develop the theory because they enable to concentrate on other important strategic aspects. In the present section, however, we present the workable concept used by the EC when it needs to apply the theory to judge whether competition is being distorted in a market. The EC concept of relevant market identifies the boundaries of competition between firms on two accounts:

- The *product* market comprises all those products which are regarded as substitutable by reason of characteristics, prices and intended use; products that could readily be put on the market by other producers without significant switching cost or by potential competitors at reasonable cost and within a limited time span also need to be taken into account. The legal view of the EC is interchangeability from the viewpoint of consumers **or** producers.
- The *geographic* market comprises the area in which concerned firms are involved in the supply and demand of products in which the conditions of competition are sufficiently homogeneous and which can be distinguished from neighboring areas, because the conditions of competition are appreciably different in those areas.

The definition of the relevant market is the object of controversy as the following example makes clear. The chemical firm DuPont used to control some 75% of the US cellophane market in the 1950s. The US government interpreted the data as evidence of monopoly while a court agreed with the company that the relevant market was much larger because cellophane could be substituted by aluminium foil, wax paper or polyethylene. In the resulting "wrapping" market, DuPont's share was only a modest 20%. It is

clear that a firm under scrutiny will always argue for a high degree of substitutability in order to mechanically decrease her market share.

Historically Cournot (1838) and Marshall (1890) argued that a perfectly competitive market would be characterized by highly correlated prices i.e., the prices need not be the same in different locations (e.g., because of transportation cost), but they should change at the same time. Also, two close substitutes should have a stable relative price because arbitrage opportunities among consumers act as a centripetal force on prices.

In practice, the EC uses own and cross-price elasticities of demand to check whether two products are close substitutes in which case they are said to belong to the same market. The methodology consists in applying a small but significant non-transitory increase in prices (SSNIP-test) for all product of the candidate market. If the joint profits of firms increase, then the candidate market can be treated as the relevant market because the set of included products are sufficiently close substitutes among themselves and sufficiently distant substitutes for the other non-included products.

Verboven (2002) applies this test to passenger cars in the EU using sales data between 1970 and 1999 from the top 5 national markets covering 75% of European sales. For that period, the permanence of important international price differentials is proof of the existence of trade barriers, thus the geographic markets for car retailing should be accordingly defined as the national markets. The candidate markets are defined using common classifications from the specialized press. Table 15.11 displays the candidate markets and the econometric estimations with the own price elasticity ϵ, the cross price elasticity inside the segment ξ, the cross-price elasticity across segments η and $\Delta\Pi$, the joint profits increase (for Germany) after a 10% joint increase of prices within one segment.

Segment	Example	ϵ	ξ	η	$\Delta\Pi$
Small	Opel Corsa	3.0	0.01	0.002	16
Compacts	VW Golf	2.4	0.04	0.002	8.4
Intermediate	Peugeot 406	2.9	0.03	0.002	8.2
Standard/luxury	Audi A6	5.9	0.11	0.001	22
Sports	Mercedes SLK	3.3	0.02	0.001	5.4
Minivans	Renault Espace	1.7	0.10	0.001	0.1

Table 15.11: Relevant Car Market in Europe

The conclusion we can read from the table is that all but the minivan segments are relevant markets. For the failing one, a further segmentation according to size and price would be needed to distinguish the now large variety of monospace vehicles.

Participants

Since substitute products may not be already on the market, the EC distinguishes two kind of actors:

- An *actual* competitor to a firm is another firm which is either currently active on the same relevant market or which is able to switch production to the relevant market in the short term without incurring significant additional costs or risks in response to a small and permanent increase in relative prices (immediate supply-side substitutability).
- A *potential* competitor is a firm that would be likely to undertake the necessary additional investments to enter the market in response to a small and permanent increase in prices. Market entry needs to take place sufficiently fast so that the threat of potential entry is a constraint on the market participants' behavior. The time period needed by companies already active on the market to adjust their capacities can be used as a yardstick to determine this period.

The EC then defines *potential competition* as the pressure exercised upon incumbent firms by the possibility that new or existing firms will enter their market. New entrants may be attracted by above normal profits made in this market by incumbent firms, possibly as a result of weak competition. Additional firms entering the market will increase the overall quantity supplied with the effect that prices fall and above normal profits disappear. Hence, potential competition has a "disciplinary effect" on the behavior of incumbents. However, this threat is relatively small when entry barriers are high.

Classifications

Products classifications are also useful to define markets as they sort out the technological characteristics of products and services.

The International Standard Industrial Classification of all Economic Activities (ISIC Rev. 3.1) of the United Nations has 60 divisions belonging to the 17 categories presented in Table 15.12.

Two widely used classification systems are:

- North American Industry Classification System (NAICS) developed by the member countries of the NAFTA agreement (USA, Canada and Mexico).
- General Industrial Classification of Economic Activities (NACE) (acronym derived from the French full name) developed by the member countries of the European Community.

A	Agriculture, hunting and forestry
B	Fishing
C	Mining and quarrying
D	Manufacturing
E	Electricity, gas and water supply
F	Construction
G	Wholesale and retail trade, household goods
H	Hotels and restaurants
I	Transport, storage and communications
J	Financial intermediation
K	Real estate, renting and business activities
L	Public administration, defense, social security
M	Education
N	Health and social work
O	Other community, social and personal services
P	Private households with employed persons
Q	Extra-territorial organizations and bodies

Table 15.12: ISIC Activities Classifications

Both systems use the ISIC categories and divisions but differ in finer levels because NAICS is more focused on the production process. A convergence project between NACE and NAICS is currently under development. A wealth of information on classifications can be found at Eurostat's Server RAMON.

Product diversity

Assessing the diversity of available products in a given market is not straightforward. Intuitively, one would use the number of products. Applying this to the market for computer operating systems we have to count all softwares available in commercial, shareware or freeware form; we end-up with a large figure but also with the feeling that the right one would be 3 since the most popular ones are Microsoft's Windows, Linux and Apple's Mac OS X. In a sense, their popularity is a demonstration of their quality and therefore an indicator of the real range of products available for a new user. At the limit, the indicator should not to jump on the day a new product is launched. Lastly, if we divide this market into personal and business computing, we would like total diversity to be the sum of product of diversity among the two classes and within them.

To build an index Υ of product diversity satisfying these properties the information at our disposal is the current number of products n and their market shares, hence Υ

is a function of the vector of market shares $\mathbf{s} = (s_i)_{i \leq n}$. The first property states that counting products is correct if the market is evenly shared: if $\forall i \leq n, s_i = \frac{1}{n}$ then $\Upsilon(\mathbf{s}) = n$. The second property is akin to continuity of Υ (small changes induce small changes). To enunciate the third stability property, we need to decompose the long list of n products into a smaller number of classes $j = 1$ to m, each one containing k_j products. The class shares are then $\forall j \leq m, \sigma_j = \sum_{i \in j} s_i$ while shares within classes are $s_i^j = s_i / \sigma_j$. The class-stability property expressed in multiplicative form then reads $\Upsilon(\mathbf{s}) = \Upsilon(\boldsymbol{\sigma}) \times \prod_{j \leq m} \Upsilon(\mathbf{s}^j)^{\sigma_j}$. As shown by Theil (1967), these axioms uniquely define our product diversity index as the *entropy* measure $\Upsilon(\mathbf{s}) = -\prod_{i \leq n} (s_i)^{s_i}$ introduced by Shannon (1948) in his theory of information.[15@]

15.3.2 Concentration Indices

Concentration arises either where two or more previously independent firms merge (merger) where a firm acquires control of another (acquisition of control), or where a joint venture is created, performing on a lasting basis all the functions of an autonomous economic entity (full-function joint venture).

Assuming that the relevant market has been defined, the active firms are identified and their sales over a meaningful period of study, say a year, are recorded. We can then rank firms according to their market shares, a number in $[0;1]$, from the largest s_1 to the smallest s_n. The *concentration* index sums the market shares of the k largest firms,

$$C_k \equiv \sum_{i=1}^{k} s_i$$

while the index proposed by Herfindahl (1950) and Hirschmann (1964) sums all squares,

$$H \equiv \sum_{i=1}^{n} s_i^2 \tag{15.12}$$

Index C_4 was first used in 1968 by the US merger guidelines and later replaced in 1982 by the Herfindhal-Hirschmann index (HHI) which is H where shares are expressed in percentage; it ranges from 10.000 in the case of a pure monopoly to a small number in the case of an atomistic market (nearby perfect competition). The EC, being a junior antitrust body, directly adopted the latter index. For practical purposes the spectrum of market concentration is divided into non concentrated for $HHI < 1000$, highly concentrated if $HHI > 2000$ (1800 in the US) and moderately concentrated otherwise. To get an idea of the meaning for these thresholds, consider n firms sharing evenly a market, we have $HHI = 1000/n$ i.e., $HHI = 1000$ for $n = 10$ and $HHI = 2000$ for $n = 5$.

As first uncovered by Gibrat (1931) and developed §15.3.4, the distribution of firm sizes is highly skewed in the sense that a few large players make up most of the output (or whatever measure of sizes we may consider). This implies the existence of a competitive fringe whose individual contribution to HHI is almost nil (due to the compressing effect of the squaring). We can thus treat the firms below 1% as having a zero contribution and fasten the analysis by focussing on the ten to twenty largest firms in the market. Applying Gibrat's Law stating that firm growth rates are independent of size, we obtain an asymmetrical distribution[16@] for which we can compute in Table 15.13 concentration as a function of the number of active firms.

n	3	4	5	6	8	10	15	20	30
HHI	4894	3790	3155	2740	2227	1917	1494	1271	1031

Table 15.13: HHI under the Gibrat's Law

In case of a merger, one considers both the post-merger market concentration and the increase in concentration resulting from the merger. The EC rule is slacker than its US counterpart and amounts more or less to accept any merger in any of the following cases:

- post-merger HHI below 1000
- post-merger HHI between 1000 and 2000 (1800 in the US) but with an increase less than 250 (100 in the US)
- post-merger HHI greater than 2000 (1800 in the US) but with an increase less than 150 (50 in the US)

Another popular measure of market power, especially among economists, is the Lerner (1934) index (cf. eq. (3.4)), $L_i \equiv \frac{p_i - C_{mi}}{p_i}$ that divides the margin by the price; it is zero in a competitive market and nearby one for a monopoly facing an inelastic demand. The Lerner index for the market is then $L \equiv \sum_{i=1}^{n} s_i L_i$ which, in a first approximation, is equal to H divided by the elasticity ϵ of the market demand as shown by Dansby and Willig (1979).. Computing the Lerner index at the Cournot equilibrium, one finds $L_i = \frac{s_i}{\epsilon}$ where $s_i = \frac{q_i}{\sum_j q_j}$ is the market share and ϵ the elasticity of consumer demand. One then checks that $\sum_j s_j L_j = \frac{1}{\epsilon} \sum_j s_i^2 = H/\epsilon$.

On analyzing a potential M&A operation between industry members i and j, the current HHI is often compared to the potential HHI that would result if all market shares remained the same because the computation involved is trivial. The mechanical increase of HHI is $(s_i + s_j)^2 - (s_i^2 + s_j^2) = 2 s_i s_j$. Yet, if no firm changes its output, the price does not change either so that welfare increases as soon as there is an economy of scale for the merged entity (i.e., the sum of insiders profits increase). It would therefore appear that HHI is not a good indicator of market power. This apparent contradiction is in fact

a reminder that the post merger HHI must be computed at a market equilibrium just like the pre-merger HHI in order that we can meaningfully compare these two figures.

15.3.3 Industry Cases

We conclude this chapter with several numerical illustrations. §9.1.2 presents the concentration figures for the famous OPEC oil cartel.

Movies

Data kindly provided by Bruce Nash allows to look at the market shares of the major film studios as reported in Table 15.14. While Disney is a clear leader, there are five other major studio that maintain concentration at a low level.

Distributor	1995	2000	2005	2009	Mean
Walt Disney	22.5	21.3	14.4	11.8	18.0
Warner Bros.	16.4	11.4	16.0	20.0	14.4
Sony Pictures	13.3	9.7	10.9	14.1	13.5
20th Century Fox	8.0	10.0	16.4	16.0	12.5
Paramount Pictures	10.0	10.7	9.6	14.3	11.7
Universal	12.6	15.8	12.4	9.9	11.5
New Line	6.2	5.0	4.7	0.0	5.0
Dreamworks SKG	0.0	10.3	5.6	0.0	3.5
MGM	6.3	1.4	2.0	0.2	2.9
Lionsgate	0.0	1.4	3.2	3.8	2.3
Weinstein Co.	0.0	0.0	1.1	1.6	0.4
Summit Entertainment	0.0	0.0	0.0	4.5	0.6
Newmarket Films	0.0	0.0	0.1	0.0	0.4
IMAX Films	1.2	1.2	0.4	0.2	0.4
Gramercy	1.1	0.0	0.0	0.0	0.2
Revenue ($bn)	**5.3**	**7.5**	**8.9**	**10.6**	**8.2**
HHI	**1356**	**1278**	**1166**	**1334**	**1191**

Table 15.14: Movie Distribution Market Shares

The situation is similar in France as shown on Table 15.15 but with changing positions as alliances are made and broken.

Distributor / 2000	*%*
Gaumont Disney	19.7
UIP	12.5
Bac Films	10.7
Pathé	9.2
UFD	9.2
Columbia	7.3
ARP	7.0
Warner Bros	6.8
Metropolitan	2.7
Pyramide	2.0
Total top ten	**87**
HHI	**1003**

Distributor / 2004	*%*
Warner Bros	15.0
UIP	11.3
Pathé	11.1
Disney	10.8
Gaumont Columbia	10.4
Mars Distribution	9.4
UFD	7.2
Metropolitan	4.1
TFM Distribution	3.9
EuropaCorp	2.7
Total top ten	**86**
HHI	**898**

Distributor / 2009	*%*
20th Century Fox	12.2
Warner Bros	9.4
Pathé	7.9
Sony	7.7
Disney	7.5
SND	6.3
TF1 / UGC	5.7
Mars Distribution	5.0
StudioCanal	4.9
Metropolitan	4.9
Total top ten	**72**
HHI	**642**

Table 15.15: Movie Distribution Market Shares

Computer

During the 1990s, Microsoft's products have gained a very large market share by outperforming competitors. For operating systems, the family of windows has gained an extremely dominant position that has only begun to shrink in the latest years; it is the ubiquitous example of monopoly. In the market for web browsers, the various versions of IE have proven less resistant to the oncoming of a variety of challengers. Lastly, the newest market of web search has allowed Google to become a new mammoth of the internet.

Table 15.16 uses recent data collected by NetApplications to illustrate this point but also the recent threats posed by aggressive new comers. Using these data, Figure 15.4 displays the evolution of the monthly HHI for the three digital markets.

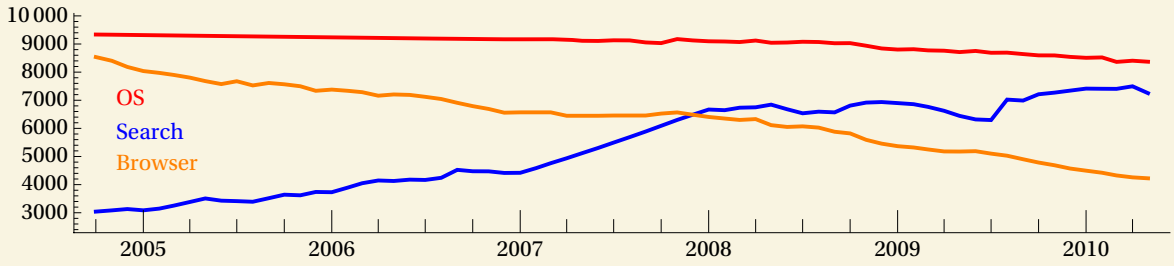


Figure 15.4: HHI in the digital markets

European Industries

In the food retail (supermarkets) sector, the EC has computed the EU wide concentration index C_5 for 1996; it obtains a low 15% which indicates an absence of concentration.

OS	Windows	Mac	Other				HHI
2004	90.3%	3.2%	6.5%				8198
2005	92.6%	3.9%	3.5%				8602
2006	92.5%	5.2%	2.3%				8580
2007	91.6%	6.2%	2.2%				8430
Browser	IE	Firefox	Safari	Netscape	Opera	Other	HHI
2004	92.3%	2.7%	1.5%	2.1%	0.5%	0.9%	8536
2005	86.5%	8.6%	2.6%	1.3%	0.5%	0.5%	7568
2006	81.3%	13.0%	3.9%	0.8%	0.6%	0.4%	6791
2007	79.8%	13.7%	4.7%	0.8%	0.7%	0.4%	6570
Search	Google	Yahoo!	MSN	Other			HHI
2004	50.2%	15.0%	14.6%	20.2%			3040
2005	57.5%	12.5%	9.6%	20.4%			3642
2006	65.2%	11.6%	5.4%	17.9%			4475
2007	64.8%	11.3%	4.8%	19.1%			4420

Table 15.16: Market Shares and HHI in the Software Industry

Nevertheless Europe was still the conjunction of 15 national markets at the time, not a unified single market; this is reflected by the fact that the average of national C_5 indexes is a much larger 44% because many member states have powerful national retail chains. Matraves (1999, 2002) in her works on European integration and market structure offers interesting data gathered in Table 15.17.

C_4	1970	1975	1980	1987	1993
DE	25.4	34.0	32.2	20.8	22.0
UK	22.1	26.2	29.0	34.9	63.2
IT	25.9	30.3	41.1	27.0	23.6
FR	66.9	70.7	76.2	63.2	63.4
EU	-	-	-	26.4	28.7

Soft Drinks

C_4	1985-7	1991-3
DE	26	28
UK	34	35
IT	17	15
FR	11	11
EU	19	16

Pharmacy

Table 15.17: Concentration in some European Industries

Automobile

Table 15.18 illustrates the European market for cars.[17@] We observe almost no concentration at the EU level but one should bear in mind that the sector benefited from a block exemption that fragmented the single market in rather isolated national markets. A recent EC directive has revised the exemption; it still permits restrictive vertical agreements between manufacturers and their dealers but only if the market shares held by the companies concerned do not surpass the 30% and 40% limits. Furthermore it eliminates from exemption the location clauses whereby the dealer is assigned a specific main

location and is prohibited from operating additional sales or delivery outlets at other locations. The aim is to facilitate multi-brand outlets (i.e., foster interbrand competition) and reduce territorial restrictions (i.e., foster intrabrand competition among dealers).

Firm / Year	**1980**	**1985**	**1990**	**1995**	**1999**
Fiat	18.7%	15.7%	13.8%	11.1%	9.5%
Volkswagen	15.6%	16.4%	15.7%	16.8%	18.8%
Peugeot-Citroën	17.5%	12.9%	12.7%	12.0%	12.1%
Ford	14.8%	16.3%	11.5%	13.7%	11.7%
Renault	16.6%	12.0%	9.7%	10.3%	11.0%
Opel (GM)	11.1%	13.3%	12.0%	13.1%	11.5%
Japan	10.9%	12.1%			
Daimler-Chrysler	4.3%	4.2%	3.2%	4.0%	5.6%
Nissan	3.8%	2.7%	2.9%	3.0%	2.6%
B.M.W.	0.0%	3.1%	5.6%	6.4%	3.2%
Toyota			2.7%	2.5%	3.2%
Mazda			2.1%	1.4%	1.4%
Other			2.1%	0.5%	0.3%
Rover	0.0%	4.3%	0.0%	0.0%	1.5%
Mitsubishi			1.3%	1.1%	1.2%
Honda			1.2%	1.5%	1.4%
Korea	0.0%	0.1%	0.1%	1.5%	3.2%
Sales (Million)	**7.4**	**9.5**	**13.5**	**12.0**	**15.1**
HHI	**1580**	**1338**	**1042**	**1092**	**1059**
$\mathbf{C_4}$	**67%**	**61%**	**54%**	**54%**	**52%**

Firm / Year	**2000**	**2003**	**2006**	**2009**
Volkswagen	18.7%	18.2%	19.8%	20.9%
Peugeot-Citroën	13.1%	14.8%	13.2%	13.3%
Ford	10.8%	11.0%	10.7%	10.4%
Opel (GM)	10.8%	9.8%	10.2%	8.9%
Renault	10.6%	10.6%	8.6%	9.1%
Fiat	10.0%	7.4%	7.6%	8.8%
Daimler-Chrysler	6.2%	6.5%	6.2%	4.9%
Toyota	3.7%	4.8%	5.8%	5.2%
BMW	3.4%	4.4%	5.3%	5.1%
Nissan	2.7%	2.8%	2.1%	2.6%
Hyundai	1.5%	1.7%	2.0%	2.2%
Honda	1.2%	1.4%	1.7%	1.6%
Mazda	1.2%	1.5%	1.7%	1.5%
Suzuki	0.9%	1.0%	1.5%	1.6%
Kia	0.5%	0.8%	1.5%	1.6%
Mitsubishi	1.1%	0.8%	0.8%	0.6%
Others	0.3%	0.3%	0.6%	1.1%
Sales (Million)	**14.7**	**14.2**	**14.8**	**13.7**
HHI	**1048**	**1038**	**1035**	**1060**
$\mathbf{C_4}$	**53%**	**54%**	**54%**	**53%**

Table 15.18: European Automobile Market (EU-15)

Regarding the US vehicle market, we use WardsAuto's data on Cars and Trucks to construct Table 15.19. We present the market share of the 16 largest sellers of the last decade ordered by their average. The continuous substitution of US historical big-three towards the Japanese big-three induces a reduction of concentration. As observed on Figure 15.5, the HHI concentration index has steadily decreased over the last 40 years and converged to the low level achieved in the EU two decades ago.

Firm/Year	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*
GM	28.0	28.0	28.3	27.7	26.9	25.6	23.9	23.2	21.9	19.6
Ford	22.6	21.6	19.9	19.2	18.0	17.0	16.0	14.6	14.2	15.3
Toyota	9.1	10.0	10.3	11.0	11.9	13.0	15.0	16.0	16.5	16.7
Chrysler	14.2	13.0	12.9	12.5	12.8	13.2	12.6	12.6	10.8	8.8
Honda	6.5	6.9	7.3	8.0	8.1	8.4	8.9	9.4	10.6	10.9
Nissan	4.2	4.0	4.3	4.7	5.7	6.2	6.0	6.5	7.1	7.3
Hyundai	1.4	2.0	2.2	2.4	2.4	2.6	2.7	2.8	3.0	4.1
Volkswagen	2.5	2.5	2.5	2.3	1.9	1.8	1.9	2.0	2.3	2.8
Daimler	1.9	1.8	1.8	1.9	2.0	2.2	2.4	2.1	2.4	2.4
BMW	1.1	1.2	1.5	1.6	1.7	1.8	1.8	2.0	2.3	2.3
Kia Motors	0.9	1.3	1.4	1.4	1.6	1.6	1.7	1.9	2.0	2.8
Mazda	1.4	1.5	1.5	1.5	1.5	1.5	1.6	1.8	2.0	2.0
Subaru	1.0	1.1	1.1	1.1	1.1	1.1	1.2	1.1	1.4	2.0
Mitsubishi	1.8	1.9	2.0	1.5	0.9	0.7	0.7	0.8	0.7	0.5
Volvo	0.7	0.7	0.7	0.8	0.8	0.7	0.7	0.7	0.5	0.6
HHI	**1610**	**1565**	**1524**	**1475**	**1421**	**1353**	**1329**	**1268**	**1198**	**1440**

Table 15.19: US Vehicle Market Shares

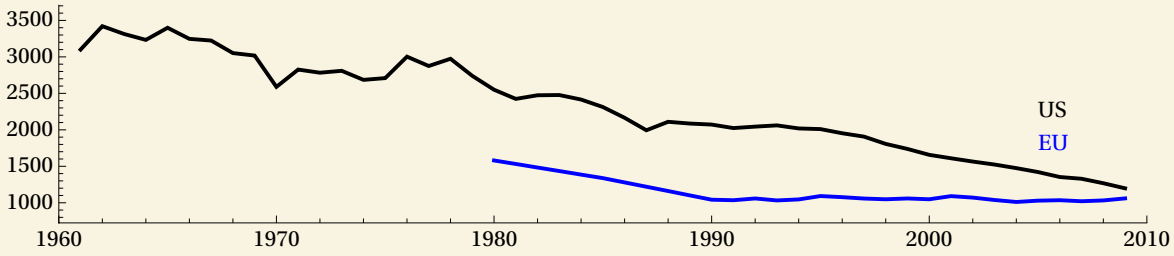


Figure 15.5: HHI evolution in the US & EU car markets

The degree of penetration of foreign cars in the EU is approximatively 45%, mostly Japanese and Korean brands. The corresponding figure for the US is 55% as European car makers have a non negligible share. On the contrary, the degree of openness of Japan and Korea is extremely low around 5% which means that the local brands have a greater market power (3 in Korean, 7 in Japan). One should not however conclude that Europeans show less national preference. Indeed, as shown on Table 15.20, member states of the EU show a marked preference for local car makers although with a twist.

Germany has 3 truly national firms but is also the place in Europe where 2 historical US firms develop cars; in the end, the national share is beyond two thirds. As a consequence concentration is about 50% higher than the European figure; there are many equally sized firms competing for the German market but the VW is really a dominant player. The French case fits the US or asian countries with a strong national preference and an extremely high concentration back in the 80s that slipped due to the decline of Renault and the progressive penetration of all foreign brands. On appearance, Spain is

Germany	2009
Volkswagen	34.2%
Opel	9.7%
Mercedes	8.3%
Ford	8.0%
BMW	6.8%
PSA	6.1%
Renault	5.9%
Fiat	4.7%
Toyota	3.9%
Mitsubishi	2.4%
Mazda	1.7%
Mitsubishi	1.6%
Nissan	1.6%
Toyota	1.5%
Suzuki	1.2%
Toyota	0.8%
Other	0.7%
Sales (Mil.)	**3.8**
HHI	**1569**
National	**67%**

France	1980	1990	2009
PSA	36.6%	33.1%	31.6%
Renault	40.5%	27.7%	25.0%
VW	5.1%	10.3%	11.3%
Ford	4.1%	7.4%	6.4%
Opel	1.8%	5.2%	4.9%
Fiat	4.6%	7.1%	4.4%
Toyota	0.7%	0.7%	4.0%
BMW	0.9%	1.3%	2.7%
Mercedes	0.8%	1.2%	2.6%
Nissan	0.9%	1.1%	2.0%
Suzuki	0.0%	0.0%	1.3%
Hyundai	0.0%	0.0%	0.9%
Kia	0.0%	0.0%	0.9%
Honda	0.4%	0.6%	0.6%
Mazda	0.7%	0.8%	0.6%
Chrysler	0.0%	0.3%	0.2%
Other	2.7%	2.7%	0.3%
Sales (Mil.)	**1.9**	**2.3**	**2.3**
HHI	**3053**	**2112**	**1874**
National	**77%**	**61%**	**57%**

Spain	2009
Volkswagen	23.1%
PSA	17.8%
Renault	10.3%
Ford	10.1%
Opel	7.7%
Toyota	5.9%
BMW	4.3%
Nissan	3.8%
Mercedes	3.2%
Fiat	2.5%
Hyundai	1.8%
Honda	1.8%
Kia	1.6%
Mazda	1.1%
Suzuki	1.0%
Mitsubishi	0.7%
Other	1.8%
Sales (Mil.)	**1.0**
HHI	**1216**
National	**73%**

Italy	2009
Fiat	32.8%
Ford	10.5%
PSA	10.4%
Volkswagen	10.2%
Opel	8.3%
Toyota	4.9%
Renault	4.3%
Mercedes	3.9%
BMW	3.4%
Nissan	2.5%
Hyundai	1.8%
Suzuki	1.6%
Kia	1.0%
Honda	0.9%
Mazda	0.7%
Chrysler	0.4%
Other	2.0%
Sales (Mil.)	**2.2**
HHI	**1549**
National	**33%**

Table 15.20: Vehicle Market Shares in european Countries

different because its sole national brand is now foreign owned; yet, the country is home to assembly plants from most European car makers so that we may compute a quasi national share at about 70%; this diversity is reflected into a concentration level only slightly higher that the EU value. Italy, lastly, hosts a unique large automobile group that dominates the market; yet concentration reaches a level similar to the German one as other automakers have shares around 10%.

Airlines

Table 15.21, using US Federal data, reports on the US market for domestic and international flights as well as revenue (for US carriers only). Tables are sorted by decreasing rank in 2009. Figure 15.6 reports the evolution of HHI over the last tow decades for domestic and international passenger volumes.

Domestic Mkt.	1990	2000	2009
Southwest	5.3	12.1	16.3
American	15.5	11.4	10.7
Delta	14.7	16.3	9.0
United	12.3	12.1	7.3
US Airways	14.0	9.4	7.2
Northwest	8.1	8.1	5.3
Continental	7.3	6.1	5.1
AirTran	0.0	1.3	3.8
SkyWest	0.0	0.0	3.2
JetBlue	0.0	0.2	3.2
American Eagle	0.0	1.8	2.4
Alaska	1.2	2.0	2.3
Atlantic Southeast	0.0	1.0	2.1
Expressjet	0.0	1.2	1.9
Pinnacle	0.0	0.0	1.7
Total (M. Pass.)	417	600	621

Int. Mkt.	1990	2000	2009
American	10.4	12.7	13.0
Delta	5.1	5.4	8.0
Continental	5.6	6.2	8.0
United	5.5	7.5	6.3
Northwest	6.4	5.8	4.5
US Airways	2.0	2.2	4.3
British Airways	4.8	4.5	3.7
Lufthansa	2.2	2.7	3.2
Air Canada	3.9	3.4	3.2
Air France	1.2	2.1	2.4
Virgin Atlantic	1.0	2.5	2.3
Mexicana	3.5	2.0	2.0
JetBlue	0.0	0.0	1.6
Korean	1.2	1.2	1.5
Westjet	0.0	0.0	1.4
Total (M. Pass.)	81	141	151

Dom. Revenue	1990	2000	2009
FedEx	10.0	12.0	12.9
American	14.5	13.9	12.8
Delta	11.5	11.8	11.6
United	14.5	14.8	10.6
Continental	6.9	7.0	8.0
US Airways	8.0	7.0	7.0
Northwest	9.6	8.4	7.0
Southwest	1.6	4.3	6.7
UPS	1.2	1.9	2.9
JetBlue	0.0	0.1	2.1
Alaska	1.2	1.4	1.9
AirTran	0.0	0.5	1.5
American Eagle	0.0	1.0	1.2
SkyWest	0.0	0.0	1.1
Hawaiian	0.4	0.5	0.8
Total ($bn)	76	130	155

Table 15.21: Airlines Market Shares in the US

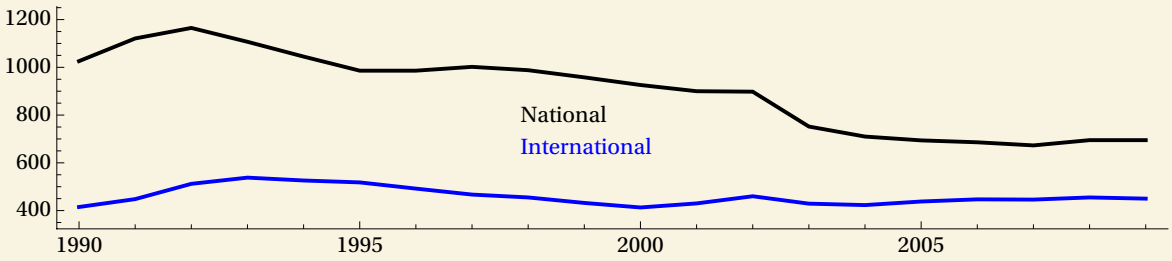


Figure 15.6: Concentration in the US airline market

15.3.4 Size Distribution

Cities (within a country), firms (within an industry), wealth (within a population) display highly unequal sizes, mostly because some have grown faster. As originally shown by Pareto (1896) for wealth distribution, Auerbach (1913) for cities population and Gibrat (1931) for firm size, 80% of the quantity being measured comes from 20% of the producing units aka. the 80-20 rule. Such an empirical regularity has been observed for a large number of social activities but also in nature. For instance, Gaffeo et al. (2003) show that the distribution of firms' size in the most industrialized countries follow closely Zipf's rank rule presented hereafter.

It is important to explain why the statistical distribution of firm sizes does not obey the predicaments of theory where firms end up being symmetrical in equilibrium. The explanation given by Gibrat (1931) is that each family in a city, each employee in a firm or each euro of capital wealth are equally able to generate an offspring, come up within an innovation or being invested in a superior project. As shown in the theory section below, such a proportional (or size independent) growth process generates a power law

statistical distribution (after enough periods) fitting adequately the empirical one. First, we offer an informal presentation based on popular items of everyday life.

Intuition: Zipf's law

Classify all your emails according to who is the sender or receiver and sort them into separate boxes for each of your m contacts. Then order the boxes from the biggest to the smallest as in Table 15.22.

Names	Ali	Bess	Chris	Doug	Ellen	...	$\#i$	Total
Count	n_a	n_b	n_c	n_d	n_e	...	n_k	N
Frequency	$f_a = \frac{n_a}{N}$	f_b	f_c	f_d	f_e	...	f_k	1
Set $\gamma \equiv f_a$ and compute	f_a/γ	f_b/γ	f_c/γ	f_d/γ	f_e/γ	...	f_k/γ	
only to find	1	$\simeq 1/2$	$\simeq 1/3$	$\simeq 1/4$	$\simeq 1/5$	...	$\simeq 1/m$	

Table 15.22: Email contacts

It is then possible to show that $\sum_{i=1}^{m} \frac{1}{i} \simeq 0.577 + \ln(m)$ with convergence for m large. Manipulating the equation $1 = \sum_{i=1}^{m} f_i$, one obtains $\gamma \simeq \frac{1}{0.577+\ln(m)}$ i.e., that in most human collections like books or friends or words in a language, the most frequent item has a frequency as given in the following table:

Collection size (m)	10	100	1000	10^4	10^5	10^6	10^7	10^8	10^9
Frequency (%)	35	19	13	10	8	7	6	5.3	4.7

Table 15.23: Frequency of preferred item

Now, the utility we derive from exchanging emails with all m contacts can be estimated by the total number of emails

$$u(m) = N = N \sum_{i=1}^{m} f_i = N\gamma \sum_{i=1}^{m} \frac{1}{i} = n_a \sum_{i=1}^{m} \frac{1}{i} \simeq n_a (0.577 + \ln(m)) \approx \log(m)$$

the base 10 logarithm since we can change the monetary unit (divide by $\frac{n_a}{\ln(10)}$) and then subtract a suitable constant without altering the representation of preferences (cf. §2.2.1). Hence, the utility of having 10 email contacts is $\log(10) = 1$ while that of 100 contacts is $\log(100) = 2$. In other words, my "top-ten" contacts are as valuable as the remaining 90 entries. This decomposition of total utility into increments can be applied to other fields using the property $\log(10^m) = m$:

- Books collection of size 1000: the 10 books I read most often are as worth to me as the next 90 I consult from time to time which are as valuable as the remaining 900 which I barely remember reading.

- Internet Music Store of size one million songs: the collection may be divided into 3 clusters of equal value to the owner (say yourself or a store): the top 100 of "hits", the next 9900 of "classics" and the remaining 990000 "never-heard-of" songs.
- US Phone Network: if the customer base is approximatively 100 millions (10^8) then the 10000 biggest clients (10^4) generate half of the network value.

This clustering shows that digital databases have a very Long Tail that can still be profitably marketed thanks to the nearly zero cost of accessing an item when there is a request for it.

Theory

An economic variable X follows a power law if the complementary cumulative distribution function [18@] has the form $\Pr(X \geq x) = bx^{-a}$ for some positive parameters a and b. The property is said to apply to the (upper) tail when it holds true for large x only. To see whether some data follow the power law, take a set of ordered observations $x_1 > .. > x_n$, estimate $\Pr(X \geq x_i)$ by $g_i = \frac{i}{n}$ and plot $(\ln x_i, \ln g_i)_{i \leq n}$ to find a straight line with slope $-a$.

Gibrat (1931)'s hypothesis of a proportional growth process is $Z_t \equiv \frac{X_{t+1} - X_t}{X_t} = \mu + \sigma\epsilon_t$ where ϵ_t is a white noise.[19@] The idea is that each family in a city, each employee in a firm or each euro of capital wealth are equally able to generate an offspring. If this is correct, we can use the fact that $Z_t \simeq \ln X_{t+1} - \ln X_t$ for small periods, to sum from a initial period up to a final time T and get $\ln X_T = \ln X_0 + \mu T + \sigma \sum_{t \leq T} \epsilon_t$. By the central limit theorem, the mean $\frac{1}{T}\sum_{t \leq T} \epsilon_t$ tends to a normal variable when T grow large, so that $\ln \frac{X_T}{X_0}$ tends to a normal variable with mean μT and variance $\sigma^2 T$ (i.e., the distribution of X_T is lognormal).

Gibrat (1931)'s model fits well empirical data except for the upper tail that is systematically fatter or longer than predicted in the sense that more large units are observed than predicted. It is then necessary to take into account the entry and exit of units from the current population. Regarding firms, we may assume that, at each period, a new market opportunity or innovation arises and can be taken up by an existing firm or by an entrepreneur that will create a new firm to exploit it. When we look today at the current population, it consists of units born at different times in the past so that we do not have a sample of realizations of the random variable X_T for an exogenous T but realizations of $X_{\hat{T}}$ where $\hat{T}$ is the endogenous age of the firm. Reed (2001) shows that if the instantaneous probability of entry λ is constant over time (exponential process) then

$Y \equiv \ln \frac{X_{\hat{T}}}{X_0}$ follows a power law in both the upper and lower tails i.e.,

$$\begin{cases} \Pr(Y > y) = \frac{\beta}{\alpha+\beta} e^{-\alpha y} & \text{if } y > 0 \\ \Pr(Y < y) = \frac{\alpha}{\alpha+\beta} e^{\beta y} & \text{if } y < 0 \end{cases} \Rightarrow \begin{cases} \Pr(X_{\hat{T}} > x) = \frac{\beta}{\alpha+\beta} x^{-\alpha} & \text{if } x > x_0 \\ \Pr(X_{\hat{T}} < x) = \frac{\alpha}{\alpha+\beta} x^{\beta} & \text{if } x < x_0 \end{cases}$$

where α and $-\beta$ the two roots of the characteristic quadratic equation $\frac{1}{2}\sigma^2 t(t-1) + \mu t = \lambda$ and using $y = \ln(\frac{x}{x_0})$.

Part G

Public Oversight

Part D has studied the interactions of firms with competition law taking the latter as exogenously given. In this part, we inquire about the motivations for public authorities to step up into the economic life. We take a shortcut by referring to the *State* as the superstructure gathering all the organizations in charge of framing the economic activity; it includes the legislative, judicial and executive bodies and within the latter, the government, its bureaucracy and regulatory agencies.

The first chapter recalls what is the State, why it exits, how it came to weight so much over economics and how it receded a little in recent decades. The next chapter takes an advanced look at regulation, the direct State intromission in an industry. We end this part with a chapter on natural resources because this sector displays natural monopoly features and externalities requiring public oversight.

Chapter 16

The State

In microeconomics, the State is foremost associated with the design and enforcement of the legal framework under which economic agents interact. Yet it is also the main economic actor being the largest employer (civil servants, health and education personnel), the largest purchaser (procurement), the largest investor (infrastructures) and, last but not least, a tight controller of most markets through price caps, quotas and regulations. Firms thus make business with the State while being overseen by him. This chapter will try to make sense of these seemingly unrelated activities, keeping an emphasis on the implications for firms.

We first present the *missions* that the State ought to perform and the leading role it is driven to assume in the economy. We then rationalize this judge-and-party duality using the economic concepts of *market failure* and its opposite, the *government failure*. We also offer a power struggle view of the formation of the State called *rent-planning*. It helps to make sense of the ubiquitous *rent-seeking* phenomenon induced by the hierarchical State-firm relationship, which is the object of the next section. Lastly, we review the recent changes in the structure of the State with the *liberalization* wave.

16.1 Missions and Means

16.1.1 Missions

The French Republic's motto of liberty, equality and fraternity summarizes neatly the objectives of advanced democracies: rule of law, provision of public services and wealth redistribution. We present them in turn and conclude with a few statistics regarding the weight of the State in the economy.[1@]

Rule of Law

The pursuit of happiness, to which we all aspire, requires living in a peaceful society and enjoying effective political and economical freedoms. The accepted definition of Weber (1922) defines a State as "a human community that (successfully) claims the monopoly on the legitimate use of physical force within a given territory". Its sovereign missions, defining a minimal State, are diplomacy, defense, justice and police. We can also add macro-economic stabilization since nation-wide crisis have the potential to generate chaos and anarchy.

The general aim of justice is to protect individuals from abuse and exploitation. The State sets limits to our freedom in some dimensions to guarantee us minimal freedoms in all dimensions; it is then the enforcer of "human rights". The transposition of "freedom as effective rights" from citizens to economic agents (firms, entrepreneurs) revolves around "property rights" which are fully developed in §8.1. The core mission of the State is thus to establish the *rule of law* i.e., that everyone, including the State itself (and its representatives), is to obey the laws enacted by the community (through its representation system). The hurdles faced in this endeavor are analyzed later in §16.2.4.

Public Services

Equality as a guiding principle of democracies is nowadays understood as equality of opportunities, not of results such as income. The State has then a mandate to "level the playing field" for individuals, in the same manner that competition law does for firms (cf. §8.2.4). To implement this equity principle, society makes value judgments and defines the public services as those which citizens "should" consume. This modern vision justifies in hindsight the avowed paternalism of economists (and more generally intellectuals) that existed before the advent of democracy and widespread education.[2@] An economic, but partial, rationalization is soft paternalism: as we suffer from bounded rationality, cognitive problems or lack of information, we make mistake so that there is a corrective role for the State (cf. also Behavioral Economics).

For out purpose in this book, public services are conveniently divided into *essential* and *network* ones. The former are education, health, sport, culture, housing and social services.[3@] The State subsidizes some of them but more often than not, organizes free universal provision (in-kind transfers) or even mandates compulsory consumption. The fact that the State is also a producer of these services has political roots but is also an optimal economizing response in the face of a lasting market failure: information asymmetry (cf. §13.1.3).

The network services are transportation (road, train, plane), communication (mail,

phone, internet) and households services (water, sewage, energy). Although society does not treat these services as "essential", they are seen as "basic". For that reason, the State oversees their development and funds their consumption, both on the supply and demand sides. At the same time, the industrial sectors supplying network services are natural monopolies involving high sunk costs, positive network externalities and a potential abuse of monopoly power. As we explain in detail in §17.1.1, these market failures warrant State oversight in order to guarantee maximal efficiency (allocative and productive).

Public services are provided either directly by state owned enterprises (SOE), indirectly by delegation to private firms or by a mix of the previous options. The specialized arm of the State supervising a private provider is called the *regulator* (so that SOEs are often referred to as self-regulated). Its tasks and the analysis of the strategic interaction with natural monopolies is the object of chapter 17 on regulation.

Welfare State

Fraternity (brotherhood) provided by the traditional circles of family, business or parish has faded with the anonymity of urban life; this social safety net has been replaced in affluent countries by a collective system known as the welfare state. It is foremost an expression of fairness that complements the equality of opportunities afforded by public services. According to Esping-Andersen and Myles (2008), the main component is social security which provides insurance in areas of pensions, health, family, labour or minimum wage. Although the very financing of economic security involves in itself some wealth redistribution, direct schemes (e.g., income or property tax) form a second component of the welfare state aimed at reducing income disparity, vertically towards the less affluent or horizontally towards deserving social groups.[4@]

Economic Perspective

We review fundamental, essential and basic services as well as the welfare state from a stricter economic point of view.

Economically speaking, the fundamental services rendered by the minimal State are public goods i.e., non marketable and consumed in the same amount by all.[5@] As such they suffer from *free riding*, namely that voluntary contribution would raise too little in order to sustain a fully functioning market economy. Thanks to its monopoly on coercion, the State is able to finance them by taxing the general economic activity (firms, households, transactions,...). The downsides are, the violation of consumer sovereignty as we are made to pay more than we wish, a sacrifice of valuable scarce resources and an

inefficient distortion of incentives (cf. marginal cost of public funds in §17.1). Those deficiencies are nevertheless acceptable because fundamental services implement the rule of law i.e., reduce the numerous market failures arising from imperfect property rights and fraud in the entire economy (cf. §8.1). Given that the rule of law is firmly established in advanced economies, we are warranted to analyze an increase of fundamental services in terms of cost vs. benefit.[6@]

Essential services and the welfare state mission involve a much stronger moral stance because they are financed by taxation, provided in-kind (instead of cash transfer) and often have a mandatory character. As recalled by Currie and Gahvari (2008), the related markets display some market failures but the overriding explanation of the State's involvement is political will (or paternalism). Now, since taxation, the main financing vehicle, is distortionary and generates welfare losses, we face the dilemma of choosing between efficiency and equity. Lindert (2007) however shows, using recent OECD evidence, that the trade-off need not be so stark: there is no growth loss from spending more on the welfare state.[7@] In a sense, the welfare state is concomitant to growth, neither inimical nor friendly.

Lastly, network services are not *essential* but only *basic*, they are marketed and not given away. Providers, whether public or private, must then cover their cost out of their receipts. At this point, the equity concern of society means that market segments with low willingness to pay and/or hight cost of supply are to be subsidized to foster access to the service (and its consumption). For sure, there is still a value judgement involved in the rates distortion but in a much weaker degree than for essential services since everyone still carry part of the underlying cost. We develop this issue in §17.3.4 on public service obligation.

16.1.2 Means

To carry on its costly missions, the State uses its exclusive dominion of coercion to tax economic units, operate monopoly over some profitable activities and seize land. Laws, decrees and regulations compel citizens and firms to behave adequately under the threat of retaliation with fines, confiscation or imprisonment. In democracies, expropriation for a right-of-way (eminent domain in US parlance) is conditional on a just compensation (cf. art. 17 of the French Human Rights declaration or the US Fifth Amendment.

In advanced economies, the weight of the State is by no means small since more than a third of all wealth creation goes through the State who thus becomes the largest employer and the biggest spender in the nation (cf. procurement in §13.3.3). Historically, Tanzi and Schuknecht (2000) report that general government expenditure (including

central, state and local governments' expenditures on goods and services, debt service and transfers to economic units) in OECD countries grew as shown in Table 16.1.[8@]

1870	1913	1920	1937	1960	1980	1990	1996
11	13	20	24	28	42	43	45

Table 16.1: OECD Government Expenditure in % of GDP

For the EU15 (resp. EU27), general government revenue[9@] averages 45.5% (resp. 44.4%) of GDP over the last 10 (resp. 5) years with a 58% maximum for Sweden. The US and Japan equivalent figures are 33% and 31%. Most revenues come from taxes, the rest being sales of public services. Currently, government tax revenue, including social security, is about 41% of GDP in Europe (EU27) ranging from 50% in Sweden to 29% in Romania. It comes in equal parts from social contributions, direct taxes (income and wealth) and indirect taxes (products and imports).

Table 16.2 displays, for some OECD countries, the distribution of expenditure according to the COFOG methodology with an aggregation fitting our three missions typology.[10@] It is noticeable that the weight of all three missions decrease over the decade (cf. §16.4 on liberalization) except for the converging Korean economy. To assess how costly is the core mission of upholding economic and political rights, we take off defense and debt service (interest) and come-up with a cost between 6 and 8% of GDP.[11@]

Regarding civil service, Table 16.2 also display public employment cost as a share of GDP (cf. labour column). Comparing with the labour force, the highest rates of public employment in 2001 are Denmark (23%), Finland (21%) and France (18%) (cf. OECD (2005)). Some Eastern Europe countries had high levels but have reduced them since 1990 to sanitize their public finances. Although the Anglo-Saxon countries profess a quite different model of the State, the figures for Australia, Canada, the UK and the US range from 16% to 14%.

The conclusion that emerges from this statistical section is that the State is a gigantic actor in all advanced economies, whatever their avowed political stance may be. It controls and manages between one third and one half of the GDP.

16.2 Rationalizing the State

Having presented what the State does, we now seek to rationalize this outcome. We thus expose the standard theories regarding the role of State in the economy and conclude with a view more focused on the power motives of leaders.

Area	Year	**Total**	*Labour*	*Social benef.*	*Core missions*	Gen. services	Defence	Public order	Environment	*Public Services*	Econ. affairs	Health	Education	Culture	Housing
EU15	96	***52.5***	*10.6*	*19.8*	*12.8*	8.6	*1.8*	*1.7*	*0.7*	*19.8*	*6.4*	*5.9*	*5.2*	*1.0*	*1.3*
	01	***45.5***	*10.3*	*18.4*	*10.6*	6.5	*1.7*	*1.7*	*0.7*	*16.6*	*3.5*	*6.0*	*5.1*	*1.0*	*1.0*
	06	***47.1***	*10.1*	*18.8*	*10.6*	6.5	*1.6*	*1.8*	*0.7*	*17.8*	*3.8*	*6.6*	*5.3*	*1.1*	*1.0*
USA	96	***36.5***	*10.2*	*7.2*	*12.3*	6.5	*3.8*	*1.9*	*0.0*	*17.0*	*3.5*	*6.7*	*5.8*	*0.3*	*0.7*
	01	***35.3***	*9.8*	*6.9*	*10.6*	5.3	*3.3*	*2.0*	*0.0*	*17.8*	*3.9*	*6.8*	*6.2*	*0.3*	*0.6*
	06	***36.6***	*10.1*	*7.0*	*11.2*	4.8	*4.2*	*2.1*	*0.0*	*18.5*	*3.7*	*7.7*	*6.2*	*0.3*	*0.6*
CAN	96	***46.6***	*13.0*	*10.7*	*15.5*	11.7	*1.3*	*1.8*	*0.6*	*20.4*	*4.0*	*6.0*	*8.5*	*1.0*	*0.9*
	01	***42.0***	*11.4*	*9.6*	*13.3*	10.0	*1.1*	*1.7*	*0.5*	*19.1*	*3.5*	*6.2*	*7.5*	*0.9*	*0.9*
	06	***39.3***	*11.6*	*9.2*	*10.4*	7.3	*1.0*	*1.6*	*0.5*	*19.7*	*3.4*	*7.3*	*7.2*	*0.9*	*0.9*
JAPAN	96	***36.7***	*6.3*	*9.7*	*10.9*	6.8	*0.9*	*1.4*	*1.8*	*16.1*	*5.2*	*5.5*	*4.1*	*0.2*	*1.0*
	01	***38.4***	*6.6*	*11.5*	*10.3*	6.3	*1.0*	*1.4*	*1.7*	*16.6*	*4.8*	*6.7*	*4.1*	*0.2*	*0.9*
	06	***36.1***	*6.2*	*12.2*	*8.6*	5.0	*0.9*	*1.4*	*1.2*	*15.3*	*3.6*	*7.1*	*3.8*	*0.2*	*0.6*
KOREA	96	***21.7***	*6.7*	*2.0*	*7.4*	2.5	*2.9*	*1.3*	*0.7*	*12.3*	*5.6*	*1.4*	*3.8*	*0.4*	*1.0*
	01	***25.0***	*6.5*	*2.6*	*7.6*	3.0	*2.5*	*1.3*	*0.8*	*14.8*	*5.8*	*3.1*	*4.3*	*0.7*	*0.9*
	06	***30.2***	*7.3*	*3.7*	*9.1*	4.0	*2.8*	*1.4*	*1.0*	*17.3*	*6.4*	*4.1*	*4.7*	*0.9*	*1.2*

Table 16.2: State expenditure in % of GDP

16.2.1 Public Interest

Following the lead of Pigou (1920), the public interest theory sees the State as having a duty to solve market failures, those market equilibria that fail to satisfy the equi-marginal conditions for Pareto efficiency (cf. first welfare theorem in Mas-Collel et al. (1995) or §2.1.1). These failures can originate from externalities, public good features, information asymmetries and monopoly like behavior. For instance, the core mission of the State is to uphold the rule of law in order to guarantee the basic political and economical freedoms. This endeavor has the character of a public good or at least displays positive externalities. Since voluntary provision is plagued by the free riding,[12@] the leadership assumed by public authorities is warranted.

According to the public interest view, the State ought to subsidize and support activities displaying positive externalities (e.g., research, education, health) and to limit or tax those displaying negative externalities (e.g., monopoly pricing, pollution, noise, con-

gestion).[13@] Asymmetric information, dealt with in Part H, is at the root of innumerable quantity and quality regulations such as construction (zoning, materials, codes), food (fertilizers, processing, hormones), retail (zoning, schedules, labels), transport (zoning, security, pollution, driving), schooling (certification, curriculum), health (diploma, licensing, drug authorization) and safety standards (cf. also §9.1.2). Most items in the list are less than a century old and were devised alongside the general increase of welfare in advanced economies. It is therefore natural to asks whether this improvement of our life standards occurred thanks to regulation or despite of it. We agree with Shleifer (2005) that regulation, as we know it, is society's optimal response to deep market failures (cf. §17.1.1 on the origins of natural monopoly). As technology and tastes evolve, so does regulation (though too slowly).

Although a market failure is a hard fact, the belief that the government can correct it at no cost indulges into what Demsetz (1969) calls a "nirvana fallacy"; in other words, a government failure is equally possible.[14@] We now proceed to discuss this possibility with two complementary theories.

16.2.2 Public Choice

Public Choice brings the utilitarian and rational behavior assumed in economic theory to the study of politics: voters, politicians and bureaucrats are viewed as self-interested.[15@] With such a starting point, leaving no space to dedication, enlightenment and cooperation, the outcome is necessarily bleak: in equilibrium of the political game, all activity serves the political elite and its supporters.[16@] The *rent-seeking* concept, studied thoroughly in §16.3, is the main contribution of public choice to economic theory.

A few important results are as follows. Downs (1957) shows that voters have few personal incentives to monitor government and politicians effectively because their vote or political activism weight so little. As a consequence, parliament is almost free to pass any law and the government almost free to pursue any policy. Olson (1965)'s theory of collective action abounds and further shows that special interest groups, with few members, have on the contrary maximal incentives to step-in the political arena. This tends to explain the prevalence of lobbies and pressure groups. Buchanan and Tullock (1962) show that the political decision-making game is one of bargaining among politicians who represent local special interests rather than their voters.[17@] In equilibrium, vote trading among politicians is observed to reach a majority and advance their particular objectives (see also quid-pro-quo or patronage, cronyism, spoils system or pork-barrel and logrolling in US parlance). This process is inefficient because most of these laws bring little local benefits but require additional distortionary taxation at the country level for financing.

In the US, mandatory cost-benefit analysis is an intent to reign in this wasteful process.

As far as industry regulation is concerned, we may say it is second-best efficient because of the constraints weighting on the political system. Although each regulation is initially designed to fix a market failure or carry a moral goal, it is transformed during the enactment process (to pass the majority threshold). As we show in §16.3, there is a lobbying cost involved. A second source of inefficiency is the "tollbooth" problem, a modern form of rent-planning, whereby politicians and bureaucrats use regulation to their own advantage (cf. §16.2.4). First, they limit competition (instead of fostering it) to create rents and then they extract the rents from the regulated firms through campaign contributions, votes and bribes (cf. §16.3.2).

The public choice literature agrees, in principle, to state intervention in order to solve market failures and finance public goods. Yet, it observes that the government tends to expand without bounds up to the point where it does more harm than good to the economy and the population. It thus advocates dealing with external economies and diseconomies through voluntary cooperation by setting up private organizations instead of governmental bureaus.[18@] Experience shows that such a construct works only for small communities. In larger settings, the private organization managing the voluntary cooperation of citizens has to establish rules to enforce cooperation and avoid free riding. Soon, it erects itself as a new (local) government. In other words, society cannot develop without government, whether at the central or local level. We must cope with the inevitable inefficiencies it will involve in the decision process.

16.2.3 Regulatory Capture

Whereas "Public Choice" focuses on political bargaining and the distortions it induces on the whole economy, "Regulatory Capture" is an offshoot focusing on regulated industries i.e., natural monopolies (cf. §17.1).

Issue

At the high point of regulation in the US during the 60s, many economists start to feel that the SCP paradigm (cf. §1.2) fails to adequately describe and model many US markets, especially those regulated by federal laws. Stigler (1971) takes a strategic approach to regulated industries and introduces the Special Interest Group as an actor reciprocating with the government, the consumers and other (rival) industries. He concludes that State interventionism serves special interests rather than the public interest. A clear example of this are the agricultural support programs prevalent in industrialized countries.[19@] His policy recommendations are thus towards deregulation and "laissez-faire".

The capture theory is re-reading of the standard Marshallian supply–demand analysis using Olson (1965)'s theory of collective action to address group behavior. On the supply side, politicians who enact regulations and bureaucrats who implement them are responsive to pressure because politicians seek support in terms of vote or campaign money while bureaucrats seek cash or in-kind transfers or future job opportunities. On the demand side, an oligopolistic industry has either much to gain or to lose from regulation, hence its members are strongly motivated to unite into a cohesive industry association to lobby their regulator or the legislator. Whenever there is a small number of firms in the sector, the free rider problem characteristic of collective action is manageable and the association becomes a powerful lobby, likely to influence legislation (cf. §2.4.4). The other main actor on the demand side is the general public comprising all the consumers of regulated goods and services. Since an individual consumer weights little over the final outcome, he has very low incentives to get informed about policy debates and act to obtain a favorable regulation. The free rider problem is then maximal for consumers; hence, their global influence over regulation is likely to be weak.

These considerations lead to conclude that regulators are captured by those they are supposed to regulate. There are plenty of examples where regulation is captured during enactment or meanwhile it is applied but obviously never has it been created by an industry to protect itself from competition (as it was sometimes claimed).[20@] The mitigation of regulatory capture is taken on in §16.3.4.

To conclude, it must be observed that regulatory capture is not a recent perversion of the antitrust framework. Rather, as we show in the next section, it is the modern form of an age old agreement between the State and the industry to limit competition, create rents and share those. Rent-planning, as we call it, was one of the mercantilist policies reviled by Smith (1776).

16.2.4 Rent Planning

To better understand the functioning of markets, we often assume the perfect legal framework described in Chapter 8. The various rationalizations of the State seen up to now adhere to this conception and disagree only with respect to the ruling elite's objective: common good or private interests. Yet, by ruling out political struggles and social conflicts, these theories fail to properly account for the recent phenomena of nationalization, liberalization, antitrust and rent-seeking. To make sense of these issues treated fully in the oncoming sections 16.3 and 16.4, we follow Acemoglu (2010) (ch. 4) and look back at the violent roots of State formation.

The War Motive

According to Tilly (1990), war is prevalent throughout the history of human kind because elites, whether rulers, oligarchies or parliaments, display a strong attraction for prestige and empire building (much like CEOs of modern firms).[21@] Paramount to successful dominance is military innovation which has two distinct consequences. On the one hand, innovations make combat more decisive which in turn leads to the consolidation of small territories into larger states.[22@] On the other hand, war becomes ever more expensive, forcing rulers to seek revenue beyond the income of their personnel assets (mostly royal land).[23@] The two main channels for this task are *taxation* and what we choose to call *rent-planning*, namely the monetization of privileges through the cartelization of the economy and the subsequent sale of the afferent monopoly rights to private investors. For instance, it is well know that central banks were created to enhance the creditworthiness of states engaged in wars and allow them to issue more debt. What Broz (1998) shows is that the political support from the financial sector was won by cartelizing the sector.

Our novel and neutral terminology emphasizes that these revenue generating practices are carried on by all types of rulers from dictatorships to democratically elected parliaments, going through oligarchies.[24@] Rent-planning is thus not directly determined by the recourse to violence or the absence of democracy.[25@] Obviously, in each case, the design of taxation and rent-planning reflects the size and scope of the elite ruling the State. Yet, the fundamental purpose remains the same which is to raise finance in order to further the objectives of this governing elite.

As recalled in §16.1.2, taxation is the main financing instrument of advanced economies today, but this is a novel evolution.[26@] Until the advent of the modern bureaucracy, it is fair to say that the administration is inefficient and corrupt which leads to low ratios of income to payments. To compensate, rates are set at very high levels and numerous taxes are created, which in turn leads to resentment, evasion and revolt. Tax revenue is thus endogenously limited and rent-planning can thus be understood as an alternative avenue to raise income that visibly hurts no-one and therefore passes relatively unnoticed and unopposed. Rulers use rent-planning to bypass the limits to taxation set by their councils (ancestors of parliaments) whose agreement is necessary to change the tax code.[27@]

Planning Rent Extraction

To extract the producer surplus of an economic sector, the State creates a private monopoly or a cartel structure for the sector; it then auctions the afferent property rights to

an entrepreneur or a clique against some compensation that is either (explicit) cash or (implicit) political support.[28@] In this arrangement, the producing side uses its expertise and organization skills to efficiently exploit the envisioned economic monopoly and maximize profits. The State uses his mastering at coercion to enforce the exclusive property rights of its partner from encroachment by unauthorized competitors whether they are nationals, aliens or even a foreign power contesting the monopoly. Since the fee paid to the State is independent of the amount of economic activity, the entrepreneur is made a residual claimant of all profits he might generate. His incentives to exploit the market, cut cost and innovate are thus maximized, in so far as the State does not renege on the initial agreement (and demands extra compensation). The rent-planning scheme is quite similar to the fixed price contract used nowadays for the delegation or concession of a regulated activity (cf. §17.1.1 & §14.4) with the major difference that the scope of the contract is the entire economic sector, not just one unit or region or infrastructure.

There is an inherent hold-up problem since the coercive power of the State allows it to renege at any time and extracts ex-post rents from the firm (cf. §14.2). The solution at a time where the rule of law does not bind the ruler is to invite the entrepreneurs to form a clique (e.g., guild) and enter the elite; this way, their property rights are more credibly upheld. This process explains the entry of the bourgeoisie into the elite which initially includes only public services providers (defense, justice, police, religion). These commercial new comers are however quite different since they have a dual role, being cartel members for wealth distribution purposes but also productive agents within the economy. This duality is at the root of the democratization process because these cliques have long lived investments to protect and have thus the greatest interest to impose the rule of law (cf. model in §3.3.3).[29@]

The most famous facet of rent-planning is that relating to trade policies known as mercantilism,[30@] protectionism[31@] and colonialism.[32@] The general thrust is to protect an entire industry with import tariffs, export subsidies and citizenship restrictions (e.g., employment, transport, capital).[33@] Whereas mercantilism emphasizes national monopolies, protectionism vie for competition at the national level which is more conducive of efficiency and innovation.[34@] The free-trade movement associated with the globalization of commerce finally aims at removing all barriers to trade.

Inefficiency

Rent-planning generates several forms of inefficiency that were clearly enunciated by Smith (1776). Firstly, cartelization leads to monopoly pricing, an allocative inefficiency (cf. §3.2). Nowadays, the reverse problem of excessive output occurs in the public services because price subsidization is used to buy political support from the masses. These

diametrically opposed price distortions send equally wrong signals for the use of factors among industries which is a source of productive inefficiency (cf. §2.1.2) . A second source of waste recalled by Baumol (1990) is that entrepreneurship and innovation are hampered by the State's systematic appropriation of all new sources of rent; this is an instance of dynamic inefficiency.[35@] This encroachment also drives part of the economic activity towards the black-market where property rights are seldom enforceable, thereby hampering further innovation.[36@]

Last but not least, rent-planning is a source of long term institutional inefficiency because of the inherent dichotomy between innovators and administrators. Our claim is rooted in the observation that inventive entrepreneurs are new comers without link to the current elite. This is best understood by looking at their offsprings' occupational choices. Entrepreneurs are risk lovers because being penniless, they have nothing to lose. The offspring of successful entrepreneurs who have entered the elite are on the contrary risk averse because they are born within the elite. It is thus rational for them to pick an administrative or political job rather than continuing the risky business of their fathers. The ruling elite thus displays a dynastic character.

Acemoglu (2008) explain the economic consequences of this dynamic as follows: when successful entrepreneurs seize political power, they pursue at first policies favorable to their trade which is efficient. Now, when a new technological wave starts, the relative prices of inputs and outputs change and this calls for a reorganization of economic activity. The ruling elite, only mastering the old technology, stands to lose in front of new comers who master the new arts; the elite then rationally blocks the necessary institutional changes, thanks to its dominion of political power.[37@] Growth, which depends so much on innovation, is thus hampered by institutional inertia. The same basic observation for industry sectors leads Schumpeter (1942) to enunciate his theory of creative destruction. As a matter of conclusion, one cannot fail to notice that countries were rent-planning is paramount have failed to introduce anything useful to the human kind over the last century.[38@]

Legacy

Section 16.3 on Rent-Seeking studies influence (lobbying) and corruption that can be seen as agreements between a clique (aka Special Interest Group) and the State represented by either the legislative or judicial bodies, the central or local government or a bureaucrat. This concept is the natural evolution of Rent-Planning within advanced industrialized democratic countries.[39@] There are two noticeable differences. Firstly, rent-seeking as the name indicates is at the initiative of the rent-seeker and secondly, protection is not based on violence (or menace) but on an abuse of the legal system to

create (and enforce) undue property rights (but not necessarily unlawful).

Section 16.4 on liberalization looks at nationalization, privatization and deregulation. Nationalization (cf. §16.4.2) involves State planning but without the rent generation motive. Historically, it is prevalent in crisis times when the government is forced to bailed out large firms to avoid further damage to the economy (e.g., banks, railways, airlines). Otherwise it follows ideology and, above all, the desire to control the economy but not to extract rents as in our original argument. The recent wave of privatization (cf. §16.4.3) is a modern case of rent-planning. Indeed, empirical studies have shown that the major motive was to refloat State finances rather than to improve the competitiveness of markets (at least not in the short term).[40@] Lastly, deregulation (cf. §16.4.4) aims at eliminating rent-planning practices that have survived in most regulated activities among which natural monopolies.

16.3 Rent-seeking

The topic of this section is best illustrated by Pareto (1906): "A protectionist measure provides large benefits to a small number of people, and causes a very great number of consumers a slight loss. This circumstance makes it easier to put it into practice."

Advanced economies have more or less succeeded to eliminate violence and most illegal behaviors but this does not mean that all human economic activity is geared toward production. The presence of the State as a protector and a potential oppressor gives rise to a host of wasteful conducts such as corruption, red-tape, log-rolling, lobbying or cronyism that we cluster under the name of rent-seeking. The strategic analysis is undertaken in §7 on contest and conflict. We shall look here at the roots of the phenomenon, distinguish between corruption and lobbying, reflect on the inefficiencies generated by rent-seekers before looking at some remedies.

16.3.1 Tipology

Non productive behavior can be geared at the State or at competitors and goes under the respective names of *rent-seeking* and *profit-seeking*. Profit-seeking through legal means is analyzed in §7 on contest and conflict while deceptive competition is thoroughly studied in Part D on antitrust. For instance, entry deterrence is a legal defense where the incumbent invests excessively into production capacities, advertising or accumulation of intellectual property rights (IPR) such as patents or trademarks (cf. §12) to discourage entry on its protected market. Predation, foreclosure and restrictive agreements are anti-competitive (illegal) defensive strategies. Collusion is the artificial creation of a

monopoly to increase industry profits.

Most rent-seeking originates from the State decisions, voluntarily or not. Any regulation imposed for general quality purposes (e.g., security, hygiene or standards) necessarily bestows an advantage upon the firms that have passed the regulatory control as opposed to those who haven't (yet). Being in limited number, often a monopoly, license holders are shielded from standard competition and thus earn economic rents. This circumstance makes the license valuable for an incumbent who fears losing it as well as for a potential entrant eager to grab a share of the cake. Incumbents, as a group, also vie to block entry by asking the State for a Numerus Clausus policy to maintain per capita profitability.[41@] Challengers, on the other hand, seek the removal of entry limitations to become new incumbents.

Likewise, wealth redistribution programs increase the recipients' utility (at the expense of the entire economy). The insiders enjoying the situation fear to lose while outsiders would like to come on board. In the same class of interactions, we find import tariffs and quotas for international trade, minimum and maximum prices for market control, taxes and subsidies for specific groups, be it an industry, a geographic area or a demographic class. One feature that makes rent-seeking a lasting entrenched phenomenon is loss aversion, a deep trait of human nature whereby one is ready to suffer much more to defend a possession than to conquer a new one.[42@] When a group benefits from a privilege or subsidy for some period of time, it starts to take it for granted. Once this belief is firmly held, the group will expand no few resources to defend it so that the State often prefers to ratify it as a right which then become inexpugnable.

Natural monopolies are more subtle to interpret. If the regulation is ideally administered, it leaves no rent to the incumbent. The insiders benefiting from the regulation are then the consumers (who will insist on low prices) while the outsiders are the current and potential providers of the service who crave for a softening the regulation (a price raise). If the regulation is laxly applied, the roles are reversed: the incumbent seeks the status-quo, consumers seek a price freeze and contenders seek a change of provider. Lastly, a public monopoly (e.g., municipal agency) is vulnerable to internal rent-seeking because the bureaucrats in charge may distort the objectives of public service towards their own; it is however invulnerable to external rent-seeking in the short run, but might become a target if the government decides to privatize the service.

In all cases, there are at least two contenders looking for an actual or potential rent (possibly with different personal valuations) that is decided upon by a State representative, most often a committee. Hence, the study of rent-seeking as a strategic interaction in §7.2 requires to treat voters, politicians and bureaucrats as rational decision makers much like firms and consumer groups (cf. public choice theory in §16.2.2).

16.3.2 Corruption and Lobbying

Just like market competition relies on price and non-price strategies, rent-seeking displays a similarly dichotomy with corruption and lobby. As we shall see, the former is a more efficient channel but as it is often prohibited or costly to use, the second one has gained prominence.

Corruption

Svensson (2005) defines it as any abuse of entrusted power for private gain (cf. Transparency International). One speaks of *bribing* when it involves a cash transfer. The recipient may be a bureaucrat, a regulator, a legislator, a manager (within a firm or any organization). The briber is any stakeholder in the decision process such as a firm in a procurement contest or in a regulatory evaluation, a lower level manager in labor tournament or anyone involved in a legal issue. Bribing is seen as disloyal competition under most legal systems and so is declared illegal per-se. This obviously limits the ability to use that instrument in environments where the law prevails.

Corruption thus turns to alternative (and more costly) channels to transfer utility towards the targeted decision makers such as in-kind gifts (e.g., lunch, travel or club membership) or legal spending in activities (e.g., charitable foundation) that bestow prestige onto the prey. Finally, the promise to give the bureaucrat or politician a lucrative job in the future can be used to transfer wealth inter-temporally. Many countries try to limit the effectiveness of this artifice by imposing to its civil servant (and alike) a waiting period before one is able to work in the industry he previously regulated. It is often quite easy to go around the limitation (e.g., being employed by a subsidiary of the company active in an altogether different sector of the economy).

In the cases that matter for industrial organization such as a telephone license or an import quota, the stakes are very large in comparison with the maximum bribe that can be realistically transferred. Indeed, bureaucrats (or anyone with a public profile) in advanced economies are neither poor nor rich, thus they can't change their lifestyle without being noticed and get into legal trouble. This puts an upper limit on the graft money they can accept whatever channel it comes through. To conclude, we offer Frye and Shleifer (1997)'s typology of corruption in governments.

Issue / Typology	*Invisible*	*Helping*	*Grabbing*
Government	under law	above law	above law
State activism	public goods	help business	extract rents
Contracts enforcement	Courts	benevolent bureaucracy	predatory bureaucracy
Regulation type	minimal equitable	targeted sectors	predatory aleatory
Alternate institutions	private litigation	lobbies guilds	mafia local barons

Table 16.3: Levels of Corruption

Lobbying

In advanced economies where the legal system is strong enough to effectively limit corruption, the ubiquitous behavior to seek rent is *lobbying* whereby cohesive minorities vie to *influence* decision makers by pressuring them. The "friendly" pressure is the exposition of one's arguments in favor of a particular course of action. For instance, a regulated monopoly asks for a price increase arguing that inputs have become more costly while the consumer group asks for a price freeze arguing that the incumbent makes fat profits. A potential entrant asks for a license to operate in a "numerus clausus" market such as taxicab or hertzian TV arguing that this will increase the supply quantitatively and qualitatively. Conversely, the association of architects asks for a limit on the number of degrees delivered by universities to maintain the highest quality (and avoid cut-throat competition).

The complementary channel of influence for lobbying, as in "carrot and stick", is hostile pressure such as (unwarranted) judicial harassment, public protests, strikes or negative media exposure (smear campaigns). This method lies at the frontier of legality since it often recurs to coercion and violence (cf. §7.3.1 for a model of such behavior).

The Lobby sector is said to employ some 15000 people in Brussels where most European institutions sit and twice as many in Washington DC in the US. The consensus based approach to EU policy-making and lobbying is often contrasted with the professional and more aggressive US lobbying style. Reasons for these differences are to be found in the well established tradition of lobbying in the US that has attracted all society stake-holders; we find at the same time professional associations defending the strengthening of copyright laws and those defending the strengthening of innovation. The relative smallness of the European institutions in comparison with the US federal equivalents and the newness of lobbying in Europe makes it a preferred target of industries representatives whose inputs are welcomed by authorities to frame new policies.[43@]

16.3.3 Inefficiency

It is well known that when rational agents (firms, consumers, countries) struggle over a given prize, they expand resources to improve their bargaining position or fighting ability. If they could agree ex-ante to an exogenous redistribution, they would save the aforementioned resources. By failing to settle their dispute and engaging in actual conflict, they generate a welfare loss.[44@] Tullock (1967), Krueger (1974) and Posner (1975) start to inquire this issue theoretically and empirically for publicly awarded rents (e.g., natural monopolies, protectionism, licensed occupations).

These authors argue that if there is perfect competition in the rent-seeking activity (including free entry) then average benefit equals average cost in the long run i.e., the entire benefit, the rent, is equal to the entire cost, the resources expanded to secure the rent. In that extreme situation, the rent is fully dissipated but this does not mean that it is entirely lost in terms of welfare. Empirical studies find welfare losses amounting to a few percentage point of GNP. Though not as large as what the initial literature lead to believe, this loss is however far greater than the traditional dead-weight loss due to the exercise of market power. As we show analytically in §7.2.1, many elements concur to reduce the rent dissipation from such a high level. Let us review the arguments exposed by Fisher (1985).

When, as in most cases, the license is to be renewed, the incumbent holds a considerable advantage. He knows better the demand and the cost, thus faces less risk which makes him a stronger rent-seeker. Furthermore, he holds up the regulator so that renewal is almost automatic (cf. §17.3.3) These features imply that contenders won't spend much effort to secure the monopoly since they mostly stand to lose; rent dissipation is then limited. By reversal of the previous arguments, we can expect the greatest amount of lobbying for the awarding of a new monopoly license. Yet a countervailing effect is the uncertainty of the rent; rent-seekers will tune their effort to the certainty equivalent of the rent which amounts to subtract a risk premium from the expected rent, hence the waste is once again reduced.

With respect to private non regulated quasi monopolies (e.g., Microsoft), it must also be noted that many monopolies are obtained through luck rather than foresight i.e., as a consequence of technological changes or the evolution of customs. If the rent holder is a lottery winner, then the optimal amount of rent-seeking effort is zero and there is no social waste. The learning curve is also a natural barrier to entry that protects some of the monopoly rent. Lastly, one can draw an analogy with traditional rent to express the idea that a monopoly rent may not be contestable and hence will not be socially wasted (cf. §6.1.7). Under perfect competition, it is possible for a firm to earn a rent if she owns an asset of value such as management talent or good location. This does not contradict

the fact that in the long run, her average cost is equal to the market price because the rent derived from the asset is incorporated into the firm's cost as an opportunity cost. The same phenomenon occurs for an incumbent monopoly for whom the valuable asset is the set of barriers to entry that he has come to enjoy (independently of how these barriers were erected).

16.3.4 Remedies

Since rent-seeking is created by the State in the first place, one simple way to eradicate this wasteful activity would be to scale down the State but this is often impossible because the rent appears as a by-product of the State's primary missions (cf. §16.1.1). However, in some sectors it has been possible to "deregulate" and insulate the regulatory agencies from pressure groups (whoever they are). The other remedy is to turn the spontaneous and wasteful rent-seeking game into an ordered welfare enhancing contest (cf. §7) that will select the most efficient firm and regulate it in a second-best fashion (cf. §17.2). The mechanism most commonly used for that task is an auction because it has the added feature of generating public income (cf. §22).[45@] Auctions are also used to sell occupational licenses for professions such as taxi-cab or architects. Under an auction, the identity of the winner does not depend on lobbying or bribing but solely on the amount pledged, hence money is best used in the bid and there is no wasteful rent-seeking.[46@] A licensee has also the right incentives afterwards to run his business efficiently (cost cutting, innovation) since the scheme is formally equivalent to a fixed price regulation. For those public services that come in limited supply and who cannot be auctioned due to equity considerations, the sole "rent-proof" allocation mechanism, that is also fair, is random allotment (e.g., kindergarten or social housing). It however suffers either from transaction cost in the re-trade market or from irremediable inefficiency when re-trade is prohibited (cf. §22.1.2).

The case of industries or sectors displaying scale economies (e.g., natural monopoly) calls for a different treatment (cf. §17.2).) For the numerous cases where *competition in the market* is likely to result in monopolization (or cartelization) and direct governmental regulation seems unsatisfactory, Demsetz (1968) proposes[47@] to switch to *competition for the market* i.e., franchise bidding. Initially, each contender is willing to bid up to the full monopoly profit for a license to exploit the market at will. This, as we know, is not welfare maximizing given that output will be restricted. The (constrained efficient) solution is to require a minimum level of quality and let candidates bid for the lowest customer price. Obviously, the regulation must provide with high penalties for failure to meet quality objectives and the provision of enough collateral by the firm to avoid ex-post non

compliance (cf. §13.3.3 on the hold-up).[48@] If the best technology is available to all firms in the industry, the winner earns no rent since the Bertrand type competition inherent in auctions drives the customer price down to the average cost. The license holder keeps the right incentives to cut cost and innovate as before.

Goldberg (1976) shows franchise bidding still suffers from capture as it relies on "cheap talk". Indeed, the awarding being based on a comparison of tentative figures, it is tempting to bid aggressively in order to win and then force a renegotiation arguing (falsely) for some "force majeure" event that has unexpectedly raised costs. The regulator faced with this claim would reduce welfare if he was to dismiss the firm and run the scheme a second time; on top of this, it would be an admission of its own failure which is the worst possible outcome for the agency's employees. The auction for the franchises of water distribution in French cities admired by Chadwick (1859) is a fine example of an efficient regulation gone captured by its clients. The historical providers have grown into three large utilities that currently share the national market on a geographical basis; the situation is quite similar to a cartel. According to a 2001 report by the national assembly, franchises are not auctioned anymore and are the object of collusive agreements between firms and municipalities to unduly tax citizens (city expenses are channeled into the water bill).[49@]

16.4 Liberalization

Chapter 17 is entirely devoted to the regulation of natural monopolies and other industries deemed crucial by the State. In the present section, we relate the liberalization process of the last decades. The term is generic and clusters two approaches geared at the specificities of the European and US traditions, namely *privatization*[50@] and *deregulation*.

Liberalization tries to establish a light handed regulation over sectors where perfect competition is not viable overall but where some segments can be made competitive while other, with natural monopoly features, remain regulated. As argued early on by Posner (1969), effective challenge but also potential competition is what prohibits a monopoly from abusing his dominant position. This is why, in theory, liberalization ought primarily to foster entry by reducing barriers (cf. §6.1.7) and, secondarily, improve direct regulation with incentives (cf. §17.3).

Before studying the liberalization trends, it is worthwhile to take a look at the past and see when and why the State decided to take an active or passive role in the productive sectors of the economy, either by delegating tasks or by controlling them closely.

16.4.1 Grants and Concessions

To raise revenue, the State sells specific rights or or delegates special tasks to private individuals or local governments, thus taking a passive role towards economic activity. Historical examples abound where the ruler (or parliament) uses his privileges to issue a letters patent granting (or leasing) an item such as a piece of land, an activity to be performed over royal land (e.g., fishing, hunting), an office (e.g., notary), a title, a pension, a patent for an invention, a right to hold market or fair, an exclusive right to some activity over some territory (e.g., mill wheat, work wool, operate a lottery), a pardon, a sinécure or a judicial commission.[51@] Another important legal act is the municipal charter awarded (i.e., sold) to a city in order to free its inhabitants from military service and acquire monopoly rights over some trade or industry (which still weights on the local economy centuries later). Even taxation can be delegated (outsourced) for a yearly fee (cf. §16.2.4).[52@] In some cases, the tax base (e.g., city, county) is rich enough to buy out the perpetual right, turning the transaction into a privatization.

The provision of public services and goods is also delegated to the private sector, especially when it involves large investments or requires specialized knowledge. The instrument is called concession whose latin meaning is "put in place of" (the prince). Since the rent extraction motive is less present in those cases, financing through voluntarily agreed taxation is generally observed. As recalled Bezançon (2005),[53@] state delegation through concession develops in stages throughout history. In Roman times, delegation pertains to public works such as thermal baths, market places, roads, ports or aqueducts. Private parties also run services of public interest such as the post office, tax collection and money minting. Concessions for public services and infrastructures, often proposed by private parties, abound. From the XVth til the XVIIth centuries, new delegations include post coach, stage coach, marsh draining, mining, canals, roads and bridges. In the XVIIIth century, waste collection, street pavement and even city districts construction or renovation are offered in concession. Contracts are refined and include a 10 years maintenance obligation for contractors. In the next century appear concessions for water distribution, street lightning, railways, gas distribution, streetcars, telephone, electricity, subway and airways. This topic is further elaborated in §17.1 on the natural monopoly.

Governments have now at their disposal a large spectrum of options for ownership and operation of public utilities. Table 16.4 presents the World Bank's glossary on public private initiatives (PPI).

Public Enterprise In-house production

Outsourcing Short term contract (< 3 years) involving no commitment for the firm and no or

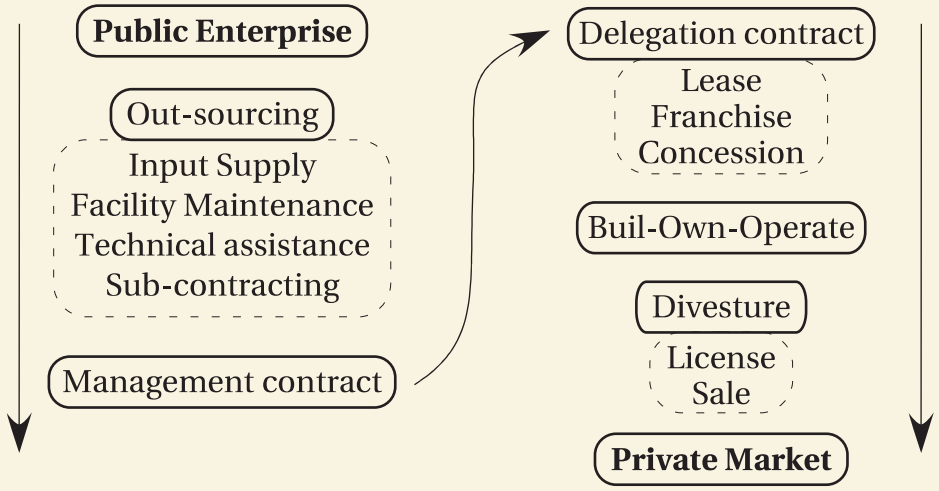


Table 16.4: Spectrum of Concession and Franchise

low risk.

• Input supply: procurement of raw materials

• Facility Maintenance: take advantage of economies of scope and scale to maintain building and other standard assets.

• Technical assistance: take advantage of the specialization of consulting firms to develop IT or management systems.

• Sub-contracting: take advantage of economies of scope and scale to delegate a standardized task such as billing to an outsider.

Management contract Short term without strategic risk (insulation from politics) but some commercial (demand related) risk.

***Delegation contract* (concession)** The government sets the duration, rights, obligations and tariff, the latter defining implicitly the degree of risk for the firm. Asset ownership remains public but the firm maintains them.

• Lease: term ≃ 5 years, no investment, commercial risk and freedom of action. Revenue is either fixed or depends on performance.

• Franchise: term ≃ 15 years, new investments, commercial risk and limited freedom of action. The franchise is generally awarded to the highest bid which can be a once-for-all fee (bonus),[54@] a royalty (percentage of future revenues) above a minimum threshold set by the franchiser or a share of future profits (cost need be audited).

• Concession: term ≃ 30 years, large investment in either greenfield (aka Build-Own-Transfer contract) or rehabilitation.

Build-Own-Operate is an example of Public Private Partnership (PPP). A design-build-maintain (DBM) with or without operation creates a lifecycle responsibility and thus provides an incentive to deliver better quality in the initial design and construction of the project because the firm will have to bear any additional maintenance and repair costs if the initial quality

is inadequate.

Divesture The government sells to the private sector the right to provide the service.

- License: limits entry for activities with network externalities. Heavy handed regulation remains.
- Sale (privatization): the light handed regulation and freedom of entry are made possible by technological developments that have eliminated the historic element of natural monopoly.

Private market The service is provided by unregulated firms freely competing.

16.4.2 Nationalization and Municipalization

In this section, the State takes an active role in the economy. Nationalization occurs when the State buys a private firm and turns it into a State Owned Enterprise (SOE).[55@] This relatively recent (about a century old) phenomenon is prevalent in crisis times when the State is forced to bailed out large firms to avoid further damage to the economy (e.g., railways and airlines). Otherwise there is an ideologic motive to control the economy and steer its development.

The direct involvement of the State with productive businesses is linked to the industrial revolution. New services like sanitation, light, heat or communications require heavy network investments to cover large geographical areas (e.g., an entire country). With the exception of the US and UK, national financial markets are underdeveloped and unable to sustain the required effort so that public finance and thus ownership comes into the picture. At the city level, the desire to avoid anarchical duplication of networks, to guarantee a high level of quality and to eliminate corruption, makes direct municipal oversight unavoidable. Either a municipal agency is created or early private firms are bought out (municipalization). Finally, in heavies industries such as mining, chemicals or petroleum, the choice made by the State of building SOEs, either directly or by purchasing private firms responds to a desire for modernization and growth promotion.

Political ideology also plays a significant role for the nationalization push during the last century. Most of the activities under concession during the liberal era are either nationalized or municipalized by leftist governments as a mean to redistribute wealth from the capitalist owners, the bourgeoisie, towards the less affluent classes of society. Likewise, authoritarian governments (e.g., soviet and fascist) being ideologically prone to direct control, nationalize large parts of the economy to advance their particular goals.[56@] The result is a State owned economy in soviet countries and a State owned conglomerate of industrial firms in fascist countries like Italy or Spain.

In the aftermath of the second world war, most Western European countries (even

the UK) nationalize large parts of their economies to accelerate reconstruction, apply long-term planning and remain master of their economic development (cf. the national interest concept). The last major episode of nationalization in advanced economies takes place in France in 1982 but is overturned from 1986 on. Lately, some latin american countries have started to (re)nationalize industries that were privatized in the 1990's arguing they failed to deliver on their promises.

Nationalization is unknown in the US where the private ownership of assets has never been called into question. Apart from the Postal Service or the Tennessee Valley Authority (power plants), there are hardly any federally owned firms. However, a quarter of municipalities own electric distribution companies and most public transit systems (tubes and buses) are municipally owned.

16.4.3 Privatization

The first modern privatization is Germany's carmaker Volkswagen in 1961, followed by the power and mining company VEBA in 1965 (now E.ON).[57@] The first wave of privatizations was launched in Chile between 1974 and 1978; two hundred SOEs were sold to private investors that later united them into six large conglomerates (cf. Yotopoulos (1989)).

However, the big push for privatization in the advanced economies is the worsening of public finances starting in the 1970s.[58@] Concomitantly, SOEs suffering from underinvestment and over-employment make losses; privatization thus appears as a remedy for two ills: improving the efficiency of these firms and refloat the treasury. The public finance motive that was at the origin of concessions is again at the heart of many privatizations and the fashionable use of concessions today. Large scale privatization is associated with the Thatcher UK government of the 1980s (and 1990s under John Major). The popular success of the British Telecom share issue privatization (SIP) in 1984 can be seen as the starting point of a worldwide wave of privatizations. In the UK, SOEs amounted to 10% of GDP and to 8% of employment in 1979; these figures were cut by privatization down to 5% and 3% in 1992 and virtually zero in 1998. In France, SOEs reached a maximum of 9% of the work force in 1985 but have since fallen down to 3% in 2004.

The goals of privatizations are numerous: raise revenue for the State, promote wider share ownership among the public, promote economic efficiency and competitiveness. The later objective is supposedly achieved by reducing government interference in the economy and by putting privatized firms under market discipline. All privatizations of natural monopolies go hand in hand with the enactment of a law establishing the

regulatory framework for the sector; it applies to the newly privatized firm but also to challengers whose entry is to be facilitated.[59@] A positive byproduct of privatizations has been the development of financial markets since large amount of money were turned into equity. Figure 16.1 and Table 16.5 taken from the Privatization Barometer display the trend of privatization in Europe over the last two decades (PS: private sale, PO: public offering).

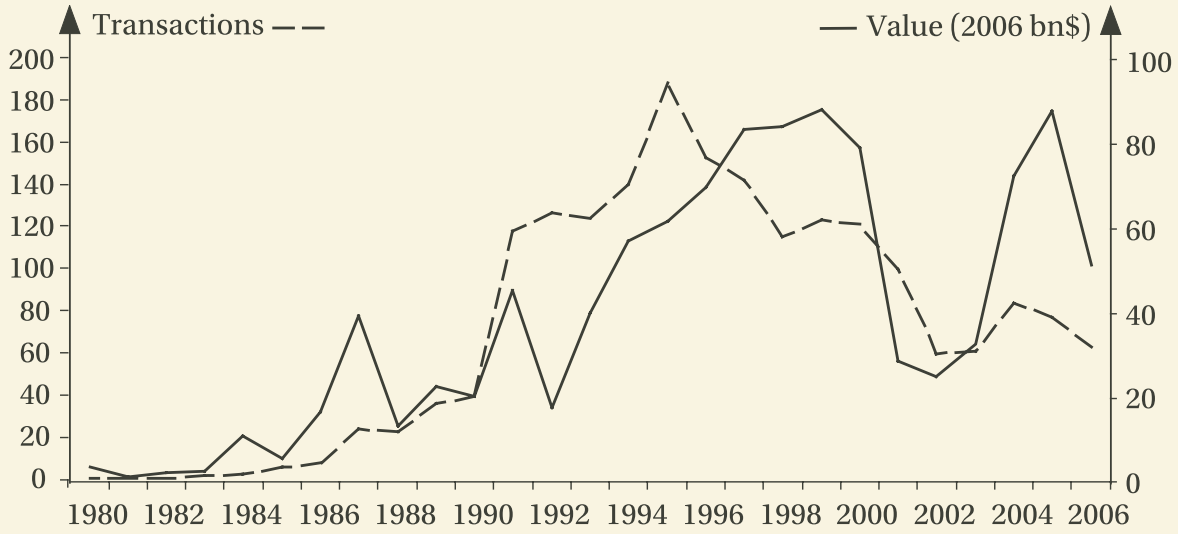


Figure 16.1: Privatization trends in Europe

class	*Transactions*	*Mean (m$)*	*Total (bn$)*
EU-15	1216	618	751
new members	752	104	78
EU-25	1968	421	829
Telecom	130	1,728	225
Oil	79	953	75
Utilities	254	671	171
Finance	254	520	132
Transport	203	461	94
Manufacture	696	145	101
Services	150	120	18

Table 16.5: Privatized European Industries

According to Vickers and Yarrow (1988), the first wave of UK privatization is politically motivated; the underpricing of shares aims at attracting votes from median income households. Tight regulation guaranteeing low prices and high quality only came with the second wave (after 1987). On the positive side, Megginson and Netter (2001) report a general empirical agreement over the fact that private firms are productively more effi-

cient than SOEs. On the negative side, their analysis suggest that governments balance competing economic, political, and financial objectives when privatizing a SOE, so that economic efficiency is only marginally enhanced. Most privatizations appear to be the outcome of a rational alliance between management and government to create a pure private monopoly out of a public monopoly. The structure of the firm is left untouched i.e., there is no separation of activities or dismantling; if necessary, the accounts are sanitized before presenting the firm to international markets.[60@] Lastly, the statutory protections that shielded the public firm from competition are more or less maintained in the privatization act on grounds of maintaining productive efficiency (the natural monopoly property of technology). All of this means that the private firm will enjoy rents in the future, so that its shares are highly valuable which translates into a large revenue for the State. To avoid internal opposition from unions, employees get shares at a preferential price. For the management, the outcome means more freedom and less supervision by the regulator (e.g., treasury or public utility commission) and also avoids falling into the iron hand of a majority shareholder since the shares are either sold to the public at large or in small batches to institutional investors. This way, the board can vote itself large salary increases, generous stock options and other assorted perks. This analysis of the recent privatization process conforms with our view of the State as a rent planner (cf. §16.2.4).

It seems that competition (and its benefits) has developed in those sectors where a European directive has forced the opening of national markets, turning all the quasi monopolies (so called national champions) into oligopolistic contenders inside a European league. This pro-entry stance, advocated by Posner (1969), is also at the root of the US deregulations (cf. Megginson (2007) for a recent overview).

16.4.4 Deregulation

The term refers mostly to the US whose public service providers are mostly private but were tightly regulated; deregulation then means introducing market competition in segments where technological and demand characteristics makes it worthwhile and change the regulation of natural monopoly segments towards a lighter mode giving better incentives to cut cost and improve quality of service. The US deregulation started during the 1970s under the combined influences of macro-economic shocks and the academic literature of the Chicago school of thought. In 1977, 17% of US GDP was produced by tightly regulated firms, a figure that was reduced to 6% after twelve years of intense deregulation in the areas of telecommunications, financial services, airlines, natural gas, oil, road shipping, railroads, banking or cable TV.

In Europe and elsewhere, regulation was designed from scratch for the formerly SOEs that were privatized. Most public services necessitate an infrastructure (e.g., network) which is essential in the sense of having the natural monopoly property. The EC has thus decided to maintain regulation on that segment only. The upstream and downstream segment, most importantly retail, are open to competition; the infrastructure owner is forced to give access to competing retailers on terms of cost and not lost profits. This principle has been applied for the liberalization of telecommunications, energy, transport and postal services.

One clear example of successful deregulation is air transport where for decades, national airlines had a monopoly over domestic routes and shared in a cartel like manner international routes through the IATA association; cabotage (right to operate within the domestic borders of another country) was prohibited, fares were high and many airports in regional capitals were underused. Deregulation started in the US[61@] with the Airline Deregulation Act and the International Air Transportation Competition Act (1979); Europe followed suit in 1997. Data from the US Bureau of Transportation Statistics show that the legacy airlines operating extensive national and international routes enjoyed in 1990 an 80% market share over domestic US routes which came down to 50% in 2006. Figure 16.2 based on data from Eurocontrol show a similar trend for Europe.

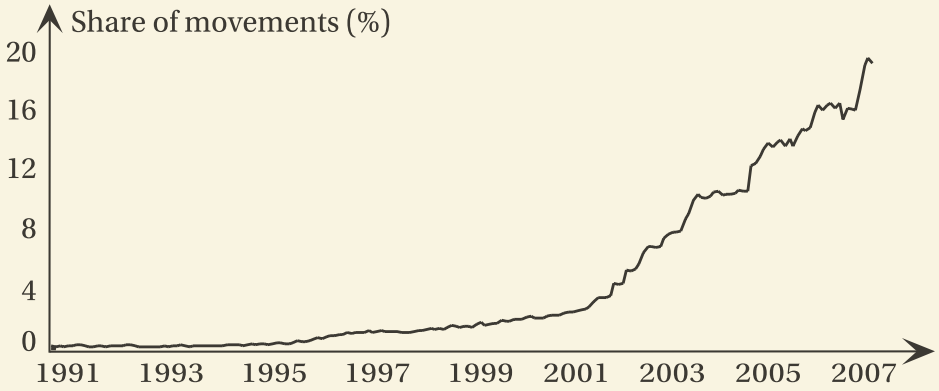


Figure 16.2: Market Share of Low-Cost Airlines in Europe

According to Morrison and Winston (2000)'s data on US airline prices regarding the 1978-98 period, deregulation brought a 25% fare reduction with respect to a regulated price that would take into account productivity increases. Since differential pricing has been permitted, segments with inelastic demand pays more than under regulation but overall, 80% of the passengers enjoyed fare reductions. Figure 16.3 illustrates this finding.

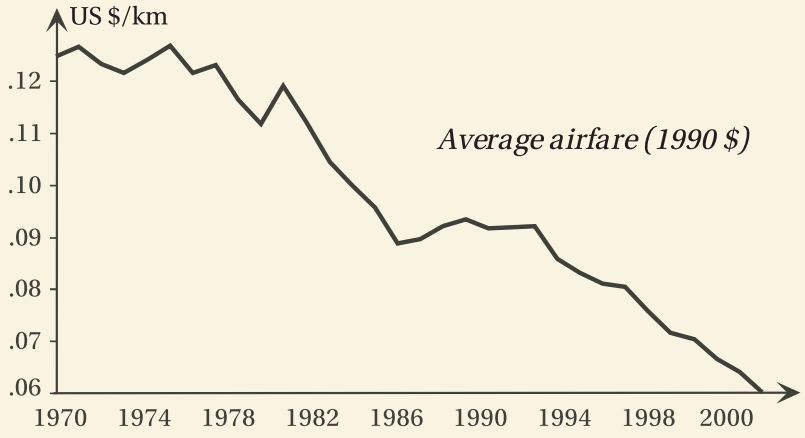


Figure 16.3: Average airfare per kilometer

Chapter 17

Regulation

Etymologically, to "govern" is to rule a country but it also means to steer a boat. This second meaning makes clear that the ruler is not supposed to row or, in the language of economics, that the government ought not participate directly in the economy. Regulation then, is that branch of public action that specifically aims to pilot the economy towards the fulfillment of social objectives. The previous chapter, dedicated to the State, has explained more thoroughly why there are social objectives and what is their nature.

In the present chapter, we study the regulation of network industries and natural monopolies from an analytical point of view. In the first section, we explain why some activities are regulated. We then analyze the objective and constraints for an ideal regulation; we thus obtain an ideal *level* of revenues for the regulated firm and an ideal price *structure* for the goods and services it produces. In the third section, we study how practical regulation evolved from cost based norms to incentive schemes and what constraints limit their effectiveness.

17.1 Why Regulate

17.1.1 Context

Transport (roads, canals, trains, airlines), telecommunications (radio, telephone, TV, internet) and utilities (electricity, gas, water, sewage) are *natural monopolies* i.e., industries characterized by very high sunk costs (cf. §2.1.3), relatively low marginal costs and increasing returns to scale (IRS) due to network externalities (cf. §25). For these sectors but also many other capital intensive activities typical of the modern era (e.g., chemicals, metals, oil, cars), the average cost of a firm is decreasing with total production (over a large range of output). As we recall in §2.1.3, cost minimization requires the pooling of production which is typically achieved by the integration of independent firms into a small number of large players. It is therefore no surprise that the development of these

activities at the end of the XIXth century lead quickly to local (in a geographic sense) or even national monopolies in all industrializing countries. Nowadays, these sectors are oligopolies i.e., competition, if any, takes place among few firms.

The economic issue at stake with this new market structure characterized by "competition among the few" is market power, the low degree of competitiveness and the allocative inefficiency it may generate. Recall indeed from the theory of monopoly (cf. §3.2) and oligopoly (cf. §5.1.3), that the fewer the competitors, the higher the prices and the lower the welfare. The dilemma posed by natural monopolies is that *productive efficiency* is achieved with a small number of firms (possibly one) while *allocative efficiency* calls for many contenders. We now recall how and why the State became closely involved in the management and regulation of natural monopolies.

Genesis of "Natural Monopoly"

Mosca (2008) recalls how the concept gradually evolved until its current day acceptation. Smith (1776) speaks upon a supra-competitive rent for high quality land citing "some vineyards in France (Bordeaux) of a peculiarly happy soil and situation" but does not relate this fact to the idea of monopoly. Malthus (1815), writing specifically on rent, introduces the dichotomy between *natural* and *artificial* monopolies i.e., arising either from nature's gift or from the ruler's benevolence (cf. §16.2.4 & §3.1.1). Like all economists of his time, he loathes the second and lauds the first, thus acquiescing to the extraordinary rents they produce (cf. §8.2.4).

Mill (1848) extends the definition further to include scarce skilled labour but more importantly adds a new category, the "practical monopoly" as arising from superior technology, capital requirements or cartelization. The first case, concretely canals and networks for water, gas and rail, is what is now understand as a *natural monopoly*. His stance is normative when stating "how great an economy of labour would be obtained if London were supplied by a single gas or water company instead of the existing plurality". Without being explicit, he seems to fear an abusive monopoly and thus calls for public oversight but oscillates between concession (cf. §16.4.1) and municipalization (cf. §16.4.2) without much of an explanation.

Cournot (1838) is the first author to take a descriptive approach to natural monopoly; he concludes that (de-facto) monopoly is the only stable market structure, because she enjoys economies of scale over the entire range of market demand (cf. §10.1 on entry barriers).[1@] Dupuit (1852) also believes this outcome ineluctable[2@] for railways and ascribes it to scale economies, capital requirement and incumbent advantage (better geographical sitting of the network) which prohibits an entrant from recouping his sunk cost. As recalled by Numa (2011), Dupuit advocates regulation as a second-best policy since he

believes the government failure (bureaucratic cost) to be a lesser evil than the market failure (monopoly pricing). Walras (1875) first uses the vocable "natural monopoly" in its modern sense and, more importantly, explicitly treats railways as a public service, thus calling for State management (cf. §16.1.1).[3@] Observing that rights-of-way for building networks can only be granted to a few firms, he concludes that these end up merging or colluding, thereby producing a monopoly.[4@]

Edgeworth (1913), without using the actual vocabulary, offers the first analytical treatment of the natural monopoly. Sraffa (1926), criticizing the logical foundations of Marshall (1890)'s theory of the firm, argues that increasing returns to scale are incompatible with perfect competition (so that oligopoly or monopoly is likely to appear). Oddly enough, the ensuing "cost controversy" remained purely theoretical and never mentioned natural monopoly as the archetypal case.

Rationalization

A definition devoid of ambiguities for "natural monopoly" would be "an industry where the socially optimal number of firms is one"; obviously it is not very instructive. We shall thus review how different schools of thought approach the natural monopoly and explain the ubiquitous presence of regulation.

Evil Monopoly ? The *public interest* theory of the State (cf. §16.2) has adopted the historical narrative and holds that an unregulated natural monopoly evolves from cut-throat competition to monopoly. Since both market structures are inefficient, the former because of price instability and looming financial crisis, the latter because of monopoly pricing, the case for intervention is clear in order to stabilize the sector (allow sustained growth) and avoid its monopolization.

Regarding the natural tendency towards monopoly, most economists agree that monopoly is inevitable but it is not per-se condemnable, only its abuse is (cf. §8.2.4). For instance, monopoly rents over land or minerals are hardly ever opposed on moral grounds as they participate to an efficient management of a scarce resources (cf. depletion in §18.1.2). Posner (1969) argues that rents arising from advanced technology (e.g., iPod) or commercial wit (e.g., Microsoft Windows) should receive the same treatment. Recall that if society aims at fostering innovation and new wealth creation, it must promise supra-competitive profits to lure entrepreneurs into such risky activities. Monopoly achieved through business excellence is thus commendable rather than reprehensible.[5@]

Hence, the *only* justification for regulating a network activity is to classify it as a *public service*: monopoly pricing would be abusive because society desires an extended level of consumption or universal access.[6@]

Evil Regulation ? The *public choice* theory of the State holds an opposite rent-seeking view: natural monopoly industries, fearing endless cutthroat competition, negotiate protection with authorities (cf. §16.3 & §16.2.4). At the same time, there are grounds to believe that Schumpeter (1942)'s "creative destruction" is at work in natural monopolies.[7@] If so, cutthroat competition is healthy as it fosters innovation and cost cutting. It is then a source of dynamic efficiency. Next, monopolies being of a temporary nature, they generate only a small allocative efficiency loss. Weighting the benefit against the loss, a clearcut conclusion emerges: the market should be left unaltered and all regulations eliminated.

This position, however, ignores that regulation in all advanced economies and all sectors has always appeared as the outcome of an intense economical, political and social conflict between opposed factions.

Contracting The intertwining of private and public interests in natural monopolies can be best understood from the perspective of transaction costs (cf. §13.3.3). Notice indeed that establishing a network service in a purely private fashion is a daunting task as it requires transacting with many individual land owners under the threat of halting the entire project if a single negotiation fails.[8@] It is thus an economizing solution for a firm to seek a right-of-way from authorities (cf. §16.1.2); this is the path followed in all sectors at all times in all countries. From this moment on, a bilateral relationship is established between a seller, the firm building and operating the network and the buyer, a public body aggregating the interest of the numerous future clients, households and businesses. In the ensuing bargaining, the regulator trades quality, geographical extension and prices against stability and protection from entry to allow the firm to recoup her high sunk cost.[9@]

What characterizes most clearly this relationship are the specific investments made by each party and the subsequent hold-up problem (cf. §13.3.3). Indeed, the firm lays down an immobile network with virtually zero value in alternative use, thus fears opportunistic price ceilings from the city (or any other costly demand). Likewise, the city, once it has permitted the building of the network, can not repossess it if she is dissatisfied with the service rendered by the firm. To avoid mutual hold-up, partners have to use detailed long-term contracts with revision clauses. The *concession* model (cf. §16.4.1) used in France for several centuries[10@] has been found by Priest (1993) to characterize the early regulation of US utilities. The ensuing evolution into a state regulation with public utility commissions (PUC) during the last century[11@] is attributed to an increase in the minimum efficient scale (improved technology) and the need for an independent body to stop municipal opportunistic behavior (cf. Knittel (2006)).

17.1.2 Pricing Quandary

We now inquire about the socially optimal pricing policy for a network firm or natural monopoly. The Pareto conditions for global economic efficiency (cf. §2.3.3) call for marginal cost pricing everywhere. Yet, applying such a policy to natural monopolies is bound to generates operating losses since the average cost remains systematically above the marginal cost (i.e., fixed cost are not covered). We look in turn at two opposing and hotly debated compromises to solve this quandary (cf. Ruggles (1949)).

Marginal Cost Pricing

The archetypal example of marginal cost pricing is found in Dupuit (1844)'s bridge: once build, it is inefficient to set a positive passage price since there is no variable cost involved, in the absence of congestion (cf. §25.3). Thus, we should "let bygones be bygones" and make it a free service. The same applies today to digital goods (cf. §12.3). Obviously, the service is available today because an investment was made in the past which raises the question of its financing and the just remuneration of the risk taken by the entrepreneur.

For a public service, the government can cover the operating losses of a natural monopoly by resorting to public funds. The ideal instrument to raise money without impinging on the economy is the *lump sum* tax, a fixed amount to be paid by every household. The problem with the practical implementation of this scheme is that the aggregate level of all fixed costs of all regulated activities can sum up to a significant share of the country's GDP which means that poor people may not be able to pay their lot. Once we start to exempt poor households from the payment, the lump sum tax has become an income tax. Thus, the financing scheme has to be a mix of income and commodity taxation which are known to be distortionary. An inefficiency is thus corrected by the creation of another inefficiency.[12@]

Vickrey (1955) summarizes the problem by introducing the concept of *marginal cost of public funds*, the fact that raising 1€ of subsidy requires withholding more than this from tax payers because the administration is a costly redistribution mechanism (e.g., wages of bureaucrats) and because taxes distort the economy (cf. §2.3.3). The US government uses a value of 25% for cost-benefit analysis and recent estimates by Kleven and Kreiner (2003) show a range from 10% to 55% for Japan, US, UK, France and Germany in increasing order.

Average Cost Pricing

The alternative pricing policy, advocated by practitioners since the inception of utilities, is that "every tube must stand on its own bottom" i.e., the price of a service should be no less than its average economic cost. The fundamental underlying concept here is the cost of capital i.e., the return needed by a potential investor to invest in that class of risk. For a private firm, regulated or not, pricing above average cost is a necessary condition to guarantee the continuity of supply for otherwise, no new investment would take place so that the service would deteriorate and possible come to an halt. Obviously, a private (profit maximizing) firm will more than agree with this view since its objective goes beyond covering cost as she seeks to obtain an extraordinary profit. For a public firm (state owned), computing economic cost using an economy wide cost of capital is a way to identify the most useful investments for society.

When the firm produces a variety of goods and services, the average cost rule boils down to find a vector of prices generating enough revenue to cover economic cost (including the remuneration of capital). The thorny question is how to set relative prices as this implicitly defines cross-subsidization among the product lines (cf. §17.3.4). Likewise, when all natural monopolies are pooled inside a State owned conglomerate such as INI in Spain or IRI in Italy, there is only one budget constraint for the whole public sector but the permanent surpluses of some activities such as petroleum are bound to compensate for the chronic deficits of others such as postal services. Socializing fixed costs over the whole economy makes it impossible to know who pays for what and check whether the general redistribution objectives of the government are met. Furthermore, this conflation of costs is opposed by almost everyone since there is always a service we do not consume but which we end up financing.[13@] In some sense, the self financing of each regulated activity through an adequate price is a transparent and fair method which has gained much political support. The only remaining problem, discussed at the end of the section, is the allocation of common cost.

Historically, it must be noted that average cost pricing has always prevailed for regulated private firms since it is politically unsustainable to channel public money to private firms (as required by marginal cost pricing). It has also gradually become the rule for state owned enterprises (SOE) since the 1970s when economic downturn started to put pressure on public finances. The logical conclusion of the dissociation of utilities from the central state is privatization (cf. §16.4.3). We thus let end-users pay the cost of service directly through prices instead of paying for the fixed cost indirectly through the State tax system. The privatized firms are nevertheless subjected to some minimum requirements regarding quality of service, geographical coverage and guaranteed end-user prices or any other public service objective (cf. §17.3.4).

Welfare Horizon

The marginal vs. average cost controversy originates in the distinction between variable and fixed cost or in the division of cost into past and current. Wiseman (1957) judges this dichotomy as artificial since for all public services we may think off, when we increase consumption by turning on a light or opening a faucet, we do not generate additional cost. Indeed, the resources have already been committed and paid hours or days in advance by the utility (who take decisions based on the statistical expectation of aggregate demand). This means that, like in the Dupuit (1844) bridge example, the service should be given for free to satisfy the short-term efficiency conditions.

At the other extreme of the lens, the fixed cost problem originates in the durable and lumpy nature of the assets needed to provide the service e.g., bridge, plant, reservoir. Recall now that the Pareto efficiency conditions (cf. §2.1.1 & eq. (2.8) or eq. (2.9)) derive from an arbitrage argument which presupposes that every factor is perfectly mobile and divisible. Applying the welfarist argument to public services and their long term investments thus requires us to consider an extremely long time period e.g., 20 years. Once we do so, and consider a large country with numerous plants, we can fine tune output by planning obsolescence and replacement with the effect that all of cost become variable and none is fixed. Over this long term, the marginal cost of service is then the cost of expanding marginally all assets and inputs to serve an additional client for that duration. An example is treated in §25.3 on Peak Load Pricing.

What these extremes show is that the marginal cost pricing controversy reduces to selecting an horizon for welfare maximization. We are thus warranted to give more weight to the present (and lower prices) if technological progress reduces the investment cost for the future (since current pricing is depressing future income for the utility) and if knowledge or some other positive externality is generated by the greater use that low pricing today generates.

Club Principle and Common cost

To maintain short-run efficiency and still cover fixed cost, we can resort to a two-part tariff with a unit price equal to marginal cost and a subscription fee equal to per-capita fixed cost or any other formula carrying a redistributive purpose inso far as the total collected over customers make up for the entire fixed cost.

If the fixed cost can be exactly individualized, there is no wealth redistribution and everyone pays his fair share which is exactly by how much the system cost is increased to serve him. An example would be production of a good plus individualized transport by scooter. Observe then that the consumer is buying two products, good and delivery.

Yet, in may instances, scale economies are achieved by bundling transport into a larger carrier that visit several clients. More generally, it is necessary to pool production in order to harness positive network externalities. The corresponding common cost must then be allocated among clients in a manner furthering the government's social criteria. Hence, regulation is required.

Libertarians offer an alternative avenue free from government intervention. Buchanan (1965)'s "club principle" is an equal sharing method that gathers potential clients of a service with a view to either produce it themselves[14@] or procure it from an independent producer (cf. §13.3.3). According to this principle, citizens also decide on the characteristics of the service such as quality as well as on its finance scheme. Because people voluntarily join the club, its coming into existence will generate an acceptable wealth redistribution even if there is price discrimination so that welfare is increased i.e., it is a Pareto improvement.

Several critics can be addressed to this rosy view. Firstly, the club operation involves a political decentralization since the central government is in effect agreeing that the group of citizens will carry a policy in accordance with its general social criteria. Such a conclusion is by no means obvious since the gathering obeys cost minimizing considerations and geographical constraints without relations to social realities that traditionally guide the government. A second problem with the club view is that the utility has the upper hand and therefore does not leave much surplus to the citizen community.[15@] Indeed, to take advantage of the underlying scale economies, a large minimum size of operation is needed which requires expertise in technology and finance. These in turn are found in profit maximizing corporations who end becoming natural monopolies. To avoid an abuse of dominant position, authorities have to step in.[16@] To conclude, except for the unlikely cases of where all dimensions of the service can be ascribed to the final users independently, the pricing of a public service (at least the subscription fee) requires allocating common cost by making interpersonal comparisons which ought to follow the general social criteria of the government as outlined in the contractual view.

17.2 Ideal Regulation of a Natural Monopoly

We shall study regulation under the proviso that *a regulated firm finances her cost out of direct revenues*. This means that the regulator, also known as the social planner in the public economics literature, maximizes welfare under the budget constraints of the regulated firm. It is customary to speak of a *second best* or *constrained efficiency* approach and reserve the label *first-best* or *full efficiency* to the hypothetical case where the deficits generated by marginal cost pricing can be financed out of lump-sum taxation.

17.2.1 Homogeneous Demand

In the simplest case of a regulated firm producing a single good, the solution to the above program, already hinted at by Edgeworth (1910), is readily identified on Figure 17.1. The first-best which maximizes welfare without any restriction is the quantity q^* equating WTP and marginal cost; since there is a fixed cost, the average cost is above the marginal cost so that selling q^* units at price p^* generate operating losses equal to the stripped area (and also to $q^* \times (AC(q^*) - p^*)$). Since welfare increases with the quantity produced (and consumed), the minimal loss is achieved by reducing the quantity from the first-best level till the level $\bar{q}$ where the budget constraint $qP(q) \geq C(q)$ becomes satisfied; this occurs when the price is equal to the average cost ($P = AC > C_m$).

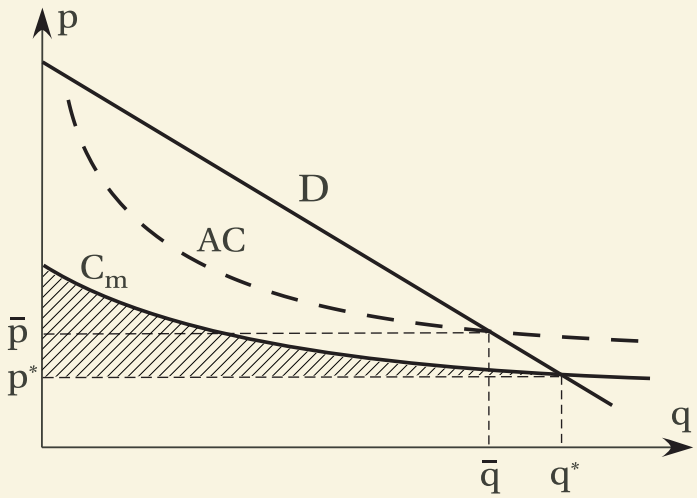


Figure 17.1: Cost Structure for a Natural Monopoly

Notice however that the first-best quantity q^* is only a candidate because one must check that the maximum welfare W^*, the area between the demand and marginal cost curves, covers the fixed cost F. If this is not the case then society is better off closing the plant, saving F and renouncing to the surplus W^*. To take the correct decision, the regulator has to compute the consumer surplus which is quite difficult as it crucially depends on the willingness to pay of people who always consume (i.e., whatever the price). To understand this, imagine that the demand drawn on Figure 17.1 corresponds to a state of high demand (D_h) while a rotation to the left around the efficient point (q^*, p^*) corresponds to a flatter and lower demand curve (D_l). Assume that the fixed cost stands between the surpluses W_l^* and W_h^*. According to the state of demand, the service should or shouldn't be produced. Notice that small price variations will never succeed to reveal the full shape of the demand curve while large price variations (governmental experimentations) are out of the question as they would almost surely trigger a popu-

lar insurrection. Vickrey (1948) argues that this uncertainty is a additional reason to avoid marginal pricing and use instead the (second best) average cost pricing because the later would almost automatically reveal whether the activity should be produced or not. Indeed, when demand is low, the average price $\bar{p}$ rises and demand shrinks thereby revealing that the service hardly produces anything worthwhile to society.

> The (second best) efficient regulation of a single activity is average cost pricing.

17.2.2 Heterogeneous Demand

The ideal simultaneous regulation of all network industries and other activities involving decreasing average cost is derived in a general equilibrium framework by Boiteux (1956) and refined by Drèze (1964). While it builds on the previous simple result, care has to be taken of heterogeneity. A regulated firm, like any other firm, usually sells its basic good or service to several types of clients, like households and professionals or to customers who have fairly different uses of the product. This disjunction brings us back to one of the everlasting problems of economics, how should we allocate the fixed cost between the various classes of customers? Abstaining for the moment from redistribution concerns (i.e., who should pay), we solve this problem on terms of efficiency alone.

Let us consider a good or service bought by households and businesses. The regulator would like to price each segment at marginal cost but must accept to raise the prices in order to allow the firm to recover its fixed cost. Raising the price in a segment involves a welfare loss and a profit increase. Let us then tune prices so that the profit increases by 100€in each segment; this goes with a welfare loss of say 130€ in the business segment as opposed to 160€ in the household segment. It is then pretty clear that more sacrifice should then be ask from the business segment since each 100€ of cost recovery involves a waste of only 30€ as compared to a waste of 60€ in the household segment. Practically, we can start from prices equal to marginal cost and then perform a series of small sacrifices (price increases) in all segments following this equi-marginal principle until the budget constraint is satisfied (cf. §2.1.1). From this arbitrage reasoning, we deduce

> The efficient pricing rule for welfare maximization of an heterogeneous demand under a single profit constraint is to equate the ratio of welfare loss to profit increase in all segments.

We now relate this finding to the theories of perfect competition and monopoly. Let q_i denote the quantity consumed by segment $\#i$ and $C_{m,i} = \frac{\partial C}{\partial q_i}$ the marginal cost of service. The willingness to pay for that good is $P_i(q_i)$ while the marginal revenue from that

segment is $R_{m,i} = p_i + q_i P_i'$. The welfare loss of a price increase in segment $\#i$ is $P_i - C_{m,i}$ while the corresponding profit gained by the firm is $C_{m,i} - R_{m,i}$. Stating that the ratio of the former over the latter is constant across segments implies that optimal quantities $q_1, ..., q_i, ..., q_n$ satisfy

$$\frac{P_1 - C_{m,1}}{C_{m,1} - R_{m,1}} = ... \frac{P_i - C_{m,i}}{C_{m,i} - R_{m,i}} = ... = \frac{P_n - C_{m,n}}{C_{m,n} - R_{m,n}} = \lambda \tag{17.1}$$

where $\lambda > 0$ is greater, the greater the overall sacrifice (fixed cost). A simple manipulation of (17.1) yields the Ramsey-Boiteux formula[17@]

$$(1+\lambda)(p_i - C_{m,i}) = -\lambda q_i P_i' \quad \Leftrightarrow \quad \mathscr{L}_i \epsilon_i = \frac{\lambda}{1+\lambda} \tag{17.2}$$

for all $i \leq n$ using the Lerner index $\mathscr{L}_i = \frac{p_i - C_{m,i}}{p_i}$ (cf. eq. (3.4)) and the own-price elasticity of demand $\epsilon_i = \frac{q_i P_i'}{P_i}$.

This efficient pricing rule is qualitatively identical to the characterization of imperfect discrimination by a monopoly seen in equation (4.1). Although the social planner and the monopoly have different objectives, they both charge the more elastic segments (large ϵ) near marginal cost while they force the more inelastic ones to bear a high share of the fixed costs of the public sector or of the monopoly rent. When the fixed cost rises from zero, the λ parameter increases above zero and prices start to be distorted away from marginal cost. A large fixed cost can be recovered only by giving leeway to the firm to raise a large producer surplus i.e., by increasing λ. At the limit when the parameter is infinite, the ratio $\frac{\lambda}{1+\lambda}$ tends to one and (17.2) becomes identical to equation (4.1) that characterizes profit maximization. The maximum fixed cost that can be recovered is the maximum producer surplus which is precisely the monopoly profit as studied in §4.2.2 (cf. Boccard (2010a)). To conclude:[18@]

> Welfare maximization of an heterogeneous demand under a single profit constraint can range from the perfect competition outcome to the monopoly outcome as the fixed cost grows from zero to the maximal sustainable level.

Notice that optimal prices are larger than marginal cost but not proportional to them, rather, the wedge is proportional to the inverse elasticity of demand. This formal analysis therefore confirms the historic rule of thumb according to which “prices ought to be set according to the value of service (perceived by users)”.

When two-way transfers between regulated firms and the government are feasible, public services become an additional avenue for taxation and public funding. If λ denotes now the cost of public funds, every € that can be taken away from the firm’s profit can be

used to reduce the general taxation burden, thus saving $1+\lambda$ € to taxpayers. This means that the social value of profits is magnified by a factor $1+\lambda$. The regulator therefore maximizes $W_C(q)+(1+\lambda)(W_P(q)-F)$ so that the efficient pricing rule is again (17.2). We can thus interpret the economy wide λ that we take as exogenous, as an average of the λ's endogenously determined in each industry that must stand on its own (budget neutrality). Observe that the application of this rule to all regulated industries implies that profit making industries end up subsidizing deficit making ones. If the government cannot give and take at will, then transfers among industries are impossible to carry out yet it remains possible to organize them inside a firm so that a profit making branches finances a deficit making one.

Wealth Effects

The partial equilibrium approach we follow in most of this book disregards wealth effects which is acceptable whenever the amount disbursed by a household or firm to procure the good or service is a small fraction of his income. Almost by definition, public services are crucial for the poorest families (cf. §16.1.1) who may thus face liquidity constraint when buying them. For instance, efficiency commands to price at marginal cost and set a (per-capita) subscription fee large enough to cover fixed cost. But this may lead some poor clients to do without the service altogether (and use an improper substitute). A sound pricing policy for regulated firms should therefore trade efficiency against (income) redistribution.

Feldstein (1972) adds a distributional concern in regulated two-part pricing. For normal goods, whose demand rises with income, the price should be set above marginal cost. This way, richer consumers carry a larger share of the fixed cost. This discriminatory outcome, akin to a reduction of income disparity, is a side benefit of a policy focused on efficiency under budget neutrality and which ignores distributional concerns. However, if the public service is an inferior good, whose demand decreases with income, it should be priced below marginal cost, so that the subscription becomes larger (in proportion of total expenses). Whereas all public services used at home like energy or telecommunications are normal goods, public transportation is probably an inferior good since most people prefer to use a car if they can afford it. The pricing policy should therefore be markedly different. Most cities correctly offer monthly subscriptions to their subway i.e., set the price of a ride at zero which in any case is close to the marginal cost.

To simplify exposition, the general approach based on utility functions is streamlined by assuming that heterogenous consumers characterized by some parameter θ differ in their (constant) marginal utility of income v_θ. We denote $\mathbb{E}[]$ the expectation operator associated with the distribution of θ. Individual welfare, a monetary measure, is the

consumer surplus $W_{d,\theta}(p)$ at unit price p minus the fixed cost contribution f. Global welfare is the mean welfare, weighted by marginal utility of income i.e.,

$$W = \mathbb{E}\left[v_\theta(W_{d,\theta}(p) - f)\right] = \mathbb{E}\left[v_\theta W_{d,\theta}(p)\right] - \mathbb{E}[v_\theta]\left(C(D(p)) - pD(p)\right)$$

since the fee is adjusted to cover the remaining fixed cost. Given that the marginal consumer surplus is the opposite of individual demand (cf. eq.(2.18)), the FOC for the maximization of W is

$$-\frac{\mathbb{E}[v_\theta d_\theta]}{\mathbb{E}[v_\theta]} = (C_m - p)D' - D \Leftrightarrow \frac{p - C_m}{p}\epsilon = \frac{\mathbb{E}[v_\theta d_\theta]}{\mathbb{E}[v_\theta]\mathbb{E}[d_\theta]} - 1 \tag{17.3}$$

introducing the (negative) elasticity of demand and using the fact that total demand is $D = \mathbb{E}[d_\theta]$. The RHS of (17.3) is the covariance between marginal utility of income and demand. For a normal good it is negative so that the price mark-up is positive but for an inferior good the reverse holds. If income distribution does not matter then $v_\theta = 1$ and (17.3) calls for marginal cost pricing.

17.2.3 Price vs. Quantities

Governments often intervene markets because of *externalities*, whether negative such as pollution (e.g., CO_2, noise) and congestion (cf. §25) or positive such as public services (cf. §16.1.1)[19@] and public goods (e.g., justice, police, defense). In all these cases, production of the main output generates a by-product, called the externality *transmitter*, impacting society adversely or favorably. For instance, a coal plant generates one ton of CO_2 per MWh of electricity.

Analytically, the activity under review is characterized by a cost $C(q)$, a market revenue[20@] $R(q)$ and, by way of the transmitter, an external *harm* $H(q)$ and/or *boon* $B(q)$. Social value is defined as $V = R + B - H$ so that welfare is $W = V - C$ whereas the firm's profit is $\pi = R - C$. In the absence of property rights regarding the transmitter, no one controls its production or consumption so that the firm produces to maximize profit. The optimal production $\bar{q}$ solves the FOC is $R_m = C_m$ and is sold at price $\bar{p}$ such that $D(\bar{p}) = \bar{q}$ (where D is the market demand for the product). In a world of perfect information, the government or regulator knows everything and can compute the welfare maximizing level q^* (solving $V_m = C_m$) which is greater (resp. lesser) than $\bar{q}$ in the case of harm (resp. boon).[21@] To bring production in line with the efficient level, the regulator can either mandate the firm to produce no more (resp. no less) than q^* in the case of harm (resp. boon). Alternatively, she can mandate a maximum (resp. minimum) price $p^* = C_m(q^*)$ in the case of harm (resp. boon). Prices and quantities are then interchangeable perfectly

efficient policy instruments.

Weitzman (1974a) shows that such a conclusion must be amended once demand and supply uncertainty bear on the regulator.[22@] Specifically, assume that cost and valuation are functions $C\left(\tilde{\theta},q\right)$ and $V\left(\tilde{\eta},q\right)$. We denote $\mathbb{E}[.]$ the expectation operator over the probabilistic distribution of the independent random variables $\tilde{\theta}$ and $\tilde{\eta}$ (upon which everyone agrees). Let $\widetilde{W}(q)\equiv V\left(\tilde{\eta},q\right)-C\left(\tilde{\theta},q\right)$.

When using the quantity instrument, the regulator can only maximize the expected welfare $\mathbb{E}\left[\widetilde{W}(q)\right]$ to arrive at an expected quantity $\hat{q}$ solving $\mathbb{E}[V_m]=\mathbb{E}[C_m]$. It is obviously ex-post inefficient since $V_m\left(\tilde{\eta},\hat{q}\right)=C_m\left(\tilde{\theta},\hat{q}\right)$ is a zero probability event. The price instrument works in a more contrived way. After being told to price at p, the firm chooses the quantity $q_\theta(p)$ solving $p=C_m\left(\theta,q\right)$. Constrained welfare is thus $\mathbb{E}\left[V\left(\tilde{\eta},q_{\tilde{\theta}}(p)\right)-C\left(\tilde{\theta},q_{\tilde{\theta}}(p)\right)\right]$ whose maximum $\hat{p}$ solves the FOC

$$\mathbb{E}\left[\left(V_m(q_{\tilde{\theta}})-C_m(q_{\tilde{\theta}})\right)\frac{\partial q_{\tilde{\theta}}}{\partial p}\right]=0 \quad\Rightarrow\quad \mathbb{E}\left[V_m(q_{\tilde{\theta}})\frac{\partial q_{\tilde{\theta}}}{\partial p}\right]=p\mathbb{E}\left[\frac{\partial q_{\tilde{\theta}}}{\partial p}\right] \tag{17.4}$$

The output $q_\theta(\hat{p})$ will then depend on the state of the world but since it ignores the harm information ($\tilde{\eta}$), it also falls short of ex-post efficiency. The criteria to compare the instruments is the advantage of prices over quantities $\Delta\equiv\mathbb{E}\left[\widetilde{W}(q_{\tilde{\theta}}(\hat{p})-\widetilde{W}(\hat{q})\right]$.

To permit an analytical derivation, we assume that around the optimal quantity $\hat{q}$, functions are linear with $C_m\left(\tilde{\theta},q\right)=c+\tilde{\theta}+\delta q$ and $V_m\left(\tilde{\eta},q\right)=v+\tilde{\eta}-\beta q$ where $\tilde{\theta}$ and $\tilde{\eta}$ are centered variables with variance σ^2 and ν^2. We have $\hat{q}=\frac{v-c}{\delta+\beta}$, $q_\theta(p)=\frac{p-c-\theta}{\delta}$ and $\frac{\partial q_{\tilde{\theta}}}{\partial p}=\frac{1}{\delta}$ which is non random. Equation (17.4) simplifies into $\hat{p}=\mathbb{E}[V_m]=\mathbb{E}\left[v-\tilde{\theta}-\frac{\beta}{\delta}(\hat{p}-c-\tilde{\theta})\right]=v-\frac{\beta}{\delta}(\hat{p}-c)\Rightarrow\hat{p}=c+\delta\hat{q}$ which links the two policies. Observe then that $q_\theta(\hat{p})=\hat{q}-\frac{\theta}{\delta}$ so that taking expectations, we have $\mathbb{E}[q_\theta(\hat{p})]=\hat{q}$ and $\mathbb{V}[q_\theta(\hat{p})]=\frac{\sigma^2}{\delta^2}$. Since $\widetilde{W}(\tilde{x})=(v-c)\tilde{x}+(\tilde{\eta}-\tilde{\theta})\tilde{x}-\frac{\beta+\delta}{2}\tilde{x}^2$,

$$\begin{array}{rcccccc}
\Delta &=& (v-c)\left(\mathbb{E}[q_\theta(\hat{p})]-\hat{q}\right) &+& \mathbb{E}\left[(\tilde{\eta}-\tilde{\theta})(q_\theta(\hat{p})-\hat{q})\right] &-& \frac{\beta+\delta}{2}\mathbb{E}\left[q_\theta(\hat{p})^2-\hat{q}^2\right]\\
&=& 0 &-& \frac{1}{\delta}\mathbb{E}\left[(\tilde{\eta}-\tilde{\theta})\tilde{\theta}\right] &-& \frac{\beta+\delta}{2}\mathbb{V}[q_\theta(\hat{p})]\\
&=& \frac{\sigma^2}{\delta}-\frac{\beta+\delta}{2}\frac{\sigma^2}{\delta^2} &=& \frac{\sigma^2}{2\delta^2}(\delta-\beta) &&
\end{array}$$

Price is the warranted control instrument when the curvature of the cost function (δ) is greater than that of the value function (β) i.e., uncertainty regarding cost generates more variability than uncertainty regarding valuation. Contrarily, quantity regulation is socially superior when the valuation can vary widely with uncertainty as with natural disasters and health outbreaks.

If instead of interacting with the producer to take advantage of his cost minimizing behavior, the regulator interacts with the consumer (to take advantage of his value maxi-

mizing behavior), all the computations go along only that labels must be exchanged, thus the advantage of price (told to the consumer) over quantity is $\frac{\nu^2}{2\beta^2}(\beta-\delta)$ so that quantity is either dominated by price told to firm or told to consumer (but dominates the other one).

If instead of a monopoly, an entire industry is regulated the case for prices is reinforced because quantity schemes involves inefficiencies across members of the industry. Specifically, when the technologies are similar in curvature (same δ) but still suffer different shocks, the β factor is multiplied by $\frac{1+(n-1)\rho}{n} \leq 1$ where $\rho \in [0;1]$ is the correlation coefficient between firms' shocks (i.e., covariance is $\sigma_{ij} = \rho\nu^2$). A larger number of industry participants and more independent cost between them (lower ρ) make prices relatively more attractive to the regulator.

17.2.4 Public (non-market) Provision

As explained in §13.1.3 on bureaucracy, the provision of many public services is provided in-house by a bureau or agency. Niskanen (1968) proposes a formal model of bureau behavior based on two premices. Firstly bureaucrats, motivated by the ideal of public service, prefer more to less budget[23@] in order to provide the public with more output. Next, a bureau negotiates with the finance ministry a bundle of service objectives against a global budget (able to cover for the cost of the aforementioned objective).

Within such a framework, the bureau is like a discriminating monopolist making a "take-it-or-leave-it" offer to society; he will thus be able to exact an amount equal to society' total willingness to pay (TWTP) for the stated objective. Since the latter sums the social WTP for all units, the bureau's ideal offer is the output at which the social WTP is nil, an amount considerably larger than what efficiency warrants since it completely ignores cost. Yet, the prohibition of running deficit may force the bureau to back down a little.

Indeed, in our linear framework $P(p) = \frac{a-q}{b}$ and $C_m(q) = c + \delta q$, the efficient output solving $P = C_m$ is $q^* = \frac{a-bc}{1+\delta b}$. The budget limit is TWTP $V(q) = \int_0^q P(x)\,\mathrm{d}x = aq - \frac{1}{2}q^2$ while total cost is $C(q) = cq + \frac{1}{2}\delta q^2$, hence the financing condition $V \geq C$ boils down to $q \leq 2q^*$ which means that the bureau will tend to operate at twice the adequate level.[24@]

Assuming that all the budget is spend to increase output (or quality) of the public service is quite strong. Obviously, the bureau has enough leeway to spend money into items and issues that bring satisfaction only to the bureau members, thus being socially unwarranted. We should thus distinguish within the budget between appropriation and production. It remains nevertheless true that production will not be lesser than q^* since a marginal increase brings about utility, thus budget at a rate equal to WTP which is

greater than marginal cost, hence also increase the free cash flow. As output rise from zero to the maximum $2q^*$ that can be financed, cash flow (profit) rises to a maximum at q^* and then decreases back to zero at $2q^*$.

Hence, a pure profiteer at the head of the bureau will produce q^* in order to enjoy the maximum cash flow; in this case, a careful audit of cost would reveal some "fat" in the budget (wasted resources). Yet, allocative efficiency is achieved since the socially adequate level of output is chosen. There is however a distributive inefficiency inso far as the welfare (difference between TWTP and cost) appropriated by the bureaucrats requires distortionary taxation to be incorporated in the budget and is then spend in unproductively.[25@]

At the opposite, a pure civil servant will choose $2q^*$ and leave no "fat" since his budget just covers his total costs.[26@] Most likely, the output decision will be intermediate between those extremes reflecting the mixed motives of the bureau managers. In any case, suppliers to the bureau (selling input at total price $C(q)$) share a common interest with the bureaucrats in expanding the budget.

It is obvious that this formalization of the bureau applies to any organization such a division within a firm or a school or an hospital which does not sell its output at unit price in a market but negotiate with a higher authority an annual objective against a budget. Take-over mechanisms (or competitive entry) being more readily available in the private sector, there should be less discretion in the use of budgets.

Williamson (1963) shows that managerial discretion has a non trivial impact on resource allocation within the firm when a manager cares for staff size and free cash flow (FCF) spend over pet projects. Given revenue R, cost C, staff S, tax rate t and the minimum profit π_0 required by shareholders, free cash flow is $\pi = (1-t)(R-C-S)-\pi_0$. If the manager utility is $u(F,\pi)$ then the FOCs for maximization are $R_m = C_m$ and $\frac{\partial R}{\partial F} = 1 - \frac{1}{1-t}\frac{u_F}{u_\pi}$. The former corresponds to standard behavior while the latter shows excessive staffing since the marginal value of staff in terms of added revenue is equated with less than its unitary cost, due to the MRS between CFC and staff expressed by the manager. If the tax rate increases then output and staff are likely to increase.

17.3 Practical Regulation

The previous section was theoretical and emphasized the ideal *price structure* assuming implicitly that average cost pricing was easy to implement and would provide the firm with adequate revenues. In this section, we delve into the more concrete problem of *revenue structure* raising up a number of practical difficulties.

17.3.1 Cost Based methods

The regulatory framework emphasizing cost aspects is diversely known as “Rate of Return Regulation” ,“Cost of Service” or “Cost Plus” ($C+$). The label “Cost Plus” is used for public services such as the local police or the building of an hospital (public procurement) while the label “Cost of Service” is used for private services like distribution utilities (water, gas, electricity) or the trash collection.

Fully Distributed Cost

The first step in the $C+$ regulation is to allocate all cost among activities of the firm whether some or all are regulated. Take the example of local (#1) and long distance (#2) telephone. Total cost can be decomposed as $C = F_0 + c_1 q_1 + F_1 + c_2 q_2 + F_2$ where c_i and F_i are the marginal and fixed cost of activity $\#i$ while F_0 is the (fixed) cost of production shared by the two activities. This encompasses the administrative cost but above all the cost of building, maintaining and operating facilities used by the two activities.

The most disputable decision to be taken by the regulator is setting the share λ_i of the join cost F_0 that falls on activity $\#i$. The revenue requirement of this business line, given the estimated demand q_i is then $C_i = c_i q_i + F_i + \lambda_i F_0$ and the final stage of the regulatory process is simply to tune the price of service $\#i$ so as to ensure that revenue meets cost i.e., to price at average cost. For a non regulated activity $\#j$, the firm sets her price to maximizes profits taking into account that the fixed cost $\lambda_j F_0$ will not be covered by regulated revenues.

Among the distribution methods, the attributable cost sharing sets $\lambda_i = \frac{c_i q_i + F_i}{C - F_0}$27@ and the output sharing sets $\lambda_i = \frac{q_i}{q_1 + q_2}$. However, the political will to support an activity or a group leads to distort the objective schemes presented above. For instance, urban areas frequently subsidy rural ones in networks. One can then speak of politically or ethically motivated cross-subsidization.

Fair Rate of Return

In the previous discussion, fixed cost are economic cost, not accounting ones, thus they incorporate a remuneration of the capital invested. In the $C+$ framework, the regulator delineates the *regulatory asset base* (RAB) consisting of all the assets necessary to perform the regulated activity and sets a fair rate of return $\bar{\rho}$ to be earn on the RAB. The notion of “fairness” has been hotly debated in the US but stabilized after a 1944 supreme court judgment stating that “the return to the equity owner should be commensurate with returns on investments in other enterprises having corresponding risks”; this is what is usually called the cost of capital.

Practically speaking, the regulatory process starts with lengthy review of all investments in order to update the RAB composition and its valuation K. The maximum allowed profit for the regulated firm is then set to be $\bar{\rho}K$. If the revenues net of operating cost fall short of this benchmark, the regulator must increase the prices otherwise prices are cut to reduce this excessive rent. We now proceed to analyze the behavioral consequences of the $C+$ regulation.

Gold Plating or Padding

Since the inception of $C+$ regulation, it has been argued that firms try to inflate their RAB to earn higher profits since the multiplier, the fair rate of return, is quite constant. This strategy can take various forms such as buying unnecessary material or designs (as if gold plating of copper wires was necessary for telephone service), increase capacity of service excessively (on grounds of security, reliability or quality) or enter non profitable markets (often unregulated). The later, more modern strategy, also enables the firm to preempt markets that may become profitable in the future.

In practice, the harm is limited because the firm's costs statement are scrutinized by regulatory commissions with the threat that imprudent investments or unnecessary expenses be left out of the RAB. The bureaucratic review process which is itself costly for the economy has to a large degree been successful in preventing grossly excessive spending by regulated companies. Another positive feature, though not intended by regulators, is that during the period separating two rate reviews, prices remain constant, thus incentives to cut cost are greatest (cf. "fixed price" regulation later on).

Perverse incentives of $C+$ regulation are nevertheless present. Wellisz (1963) shows that a utility confronted to alternating peak and off peak demand (cf. §25.3) has a clear incentive to lower peak prices or at least to oppose the more efficient but higher peak load prices. If successful, this strategy boosts the peak demand beyond the efficient level with the consequence that to maintain service the firm has to increase capacity. In the process, the firm increases her RAB and thus her profits.

Averch-Johnson Effect

In the first mathematical model of firm behavior in a regulatory framework, Averch and Johnson (1962) show that the rate of return regulation reduces the perceived price of capital so that the firm is lead to use a relatively excessive amount of capital in comparison to other inputs such a labor; it might even be the case that the workforce is reduced through "downsizing". We study two flavors of this model, firstly under a strict $C+$ regulation where the price is set by the regulator and then when the firm is free to set the

price.

Consider a regulated monopoly whose production function is $\Phi(K,L)$. The market willingness to pay is denoted $P(q)$ and $R(q) = qP(q)$ is the revenue function; profit is $\Pi(K,L) = R(\Phi(K,L)) - wL - rK$ where r is the cost of capital and w the wage rate. The cap imposed by the regulator on the rate of return (`ror`) on capital (asset base) is denoted $\bar{\rho}$; it leads to the constraint

$$\rho = \frac{R(\Phi(K,L)) - \omega L}{K} = \frac{\Pi}{K} + r \leq \bar{\rho} \tag{17.5}$$

The `ror` at the unconstrained monopoly optimum is denoted ρ^*. A meaningful fair rate $\bar{\rho}$ should satisfy

$$r < \bar{\rho} < \rho^* \tag{17.6}$$

for if the limit is set below the cost of capital r, no investor will remain in that activity and this will put the service at risk.[28@] Likewise if the limit is set above the unrestrained rate of return, the firm will happily continues to behave as a monopoly under a useless regulation.[29@]

Under a strict $C+$ regulation, the regulator sets the price p which in turn determines a demand $q = D(p)$ but also the total revenue of the firm; profit maximization then amounts to cost minimization. The only task of the firm is to choose inputs K and L so as to meet demand i.e., $\Phi(K,L) = q$ and its optimal choice is thus the cost minimizing pair (K^*, L^*). If the resulting `ror` ρ^* is greater than $\bar{\rho}$, then the firm must find a way to meet constraint (17.5) with equality without losing too much profits. Equality in (17.5) yields $\rho = \frac{R-\omega L}{K}$. We now use the fact that $\Phi(K,L) = q$ (inputs meet demand) to deduce that L is a decreasing implicit function of K. It is now obvious that to reduce ρ one cannot do otherwise than increasing K until $\rho = \bar{\rho}$.[30@]

> The capital to labor ratio chosen under rate of return regulation is excessive with respect to the efficient one (which minimizes production cost).

In the original setting of Averch and Johnson (1962), the firm is less tightly regulated since it is free to set the price of the regulated service (or equivalently the level of production). The previous result continues to hold: for a given production and associated revenue, the firm over-invests in capital (with respect to the cost minimizing mix). What is then the optimal output for a `ror` regulated monopoly? Starting from the unconditional profit maximizer which violates the `ror` constraint, a capital increment loosens the constraint as can be seen from (17.5) and increases production, thus decreases profits which further loosens the constraint; we have thus shown that the optimal production under `ror` regulation is larger than without regulation i.e., $\hat{q} > q^*$. We can then state the original conclusion of Averch and Johnson (1962)

A profit-maximizing firm regulated by a cap on its `ror` increases production without minimizing cost and by overinvesting in capital with respect to other inputs such as labor.

The previous reasoning can be illustrated with the help of Figure 17.2. The expansion path of cost minimizing inputs solves $r/\Phi_K = \omega/\Phi_L (= C_m)$; it is an increasing curve in the (K, L) plane. The `ror` constraint $\bar{\rho}$ defines an "eye-drop" shaped curve such that inside points are prohibited.[31@] Since $\rho^* > \bar{\rho}$, the unconstrained optimum is a point on the expansion path strictly inside the "eye-drop" (it uses capital K^*). The constrained optimum is a point on the "eye-drop" that maximizes profits. Since (17.5) holds with equality, profit is simply $(\bar{\rho} - r)K$, hence the maximum while staying on the "eye-drop" is reached at the right most point $\bar{K}$.

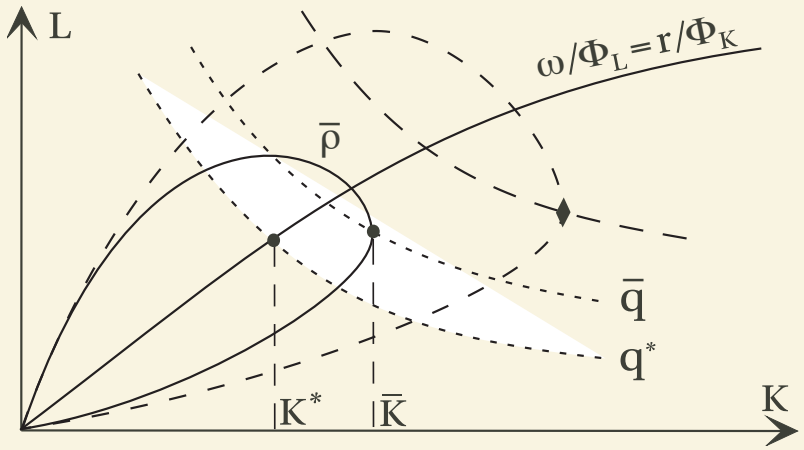


Figure 17.2: Rate of Return Regulation

Lastly, we assess whether it is welfare enhancing to set the cap close to the cost of capital (recall that it cannot be lower if one wants to maintain service). Strengthening the regulation by lowering the cap, amounts to inflate the "eye-drop". The new optimum is displayed as a diamond on Figure 17.2: on the one hand production is increased which is positive but on the other hand, the input mix is pushed further away from the cost minimizing expansion path which is negative. To disentangle this dilemma recall that the second best solves $P = AC$ and that for a natural monopoly we have $AC > C_m$ everywhere (scale economies). Setting $\bar{\rho}$ nearby r forces the firm to produce an output that nearly equates P and $\widehat{AC}$, the distorted average cost. As it very unlikely that the distortion of input mix invalidates the inequality $\widehat{AC} > \widehat{C}_m$, we would still have $P > \widehat{C}_m$ which is the sign that the additional production so obtained was welfare enhancing.

Strict equality of $\bar{\rho}$ with the cost of capital is not advisable. As we show in §19.3, whenever there is some slight uncertainty regarding future demand or cost, it is neces-

sary to leave a minimal rent to the firm in order to guarantee her presence in the market. We can therefore conclude

> The socially optimal `ror` regulation of a natural monopoly should be set nearby the cost of capital.

Differing Objective

As shown by Bailey and Malone (1970), the `ror` regulation yields an opposite result if the objective of the firm is either revenue or output maximization which is a quite plausible objective for public utilities having local elected representatives on their board of directors (this issue is discussed in §15.1).

To catch the intuition, recall first that the `ror` constraint is $R(q) \leq wL + \bar{\rho}K$. To maximize revenue $R(q)$, the manager can start from very large amounts of inputs whose cost is large; this enables a large production but a small revenue since the price has to be adjusted downward to sell such a quantity. There is thus considerable slack in the `ror` constraint. Consider now reducing capital K, the revenue increases at speed $\frac{\partial R}{\partial K} = \Phi_K R_m$ while the cost element in the constraint reduces at speed $\bar{\rho}$; the slack in the constraint is reduced at rate $\Phi_K R_m - \bar{\rho}$. A reduction of labor L reduces the slack at rate $\Phi_L R_m - w$. When the constraint is violated at $\bar{q}$, the reductions in capital and labor are tuned so that all the available slack is used up; at the optimum, it must be true that $\Phi_K R_m - \bar{\rho} = 0 = \Phi_L R_m - w$. A simple algebraic manipulation yields $\frac{\Phi_K}{\Phi_L} = \frac{\bar{\rho}}{w} > \frac{r}{w}$ as if capital was more expensive for that firm; it is then obvious that the optimal mix involves under-capitalization (with respect to the cost minimizing way to produce the optimal quantity).

> A revenue-maximizing firm regulated by `ror` restricts output (with respect to her ideal choice) but underinvests in capital and oversizes its workforce (with respect to the cost minimizing choices).

17.3.2 Incentive Regulation

Beyond the issues previously debated, the $C+$ regulation's main defect is that the firm has no incentive whatsoever to reduce its cost structure, improve quality or to take risk at introducing new technologies. The idea of *incentive regulation* is to thwart the "business-as-usal" behavior of managers of firm acting under a $C+$ regulation. The basic and theoretically most powerful scheme is the *Fixed Price* (FP) aka *Price Cap* regulation. For a public service such as the building of a public library, the regulator sets a price equal to the average *expected* cost of service. For a private service such as trash collection, he computes the fair price to be charged to end users.

Under FP, the firm is the residual claimant of any reduction of its production costs; this gives her the maximum incentive but also forces her to bear maximum risk. A frequently studied example is the US Medicare program that repays the hospital a fixed amount for each patient according to his diagnostic group.

Sliding Scale

More than a century ago, English gas utilities were regulated by a sliding scale; starting from an adequate initial price and profit, the firm is allowed to increase profits if she lowers prices so that consumers also benefit the change. For instance, if a new production technique allows a 2€ unit cost reduction, then the price has to drop by 1€ so that the unit margin increases also by 1€. The sliding scale is in fact a combination of $C+$ and FP whereby the firm is allowed a price increase if her cost rises but at a less than one-to-one rate. $C+$ occurs when the rate is unity (price tracks exactly cost) while under FP it is zero (price ignores cost). A sliding scale thus provides incentives to cost reduction and risk-sharing at the same time.

Revenue Cap

some energy utilities used to be regulated trough a global revenue cap in order to allow some flexibility in the setting of prices across market segments. A defect of this regulation identified by Crew and Kleindorfer (1996) is the tendency for firms to overprice their services. In the traditional price-quantity space, a revenue cap corresponds to an iso-revenue hyperbola below which the firm can choose her price-quantity pair as shown on Figure 17.3. The highest conceivable is the one passing through the unregulated monopoly optimum (q^M, p^M). The regulation will make sense only if the cap is stricter than $q^M \times p^M$. In that case the new limiting revenue hyperbola will cut the demand at two prices above and below p^M.

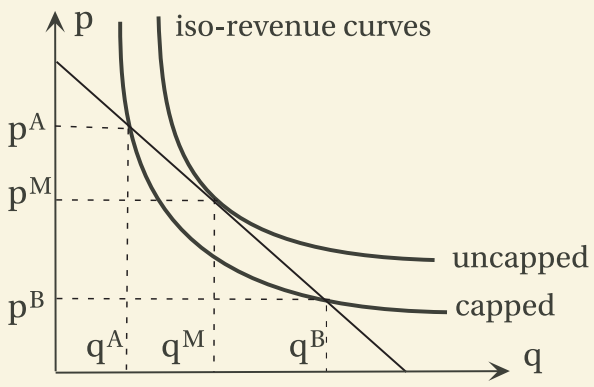


Figure 17.3: Revenue Cap Regulation

Since the pair has to lie on the demand curve, there are in fact only two possibles prices that lie around the unconstrained monopoly price. The firm will elicit the larger price p^A because it yield less sales, thus less cost but the same revenue which ultimately means more profits than the lower price p^B.

Price Cap

The now most frequently used form of incentive regulation is the *price cap*, proposed originally by Littlechild (1983) for the regulation of UK's former telecommunication monopoly at the time of its privatization. This framework amounts to fix the maximum output price that can be charged to end-users, thereby generating a minimal consumer surplus. Since the firm's profit is an increasing function of price from zero to the monopoly's optimum, the firm will price at the cap whenever it is set below the monopoly price by the regulator.[32@]

Obviously, the lower the cap, the greater the production and the consumer surplus but the lower the firm's surplus. This might become problematic because if the firm cannot cover her fixed cost out of this left-over, shareholder are getting a below average return on equity and may therefore force the firm to exit the market. To avoid this dreadful eventuality, the regulator is forced to set the cap strictly above the expected average cost of the firm so that the efficiency gains from setting a maximum price lower than the unconstrained monopoly choice are somehow wasted away. In that respect, FP fares worse than $C+$ which is an ex-post regulation able to compensate the firm for any unexpected event that lowered profits.

The positive side of FP regulation is its ability to maintain the firm on the track of technological progress. Indeed, the price cap leaves the firm free to pocket the full benefit of any cost reduction she might succeed to generate; it therefore fosters the adoption of innovative technologies whereas, in this respect, the $C+$ framework leaves the firm indifferent. Obviously, the downgrading of quality is an easy way to achieve cost reductions which is why regulators explicitly include quality objectives in the service contracts of utilities. We may therefore conclude

> FP regulation fares worse than $C+$ with respect to allocative efficiency but better with respect to dynamic efficiency understood as the ability to incorporate technical innovations.

Since the economy is constantly evolving, the price cap has to be adjusted periodically. Firstly, one accounts for the general increase of input prices measured imperfectly by the consumer price index (CPI) (retail price index (RPI) in the UK). Secondly, the increase of demand and the technological progress in the economy or within the sector have to be

integrated in the formula; this leads to the famous "$CPI - x$" or "$RPI - x$" formula stating that the price in the next review will be increased by $CPI\%$ minus $x\%$, where x measures the savings due to the exogenous rate of technical progress and the increase of demand that enables to take advantage of scale economies.

Yardstick Competition

The price observed in a competitive market is the ideal yardstick to measure the level of unit cost in the industry since individual output realizes the equality of price and marginal cost. This public information is quite useful for a firm owner as it enables him to check whether her manager runs properly the business. Whenever a firm is shielded from competition as in the case of a regulated natural monopoly, one loses this reference. Mead (1944) proposes to compare the performance of the regulated firm to that of all the other firms producing the same good or service in similar conditions; this way one builds a yardstick that can be used to discipline the manager. Nowadays, regulators performs statistical productivity analysis upon samples of comparable firms to determine indexes of efficiency that later serves to set the x factor in $CPI - x$ price cap regulations; for instance one can set it to half the distance between the worse and the best firm in the sample. Shleifer (1985) calls this form of regulation *yardstick competition* and shows it can induce all firms to behave efficiently in terms of cost reducing innovations.

Comparison of FP and $C+$

Table 17.1 summarizes the main features of the Rate of Return and Price Cap regulatory frameworks.

Dimensions	Price Cap	Rate of Return
1 - Prices flexibility	Yes	No
2 - Regulatory lag	Long	Short
3 - Sensitivity to realized costs	Low	High
4 - Regulatory discretion	Yes	No

Table 17.1: Regulatory Regimes

Some comments are:

A Price flexibility: if there were no asymmetry of information there would be no point in letting the firm set (potentially harmful) prices. Otherwise flexibility may allow the firm to employ its superior information to design prices that generate higher levels of welfare. The usual forms are average revenue regulation and tariff basket regulation; in the later, the firm can chooses the prices it sees fit as long as total

revenue does not increase with respect to the previous period. The problem is that the firm may cross subsidy a segment to attack a competitor and may also cease to cross subsidy the poor. A partial remedy for the latter is to include a public service obligation.

B A regulatory policy may be unable to secure substantial surplus for consumers when it is first implemented, but repeated application of the policy may serve consumers well. Suppose that the X factor in a CPI-X regulation is updated periodically to eliminate the firm's expected future rents and is based upon the current realized revenues and costs. Even though the firm can retain all yearly profits, it recognizes that larger present profits-generated for instance by efficiency gains may result in smaller future earnings. Consequently, the incentive to reduce operating costs is weakened (cf. ratchet effect §17.3.3). Frequent revisions are optimal if costs are exogenously driven and demand is elastic because the deadweight loss matters. Otherwise reviews should be infrequent to motivate the firm to reduce costs. Fixing the X factor is the most difficult and controversial part of the regulatory process (firms often sue their regulator delaying and weakening the regulation process).

C Although the X factor is supposed to track possible cost reduction due to technological progress, the regulator can renounce to use cost information because it is difficult to decipher from the firm accounting statements. Another reason comes from the fact that lump-sum transfers to the regulated firm are prohibited by the EC so that the price cap is the only instrument available to achieve two objectives: provide incentives to reduce costs and make sure that prices follow realized cost in order to promote allocative efficiency (deadweight loss reduction). A compromise is thus required.

D Lastly we consider how much policy discretion to afford the regulator himself. If granted flexibility there is the risk that he acts opportunistically (ratchet effect); he won't be able to refrain from maximizing welfare ex-post (setting a large X) and this will badly distort the ex-ante incentives of the firm toward cost reducing investments. Alternatively, the regulator might succumb to industry pressure.

Given the practical difficulties for the regulator to characterize the missing information, the relevant constraints, the correct objective and the most reliable instruments, he is often lead to use simple regulatory rules. The `ror` regulation is interventionist in that all prices are set by the regulator and constantly revised using cost statements to ensure closeness to the `ror` objective which at the same time is guaranteed to the regu-

lated firm. Under a price cap regulation the firm is free to discriminate among product lines or clients and for a longer time before revision. This is so because the regulation ceases to monitor cost. Nonetheless the regulatory regime being more flexible it is also susceptible of important changes when revision comes.

17.3.3 Regulatory Constraints

The previous analysis completely ignored the constraints on information, contracts and policy that the regulator faces whenever trying to *implement* a regulation. These issues are related to the agency theory developed in part H and to firm theory that deals with authority relationships inside firms (cf. §13).

Transaction Costs

The first constraint faced by a regulator is related to transaction costs (cf. §13.3.3), the difficulty to write and enforce contracts, between himself and the firm. Incumbents of regulated services or challengers willing to replace them often make promises that look attractive for the public but which are difficult to monitor and enforce. Regarding enforcement the regulator has limited powers. If the firm fails to deliver on promises, the regulator can hardly terminate the contract on grounds of "failure to meet contractual performances" because this is an implicit confession of the regulator's own failure to select an able candidate. Furthermore, the lag of time necessary to conduct another selection process is likely to result in deteriorated service which is very costly in political terms.

In the same line, the effectiveness of regulation is limited by the law establishing the regulatory agency, by the laws of higher order that limit the available instruments, by government meddling (for electoral purposes) and by the integrity of the regulator itself. Regarding the last issue, lobby groups representing consumers, but above all, individual firms and their professional associations often step in the regulatory process and try to influence the regulator; when successful one speaks of "regulatory capture", a topic we already studied in §16.3. A remedy is to isolate the regulator from all interferences through the original legislative act but it comes at the cost of inflexibility. A rigidity often ascribed to regulators is their obsession with stable prices and their dislike of discriminatory pricing and Ramsey-Boiteux formulas. It can be understood as a guarantee for both producers and consumers that prices will evolve smoothly. Such an insurance fosters investments by both sides into the network and into devices to use the network service, respectively.

Hidden Information

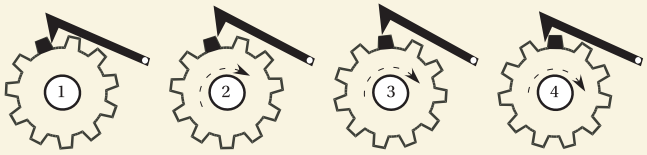
The ideal (second best) regulation for a public service is a per unit price (close to marginal cost) coupled with a subscription (covering fixed cost). The correct values are not easily computed because the regulator has to use the cost statements handed over by the firm who might therefore have an incentive to overstate her true cost. This possibility is a real issue given her better knowledge of production technology and demand. In the absence of a costly monitoring, the manager of a private firm is likely to behave as a pure monopolist simply claiming he faces tough conditions to justify the high prices that maximize his profits. Placed in a similar situation, the manager of a public firm would pursue personal objectives at odd with public interest.

The distortion brought about by this asymmetry of information is studied in §21.2.3. There is a traditional distinction between variables that are *endogenous* like the intensity of effort or the favoring of one activity over others and those variables that are *exogenous* like the knowledge of underlying cost or future demand. In the first case, one speaks of hidden action or moral hazard (cf. §20) while the second case is referred to as hidden knowledge or adverse selection (cf. §21). In a one-shot (not repeated) relationship, it is possible for the regulator to design an incentive contract that extracts the precious information initially possessed by the firm regarding costs or demand. Yet, as public funds (cf. §17.1.2) involve a loss of efficiency, the optimal regulatory policy will be distorted as if there was an additional marginal cost of eliciting the private information of the firm.

Ratchet Effect and Commitment

Yet, if the firm reveals her information too quickly, the regulator will start the next regulatory round in a much better position and will squeeze her future informational rent by asking for higher effort or lower costs. This phenomenon, due to the repeated nature of the interaction, is called the *ratchet effect*. The term was coined by Berliner (1957) who observed that, for political reasons, production objectives in the soviet plan were always to be set at higher and higher levels ignoring the existence of economic cycles or exogenous constraints.

It was as if the russian economy was a toothed wheel engaged with a lever forcing it to turn in only one direction, growth as displayed here.



This mechanical description is that of a ratchet which explains the analogy. Hence, in a repeated contracting environment, the regulated firm will not react to incentive

schemes as we might hope because she needs to conceal her private information. This problem occurs even if the regulator adopts a low powered incentive scheme because he cannot commit not to change it the day he will discover the true characteristic of the firm. Lengthening the duration of contracts would seem attractive but is often an impossible commitment given political turnover. One solution is to pass a law of constitutional nature (hard to repel) creating a regulatory commission free from political influence and with clear objectives such as the European Central Bank or the US federal commissions. The major drawback is then the rigidity of the regulator who, by the very nature of its charter, cannot choose flexible policies adapted to the current economic environment.

The root of the problem with the ratchet effect is the inability to commit for future behavior. A similar issue is at work with governments (democratic or not) who can always force a renegotiation after elections on the ground that they have a popular mandate for it. Especially, when the government control the judiciary, it does act opportunistically against private contractors and concessionaires, especially when their investments are immobile such as with infrastructures (e.g., network, airport).

As an application, consider the issue of reducing carbon emissions. An industry may fear that its early curbing efforts may be built (ratcheted) into future, differentiated, requirements, that will be disadvantageous. Consistently with the incentives of the ratchet effect, the industry may minimize those early efforts, to the cost of the national program. This effect may be a further argument for economic measures, such as transferable permits. Permits involve a greater degree of transparency and hence commitment by the national government and the international community. They do not affect individual industries in a differential way after the revelation of information.

Stranded Cost

Regulated firms in energy sectors continuously invest in long lived assets. At the onset of a deregulation, large cost are already incurred (sunk) but not yet fully depreciated. The recovery of these embedded costs becomes uncertain for regulated incumbents with the advent of entry that will likely skim their most profitable clients. Entry is either due to deregulation (removal of statutory barriers) or to technological developments that lower entry costs (in which case the natural monopoly rationale vanishes). Sidak and Baumol (1995) explain the dilemma faced by the regulator. The benefit of withholding the implicit contract (aka regulatory compact) with the incumbent is stability and the avoidance of "regulatory uncertainty" that investors fear so much; this attitude is positive for future investments. The drawback of allowing the recovery of stranded cost is to derail deregulation. Indeed, if a surcharge is applied to end-users, prices increase in the short term while if the surcharge is applied to inputs (e.g., network access) needed

by entrants, it forestalls entry.

A simple theoretical answer to the dilemma would be to judge according to the "prudent investment" criteria. If the coming of competition was foreseeable, then no recovery should be allowed whereas if increased competition was unforeseen, partial recovery is optimal to avoid giving firms incentives to make exceedingly risky choices (cf. §23.2.1 on free cash flow). However, the situation observed often involved an incoherence, traducing a successful regulatory capture. Several years before the actual deregulation would materialize by entry of new players, a government would prepare the future law with the hope of bringing down the price for end-users either through a reduction of average cost or a lessened exercise of market power. At the same time, the government would negotiate stranded cost with incumbents. The problem is that if entry or foreseeable technological improvements make it almost certain that the price will go down, then all the authorized investments under the old regulatory compact reveal themselves to be imprudent or undue; indeed, the regulator had the knowledge that either incumbents were exercising market power or that new technologies were readily available.

17.3.4 Public Service Obligation

A legal introduction to this topic is Sauter (2008) while an economic one is Cremer et al. (2001).

Concept

Public services, or services of general economic interest (SGEI) in European parlance, are a set of products and services deemed essential by the community in order to promote social and political objectives, such as equity, participation, cohesion or solidarity. The practical goal is to ensure the availability of a minimum set of *essential* services to all users at an *affordable* price and a minimum quality. For these activities, there is the presumption (often based on past experience) that private provision would be inadequate. More precisely, it could be the case that some groups of individual would not be served at all or would be offered the service at an exorbitant price either because the cost is high or because the provider has market power.

A Public Service Obligation (PSO) or Universal Service Obligation (USO) in US parlance[33@] is then a duty imposed by the State on an industry to supply the SGEI to all customers, irrespective of the fact that some, by virtue of their wealth, usage characteristics or location, might be more costly to serve then others. For instance, the telephone subscription is regulated at country level although the connection to the network is more costly for an isolated house in a village than for an apartment in a urban area. Likewise,

a local call is costlier to carry than a long distance one but is often billed at a lower rate. The PSO might then be defined as the difference between the cost of serving those protected customers and the revenue they generate (more precisely the forgone revenues from discontinuing the service). It may also involve making financial provisions so that the service can be made available to users with a disability or in financial distress.

The original EC treaty defines SGEI vaguely and implicitly leaves the principles and conditions of supply of these services to the appreciation of member states. Since a PSO involves a compensation, it potentially infringes the State Aid legislation on distortion of competition (cf. §9.2.4 for some data).[34@] The ECJ has stated in the Altmark judgment of 2003 that the PSO must be clearly defined in the national law with objective and transparent parameters for compensation. The latter covers cost and a reasonable profit and should not confer a competitive advantage. Whenever the service provider has been selected through open bidding or a public procurement process, it is assumed by the EC that these conditions are satisfied. Member states, however have leeway to define exactly what services are "essential" for the collectivity and what is an "affordable" price level. Since 1985, directives implementing these concepts have been adopted in the sectors of post, mobile telecommunications, airports, ports and maritime transport, insurance, broadcasting, energy and railways.

Milne (1997) refines the goals of USO according to the state of development of the underlying market:

- network establishment: Acquire new technology, set up long-distance links between cities, public telephones.
- wide geographic reach: Reach regional parity, connect all urban areas with widespread adoption in business.
- mass market take-up: Stimulate economy with residential take-up.
- network completion: Achieve political cohesion: affordable to all and available to disabled.
- services to individuals: Individual right to communicate and public access for education and health institutions.

Table 17.2: Stages of Universal Access/Service

Cross-Subsidization

Since a PSO usually runs a deficit, its provider is authorized to *cross-subsidize* the costly segment by applying market power upon the more profitable segments. When non discriminatory pricing is further present, this amounts to raise either the unit price or

the subscription or to raise the price of another regulated service on top of its average (economic) cost.

In a deregulated market where the PSO provider competes with other firms, this strategy ceases to be viable because of the cream skimming threat i.e., an entrant can capture the profitable segment where the incumbent is currently pricing above average cost by undercutting him. The law can therefore establish a fund where all firms contribute according to their market shares in order to finance the extra cost of the PSO provider. Still, the presence of asymmetric information between the firm and the regulator make cross-subsidization unavoidable because the PSO provider will artificially transfer cost from the unregulated side to the regulated side in order to gain an edge over competitors in the deregulated area.

The modern implementation of a PSO therefore aims at making the subsidies "explicit" rather than "implicit" to ensure competitive neutrality i.e., they should be quantifiable and their distribution clear enough to avoid favoring one competitor or one type of technology. Recent implementations have demonstrated that PSOs are not costly, typically a few percent of the sales volume of the company in charge; adding on top the goodwill amassed from the general public by the provider (he is seen as a benefactor), the net cost of a PSO for the economy appears to be much smaller than what debates in the 1990s would have lead one to think. For instance, Ofcom computes the PSO cost of British Telecom (BT) as one pound per year per capita and nearly zero if one accounts for the added goodwill that the PSO bestow upon BT.

Origin of the PSO

The first PSO came from the US and is said to originate from regulatory capture. Universal service was originally a commercial strategy of extensive geographical coverage focussing on major cities imagined by AT&T in the 1880s for its new telephone service (the Bell system) in its competition against the older telegraph service of Western Union. With the advent of competition on telephony after the expiration of Bell's patents, the AT&T motto "a telephone in every home connected to every other telephone in the country" was clearly aimed at taking advantage of positive network effects and eliminating competitors by refusing interoperability with their system. To avoid costly duplication of networks, telephone was granted protection from antitrust laws in 1921 so that mergers and technical convergence could take place. Later on, with the federal communication commission (FCC) created in 1934 and the Telecommunications Act of 1936, the Bell monopoly became regulated but on costs only. Except for the preamble of the law, there was no mention of affordable rates, a notion linked to willingness to pay of poor households.

Mueller (1997) claims that universal service and the underlying cross-subsidization was made popular by AT&T in its defense against antitrust allegations by claiming that universal household telephone penetration never would have existed without monopoly and regulatory subsidies. A close examination reveals that rate subsidization started in 1965 when more than 70% of residential households had been connected without any help from the government. The practice evolved out of the debate over how to properly separate costs from long distance that are under federal regulation from those of local services which are under state regulation. This author thus holds a "regulatory capture" view of universal service and conclude that one could dispense with it. This run opposite to the general view in continental Europe where universal service is seen as an atrophied version of the traditional public service.

Conclusion

At the outset of this section on practical regulation, the reader may get the feeling that regulation does not work well. This is entirely correct but it does not constitute a reason for its removal for we miss the counterfactual: what prices and business behavior would we observe if a public service (without close substitute) was completely deregulated ? The answer is still unknown since all governments at local, regional and country levels have always regulated those activities even in the UK or US, the staunchest supporters of unbridled competition.

The one exception we may cite is telephone. The decrease in the cost of setting up a mobile phone network over the last decades has allowed entry of many players and generated an intense competition which in turn has spilled over the landline monopoly. In the UK, the regulator Ofcom, after forcing a vertical separation between landlines maintenance and retail upon the former monopoly, has succeeded to create a competitive market. The only remaining restraint is a PSO whose cost is in line with the benefit it bestows upon its holder.

Chapter 18

Natural Resources

Natural Resources, whether renewable like forests and fisheries or exhaustible like oil and minerals are commodities of great importance to our material well being. Even though advanced economies rely more and more on services (cf. Figure 1.1), they have not yet managed to dispense with natural resources to build the basic staples of modern life (car, plane, TV, PC, iPod, games, ...). Technological progress allows a more efficient use of natural resources and thus a smaller per capita demand; but in the aggregate, this saving is thwarted by population growth. Because natural resources offer a limited supply over the long run, they are deemed strategic by economic actors, private and public alike. They are thus worthy of our attention.

Another interesting feature of natural resources is how they differ from traditional private goods such as cars, electricity or pasta. Indeed, nature is directly involved in their production whereas the former use only human made factors. This idiosyncrasy requires a specific economic analysis taking into account how nature is involved.

We are ultimately warranted to include natural resources in this text for two reasons. The involved industries display either scale economiesor externalities so that we either find market power or the need for regulation. EC (2008) reports, p12, that, out of 4000 (non-fuel) mining companies with production facilities including mining, smelting and refining, the 10 majors make about 83% of the total value whilst the remaining 17% is accounted for by about 1000 medium sized and small companies. On the concentration front, PWC (2008) reports on the numerous M&A operations in the sector.

18.1 Exhaustible Resources

Exhaustibility of crude oil or natural gas is a concern insofar as the depletion of low-cost deposits could become soon a reality and then restrict the long-run growth potential of the economy. Other exhaustible resources are underground aquifers and the earth's capacity of the atmosphere to absorb emissions and waste. Stocks are classified into

proven, probable and possible reserves as a way to account for increasing costs of extraction and the uncertainty surrounding their scale. As a matter of fact, the size of the total reserves of most resources is still unknown.

18.1.1 Identities

Intertemporal Optimization

Resources, whether exhaustible like oil or renewable like a fishery, involve stock and flow, thus time is of the essence for their management. To solve for the intertemporal conflict between present and future use, we follow Weitzman (2003)'s canonical model of intertemporal capital accumulation inspired by Gray (1914) and Ramsey (1928).[1@]

Consider a decision maker managing a lasting resource called capital, of size k, through the savings rate $s = \frac{dk}{dt} = \dot{k}$.[2@] The latter may be investment in physical or human capital, biomass natural growth, extraction, catch or retirement. Capital is required to remain positive and cannot be destroyed although obsolescence or depreciation tends to shrink its stock.

The objective of the decision maker in the current period is the value function $v(k, s)$. Capital is assumed to be worthwhile, thus $\frac{dv}{dk} > 0$ but to obtain more capital in the future, we must agree to a current sacrifice, hence, $\frac{dv}{ds} < 0$. The decision maker's time preferences are characterized by the discount rate r (cf. §19.1.2 on time preferences). The intertemporal objective is thus the sum of discounted values to be received in the future: $V = \int_0^\infty v(k_t, s_t) e^{-rt} \, dt$. The FOC characterizing the optimal path is the Euler-Lagrange equation (derived below):

$$\frac{dv}{dk} = \frac{d}{dt}\left(\frac{dv}{ds}\right) - r\frac{dv}{ds} \quad \Leftrightarrow \quad v_k = \dot{v}_s - r v_s \tag{18.1}$$

Applications

In all our applications, human or natural forces make capital grow at the rate $\Phi(k)$, net of depreciation (natural or technical obsolescence). Investment is then growth net of extraction or output i.e.,

$$s = \Phi(k) - q \tag{18.2}$$

Output $q = s - \Phi(k)$, in turn, is the main source of value for the agent although it may be influenced by the stock of capital, hence we have $v = \pi(k, q)$. Combining with (18.1)

and (18.2), we obtain[3@]

$$\underset{\substack{\text{technical}\\\text{discount}}}{\Phi_k} = \underset{\substack{\text{subjective}\\\text{discount}}}{r} - \underset{\substack{\text{stock}\\\text{effect}}}{\frac{\pi_k}{\pi_q}} - \underset{\substack{\text{capital}\\\text{appreciation}}}{\frac{\dot{\pi}_q}{\pi_q}} \tag{18.3}$$

Table 18.1 gathers several applications that are commented afterwards.

Variable	Firm ①	Growth ②	Exhaustible ③	Renewable ④
value	profit	utility	profit	profit
k	capital	capital	reserve	biomass
q	profit	consumption	extraction	harvest
$\Phi(.)$	output	production	discoveries	recruitment
π	q	$u(q)$	$(p-c(k))q$	$R(q)-C(k,q)$
$v(k,s)$		$u(\Phi(k)-s)$	$\pi(k,\Phi(k)-s)$	$\pi(k,\Phi(k)-s)$
L		labour	...	fishing effort
FOC	$\Phi_m = r$	$\Phi_m = r+\eta\frac{\dot{q}}{q}$	$r\pi_q = \dot{\pi}_q$	$r\pi_q = \pi_k + \pi_q\Phi_m$

Table 18.1: Nomenclature for growth models

① An entrepreneur owns a technology with obsolescence factor α generating net revenue $R(k)$ (itself arising from some profit maximizing market behavior), thus profit is $\Phi(k) = R(k) - \alpha k$.

② In the neoclassical optimal growth model, $\dot{\pi}_q = \frac{du'(q_t)}{dt} = u''\dot{q}$, hence $\frac{\dot{\pi}_q}{\pi_q} = \eta\frac{\dot{q}}{q}$ where $\eta \equiv \frac{-qu''}{u'}$ is the elasticity of marginal utility with respect to consumption and $\frac{\dot{q}}{q}$ is the growth of consumption along an optimal trajectory. In the Ramsey-Keynes approach, philosophical reasons lead to pick $r = 0$. If we further look at consumption and not utility (utilitarianism), then $u(q) = q$ and we fall back on ①.[4@]

③ For exhaustible, discoveries are often nil (case $\Phi = 0$) and the FOC involves time because the price must grow i.e., $p = p_t$ (see later).

④ For renewables, the FOC has no time derivative because demand and extraction technology are assumed steady.

Optimal Path

In the long run, the optimal path tends to a steady state found by dropping the last term from (18.3) (cf. below for an intuitive characterization). When the objective is linear in

investment as in problems ①, ③, ④, a most rapid approach (MRA) is optimal because every moment spent in a state where the "steady" FOC is violated yields a lower NPV than might otherwise be obtained. One should thus build the asset (or stop extraction) when the stock is below the optimum because the asset value is larger than the opportunity cost r. On the contrary, if the stock is above the optimum, extraction at full capacity or non renewal of obsolete equipments is warranted i.e., the asset should be realized (sold) rapidly. Insofar as there are no severe limits on the maximum rates of investment and divestment, we may directly focus on the optimal steady state (and its "steady" FOC). In practical terms, only transaction costs relative to the implementation of the optimal steady policy that will slow its advent.

For the Ramsey optimal growth problem ②, the diminishing returns to consumption imply that we refuse to give away much consumption now in favor of having more later;[5@] because of these adjustment costs, the optimal path approaches the steady state in a slower fashion as shown in Weitzman (2003).

Euler Equation

Following Sun (2005), we derive intuitively the Euler equation (18.1) using an intertemporal version of the equi-marginal principle seen in §2.1.1 (cf. formal proof in appendix).

At some time τ, we consider whether to "consume" a small capital amount ϵ by reducing investment at rate Υ during a fraction of time ϵ/Υ. The direct effect is a gain, in NPV terms, of $\beta = -e^{-r\tau} v_s \epsilon$. The indirect effect is due to the permanent stock decrease, hence a loss of $\alpha = \int_\tau^\infty e^{-rt} v_k \epsilon \, \mathrm{d}t$. When optimizing the entire investment path, we ought to perform such small changes at the moment τ where the combined effect $\alpha + \beta$ is maximum. By definition of an optimal path, this effect becomes time-independent and the time derivative of $\alpha + \beta$ must be nil. Since $\frac{d\alpha}{d\tau} = -e^{-r\tau} v_k \epsilon$ while $\frac{d\beta}{d\tau} = -re^{-r\tau} v_s \epsilon + e^{-r\tau} \dot{v}_s \epsilon$, we obtain $0 = \epsilon e^{-rt} (\dot{v}_s - r v_s - v_k)$ which simplifies into (18.1).

The optimal steady stock is also very easy to characterize because the NPV of a steady stock simplifies into $v(k,0)/r$. If we choose to increase k by ξ, we suffer an immediate loss of $v_s \xi$ but value, from then on, will be increased by $v_k \xi$. The NPV of this steady increment is $v_k \xi / r$. The optimal steady stock thus equates these margins i.e., $r = \phi(k) \equiv \frac{v_k(k,0)}{-v_s(k,0)}$, the stationary rate of return on capital (assumed decreasing); it represents the rate at which present value (money) can be transformed into perpetual flows of value.[6@] Weitzman (2003) further proves that if v is concave then starting from the unique stationary solution, it is optimal to remain forever in the stationary state i.e., there is no path (stationary or not) that can improve the NPV.

18.1.2 Optimal Resource Extraction

Hotelling (1931) studies optimal resource extraction from three points of view, a competitive firm, a monopolist, a social planner. Let us denote r the risk-free rate paid by financial markets, c_0 the current average extraction cost and c_t the anticipated average extraction cost for time t.

Hotelling rule of arbitrage

Imagine that a single unit can be extracted and sold. If the price is constant (or not expected to change in the future), sale is immediate if and only if the cost of extraction is lesser than the price. When the price tends to rise with time, extraction takes place at a time determined by the following basic financial arbitrage argument:

- Selling now (at time t) and investing in the financial market generate a cash flow $(1+r)(p_t - c)$ in the next period.
- Postponing extraction for one period yields $p_{t+1} - c$.

The optimal time for selling is thus when these two options yield identical returns or alternatively, when the external interest rate is equal to the internal marginal net return on postponing extraction: $r = \frac{p_{t+1}-p_t}{p_t-c}$. If, for instance, the price increases at rate $\alpha < r$, then sale occurs at the trigger price $p^* = \frac{rc}{r-\alpha} > c$.

More generally, the instantaneous profit from extracting an amount q_t with technology $C(.)$ and selling (competitively) in the market at price p_t is $\pi(q_t) = p_t q_t - C(q_t)$. The previous arbitrage remains valid for the owners of the firm. At time t, we can extract an additional unit to earn the marginal profit π_q and invest the proceed in the financial market in order to receive the cash flow $(1+r)\pi_q$ in the next period. If instead, we wait for the next period to realize this unit, we earn $\pi_q + \dot{\pi}_q$, where the second term denotes the appreciation (or depreciation) of the underlying asset with time. Profit maximization requires the absence of intertemporal arbitrage, thus that the financial return on a marginal additional extraction $r\pi_q$ be equal to the capital appreciation $\dot{\pi}_q$ i.e.,

$$r = \frac{\dot{\pi}_q}{\pi_q} \tag{18.4}$$

Given that extraction tends to display decreasing returns to scale around the optimal level of activity ($\pi_{qq} < 0$), the firm is able to continuously fine tune extraction in order to achieve (18.4) which a particular case of (18.3).

In our particular case, we may solve this equation to find $\pi_q = \lambda e^{rt}$ where λ is the shadow price (in present value) of the initial stock of the resource. This is the Hotelling

rule of extraction: *marginal profit grows at the interest rate*

This rule can also be derived using the "manna" trick (cf. §2.1.1): if the resource size is increased by one unit, the latter should be extracted at the time yielding the largest NPV. In the presence of decreasing returns to scale, the NPVs tend to become equated by the application of this rule, hence when extraction is spread over time, an equi-marginal intertemporal principle should hold: the marginal profit, in NPV terms, must be constant across periods. Now, given an extraction path q_t, the latter is $\pi_q(q_t)e^{-rt}$, hence π_q must grow at the rate r.

Solving (18.4) analytically requires a rather crude specification of either the extraction cost or the market price process which we take in turn hereafter.

Competition

A competitive firm takes the current price p_0 and the future price p_t as given. Since a unit of sale made today yields an average profit $p_0 - c_0$, it can be invested at the risk free rate to yield $(p_0 - c_0)e^{rt}$ at time t; the firm is indifferent between selling an additional unit now or later (at time t) if and only if $p_t - c_t = (p_0 - c_0)e^{rt}$. From this basic arbitrage argument, we deduce that the most easily accessible resources are exploited first. Indeed, their associated average cost is smaller, thus yield a greater current profit which commands immediate extraction.[7@]

If the resource owner is a small entrepreneur, it makes sense to replace the market rate r by the individual's rate of time preference. The Hotelling rule now tells us that people who favor present consumption exploit their resources at a faster rate more than those who favor the future.

Efficiency

The instantaneous welfare is the concave increasing function $W(q) = U(q) - C(q)$ where $U(q) = \int_0^q P(x)\,\mathrm{d}x$ and $P(.)$ is the aggregate WTP, deduced from market demand. The discounted welfare associated to an extraction path q_t is thus $V \equiv \int_0^T W(q_t)e^{-r_0 t}\,\mathrm{d}t$. If the resource is overabundant, we can sustain forever the traditional efficient level q^* solving $P = C_m$. If, as is most often the case, the resource comes in a limited supply k, then the consumption path must satisfy $\int_0^T q_t\,\mathrm{d}t = k$ where T is the final date of consumption i.e., such that $q_T = 0$. The "manna" trick (cf. §2.1.1) tells us that the marginal welfare in NPV terms $v_t \equiv \big(P(q_t) - C_m(q_t)\big)\,e^{-r_0 t}$ must be constant across time. Denoting $p_0 - c_0$ this initial value, we obtain $p_t - c_t = (p_0 - c_0)e^{r_0 t}$ where $p_t \equiv P(q_t)$ and $c_t = C_m(q_t)$. We have thus shown that perfect competition is conducive of efficient management of the resource.

The optimal duration T is determined as follows. From the final condition $P(0) -$

$C_m(0) = p_T - c_T = (p_0 - c_0)e^{r_0 T}$, we derive $p_0 - c_0 = (P(0) - C_m(0))\, e^{-r_0 T}$ thus $\frac{p_t - c_t}{P(0) - C_m(0)} = e^{r_0(t-T)}$. Since $p_t - c_t$ increases with time and $P - C_m$ is decreasing with quantity, the efficient consumption diminishes with time. Another proof goes by contradiction: if quantities q and $q' > q$ were consumed at times t and $\tau > t$, they could be switched thereby generating an increase in value because more is consumed sooner.

This optimal path is parametrized by T. Integrating it over $[0; T]$ yields the total extraction as a function of T; equating with the limited resource k, we derive the optimal duration T.[8@]

Monopoly and Oligopoly

The monopoly maximizes $V \equiv \int_0^T \Pi(q) e^{-r_0 t}\, \mathrm{d}t$ under the restriction $\int_0^T q_t\, \mathrm{d}t = k$. By the intertemporal arbitrage argument seen above, $\Pi_m(q_t) = \lambda e^{r_0 t}$ for some $\lambda > 0$. If Π_m is decreasing then the previous study applies: optimal extraction decreases with time and the optimal duration T is determined by the method exposed before.

Discussion

Most mineral resources are exploited in a similar rapid fashion. The owner is either a private firm whose stockholders are foremost interested by short term dividends or the public monopoly from a country ruled by a despot.[9@] In both cases, the decision maker tends to severely discount the future. If furthermore, he is optimistic, he believes that new discoveries will be made and that substitutes will be discovered, so that this particular resource will never become scarce. If this was to be true, the price would never rise to the choke level (it could even decrease). Under such circumstances, it makes sense for the myopic decision maker to extract his resource as fast as possible and invest the returns in financial markets or use them for immediate consumption.[10@] When such a short-sighted behavior is generalized, aggregate supply tends to be large so that the price of the commodity remains low which rationalizes the original belief (until we hit the wall if the discoveries or substitute technologies fail to sprout).

18.1.3 Examples

Constant Returns to Scale for Extraction

If the price is time varying while marginal cost is constant, (18.4) implies that the profit of a single extractor satisfies

$$p_t - c \propto e^{rt} \quad \Rightarrow \quad \frac{p_t - c}{p_0 - c} = e^{rt} \Leftrightarrow \frac{p_t - c}{\bar{p} - c} = e^{r(t-T)} \tag{18.5}$$

where T is the terminal time where the last remaining unit is sold (infinite if there is no choke price $\bar{p}$ such that $D(\bar{p})=0$). Letting k denotes the aggregate stock of the resources (proven reserves), we can derive T by the intermediate value theorem using the equality between total sales and total purchases.

$$k=\int_0^T q_t\,\mathrm{d}t=\int_0^T D(p_t)\,\mathrm{d}t \tag{18.6}$$

We can solve (18.5-18.6) for the linear demand $D(p)=a-bp$ with associated efficient instantaneous consumption $q^*=a-bc$. Since WTP is $P(q)=\frac{a-q}{b}$, we have $p_t-c=\frac{q^*-q_t}{b}$, thus (18.5) $\Leftrightarrow \frac{q^*-q_t}{q^*}=e^{r(t-T)}\Rightarrow q_t=q^*(1-e^{r(t-T)})$. Plugging into (18.6) yields

$$\frac{k}{q^*}=\int_0^T\left(1-e^{r(t-T)}\right)\mathrm{d}t=T-\frac{1-e^{-rT}}{r} \tag{18.7}$$

where the LHS is the myopic time left to enjoy the constant stream that society would extract ignoring the limitedness of the resource. To enable numerical approximation, observe that the foresighted duration T^* solving (18.7) is larger than $\underline{T}\equiv\frac{k}{q^*}+\frac{1-e^{-r\frac{k}{q^*}}}{r}\simeq\frac{k}{q^*}+\frac{1}{r}$, a rough aproximation when $rk\gg q^*$.[11@] For instance, foresight asks to add another 20 years when the myopic time is a century and the rate is 5%. Also, the efficient fraction to extract at time $\underline{T}$ is still 63%; it then decreases rapidly until 5% for the last year.

The monopolist, unexpectedly, can be deemed "the friend of the earth" because he would extract the resource at half the ideal speed for a social planner. Indeed, (18.4) for the same linear demand uses $R_m-c=\frac{q^*-2q}{b}$ instead of $p-c$ in (18.5). Solving yields $q_t=\frac{1}{2}q^*\left(1-e^{r(t-T)}\right)$, thus (18.7) becomes $\frac{2k}{q^*}=T-\frac{1-e^{-rT}}{r}$, hence the approximation $\underline{T}^m$ for the monopolist's optimal duration T^m becomes $\underline{T}^m=\frac{2k}{q^*}+\frac{1}{r}$ which is quite larger than the efficient level.[12@]

Decreasing Returns to Scale for Extraction

If the price is expected to remain constant at level p but extraction displays DRS with $C(q)=F+\frac{1}{2}cq^2$, the FOC (18.4) becomes $-c\dot{q}=r(p-cq)\Rightarrow\frac{-\dot{q}}{p/c-q}=r\Rightarrow\frac{p}{c}-q_t=\lambda e^{rt}\Rightarrow q_t=\frac{p}{c}-\lambda e^{rt}$. To find the closing conditions, we use $k=\int_0^T q_t\,\mathrm{d}t=\frac{pT}{c}-\frac{\lambda}{r}(e^{rT}-1)$. We also have q_T equal to the minimum efficient scale (when marginal and average cost are equal) If there is a last time of extraction T where $q_T=k_T$, then the NPV of final extraction $\left(pq_T-C(q_T)\right)e^{-rT}$ must be equal to the NPV λk_T of the remaining stock. Combining with the Hotelling rule implies that $C(q_T)/q_T=C'(q_T)$ i.e., the last extraction is performed at the minimum efficient scale.

If we now look at the efficient extraction path, we maximize $W(\mathbf{q})=\int_0^T\left(U(q_t)-C(q_t)\right)e^{-rt}\,\mathrm{d}$

under the resource constraint $\int_0^T q_t \,\mathrm{d}t \leq k_0$ where the marginal utility $U_m(.) = P(.)$ is the market WTP for the resource. Working out the same arbitrage as before yields $U_m(q_t) - C_m(q_t) \propto e^{rt}$ and since there is equilibrium on the market for the extracted resource, we have $p_t = P(q_t)$ i.e., the Hotelling rule for private competitive ownership.

CES demand

Assume zero production cost and a CES demand $D(p) = p^{-\varepsilon}$ leading to WTP $P(q) = q^{-1/\varepsilon}$. The previous method does not apply since no finite T is feasible, thus we work with $T = +\infty$ and we solve $p_t = p_0 e^{r_0 t} \Rightarrow q_t = p_0^{-\varepsilon} e^{-\varepsilon r_0 t} \Rightarrow k = \int_0^{+\infty} q_t \,\mathrm{d}t = \frac{p_0^{-\varepsilon}}{\varepsilon r_0} \Rightarrow q_0 = p_0^{-\varepsilon} = \varepsilon r_0 k$ so that at each moment, a fraction $\frac{q_t}{k} = \varepsilon r_0 e^{-\varepsilon r_0 t}$ of the reserves is extracted. Notice that a monopoly would apply the same rule because his marginal revenue is a multiple of P.

18.1.4 Commons

The economic exploitation of earth's assets all share a fundamental characteristic, their limited size.[13@] Meanwhile these resources were thought to be infinite or much larger than aggregate human demand, they were treated as free inputs for the production of intermediate goods and thus ignored by economic analysis.

The modern epoch is however one where the natural limits have started to bind because technology enable intensive extraction and population growth fuels demand for these valuable assets, nowadays called common pool resources (CPR or *commons* for short). The fact that open access or free entry leads to excessive exploitation (wrt. an efficient value maximizing use) has been known in the field of agriculture to classical Greek economists ; nowadays, the dramatic label tragedy of the commons captures the issue.[14@]

Environmental and resource economics developed to address the many problems regarding the management of earth's commons. Within the realm of industrial organization, commons are the fruit of human effort with for instance, the current stock of knowledge, transportation networks (e.g., highway, railway, pipeline) or public buildings (e.g., hospital, school, stadium, concert-hall). Solving for their inefficient over-exploitation is difficult as the issue bears much similarity with the inherent instability of a cartel (cf. §9.1). Both situations display a payoff structure similar to the prisoner's dilemma: if several firms cooperate not to over exploit the common then all get a large payoff but it is always tempting to extract a little more than partners so that in the end everybody overuses the common!

Open Access Equilibrium

The *tragedy of the commons* is given an analytical treatment in §7.2.3; the equilibrium is tilted away from the efficient management of a monopolist as new firms enter the Commons. As $n \nearrow$, individual effort, extraction and individual profits diminish but aggregate effort and extraction both increase. An obvious consequence for the future (not considered in the present model) is a worsening of the extraction technology since the resource becomes scarcer (better sources are exploited first). This process last until the entry cost starts to bind or when the resource become so crowded that marginal extraction becomes negative.

Privatized vs. Free Access

Weitzman (1974b) shows that economic agents enjoying free access to a common derive a higher income than when the resource is privatized. Whenever the proceeds of privatization are unevenly distributed, a frequent occurrence, users of commons end up worse, thus politically oppose privatization. This transaction cost may well explain the pervasiveness of free access for the management of commons.

At the root of a resource's over-exploitation is the fact that free entry occurs until the average return is equated with cost whereas efficient use commands to equate cost with marginal return. This is the dual of monopoly pricing where the monopoly cares for marginal revenue whereas society cares for the average. Technically, two conditions are required for the tragedy to take place. Firstly, the good or service under consideration requires two complementary production factors, capital that comes in fixed supply (in the short term) and labour that is elastically supplied according to wage or return. Next, the technology must display decreasing returns to scale.[15@]

The production technology at resource $i = A, B$ is $q_i = \phi_i(l_i)$ where l_i is the labour input employed at resource i. The market clearing conditions for input and output read $S(w) = l_A + l_B$ and $D(p) = q_A + q_B$ where labour supply $S(w)$ is increasing while product demand $D(p)$ is decreasing.

Under free access, labour is free to go to any resource, thus the law of one price applies and, in equilibrium, there is a common average return w from working in any resource. Since labor is an homogeneous production factor, its return is the same for all units, thus it is the average return $p\frac{\phi_i(l_i)}{l_i}$. Combining these two observations, we obtain the equilibrium condition $\frac{w}{p} = \frac{\phi_i(l_i)}{l_i}$. It then remains to plug this condition into the market clearing equations to derive the equilibrium price $\hat{p}$, wage $\hat{w}$ and labour use $\hat{l}_i$ for $i = A, B$.

In the privatization equilibrium, each resource is managed to maximize profit $\pi_i = p\phi_i(l_i) - wl_i$, thus optimal labour contracting at resource i solves $\frac{w}{p} = \phi_i'(l_i) < \frac{\phi_i(l_i)}{l_i}$ (by the

DRS property). It then remains to plug this condition into the market clearing equations to derive the equilibrium price p^*, wage w^* and labour use l_i^* for $i = A, B$. Since marginal cost equates marginal return, a positive rent $\pi_i^* = p^* \phi_i(l_i^*) - w^* l_i^*$ obtains for each resource. An alternative way to implement this allocation is to set an access fee $\tau_i = \pi_i^* / l_i^*$ for resource $i = A, B$ and leave laborers choose where to employ themselves (after paying the fee).

The privatization equilibrium coincides with the efficient allocation. Indeed, welfare is $W = \int_0^{q_0} D^{-1}(x)\,\mathrm{d}x - \int_0^{l_0} S^{-1}(x)\,\mathrm{d}x$ i.e., the aggregate WTP of produced units minus the aggregate opportunity cost of used labour. Given the market clearing equations, the FOC of efficient labour input for resource i is $D^{-1}(q_0)\frac{\partial q_i}{\partial l_i} = S^{-1}(L)$ which implies the previous FOC $\phi_i^{'}(l_i) = \frac{w}{p}$ since $S(w) = l_0$ and $D(p) = q_0$. Notice that total output is maximum at the efficient allocation. If not, there would exist a labor vector (l_A, l_B) such that $l_A + l_B \leq l_A^* + l_B^*$ (weakly cheaper) while output would be greater i.e., $q_A + q_B > q_A^* + q_B^*$. If so, in at least one resource, we have $\pi_i = p^* \phi_i(l_i) - w^* l_i > p^* \phi_i(l_i^*) - w^* l_i^*$ which is a contradiction to profit maximization.

We now show that laborers working at a resource earn less under privatization than under free access i.e., $w^* \leq \hat{w}$. If, on the contrary, $\hat{w} < w^*$ is true, then labour supply is lesser i.e., $S(\hat{w}) < S(w^*)$, thus, in at least one resource i, we have less employment i.e., $\hat{l}_i < l_i^*$ and as a consequence of DRS a greater average return i.e., $\frac{\phi_i(l_i^*)}{l_i^*} \leq \frac{\phi_i(\hat{l}_i)}{\hat{l}_i}$. As shown before, the privatization equilibrium maximizes total output, thus involves a lower price i.e., $p^* \leq \hat{p}$. Since resource i fetches a rent under privatization, we have $w^* \leq p^* \frac{\phi_i(l_i^*)}{l_i^*} \leq \hat{p}\frac{\phi_i(\hat{l}_i)}{\hat{l}_i} = \hat{w}$ because profit is nil under open access; this is a contradiction so that $\hat{w} \geq w^*$ must hold true.

We have thus shown that free access guarantees a higher wage to every laborer working in the resources and also attracts more laborers to the commons.[16@] It is no surprise then that if the proceeds of privatization are captured or unevenly distributed, there will be a majority of laborers opposed to the reform even though it creates wealth.[17@] In the case of congested networks, this political economy result indicates that access should never be free at times of peak demand (in order to eliminate congestion) and that all potential users should get free passes or rebates in order to compensate them for the lower utility they derive from the positive peak price.

Restoring Efficiency

We have seen that in a regime of open-access, commons are over-exploited. To restore efficiency, we must devise ways to limit the activity of firms. One well known possibility is to tax output or edict constraining regulations.[18@] This way, the extractive technology

worsens and become dearer to operate; firms are thus rationally lead to extract less. This, in turn, increases the economic benefit generated by the CPR (higher efficiency). Unless there is an explicit mechanism to redistribute the tax revenue, profits are likely to fall with the tax. Indeed, the loss of revenues due to taxation is a direct effect while the gain due to higher efficiency is an indirect one and it is rare to see the latter overpower the former.[19@] It is thus necessary to channel some the newly created wealth to firms in order to compensate them and win their support for the tax policy.

An example is the grandfathering of tradable quotas for carbon emissions.[20@] In the same context of carbon taxation, Stoft (2008) proposes the "untax" whereby the entire tax revenue is refunded to firms in a lump sum manner (or at least in a way that is non correlated with the extraction rate of each firm). The formal mechanism achieving this objective is presented in §7.2.4.

18.2 Renewables Resources

18.2.1 Introduction

Sustainable Development

Our everyday actions have a very small impact on the environment and the global amount of natural resources. Yet, once aggregated at the earth level, they become significant. These phenomena were not relevant in the past but it is now clear that pollution and waste of renewable resources or the extraction of exhaustible ones have very costly consequences both in monetary and utilitarian terms. After ignoring them altogether, economic theory has started to address these issues by treating them as negative externalities.

Sustainable development thus refers to economic development that would not endanger the development of our ecological container, the earth. The constraint being difficult to measure, one generally associate an ad-hoc external cost to every damage done to the biosphere.

Biological and Flow Resources

In economics, a natural resource is said to be renewable if its stock has the ability to grow (regenerate) significantly within a human generation.[21@] We distinguish two categories: biological as with animals and plants or flow resources such as solar, wind, waves, tides, streams all of whom derive from the sun's activity. A renewable resource thus grow by

"the gift of nature" whereas physical capital (buildings, machinery) or human capital depreciates with time and can only grow by applying human effort.

Whenever a biological resource (crop or animal) grows on a piece of land under the private property regime, it has a well defined owner who will realize the asset (harvest or slaughter) when it reaches commercial maturity which may differ from the biological one. In that case, the asset is akin to a standard good produced using capital, labour and consumables. The one exception is forest because trees grow slowly so that it may be more profitable to cut a young tree rather than wait more years to harvest more wood out of it.

Except for fish farms, fish (and other aquatic beings) live in the open sea (or rivers) so that the private property regime does not applies since it is quite hard to delineate frontiers on sea.[22@] There is thus open-access in the sense that anyone is free to extract from the population at any level of effort. Gordon (1954) relates this characteristic of fisheries to the problem of excessive extraction and stock depletion (cf. over-crowding in §18.1.4).

18.2.2 Stock, Flow, Extraction

We view a biological resource as an asset and its stock as capital. We consider the law that governs the evolution of its stock as well as the interaction with human extraction from the stock.

Natural Evolution and Equilibrium

The canonical model of intertemporal capital accumulation evolves according to (18.2) where k is now the stock of the natural resource, s the net periodic change, Φ the absolute natural net[23@] growth and q extraction.

We define the natural net rate of growth as $\rho(k) \equiv \frac{\Phi(k)}{k}$. If left to itself (i.e., $q = 0$), the stock evolves according to

$$\dot{k} = s = \Phi(k) \quad \Leftrightarrow \quad \frac{\dot{k}}{k} = \rho(k) \tag{18.8}$$

When the rate of growth is a constant ρ, the solution to (18.8) is $k_t = k_0 e^{\rho t}$ where k_0 is the initial stock size (indeed, $e^0 = 1$).[24@] The problem with this approximation when $\rho > 0$ is divergence towards infinity in the long run.[25@] It is obvious that since the earth is of limited size, no living population can extend indefinitely. For instance, a fishery cannot grow infinitely because the lake or sea has a finite size and is therefore home to a finite amount of food (for the fishes). The net growth rate must therefore be decreasing with the total stock k to reflect the environmental limitations to growth; we speak of

compensation.

A frequently used model is Verhulst (1838)'s logistic equation stating that the ratio of current to maximum growth rate is the degree of freedom of the population i.e., $\frac{\rho(k)}{\rho} = \frac{\bar{k}-k}{\bar{k}}$ where ρ is the underlying (or maximum) growth rate and $\bar{k}$ the carrying capacity. Normalizing the unit of measure for the stock such that $\bar{k} = 1$, we have

$$\rho(k) = (1-k)\rho \tag{18.9}$$

so that (18.8) has now two obvious long term population equilibria (steady states solving $\dot{k} = 0$), namely disappearance and the carrying capacity, the latter being asymptotically stable while the former is unstable. The exact solution to (18.8-18.9) is[26@]

$$\frac{1}{k_t} - 1 = \left(\frac{1}{k_0} - 1\right) e^{-\rho t} \Rightarrow k_t = \frac{1}{1 + \frac{1-k_0}{k_0} e^{-\rho t}} \tag{18.10}$$

The curves for constant growth rate and compensation are shown on Figure 18.1.

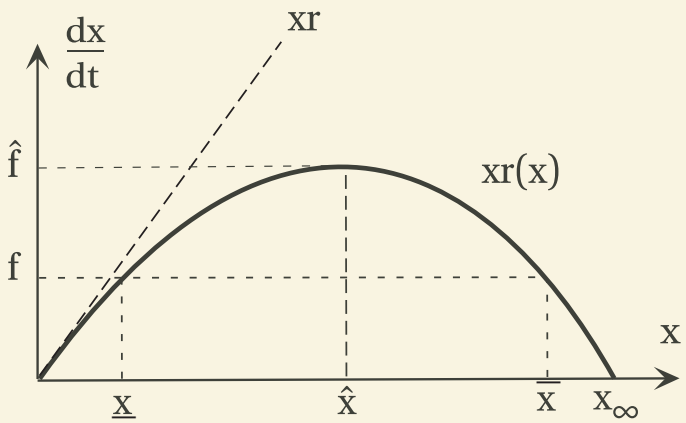


Figure 18.1: Population Dynamic and Extraction

Extraction and Effort

Let us now introduce extraction q from the stock as the aggregate extraction of all firms exploiting the resource. A steady extraction solves $s = \dot{k} = 0$ in (18.2), thus $q = \Phi(k)$.

In the logistic case (18.9), we find

$$q = \rho(k)k = (1-k)\rho k \quad \Rightarrow \quad k = \frac{1}{2}\left(1 \pm \sqrt{1 - 4q/\rho}\right) \tag{18.11}$$

Of these two solutions $\underline{k} < \bar{k}$ shown on Figure 18.1, the former is unstable and the latter asymptotically stable. Indeed, $\frac{dk}{dt} < 0$ as soon as k_t is outside the roots. Hence, if at some

point in time, for an unknown reason, we have $k_t < \underline{k}$, then population will decrease and irremediably converges to zero. In the middle range, $\underline{k} < k_t < \bar{k} \Rightarrow \frac{dk}{dt} > 0$ so that population tends to the larger root ($\lim_{+\infty} k_t = \bar{k}$). Lastly, if $k_t > \bar{k}$ happens, then the fact that $\frac{dk}{dt} < 0$ holds true drives back the population towards the upper root ($\lim_{+\infty} k_t = \bar{k}$). Summarizing, the long term population equilibria are extinction and a positive steady level.

A sustainable yield or steady output is an extraction rate $q = \Phi(k)$ associated with a steady stock k i.e., small enough to enable the stock to maintain itself at a steady level forever. The maximum sustainable yield (MSY) is $\hat{q} \equiv \max_k \Phi(k) = \frac{\rho}{4}$ which stabilizes population at $\hat{k} = \frac{1}{2}$ (50% of the carrying capacity). As can be seen on Figure 18.1 and eq. (18.11), when q increases toward $\hat{q}$, the roots $\underline{k}$ and $\bar{k}$ move toward $\hat{k}$. This also implies that an extraction rate larger than $\hat{q}$ (however small is the difference) leads to the total *depletion* of the stock.[27@]

We now account for the fact that extraction is a human decision made to fulfill some objective. It makes sense to assume that the extraction rate q depends on both the stock size k and the aggregate extraction effort L. Schaefer (1957), in the context of fishing, proposes $q = kL$ (after normalizing the effort unit). A steady effort L is sustainable only if $\dot{k} = 0$ in (18.2) i.e.,

$$kL = q = \Phi(k) = k\rho(k) = k(1-k)\rho \quad \Leftrightarrow \quad k_L \equiv 1 - \frac{L}{\rho} \tag{18.12}$$

with $L \leq \rho$ to guarantee a positive solution. The sustainable output is then

$$q_L = k_L L = L\left(1 - \frac{L}{\rho}\right) \tag{18.13}$$

reaching its maximum, the MSY, at $\hat{L} \equiv \frac{\rho}{2}$. Notice that $L > \rho$ does yield more than the MSY in the short term, but over the long term, such a policy irremediably depletes the entire stock.

Depensation The reproductive technology of the population displays *depensation* if $\rho(k)$ is increasing over some initial interval as shown on Figure 18.2; it means that the underlying population reproduces itself faster and faster (in this range). The population equation (18.12) may then have a large stable solution k_2 and a smaller unstable one k_1 such that if population drops (for whatever reason) to some level $k_0 < k_1$, the population will converge toward zero.

This model also depicts an *hysteresis* effect. If effort $L > \hat{L}$ is achieved (in the absence of regulation) then the population will quickly decrease. A regulation may reduce extrac-

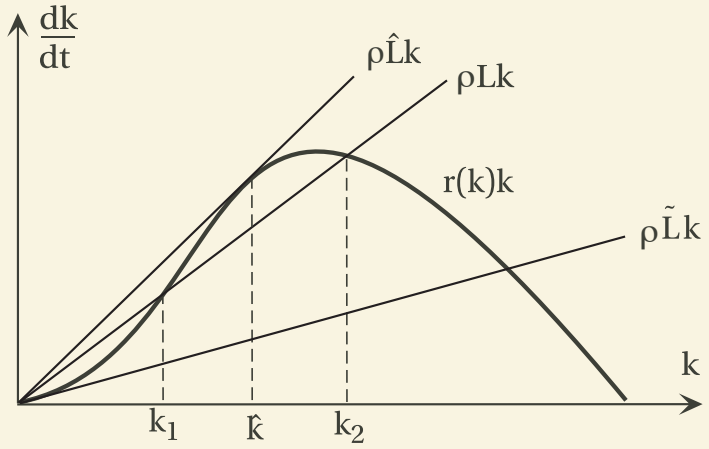


Figure 18.2: Population Dynamic and Depensation

tion effort back to a lower L but this is not sufficient to ensure that the system returns to the level k_2 that enables a sustainable yield; an additional reduction in effort down to $\tilde{L}$ is required because the population has fallen into a perishing trap (below k_1). If the curve binds too much at zero, then once the population has fallen below the critical depensation, it will inevitably come to extinction.

18.2.3 Static Equilibrium

We now use the previous steps to analyze the equilibrium stock according to a variety of human behaviors i.e., whether extraction is a competitive, monopolistic, oligopolistic or socially planned activity. We first look at a crude static optimization where all actors seek to maximize a steady income (net annual revenues).

We assume that extraction efforts provided by active firms are perfect substitutes, so that only the industry aggregate matters for assessing the evolution of the stock. Using the previous characterization of the extraction technology, the industry's cost function is $C(k,q) = wL = \frac{wq}{k}$ for $q \leq \hat{q}$ where w is the (non normalizable) opportunity cost of a firm. Denoting p_0 the market price, profit is

$$\pi = \left(p_0 - \frac{w}{k}\right) q \propto \left(p - \frac{1}{k}\right) q \tag{18.14}$$

where $p \equiv \frac{p_0}{w}$ is the relative price of the output.

Efficiency

Welfare is $W(k,q) = U(q) - c(k,q)$ for $q \leq \hat{q}$ where $U(q) = \int_0^q P(x)\,\mathrm{d}x$ and $P(q)$ is the social WTP for output usually derived from the market demand. Since each (steady) output is

associated with a small and a large (steady) population, it is preferable to go with the large one to reduce cost. Now, if the output of the resource under study is small wrt. the world output then WTP computed at the local MSY $P(\hat{q})$ is likely to be greater than unit cost, hence welfare is an increasing function of output. Insofar as we aim at maximizing income while maintaining stock constant, efficiency commands to extract at the MSY.

Oligopoly

The oligopolistic interaction over a common pool resource is a contest treated in §7.2.3. The FOC for profit maximization, (7.8), is a weighted average between marginal and average benefit.

$$0 = \frac{1}{n}\pi_m + \frac{n-1}{n}\frac{\pi}{L} \tag{18.15}$$

This enables us to identify the monopoly with $n = 1$ (maximizing π) and open-access with $n = +\infty$ (solving for $\pi = 0$, the free entry condition). Using (18.14), the solution to (18.15) is $L_i^n = \frac{1}{n+1}\frac{p-1}{p}\rho$ and total effort at the Nash equilibrium is

$$L^n = \frac{n}{n+1}\frac{p-1}{p}\rho \tag{18.16}$$

leading to steady stock $k^n = 1 - \frac{L^n}{\rho}$ using (18.12). Lasltly, using (18.14), the equilibrium profit simplifies into $\pi^n = (pk^n - 1)L^n = \frac{(p-1)^2\rho}{(n+1)^2 p}$. Industry profit is thus $\Pi^n = n\pi^n$ and it is a matter of algebra to check that the ratio of oligopoly to monopoly profits is $\frac{4n}{(n+1)^2}$ as in the standard Cournot model (cf. §5.12).

Notice that no extraction takes place if $p < 1$ since it is not worthwhile, independently of the industry structure. If the absolute output price p_0 rises or technological innovation reduces cost w, then the equilibrium extraction increases.

Monopoly

When the resource is appropriated and successfully managed by a single decision maker, its economic value can be maximized. As can be checked from (18.16) for $n = 1$, the monopolist's optimal effort is $L^s = \frac{p-1}{2p}\rho$ leading to a steady stock of $k^s = \frac{p+1}{2p}$ and a maximum economic yield (MEY) $q^s = \frac{p^2-1}{4p^2}\rho < \hat{q} = \frac{\rho}{4}$, the MSY. We may thus conclude that the static monopolist refrains from extracting too much. Notice that when either the output price p_0 rises to infinity or effort cost w vanishes, the MEY converges towards the MSY from below (since p diverges).

Open-Access

When the resource is under an open access regime, any firm or entrepreneur is free to extract from it. If furthermore, there are many small firms then we can assume competitive behavior. The process of entry and exit will then drive the individual firm profit towards zero.[28@] As can be checked from (18.16) for $n = +\infty$, total effort[29@] is twice the monopolist choice i.e., $L^o = 2L^s$; it leads to a steady stock $k^o = \frac{1}{p}$. Over-extraction i.e., $L^o > \hat{L}$, is bound to happen when the relative output price increases indefinitely because $L^o \underset{p\to\infty}{\to} 2\hat{L}$; this has been the case in the history of most fisheries as shown early on by Gordon (1954).

To counter such a negative outcome, a cooperative reduction in effort from $L = A$ to $L = B$, as shown on Figure 18.3 would lessen cost and increase yield in the long term. The problem is that the short term effect of a decrease in effort is a fall of the catch; it lasts for the time necessary to rebuild the fish population that will later enable a more comfortable catch for less effort. This poses the problem of weighting the future (discount factor) to compensate the fishermen for their temporary losses.

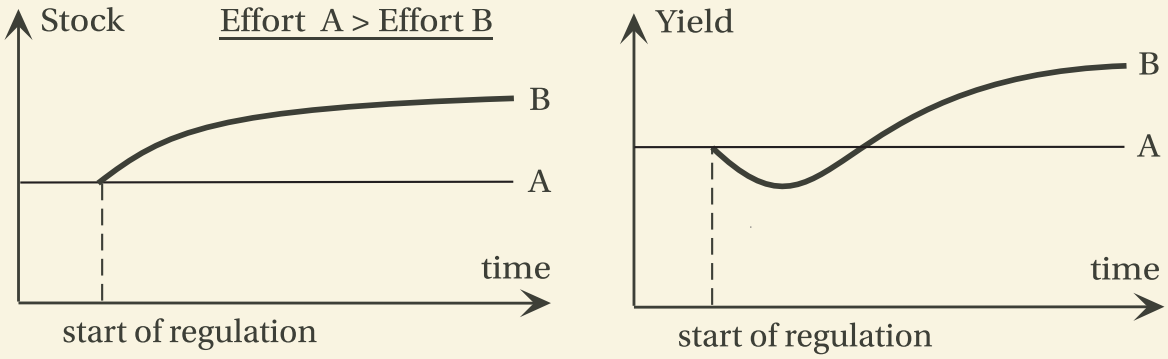


Figure 18.3: Effect of a regulation

The solutions to restore efficiency in the use of the resource are difficult to implement. If the government limits entry using quotas for fishermen, it may reduce catch in the short-term. For once, this benefits only incumbents because the price is driven up by the limited supply. The main problem thought is that incumbents will quickly increase their fishing capacity because the forces that drive over-fishing have not changed at all. Thus, an outcome identical to the open-access equilibrium takes place. Furthermore, incumbents are likely to waste financial and human resources to lobby the government to maintain the barriers to entry in their activity. The quota must therefore apply to boats (measured by their weight) rather than businesses.

An important policy implication of the open-access regime is that firms (or individuals) who stay in the business have the lowest costs, thus the lowest outside opportunity

for doing another job leading to significant social problem when trying to regulate this activity. Notice that only infra marginal firms (those with lower cost) obtain an economic rent from extraction.

18.2.4 Dynamic Equilibrium

The analysis so far looked at steady incomes obtained by maintaining a steady stock. Any economic agent exploiting the resource prefers current to future income and is thus tempted to extract more right away. Following the initial effort of Scott (1955) and Clark and Munro (1975), we flesh out this intuition in our simple set-up.

Monopoly

Since the objective is linear in stock, we can concentrate on the steady state version of (18.3) stating that the discount rate of return must equate the renewal rate plus the stock effect:

$$r \quad = \quad \Phi_k + \frac{\pi_k}{\pi_q} \tag{18.17}$$

Using $\pi = \left(p - \frac{1}{k}\right) q$ from (18.14) and $q = \Phi(k) = k\rho(k) = k(1-k)\rho$ (cf. 18.9) at the steady state, (18.17) becomes

$$r = \rho(1-2k) + \frac{(1-k)\rho/k}{p - 1/k} \quad \Leftrightarrow \quad \eta = 1 - 2k + \frac{1-k}{pk-1} \tag{18.18}$$

where $\eta \equiv \frac{r}{\rho}$ is the bionomic growth ratio. Re-arranging (18.18), we obtain the following quadratic equation in k

$$\frac{1}{\eta}\left(1 + \frac{1}{p} - 2k\right) = 1 - \frac{1}{pk} \tag{18.19}$$

where the LHS is decreasing linear while the RHS is an increasing hyperbola. The intermediate value theorem then applies to yield a single positive solution k_η since $LHS|_0 > 0 > -\infty = RHS|_0$ and $LHS|_\infty = -\infty < 1 = RHS|_\infty$.[30@] Note also that the price must be greater than unity for otherwise the RHS and LHS would have opposite signs. Given this geometrical characterization, the comparative statics are quite obvious: the optimal stock is decreasing in both the bionomic growth ratio and price since the LHS falls as either η or p rise while a rise in p pushes the RHS towards its limit.

We can relate the optimal dynamic rule k_η to the static one, k^s and the open-access one, k^o by observing that (18.19) $\Leftrightarrow 2k(k^s - k) = \eta(k - k^o)$; thus we have $k^s = k_0 > k_\eta > k_{+\infty} = k^o$. Hence, the extremes are when one cares exclusively for the future ($r = 0$) or for the present ($r = +\infty$). This identification means that the optimal policy in a dynamic setting

gears towards one or the other according to the discount rate of the decision maker.[31@]

The long term yield $q_\eta = \Phi(k_\eta)$ is depicted on Figure 18.4. Several curves are depicted according to the bionomic ratio. For $r \ll \rho$ (i.e., $\eta \ll 1$), extraction increases with price. At the limit where $r = 0$, it converges to the MSY for an infinite price. Now for r close or larger than ρ, extraction rises with price up to the MSY at $k_\eta = \frac{1}{2} \Leftrightarrow p = 2 + \frac{1}{\eta}$ (cf. eq. (18.19)) and then start to decrease as a consequence of over-extraction triggered by a too attractive price. If the demand is elastic and large as with D_1 then the equilibrium price is high the consumer surplus small because there is over-extraction. It is only for a low demand D_2 that a more traditional and efficient equilibrium is obtained.

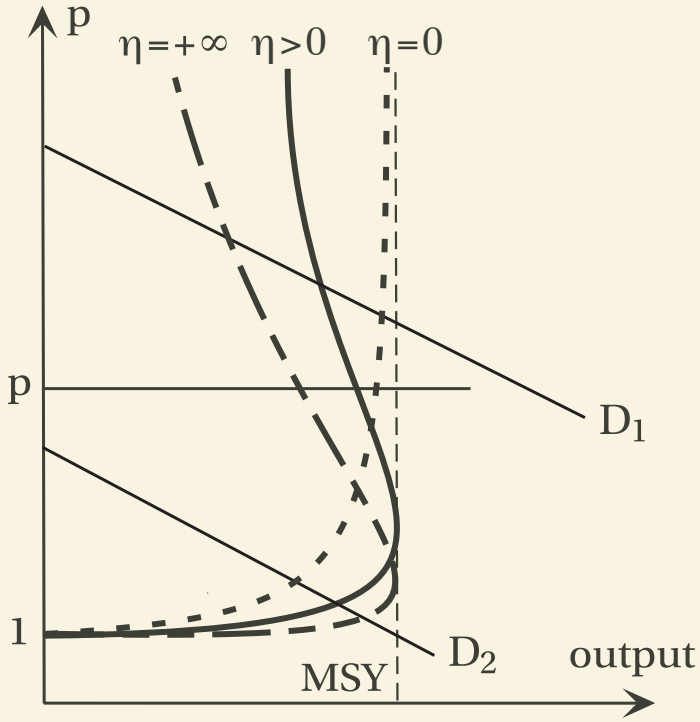


Figure 18.4: Economic equilibrium

Public Management

If we can neglect the externalities linked to the extractive process, welfare in this setting is the gross utility from output minus the extraction cost i.e., $W(k,q) = U(q) - c(k,q)$ where $U(q) = \int_0^q P(x)\,\mathrm{d}x$ and $P(q)$ is the social WTP for output derived from the market demand. Since the social planner maximizes the NPV of W instead of π, (18.17) becomes $r = \Phi_m + \frac{W_k}{W_q}$. Now, we have $W_q = P(q) - c(k)$ as $U'(q) = P(q)$ and $W_k = \pi_k$, hence the only change wrt. (18.18) is to replace the market price p by the WTP $P(\Phi(k))$. Graphically, the social optimum is found on Figure 18.4 by intersecting the plain curve ($\eta > 0$) with the demand

curve.

Regulation of the open-access equilibrium through a Pigouvian tax or transferable quotas is possible in theory but hard to compute as it draws on the knowledge of the natural growth rate ρ. Furthermore it is hard to implement at the political level since the government captures all the economic rent generated by the tax or the auction of quotas.[32@] Technically, the efficient steady output $q^* = \Phi(k^*)$ (cf. (18.2)), if computable, is the global amount of transferable quotas that ought to be awarded or auctioned to the industry. Since the demand for quotas decreases with their price, the intermediate value theorem guarantees that there is a quota price making the overall demand equal to q^*; this price is the sought after Pigouvian tax.

Depletion

Observe that the stock is never fully depleted in the current model. Yet, as soon as the cost of extracting the last unit is finite, open access leads to this fate for a large enough price as shown by Clark (1973). Indeed, for the unit cost $c(k) = \frac{1}{\epsilon+k}$, the free entry equation equation $\pi = 0 \Leftrightarrow p = c(k)$ has solution $k = \frac{1}{p} - \epsilon$ which becomes negative when the price overshoots $1/\epsilon$. The same fate occurs if the rational forward-looking monopolist has a strong preference for the present (case where the LHS of (18.19) tends to zero). Depletion however never takes place under rent maximization because the optimal choice solves $\frac{k}{2k-1} = p(k+\epsilon)$ and tends to one half of the carrying capacity when the price grows indefinitely.

Dynamic Oligopoly

Levhari and Mirman (1980) analyze a dynamic duopoly interaction under the restriction that firms play Markovian strategies i.e., at each period, their extraction depends only on the remaining stock and extraction choice of the competitor. These authors consider a mathematically convenient model: the natural resource grows geometrically at rate $\alpha < 1$ around its long run natural equilibrium (normalized to unity) i.e., $K_{t+1} = K_t^\alpha$; this ad-hoc specification has an embedded stability i.e., when the stock leaves its long term level, it is drawn back to it. We first take a look at the monopoly case.

Monopoly A firm with discount factor δ is granted the right to exploit the resource for T periods. Assuming decreasing returns to scale for the market where the output is sold, we can use the logarithm ln as a valuation function.

Consider first $T = 2$ and denote K the initial stock, a the immediate extraction and b the later one. The NPV of this extraction strategy is $U = \ln a + \delta \ln b \Leftrightarrow e^U = ab^\delta$. Given a,

the stock for the last period is $(K-a)^\alpha$ and it is obviously optimal to extract everything in the last period, hence $b=(K-a)^\alpha$. We then have $e^U = a(K-a)^{\alpha\delta}$, so that the optimal initial extraction is $a^* = \frac{K}{\gamma}$ where $\gamma \equiv 1+\alpha\delta$; we then recover $e^{U^*} \propto \left(\frac{K}{\gamma}\right)^\gamma$.

The result readily extends to any finite number of periods. Indeed, define $\gamma_T \equiv 1$ (complete extraction is optimal in the last period), $\gamma_t \equiv 1+\alpha\delta\gamma_{t+1}$ for $t<T$ and make the induction hypothesis that $a_t^* = \frac{K}{\gamma_t}$ and that $e^{U_t} \propto \left(\frac{K}{\gamma_t}\right)^{\gamma_t}$ for $t>1$. The stock at the beginning of the second period is $b=(K-a)^\alpha$ and by the induction hypothesis $e^{U_2(b)} \propto \left(\frac{b}{\gamma_2}\right)^{\gamma_2}$. Since $U_1 = \ln a + \delta U_2(b)$, we have $e^{U_1} \propto a_1\left(\frac{b}{\gamma_2}\right)^{\delta\gamma_2} \propto a(K-a)^{\alpha\delta\gamma_2}$ so that the optimal extraction is $a^* = \frac{K}{1+\alpha\delta\gamma_2} = \frac{K}{\gamma_1}$ which proves our claim.

We perform the economic analysis of the optimal behavior for a large duration T. Solving for the recursive equation, we obtain the periodic optimal rate of extraction $\frac{a_t}{K} = \frac{1}{\gamma_t} = \frac{1-\alpha\delta}{1-(\alpha\delta)^{T-t+1}}$ whose behavior for large T depends on the sign of $1-\alpha\delta$. The capitalist owner of the monopoly sets $\delta = \frac{1}{1+r}$ where r is the market interest rate. For a biological resource like a fishery, we have $\alpha = 1+\rho$ where ρ is the growth rate of the population. If the natural resources grows so fast to the point that $\rho > r$ holds true, then $\alpha\delta > 1$ meaning that $\frac{1}{\gamma_t}$ is initially very small i.e., it is better to wait for the stock to grow and extract a lot in the future. If the resource is fossil $\rho = 0$ or grows slowly ($\rho < r$), then $\alpha\delta < 1$ and in that case the optimal extraction behavior is to extract a percentage $\frac{1}{\gamma_t} \simeq r-\rho$ of the stock at every period. It is only when the final period approaches that depletion starts i.e., the extraction rate comes close to unity.

Duopoly Since in the last period everybody wants to extract everything, we need to set a boundary condition. The most natural one is equal sharing with $a_{i,T} = K_T/2$. Given stock K_t and challenger extraction $a_{j,t}$, the available stock is $K_t - a_{j,t}$, thus we may use the previous induction reasoning to prove that the best reply is $a_{i,t} = \frac{K_t - a_{j,t}}{1+\alpha\delta_i\gamma_{i,t+1}}$ where δ_i is the individual discount rate used to compute $\gamma_{i,t}$. Using the symmetric formula for the other firm, we obtain a system characterizing the equilibrium; its solution is

$$\frac{a_{i,t}^*}{K} = \frac{\alpha\delta_j\gamma_{j,t+1}}{\left(1+\alpha\delta_i\gamma_{i,t+1}\right)\left(1+\alpha\delta_j\gamma_{j,t+1}\right)-1}$$

$$= \frac{\gamma_{j,t}-1}{\gamma_{i,t}\gamma_{j,t}-1} \simeq \frac{\alpha\delta_j(1-\alpha\delta_i)}{1-(1-\alpha\delta_i)\left(1-\alpha\delta_j\right)}$$

for T large.

The import upshot here is that the standard Cournot result carries one: the resource is over-exploited with respect to the behavior of a monopoly. For $\delta_i = \delta_j$, $\frac{a_{i,T}^* + a_{j,T}^*}{K} =$

$\frac{2\alpha\delta(1-\alpha\delta)}{1-(1-\alpha\delta)^2} > 1-\alpha\delta \Leftrightarrow (1-\alpha\delta)^2 > 1-2\alpha\beta$ which is true. Obviously, the two firms may form a cartel and sustain the efficient extraction path by the threat of reversion to the above stationary strategies in case of defection (aka. trigger strategy).

Part H

Incentives and Information

Information Economics

As recalled by Laffont and Martimort (2002), for centuries, farmers have labored the fields belonging to their landlords, apprentices have worked under the orders of their master craftsman. The motivations behind these behaviors are not obvious to guess. For long, authority was seen as the major explanation but with the development of trade and industry, the economic motivation came to play a role. The theoretical debate is launched by Smith (1776) who criticizes sharecropping, the metayage system, where the farmer and the landlord share evenly the harvest because it acts like a tax that inefficiently reduces the *incentive* of the farmer to apply labour on the land. Hence, whenever the landlord can not monitor the work of the farmer to enforce the adequate amount of effort, we have a clear example of what is now called "moral hazard" because the farmer will shirk or cultivate his personal vegetable garden.

Task delegation occurs because the landlord has a limited time and ability but above all because the division of labour and the specialization enable great economies of scale. Given the complexity of most productive activities today, delegation goes along with the loss of supervision, the inability for the landlord to monitor or control the activities of the farmer. Being left alone performing his task, the farmer learns more than the landlord about all economic aspects of the productive activity he is involved in; he acquires *private information*. Furthermore, the very existence of delegation, means that the farmer is more or less able to orient the activity as he wishes.

The modern theory calls *principals* those who pay, *agents* those who receive payment in exchange for their effort. To ease exposition, we systemically refer to the principal as "she" and to the agent as "he". Their relationship form an *agency* and whenever one party holds (relevant) private information, there is asymmetric information. The landlord–farmer case is only one among the many situations fitting the formalization as shown in the table Table 18.2.

Principal	Agent	Activity	Information
landlord	farmer	cultivation	weather
client	attorney	defense	relation of the case to law
saver	broker	investment	market opportunities
stockholder	manager	industrial policy	market conditions
government	builder	public work	cost
city	utility	water	technology
insurer	policyholder	health	genetic risk

Table 18.2: Examples of Agency Relationships

Classical authors have long noted incentives as an important issue of social behavior but have not made the connection with economics. At the same time, economists viewed information as irrelevant to their domain. It is only recently that incentives have been understood as an information problem that hinges strongly on economic efficiency in many areas. To make sure that the efficient action will be carried out by the agent, the principal must elicit his private information and then convince the agent to perform that action. The latter part is by no means trivial; as we can check from the examples in Table 18.2, the principal can hardly control the effort (e.g., time, money) expanded by the agent in the activity he is getting paid for. From an analytical point of view, we are facing a problem of *hidden information* and/or *hidden action*. The usual terminology nevertheless follows the insurance vocabulary because the first theoretical works were developed in that field.

The classical case of hidden information is when risky people try to get maximum coverage; insurers speak of an *adverse selection* because all premiums will have to be raised and may discourage those at less risk from insuring themselves. Akerlof (1970) introduces the term *Adverse Selection* when citing a 1964 textbook on insurance "There is potential adverse selection in the fact that healthy term insurance policy holders may decide to terminate their coverage when they become older and premiums mount. This action could leave an insurer with an undue proportion of below average risks and claims might be higher than anticipated."

The typical circumstance for hidden action is when drivers could drive with care to avoid accidents but fail to do so once they are fully insured; this behavior is known as *moral hazard*. Health insurance professionals defined moral hazard in the 1960s as the "intangible loss-producing propensity of the insured individual" or the "hazard that arises from the failure of individuals who are affected by insurance to uphold the accepted moral qualities". Pauly (1968) accurately showed that this was no morality problem but a simple consequence of rational economic behavior: by spreading the cost of my health-care over the entire population (socializing), medical insurance makes this

service cheaper to me, thus my demand for it increases.

Chapter 19

Risk and Uncertainty

North et al. (2006) recall that humans have always faced and feared risk, to the point that many of the practices adopted by primitive tribes can be interpreted as insurance mechanisms. Societies have then build market and State mechanisms to smooth risk. In this chapter, we review quickly how one models risk and uncertainty with probabilities and how the standard theory of demand can be extended to account for randomness. We then introduce a basic measure of risk and characterize the way firms and individuals adjust their behavior when exposed to risk. We end with some more advanced results that are useful for the following chapters.

19.1 A Framework for Uncertainty

19.1.1 Introduction

The concept of asymmetric information builds on the more general idea of incomplete information, the fact that economic agents ignore some relevant features and are therefore faced with uncertainty and exposed to risk. To illustrate the importance and ubiquity of risk and uncertainty, Table 19.1 uses data from the US Energy Information Administration regarding the volatility of important commodities. Volatility is a statistical measure of how often the observed price is far away from its mean i.e., it is an assessment of the lack of regularity of the commodity's price.

Commodity	Electricity	Gas	Petroleum	S&P 500	Treasury Bonds	Copper	Gold	Coffe	Sugar	Corn	Cotton	Pork Bellies	Cattle
Volatility (%)	300	80	40	15	13	32	12	40	100	40	80	72	13

Table 19.1: Spot Market Prices Annual Volatility

Let us develop the risk concept with a simple example derived from our typical oligopoly framework: when a manager fails to observe the marketing strategy of his competitor, he lacks a crucial *information* and therefore faces *risk*; his strategy will have to account for this *uncertainty*. Although the market outcome is a deterministic function of the two strategies used by firm A and B, it is random from manager A's point of view. Take for instance the market equation of the Cournot model $p = 1 - q_A - q_B$. If manager A is unsure of what quantity q_B was decided upon by manager B, then the price becomes uncertain since his own decision q_A does not determine completely the final price (it only limits the range of possibilities).

To cope with risk, most people buy insurance e.g., life, car or housing insurance or pension plans; the importance of the insurance sector in developed economies attests of this general desire to diminish risk (cf. Table 1.1). To assess how the presence of risk, uncertainty and asymmetric information impinges on economic choices, it is necessary to extend the standard theory of demand.

19.1.2 Time and Money

Because uncertainty relates to the future and "time is money", we have to account for the opportunity cost of time. We can then build the net present value (NPV) to aggregate

a stream of revenue (or disbursement). This concepts builds on the *subjective* preference for present consumption also called time preference of the investor.

Once we make the mental experiment of viewing today's good as different from tomorrow's, we can transpose the intertemporal allocation of one good between periods into the standard microeconomic analysis of choice (cf. §2.2.1). Thus, the RMS from today to tomorrow should be equated to the price ratio of money at the successive times. Now, if money could flow freely between periods, then the latter ratio ought to be unity. Yet, in most cases, we require compensation to forgo today's consumption which means that one unit today is worth more tomorrow. Note that no market is involved here since the consumer is solving a cake-eating problem i.e., he allocates some amount between the two periods. This means that the relative price expresses his subjective preference for present consumption. This principle valid for two periods readily extend to many using a chain reasoning.

The transposition to finance is straightforward. When an economic agent, whether a consumer or a firm, lends or invests 1€ for a period of time, she forgoes an opportunity for immediate consumption or alternative use; therefore, she requires a compensation in the form of an interest at repayment time i.e., she receives $1+r$ where r is the discount rate. The market interest rate r_0 is an (objective) average of the (subjective) discount rates of all investors.[1@] Typically, an individual will lend (resp. borrow) if $r < r_0$ (resp. $>$); if there is an excess of lending or borrowing, the market rate adjusts to restore equilibrium.

The discount factor $\delta \equiv \frac{1}{1+r} < 1$ is the *present value* (PV) of 1€ to be paid within one period of time. The PV of 1€ to paid within t periods of time is $\frac{1}{(1+r)^t}$; it is found by *compounding* the periodic discount factor. The time period can be a year, month, week, day, hour or minute. The relation between the corresponding rates is found by applying compounding. If r_a is an annual rate, the associated monthly rate r_m is the solution of $(1+r_m)^{12} = 1+r_a$ which, in a first approximation can be taken to be $r_m \simeq r_a/12$.[2@] It is often convenient to consider time as a continuous variable. Notice from the previous calculation that the "per second" interest rate r is so small (in comparison with the annual one) that the formula $(1+r)^t = e^{t\ln(1+r)} \simeq e^{rt}$ becomes exact since $\ln(1+r) \simeq r$ for small r as shown by Hotelling (1925) (cf. proof).

The net present value (NPV) of a series of cash flows $\mathbf{y} = (y_0, y_1, y_2, ..., y_T)$ occuring at periodic intervals is then

$$NPV(\mathbf{y}) \equiv \sum_{t=1}^{T} \frac{y_t}{(1+r)^t} = \int_0^{+\infty} y(t) e^{-rt}\,\mathrm{d}t \tag{19.1}$$

in the continuous time version.

19.1.3 Probability Theory

Up to now in this book, information was distributed evenly among actors i.e., a relevant piece of information was either known to everyone that might be interested by its contents or known to no-one. The key novelty of this Part on information is that some people have *private information* i.e., know more about some relevant facts than others. From the point of view of those in the dark, several alternatives are to be considered and in order to form a synthetic opinion, the individual must balance or weight them according to some criteria.

Probability theory is the mathematical tool allowing the extension of economic concept to this larger world. Acquainted readers can skip this section.

Experiment

Consider Ann (α) and Bill (β) playing a dice game. When the dice rolls on the table, both observe its realization, an integer between 1 and 6. However, if the dice falls from the table on Bill's side, he will be the only only to know; Ann will ignore the true realization and will have to form a belief. The most natural one is to attribute a probability of $\frac{1}{6}$ on each possible outcome. The rational for this belief is the following experiment: throw the dice $n = 100$ times, count n_i the number of times the integer i came out and compute the empirical frequency $h_i \equiv \frac{n_i}{n}$. The law of large number tells us that if the dice is perfectly symmetrical then all six frequencies h_i will converges to $\frac{1}{6}$ as the experiment size grows large ($n \to +\infty$).

Model of the World

The mathematical modeling of the dice experiment speaks of a random variable $\tilde{x}$. The basic events we consider are the possible outcomes $\{\tilde{x} = 1\}, \{\tilde{x} = 2\}, ..., \{\tilde{x} = 6\}$ of the dice draw. A more complex event is $\{\tilde{x} \leq 3\}$ which is the union (denoted $\cup$) of the basic events $\{\tilde{x} = 1\}$, $\{\tilde{x} = 2\}$ and $\{\tilde{x} = 3\}$. The event $\{\tilde{x}$ is odd$\}$ is $\{\tilde{x} = 1\} \cup \{\tilde{x} = 3\} \cup \{\tilde{x} = 5\}$. With the AND condition (denoted $\cap$), we can form the event $\{\tilde{x}$ is even and more than $3\} = \{\tilde{x} \in \{2,4,6\}\} \cap \{\tilde{x} \geq 3\} = \{\tilde{x} \in \{4,6\}\}$. More complex events combine the OR ($\cup$) and AND ($\cap$) conditions with $\{\tilde{x} < 3\} \cap \{\tilde{x} \geq 4\} = \emptyset$, for an impossible event or $\{\tilde{x} \in \mathbb{N}\} = \Omega$, an obvious event, which is always true.

The *objective* probability that the dice shows the integer $i = 1$ to 6, denoted $h_i \equiv Pr(\tilde{x} = i)$ is $\frac{1}{6}$ in this particular example. More generally, a probability distribution is any vector $\mathbf{h} = (h_1, ..., h_6) = (h_i)_{i \leq 6}$ such that $h_i \geq 0$ and $\sum_{i \leq 6} h_i = 1$. The cumulative of $\mathbf{h}$ is $\mathbf{H} = (H_i)_{i \leq 6}$ where $H_i \equiv Pr(\tilde{x} \leq i) = \sum_{j \leq i} h_j$; for instance H_3 is the probability that the dice outcome is

lesser or equal to 3. We speak indifferently of $\mathbf{h}$ or $\mathbf{H}$ as the probability distribution or law of $\tilde{x}$ since $h_i = H_i - H_{i-1}$ i.e., each can be recovered from the other.

Belief

A probability distribution can be used to model an objective phenomenon but also the *subjective* feeling of those who observe the phenomenon with more or less accuracy. Because Ann has no information regarding the dice outcome (the dice fell out of her sight), her belief $\mathbf{h}^\alpha$ is the objective law of $\tilde{x}$ i.e., $h_i^\alpha = \frac{1}{6}$ for all $i \leq 6$ (this is an application of the principle of insufficient reason). If the dice shows number 2, then Bill knows that $\tilde{x} = 2$ occurred; this leads us to say that he holds the belief $\mathbf{h}^\beta$ degenerated at 2 i.e., $h_2^\beta = 1$ and $\forall i \neq 2, h_i^\beta = 0$. Imagine now that the faces of the dice are respectively red and green for even and odd numbers. If a third person Gail (γ) manages to see the color of the dice, say red, but not the figure written on it, her belief will be a refinement of Ann's using Bayes's rule $\forall i \leq 6, h_i^\gamma = \frac{Pr(\tilde{x}=i \cap \tilde{x} \text{ even})}{Pr(\tilde{x} \text{ even})}$ i.e., $h_2^\gamma = h_4^\gamma = h_6^\gamma = 1/3$ and $h_1^\gamma = h_3^\gamma = h_5^\gamma = 0$. The formulation we have adopted is therefore flexible enough to describe any amount of knowledge regarding the outcome of the random variable, from none to all.

Expectation

The expected value taken by the dice is $\mathbb{E}[\tilde{x}] = \sum_{i \leq 6} i \times h_i = \frac{7}{2}$; it is objective because computed using the probability distribution of the random variable (itself determined by the shape of the dice and the laws of physics). Each person present in the room where the dice experiment takes place forms an expectation of the value conditional on his/her belief and private information. For instance, Ann has no particular information (denoted $\emptyset$) and holds the belief $\mathbf{h}^\alpha$, thus she expects $\mathbb{E}[\tilde{x}|\emptyset] = \sum_{i \leq 6} i \times h_i^\alpha = \frac{7}{2}$. Bill, who knows the exact value taken by the dice, expects $\mathbb{E}[\tilde{x}|\tilde{x} = 2] = \sum_{i \leq 6} i \times h_i^\beta = 2$ and Gail who knows that the value is even expects $\mathbb{E}[\tilde{x}|\tilde{x} \text{ even}] = \sum_{i \leq 6} i \times h_i^\gamma = 4$.

Beyond the simple dice example, we may study a discrete random variable $\tilde{x} = (x_i, h_i)_{i \leq n}$ whereby the value x_i is drawn with probability h_i; its expected value is $\mathbb{E}[\tilde{x}] = \sum_{i \leq n} x_i h_i$. One speaks of a gamble when the values are monetary rewards (or penalties). Lastly, note that when the dice is rolled, it only reveals a face which may be interpreted differently by each participant. The same holds true for a gamble since money is not appreciated in the same manner by all. To be able to use the calculus of probabilities in this more general uncertainty framework, we need to assume that each person associates a numerical value x_i to each outcome (face $\#i$) i.e., holds complete preferences over the set of potential outcomes.

Economics

Let us now generalize the probability calculus to economic phenomena. It is fair to say that the price of a stock tomorrow is a random variable $\tilde{x}$ from today's point of view because the interaction of demand and supply in the stock market is the result of so many people that it seems impossible to compute exactly how today's price will be affected to form tomorrow's. The space of outcomes ranges now from 0 to some large number, say 1000€. Since the minimum monetary unit is the cent, there are a hundred thousand possible quotations; for instance $h_{8,23}$ is the probability (from today's point of view) that the stock value tomorrow is $8,23$ €. Given this very large number of possible realizations, it is more handy to assume that $\tilde{x}$ is a continuous variable taking values in $\mathbb{R}_+ = [0;+\infty[$. In that case, $h_{8,23}$ will be written $h(8,23)$ and becomes the probability that the value lies in a small interval centered around $8,23$.

Given our change of perspective, the mathematical object describing a random variable $\tilde{x}$ becomes its *cumulative distribution* H where $H(x) = Pr(\tilde{x} \leq x)$ is the probability that the realization of the random variable $\tilde{x}$ is not greater than x. This definition exactly coincides with that of H_i seen before. More generally, a random variable takes values in $\mathbb{R}$ and its distribution function H need only be positive, increasing and satisfy $H(-\infty) = 0$ and $H(+\infty) = 1$.[3@] For instance, the staircase function $H(.) = 0$ over $]-\infty;0[$, $H(x) = \frac{i}{6}$ for $x \in \left[\frac{i}{6};\frac{i+1}{6}\right[$ and $i = 0$ to 5 and $H(.) = 1$ over $]6;+\infty[$ is the cumulative of the six-face dice throwing. If H is differentiable (which we shall always assume in this book) then $h(x) \equiv H'(x)$ is called the density of the distribution. The expectation of $\tilde{x}$ is then[4@]

$$\mathbb{E}[\tilde{x}] = \int_{-\infty}^{+\infty} x h(x)\,\mathrm{d}x = \int x\,\mathrm{d}H(x)$$

Given any utility function u, $u(\tilde{x})$ is a random variable whose expectation is

$$\mathbb{E}[u(\tilde{x})] = \int_{-\infty}^{+\infty} u(x) h(x)\,\mathrm{d}x = \int u(x)\,\mathrm{d}H(x)$$

We say that an agent holds information θ with respect to the underlying phenomenon $\tilde{x}$ if he believes the random phenomenon to be distributed according to the subjective law H_θ; the latter can be interpreted as an updating of the original objective distribution H upon learning θ. The expectation of $u(\tilde{x})$ conditional on the knowledge of θ is $\mathbb{E}[u(\tilde{x})\,|\,\theta] = \int u(x)\,\mathrm{d}H_\theta(x)$ i.e., the objective distribution H is replaced by the subjective one H_θ.

19.2 Choice under Uncertainty

19.2.1 Expected Utility and Risk

In many instances, there is uncertainty with respect to the future. For instance, one of the most important decisions in our life is the occupation we engage in; shall we try to become an accountant, a broker, an actor, an astronaut or a civil servant? At the time we take the decision, we ignore if we shall succeed, if we shall persist in our choice or if we shall end up in an altogether different activity. Likewise, a firm keeps asking itself, where to invest, how much to invest or whom to employ. In economic terms, the uncertainty that will loom up in the future means that most financial payments received by economic agents are of a random nature. Let us start with the first formal study of decision taking under conditions of uncertainty.

St Petersburg Paradox

The St Petersburg paradox arises from the following game: a fair coin will be tossed until a head appears, say at toss $\#k$, the player then receives 2^k€. How much would you pay to play that game? Although the stakes are very high, people surprisingly bet a small share of their wealth which, at first sight seems paradoxical.

To see that the stakes are high, let us describe the occurrence of the first head toss as a random variable $\tilde{k}$ (it takes integer values). The law of $\tilde{k}$ is given by $h_k = (\frac{1}{2})^{k-1} \times \frac{1}{2} = \frac{1}{2^k}$ as it involves $k-1$ tails followed by one head; it satisfies $\sum_{k\geq 1} \frac{1}{2^k} = 1$ as required to be a probability distribution.[5@] The expected gain is $\mathbb{E}\left[\tilde{k}\right] = \sum_{k\geq 1} 2^k h_k = \sum_{k\geq 1} 1 = +\infty$, yet no sensible person would bid all his/her wealth to play that game!

To solve this conundrum, Bernoulli (1738) follows common sense in positing that that any increment of satisfaction Δu is proportional to the wealth increment Δw i.e., $\Delta u = \beta \Delta w$ for some constant β. Resorting to Aristotle, he also argues that incremental satisfaction is inversely proportional to current wealth which is an expression of the law of diminishing marginal utility. That is to say, the β parameter is not constant across wealth levels, it must be inversely proportional to the current wealth w i.e., $\beta = \frac{\alpha}{w}$ for some constant α. We therefore arrive at the ancestor of all first-order-conditions (FOC) used in economics: $\Delta u = \alpha \frac{\Delta w}{w}$. Assuming that u is a differentiable function of wealth, it reads $u'(w) = \frac{\alpha}{w}$ and after integration, we obtain $u(w) = \alpha \ln(\gamma w)$ where γ is an integration constant that can be eliminated by re-scaling the monetary unit. The plot of u shown on Figure 19.1 is also the first graphical representation made in economics; it emphasizes the relation between income and utility (only the concave part is relevant).

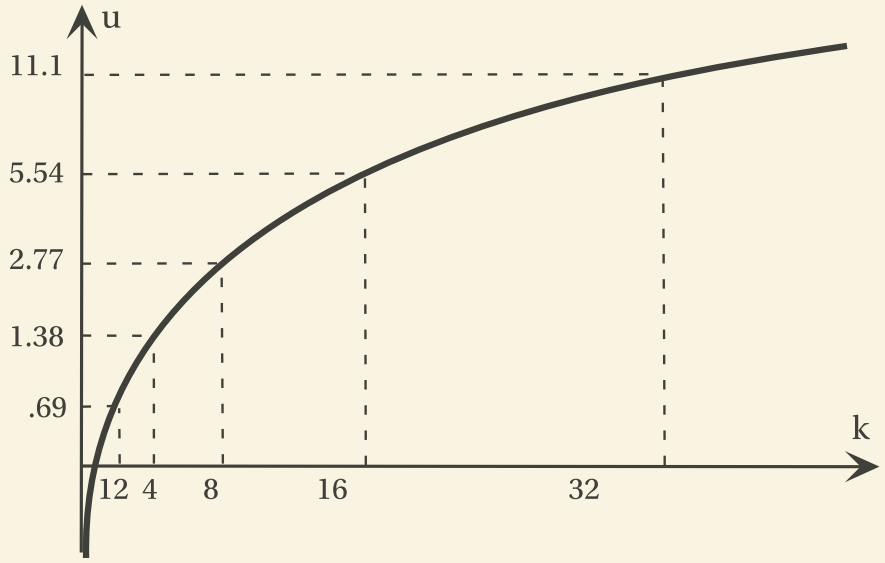


Figure 19.1: Utility of wealth

Moral Expectation

Now, just like the *monetary* value of the gamble $\tilde{k}$ is the sum of prizes weighted by the probabilities of these outcomes $\mathbb{E}[\tilde{k}]$, Bernoulli (1738) proposes the *moral* value of $\tilde{k}$ to be the sum of prizes satisfaction weighted by the probabilities of these outcomes. Thus, an individual with initial wealth w_0 who pays p to play the St Petersburg gamble will obtain a final wealth of $\tilde{w}_1 = w_0 - p + 2^{\tilde{k}}$ (it is also a random variable) and his utility will be $u(\tilde{w}_1) = \alpha \ln(w_0 - p + 2^{\tilde{k}})$; his moral expectation is thus $U(w_0 - p) \equiv \mathbb{E}[u(\tilde{w}_1)] = \sum_{k\geq 1} \frac{\alpha}{2^k} \ln(w_0 - p + 2^k)$. Clearly, if the game was free ($p = 0$) then $U(w_0) > \sum_{k\geq 1} \frac{\alpha}{2^k} \ln(w_0) = \alpha \ln(w_0) \sum_{k\geq 1} \frac{1}{2^k} = u(w_0)$ (since the probabilities sum to unity) i.e., anyone would love to play for free. Conversely, if the player spends all his fortune to play ($p = w_0$) then he expects $U(0) = \sum_{k\geq 1} \frac{\alpha}{2^k} \ln(2^k) = \alpha \ln(2) \sum_{k\geq 1} \frac{k}{2^k} = \alpha \ln(4) = u(4)$ i.e., no one with a fortune greater than 4 would give up everything to play the game.[6@] By the intermediate value theorem, there exists a price p^* such that $U(w_0 - p^*) = u(w_0)$. This is the satisfaction one gets when abstaining from gambling or the willingness to play the game. The maximum fraction of his wealth an individual would rationally agree to pay in order to play the gamble is $\lambda(w_0) \equiv \frac{p^*}{w_0}$. Using a mathematical software, we find $\lambda(4) = 100\%, \lambda(40) \simeq 16\%$ and $\lambda(400) \simeq 2\%$ i.e., wealthy people refuse to risk much of their fortune in order to increase it further.

Von Neumann and Morgenstern (1944) show that expected utility is an adequate manner of extending the rational choice paradigm to uncertainty.

19.2.2 Subjective Risk Measure

Insurance vs. Gamble

When consumers display decreasing marginal utility of income, they are risk averse i.e., we have $\mathbb{E}[u(\tilde{x})] < u(\mathbb{E}[\tilde{x}])$.[7@] For instance, if you have wealth w, you'll refuse a *fair* gamble $\tilde{g}$ i.e., satisfying $\mathbb{E}[\tilde{g}] = 0$ (your future wealth is the random variable $\tilde{x} = w + \tilde{g}$). This behavior constitutes the main explanation to the widespread demand for insurance. Nonetheless, people continue to bet on sports, buy lottery tickets and invest in speculative shares, all of which are unfair gambles. To reconcile this "risk-loving" gambling behavior with the "risk-averse" insuring one, Friedman and Savage (1948) propose to draw the utility of income as shown on Figure 19.2: it displays risk-aversion (concavity) from zero until some level significantly greater than the current income of the individual (say twice); from then on, the curve displays a risk-loving attitude towards risk (convexity).

As can be checked on Figure 19.2, the consumer prefers the certain income x to the fair gamble $\tilde{x}$ that amounts to win or lose δ with identical probability on top of x (so that $\mathbb{E}[\tilde{x}] = x$). To understand why a lottery ticket might appear attractive, recall that lotteries are gambles with a quasi certain small loss (the ticket price) and a small probability of winning a big prize (z on Figure 19.2) that dramatically changes life's opportunities for the better which is why the utility at z is so large. It is then possible that the expected utility of the lottery ticket draws a point on the chord that lies above the utility curve i.e., it was rational to buy it in the first place. This occurs if either the probability of winning is objectively not too small or subjectively inflated. This is indeed the most frequent case: we tend to place a very high value on his new life style upon winning the prize. Since risk loving behavior involves psychological elements, we concentrate on risk-aversion which is more amenable to economic treatment (although the two are mathematically symmetrical one from another).

Risk Premium

Since utility is intangible (the utility function u is not unique), the positive difference $u(\mathbb{E}[\tilde{x}]) - \mathbb{E}[u(\tilde{x})]$ we observe on Figure 19.2 has no particular meaning and fail to measure adequately the amount of risk faced by the individual. We therefore develop a monetary value to express it. As we can observe, the random income $\tilde{x}$ has a *certainty equivalent* $\bar{x} < \mathbb{E}[\tilde{x}]$ solving the equation $u(\bar{x}) = \mathbb{E}[u(\tilde{x})]$; the difference $\mu \equiv \mathbb{E}[\tilde{x}] - \bar{x}$ is the *risk premium* that the investor would agree to pay in order to avoid the risk associated with the random income $\tilde{x}$. This is a subjective value different from an insurance risk premium which is an objective market value.

If we want to emphasize the initial wealth w wrt. the acceptance or refusal of a

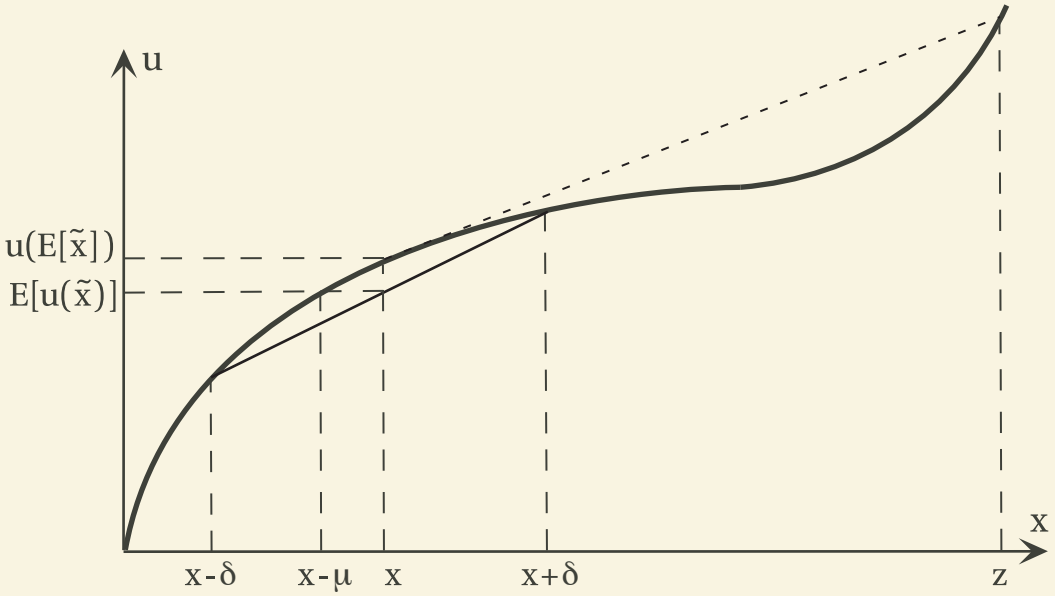


Figure 19.2: Utility of a Gamble

gamble $\tilde{g}$, we use $\tilde{x} = w + \tilde{g}$. Upon sliding the wealth by $\mathbb{E}[\tilde{g}]$, we may assume the gamble to be fair. The solution of $\mathbb{E}\left[u(w+\tilde{g})\right] = u\left(w-\mu\right)$ is then the risk-premium μ associated to the gamble (conditional on the current wealth). Note that our focus on gambles is in fact one about insurance since the act of buying insurance is a rejection of a gamble. Indeed, with insurance one pays the premium for sure and get full coverage for sure. Without insurance, one faces the risk of a large loss or saving the premium.

Degree of Risk Aversion

de Finetti (1952), Arrow (1965) and Pratt (1964) define an individual to be *more risk averse* than another if he refuses all gambles that the former refuses. An agent with monetary utility function u is called Mister u. Let the index of *absolute risk aversion* (ARA)[8@] be $\rho_u(x) \equiv -\frac{u''(x)}{u'(x)} > 0$. The following assertions are equivalent (cf. proof):

① Mister v is more risk averse than Mister u.

② v is a concave transformation of u.

③ Mister v's ARA index is larger than Mister u's ($\rho_v > \rho_u$).

④ Mister v's risk premium is larger than Mister u's ($\mu_v > \mu_u$).

A DARA is a person whose ARA index is decreasing i.e., $\rho'_u \leq 0$. It can be shown that $\rho'_u \leq 0 \Leftrightarrow \mu'_u \leq 0$ (cf. proof). A CARA is person with constant ARA index ρ; her utility function is $u_\rho(x) = -e^{-\rho x}$ for $\rho > 0$ (up to an affine transformation). She is insensitive to wealth because $\mathbb{E}\left[-e^{-\rho(w+\tilde{g})}\right] \geq \mathbb{E}\left[-e^{-\rho w}\right] \Leftrightarrow \mathbb{E}\left[e^{-\rho\tilde{g}}\right] \leq 1$.

The relation between the index of risk aversion and the risk premium can be pinpointed for a CARA person if the random income is normally distributed with mean m and variance σ^2, denoted $\tilde{x} \rightsquigarrow \mathcal{N}(m,\sigma)$. An important property of the normal law is that $\mathbb{E}\left[e^{\tilde{x}}\right] = e^{m+\frac{1}{2}\sigma^2}$. Upon observing that $-\rho\tilde{x} \rightsquigarrow \mathcal{N}\left(-\rho m, \rho\sigma\right)$, we deduce that

$$\mathbb{E}\left[u_\rho(\tilde{x})\right] = -\mathbb{E}\left[e^{-\rho\tilde{x}}\right] = -e^{-m\rho+\frac{\rho^2\sigma^2}{2}} = u_\rho\left(m-\tfrac{1}{2}\rho\sigma^2\right) \tag{19.2}$$

meaning that the risk premium is exactly $\frac{1}{2}\rho\sigma^2$. The ARA index ρ therefore represents twice the "risk premium per unit of variance" (for infinitesimal risk). The previous formula also tells us that an agent with CARA preferences facing risk normally distributed aims to maximize $\mu - \frac{1}{2}\rho\sigma^2$.

> An individual displaying constant absolute risk aversion ρ and whose final income is a random normal variable with expectation m and standard deviation σ act so as to maximize $m - \frac{1}{2}\rho\sigma^2$.

An IRRA person has an increasing relative risk aversion index i.e., $x\rho_u(x) \nearrow \Leftrightarrow -\frac{\rho_u}{x} \leq \rho'_u$. Saturating the constraint defines the CRRA minority; their generic utility is $u_\gamma(x) = \frac{x^{1-\gamma}}{1-\gamma}$ for $\gamma \neq 1$ and $u_1(x) = \log(x)$. In order that the individual displays risk-aversion, u must be concave i.e., $\gamma \geq 1$. A larger parameter indicates more risk aversion.

19.2.3 Objective Risk Measure

Clearly, risk attitude depends on one's risk aversion but also on the underlying risk of the proposed gamble or investment. The notion of "risky investment" is central in the financial sphere, both to private decision makers and public regulators. Yet, until lately, the concept has escaped a precise definition for existing candidates such as *value-at-risk* (VaR) fail one or several coherence tests.

Rothschild and Stiglitz (1970) introduce the concept of first and second order stochastic dominance (resp. FOD and SOD) as follow: $\hat{\theta} \underset{FOD}{\succ} \theta$ if a large result is more probable under $\hat{\theta}$ than under θ and $\hat{\theta} \underset{SOD}{\succ} \theta$ if θ is constructed out of $\hat{\theta}$ by replacing a value with a random variable whose mean is that value. The authors show that FOD amounts to unanimous ranking of gambles by all increasing utility functions while SOD amounts to unanimous ranking of gambles by all increasing *concave* utility functions (cf. proof). In other words, any person with monotonic preferences (the more, the better) would prefer an investment that FOD another and among those risk-averters would all prefer an investment that SOD another. These comparators are thus blameless but at the same time so demanding that few distributions end being comparable.[9@] The recent works of

Aumann and Serrano (2008) and Foster and Hart (2009) solve this issue. We present in an intuitive fashion their riskiness measures we choose to call after *de Finetti* and *Bernoulli* for reasons that shall soon become clear.[10@] Both are homogeneous (the riskiness of twice the gamble is twice larger), sub-additive (the riskiness of the sum of two gambles is less than the sum of their riskiness), convex and monotonous wrt. stochastic dominance (FOD and SOD) i.e., a gamble paying more or with less dispersion is less risky (cf. proofs).

In connection with the above comparability problem, Hart (2010) proves that the two riskiness measures each define a complete order over gambles that extends the partial orders of stochastic dominance when investors/gamblers belong to a natural family of risk-averters defined by Arrow (1965) as follows: acceptance of a gamble increases with wealth but decreases with relative wealth (if both gamble and wealth increase by the same percentage).[11@] Our intuitive presentation will use the fact that the first property is equivalent to DARA and the second to IRRA (cf. proof) i.e., risk-averters descend from DARA fathers and IRRA mothers.

The *de Finetti* measure of riskiness gives precedence to the DARA property and focuses on the people insensitive to wealth, the CARA minority (first investigated by de Finetti (1952)). Given a gamble $\tilde{g}$ with losses and gains,[12@] the solution ρ_g to $\mathbb{E}[e^{-\tilde{g}\rho_g}] = 1$ is, by construction, a CARA index.[13@] Say that Spirou, Gaston and Fantasio are CARA guys with ARA indexes smaller, equal and larger than ρ_g. Then, Fantasio refuses the gamble, Gaston is indifferent and Spirou accepts it, independently of their wealth.

Now consider Seccotine whose ARA index is $\rho_u(.)$. When her wealth is w^* solving $\rho_u(w^*) = \rho_g$, she is locally like Gaston. If $w < w^*$, Seccotine is locally similar to Spirou, while for $w > w^*$, she is more like Fantasio. So, may be, Seccotine should behave like hes alter-egos i.e., refuse the gamble when poor, accept it when rich and be indifferent when her wealth is exactly w^*. Finally, let us examine the people most sensitive to wealth i.e., the CRRA minority.[14@] Member Chiara has an ARA index $\rho_\gamma(w) = \gamma/w$. According to the previous choice rule, Chiara ought to refuse the gamble when her wealth is $w < \gamma/\rho_g$. So, a prudent rule, one that all the Chiaras would abide too ($\forall \gamma \geq 1$), will be to refuse the gamble if wealth does not reach the \$ value $\varphi_g \equiv 1/\rho_g$; it is in this sense that we may deem it an objective measure of the gamble's riskiness. In the case of a normal risk, an extremely simple formula arises: $\varphi_g = \frac{\sigma^2}{2\mu}$ as seen from eq. (19.2). Observe lastly an interesting independence property: if two independent gambles have the same risk, their sum also.[15@] In other words, if you go to a restaurant and consider two wine bottles from distinct regions whose characteristics are unrelated (uncorrelated) and which happen to be equally risky, then if you are ready to go for one, you may order both without taking more risk.

The *Bernoulli* measure of riskiness gives precedence to the IRRA property and focuses on the people most sensitive to wealth, the CRRA minority. Among these, the boldest is Daniel ($\gamma = 1$) who refuses the gamble $\tilde{g}$ as soon as his wealth is lesser than ϕ_g solving $\mathbb{E}\left[\log[\phi_g + \tilde{g}]\right] = \log[\phi_g]$.[16@] Other CRRAs, being more worrisome, will thus refuse the gamble if their wealth is inferior to ϕ_g. To say something about the other people's attitude toward the gamble, note that they are less wealth sensitive i.e., when their wealth increase, they embolden themselves but not so much as CRRA people. So make the following mind experiment: CRRAs used to be poor and have now achieved wealth ϕ_g, yet they still prudently refuse the gamble. Any other person that was equally poor and has become equally rich has nevertheless kept an higher ARA index, thus she remains more risk averse than a CRRA fellow; therefore she will refuse the gamble too.[17@] The *Bernoulli* measure can be computed exactly for binary gambles where one can either lose L or win the larger B. It solves $\left(1 + \frac{B}{\phi}\right)\left(1 + \frac{L}{\phi}\right) = 1$ i.e., $\phi_g = \frac{LB}{B-L}$. If we reinterpret this gamble in term of its size S and mean ϵS, then $L = (1-\epsilon)S$ and $B = (1+\epsilon)S$ so that $\phi_g = S\frac{1-\epsilon^2}{2\epsilon}$ i.e., is inversely proportional to the percentage wedge between loss and gain. When the gamble's distribution has a thin negative tail, the Bernoulli measure tends to the maximum loss.[18@] This makes the measure difficult to apply for laws such as the Normal who have no lower bounds.

Figure 19.3 provides a numerical comparison of the two risk measures: consider a 1$ lottery ticket to win a prize. For a large prize, riskiness is a decreasing concave function of the winning odd, slowly converging towards zero. Conversely, given a small winning odd λ, riskiness is decreasing concave with the prize; it rapidly converges. The bold upper curve is the *de Finetti* measure, the lower one is the *Bernoulli* while the expected gain is the dashed line. The two measures behave similarly although the *de Finetti* one displays a greater prudence towards risk.[19@]

From an investment point of view, the greatest risk is losing the entire capital while unexpectedly high rates of return are treated more like bonanzas. Figure 19.4 shows the comparative risks of an investment of 100€ which yields on average a 50€ return although there is a probability $\epsilon \in [0; 1/2]$ of either doubling the initial outlay or losing all of it. We observe that, ϵ vanishes, the *Bernoulli* risk measure quickly converges towards the initial outlay whereas the *de Finetti* one keeps decreasing as risk vanishes; it even becomes lesser than the expected return although never nil. It thus appears that the *Bernoulli* measure captures better the risk of gambles[20@] while the de Finetti one is more informative for risky investments.

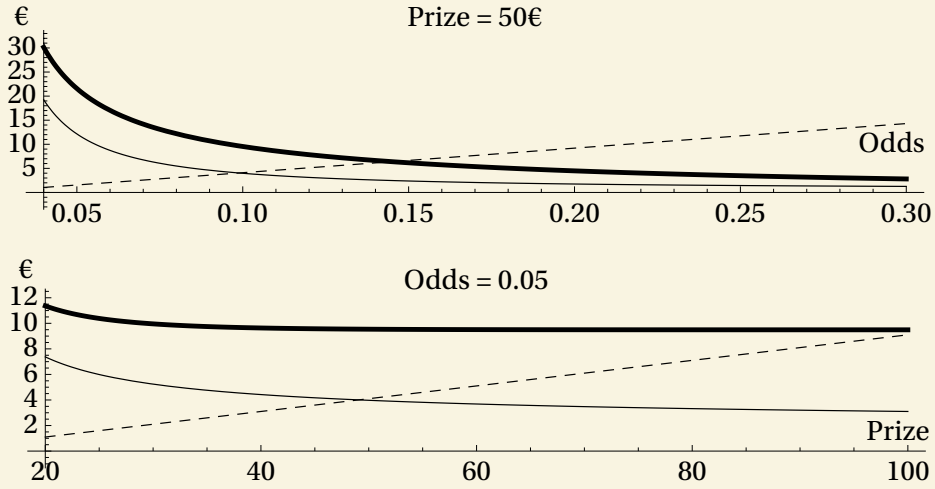


Figure 19.3: Comparing Riskiness measures over a gamble

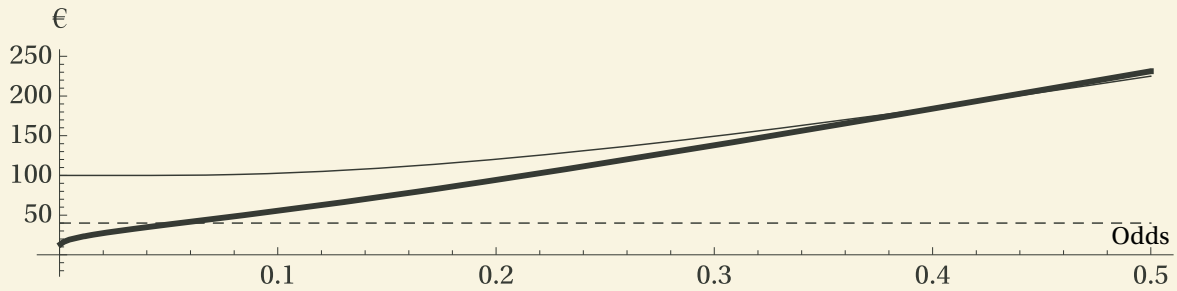


Figure 19.4: Comparing Riskiness measures over an investment

19.3 Firm Behavior under Risk

Firm and Risk

Large profits are highly valued by the stockholders of a firm but they do not necessarily lead to great monetary rewards for the managers or great burst of pride. On the contrary, great losses can lead to destitution or bankruptcy which are very negative outcomes for managers. It is therefore plausible to assume that the representative manager of a firm has a utility function that is increasing with profits but concave i.e., displaying risk-aversion. On top of maximizing profits, one would expect managers to act so as to reduce their risk exposure. We now explain why this is hardly feasible.

The two classical instruments to eliminate risk are negative correlation (do not put all your eggs in the same basket) and the law of large numbers (try to manage many baskets). Regarding correlation, the very existence of a firm originates in the expertise and know-how of its creator regarding a particular sector of the economy. Its activities take place in the same region or the same markets; thus, they are subject to the same

exogenous shocks. In statistical terms, all the revenue generating activities of the firm are positively correlated which means that profits will be highly variable across time. As for the law of large number, non financial firms tend to engage into a limited number of projects that are commensurate with their current size (as measured for instance by liquid assets); this is done to take advantage of scale economies and achieve a high profitability.[21@]

Behavioral Consequences

Sandmo (1971) studies the consequences of the manager's risk-aversion upon his productive behavior within the framework of the neoclassical theory of the competitive firm. Firm's profit is $\pi = pq - C(q)$ while the manager's objective is $u(\pi)$ where the utility function satisfies $u'' < 0 < u'$. In a risky world, the market price is a random variable $\tilde{p}$ with mean p_0 that is determined after the production choice q.

The benchmark corresponds to either risk neutrality (u linear) or absence of risk (constant price); in that case the optimal production is q^* solving $p_0 = C_m(q)$.[22@] In the general case, the manager maximizes $\mathbb{E}[u(\tilde{\pi})]$ where $\mathbb{E}$ denotes the expectation operator corresponding to the manager's belief regarding the distribution of $\tilde{p}$. The FOC of maximization in quantity is

$$\begin{aligned} 0 &= \mathbb{E}\left[u'(\tilde{\pi})(\tilde{p} - C_m(q))\right] \\ \Leftrightarrow \mathbb{E}\left[u'(\tilde{\pi})\tilde{p}\right] &= \mathbb{E}\left[u'(\tilde{\pi})C_m(q)\right] = C_m(q)\mathbb{E}\left[u'(\tilde{\pi})\right] \end{aligned} \tag{19.3}$$

because the marginal cost depends on the deterministic quantity chosen by the manager. Observe now that the covariance of $\tilde{p}$ and $u'(\pi(\tilde{p}))$ is negative[23@] because π is increasing and u' is decreasing. If so then the expectation of the product is less than the product of expectations: $p_0\mathbb{E}\left[u'(\tilde{\pi})\right] \geq \mathbb{E}\left[u'(\tilde{\pi})\tilde{p}\right] = C_m(q)\mathbb{E}\left[u'(\tilde{\pi})\right]$ by (19.3). We have thus shown that $p_0 > C_m$ at the optimum which proves that the optimal quantity under uncertainty is lesser than the certainty equivalent q^*. To conclude

> Risk exposure induces a competitive firm to reduce her output.

A more definite conclusion can reached for the case where the manager's risk aversion is constant (ρ) and the market price is a normal random variable with mean p_0 and variance σ^2. As we already showed in formula (19.2), the manager then maximizes $U(q) \equiv \mathbb{E}[\tilde{\pi}] - \frac{1}{2}\rho\mathbb{V}[\tilde{\pi}]$. We already know that $\mathbb{E}[\tilde{\pi}] = p_0 q - C(q)$ while it is easy to check that $\mathbb{V}[\tilde{\pi}] = \sigma^2 q^2$, thus $U(q) = p_0 q - C(q) - \frac{1}{2}\rho\sigma^2 q^2$ and the FOC for maximization is

$$p_0 = C_m(q) + \rho\sigma^2 \tag{19.4}$$

as if the real marginal cost was inflated by the risk premium $\rho\sigma^2$.

A last observation is that production will really takes place only if the expected utility $\mathbb{E}[u(\tilde{\pi})]$ is greater than $u(-F)$, the utility level in case of zero production. Now, since u is concave, Jensen's inequality (cf. footnote 26.3) tells us that $\mathbb{E}[u(\tilde{\pi})] < u(\mathbb{E}[\tilde{\pi}]) = u(\bar{\pi})$ thus, $\bar{\pi} > -F \Rightarrow p_0 > \frac{C(q)}{q}$. We conclude that

> A competitive risk averse firm to enter a risky market only if it obtains some economic or extraordinary profits which can be interpreted as a required risk premium.

19.4 Advanced Topics†

Optimal Amount of Risk

We would like to know how a consumer reacts to changes in wealth and what distinguish the behavior of people with different risk attitudes. For instance, what is the optimal amount of risk for a risk averse agent when building his portfolio? Shall a very risk averse person avoid all forms of risk? Contrarily to intuition, Arrow (1965) answers negatively whatever the riskiness and the risk aversion. Indeed, everyone is risk neutral with respect to very small changes and the only thing that matters in that case is the expected return. Hence, anyone will invest at least a small amount into a risky asset which is, on average, more profitable than a risk-less asset such as a treasury bond.

To prove formally this claim, consider investing 1€ into a combination of a risky asset whose return is the random variable $\tilde{r}$ and the risk-free asset whose sure return is r_0. We ought to show that the optimal share λ of risky asset is positive. Observe that the final wealth of the individual is $\tilde{w} = (1-\lambda)(1+r_0) + \lambda(1+\tilde{r}) = 1 + r_0 + \lambda(\tilde{r} - r_0)$. Letting H denote the law of $\tilde{r}$, the expected utility is

$$U(\lambda) \equiv \mathbb{E}[u(\tilde{w})] = \int u(1 + r_0 + \lambda(r - r_0))\, \mathrm{d}H(r)$$

The FOC of maximization is

$$0 = U'(\lambda) = \int (r - r_0) u'(1 + r_0 + \lambda(r - r_0))(r)\, \mathrm{d}H(r)$$

and since $U'(0) = u'(1+r_0)\int (r - r_0)\, \mathrm{d}H(r) = u'(1+r_0)(\mathbb{E}[\tilde{r}] - r_0)$, there is an incentive to buy some of the risky asset ($\lambda > 0$) as soon as the risky asset is on average more profitable than the risk-less one.

Arrow (1965) also demonstrates that a less risk averse person invests more into the risky asset (cf. proof). Intuition would suggest that richer people take on more risk; this

is correct for an agent with decreasing absolute risk aversion (DARA) since becoming richer turns him into a less risk averse person (and we can apply the previous result).

Mixing Action and Chance

When an action deterministically generates an outcome, the simple observation of the later enables to infer with exactitude the former. An example is cultivation inside a greenhouse: yield q is directly related to the effort e expanded by the farmer by a relation such as $q = \sqrt{3e}$. Upon observing q, we infer that effort was $e = q^2/3$.

Outdoor cultivation, on the other hand, is subject to weather variations so that a high yield can either reflect fair conditions or hard work and similarly, a poor yield can either reflect laziness or the losses due to a storm. Generally speaking, the action (input) is inaccurately reflected by the result (output). A simple formalization would be $\tilde{q} = e + \tilde{z}$ where e is the effort, amount of daily time spend on the field while the random variable $\tilde{z}$ captures the effect of weather variability; as a consequence, yield is also a random variable $\tilde{q}$.

To capture the interaction of effort and chance, we write $H(q|e) = Pr(\tilde{q} \leq q|e)$ the probability to observe a result lesser than q given that action e was taken. The expected value of $\tilde{q}$ conditional on the action e is then written $\mathbb{E}\left[\tilde{q}|e\right] = \int qh(q|e)\,\mathrm{d}q$ where $h(.|e)$ is the density associated to the distribution $H(.|e)$. Likewise the expectation of the random variable $f(\tilde{q})$ is denoted $\mathbb{E}\left[f(\tilde{q})|e\right]$.

The previous modeling was adapted to moral hazard where the action of the decision maker, called the agent, is hidden to another party, called the principal. In problems of adverse selection, it is a piece of information known to the agent that is hidden from the principal. In that case, the notation is θ instead of e.

Informativeness

In matters of asymmetric information, *inference* is fundamental and can be presented as follows: "given the observation of result x, what is the probability that the action undertaken was θ?"; it will obviously depend on the initial belief we hold regarding the unknown action.

Milgrom (1981) defines result x to be *more informative* than result y about a higher action if the observer believes higher actions more probable upon observing x than upon observing y. This notion of good news can be related to a property pervasive in information economics: the action change from θ to $\hat{\theta}$ satisfies the *monotone likelihood ratio property* (MLRP) if

$$\hat{\theta} > \theta \Rightarrow \frac{h(x|\hat{\theta})}{h(x|\theta)} \nearrow \text{ in } x \tag{19.5}$$

The family of distributions $h(.|\theta)$ is said to satisfy MLRP if (19.5) is true for all parameter values. It is noticeable that MLRP is stronger than FSD i.e., $\hat{\theta} > \theta \Rightarrow H(.|\hat{\theta}) \leq H(.|\theta)$ (cf. proof). To understand the MLRP concept, imagine that the likelihood of observing result y is the same after the two actions i.e., $h(y|\hat{\theta}) = h(y|\theta)$, then result $x > y$ is more likely to appear after the higher action i.e., $h(x|\hat{\theta}) > h(x|\theta)$.

> When a change towards a higher action satisfies the MLRP, greater outputs are signals of greater inputs (but not a proof).

To apply the MLRP concept of informativeness, we define a function f to satisfy the *single-crossing property* (SCP) if it crosses the axis only once and from below i.e.,

$$\exists y, \forall x, (x - y) f(x) \geq 0 \tag{19.6}$$

A fundamental theorem follows: if the family of distributions $h(.|\theta)$ satisfies the MLRP and f satisfies SCP, then (cf. proof)

$$\hat{\theta} > \theta \Rightarrow \mathbb{E}\left[f(\tilde{x})|\hat{\theta}\right] > \mathbb{E}\left[f(\tilde{x})|\theta\right] \tag{19.7}$$

Another characterization (cf. proof) is that a change in distribution of the risky asset increases its demand whatever the risk aversion (u) and whatever the risk-free rate (r_0) if and only if the change satisfies the MLRP.

Chapter 20

Moral Hazard

One of the oldest contractual relationship in agriculture is tenancy. Under *fixed rent*, the farmer pays the landlord a monetary rent every year for using of land while under *sharecropping* he shares the crop with the landlord; alternatively, the farmer can become a laborer to earn a *fixed wage*. The latter formula works well when parties work in team but if the laborer is left without monitoring, he will shirk and the yield will be very low. At the other extreme, the fixed rent motivates the farmer to exploit optimally the fields because he gets to keep all the crop; the landlord can thus ask a high rent because on average the yield will be high. Everything's fine when the weather is good but if the winter is very cold or the summer very dry, the yield might severely drop and leave the farmer without enough seeds to replant, feed his family and pay the rent at the same time i.e, agriculture is a risky activity. This may well be the reason why some people prefer to be laborer than farmers.

The landlord faces a trade-off when dealing with the farmer: *incentives* vs. *insurance*. Sharecropping therefore appears as a solution mixing both features. On the one hand, any additional effort expanded by the farmer will generate a higher yield (on average), half of which goes into his pocket; this is a good motivation for hard work. On the other hand, the farmer does not have to pay a monetary rent; he is less at risk in case of a bad harvest. The landlord is now sharing risk with the farmer.

The plan of the chapter is as follows. We first detail the agency relationship, the basic incentive problem it faces and various remedies to it. We then use managerial incentives to introduce asymmetric information and the resulting inefficiency. The next section is more formal and emphasizes unexpected contingencies. The last section presents some extensions regarding the possible renegotiation of contracts. We use material borrowed from Rasmusen (2006), Macho-Stadler and Pérez-Castrillo (1996) and Boccard (2002).

20.1 The Agency Relationship

Profit maximization calls for the pursuit of many goals such as cost minimization, quality enhancement or good customer relationships. Consider then a manager in charge of one such activity; the attainment of the goal requires work dedication, effort and a personal investment. To induce the manager to perform a significant effort, the owner can link a wage bonus to the attainment of an (explicit) objective. Now, to generate an additional improvement, an additional effort is required which can be implemented by increasing the bonus.

20.1.1 Framework

A fundamental assumption underlying the contract literature is that economic relations take place within a well defined and smoothly functioning legal framework. When economic agents sign a contract, they are bound to respect its terms whenever a court of law is able to understand them. This has two consequences; on the one hand, the contract can only include clauses and obligations that are *verifiable* by the judge or arbitrator but on the other hand, there cannot be any haggling over a verifiable item.[1@]

Preferences

The principal (e.g., landlord) hires an agent (e.g., farmer) to exert an effort e (e.g., daily hours of work). Applied into the production technology, this input gives rise to an output of value $q(e)$ (e.g., crop) where $q(.)$ is increasing and concave i.e., we assume decreasing returns to scale. The profit for the principal is $\pi \equiv q(e) - w$ where the wage w, paid to the agent, can be contingent on *verifiable* events. To simplify the analysis, we assume that the preferences of the agent for effort and income are separable with $U(w,e) = u(w) - c(e)$ where u is increasing concave (decreasing returns to wealth) and c is increasing convex (increasing value of forgone leisure time). This specification has the great advantage that money can be transferred between the principal and the agent to reach any participation constraint. We normalize the utility function so that $u(0) - c(0) = 0$.

Verifiable effort

When the effort can be contracted upon, a simple contract is $(\hat{e}, \hat{w})$ meaning "do at least $\hat{e}$ and you'll get $\hat{w}$ (otherwise nothing)". Recalling, that the moral hazard situation refers to the fact that effort is undertaken by the agent, the latter can either choose some $e \geq \hat{e}$ and derive utility $u(\hat{w}) - c(e)$ or choose some $e < \hat{e}$ and derive utility $u(0) - c(e)$. His best options under the two broad alternatives are thus $\hat{e}$ and 0, as he dislikes effort. The principal will

succeed with his objective of having the agent expand effort $\hat{e}$ if $u(\hat{w}) - c(\hat{e}) \geq u(0) - c(0) = 0$. A first result is thus the obvious observation that the salary must compensate the agent for the toil of working.

Individual Rationality

To simplify matters we assume that a single *principal* is facing a multitude of potential *agents* to whom she makes a "take-it-or-leave-it" offer. This extreme formulation is justified by the excess demand for the position that turns potential agents into Bertrand competitors ready to accept no more than their opportunity cost $\underline{u}$ to get the job; one can think of $\underline{u}$ as a minimum wage below which agents prefer to stay at home. This setting apparently opposes an all-mighty capitalist to an harmless worker; this is only done to ease the mathematical analysis. As we shall see later on, the opportunity cost acts as a slider that enable to share the benefits of the relationship in any proportion between the two parties (cf. §20.1.2 on bargaining). Hence, when $\underline{u}$ is large, it is the agent who has most bargaining power.

To convince the agent to sign the contract in the first place, the principal must offer him an expected utility at least as great as his opportunity cost $\underline{u}$. This *individual rationality* (IR) condition is in fact a participation constraint:

$$u(w) - c(e) \geq \underline{u} \Leftrightarrow w \geq \underline{w}(e) \equiv u^{-1}\left(\underline{u} + c(e)\right) \tag{20.1}$$

Since u and c are increasing, so is $\underline{w}$, meaning that the higher the effort one wishes to implement, the higher must be the compensation.

First Best

The principal can now maximize her objective $\pi = q(e) - w$ over contracts (e, w) satisfying the participation constraint (IR). Since she likes money too, it is optimal for her to saturate the participation constraint (20.1) by restricting attention to contracts $\left(e, \underline{w}(e)\right)$ i.e., pay the minimum acceptable wage. The principal's program thus becomes

$$\max_e q(e) - \underline{w}(e). \tag{20.2}$$

The FOC for (20.2) is

$$q' = \underline{w}' = \frac{c'}{u'} \Leftrightarrow c' = q'u' \tag{20.3}$$

i.e., the marginal utility of taking one minute of rest c' equals the marginal value of money u' times the additional production q' of one additional minute of effort. Since q

and u are both concave, the RHS of (20.3) is decreasing while $c$being convex implies that the LHS is increasing, there is thus a unique solution e^*called the first-best effort. The *first-best* contract is $\left(e^*, \underline{w}(e^*)\right)$ and yields the final profit $\pi^* \equiv q(e^*) - \underline{w}(e^*)$.

20.1.2 Moral Hazard

Most often, effort is too complex to be monitored closely[2@] but there does exists a minimal effort $\underline{e}$ that can be required from an employee i.e., failure to perform at this level is a juridically acceptable reason for firing him without salary.[3@] In such a situation, the wage scheme becomes flat i.e., the agent is paid a salary w.Since his utility is $u(w) - c(e)$, he has no incentive to work harder than the contractual minimum, thus he will perform $\underline{e}$. The minimal wage satisfying the participation constraint is then $\underline{w}(\underline{e})$ and the principal earns $\underline{\pi} \equiv q(\underline{e}) - \underline{w}(\underline{e})$. The literature often assumes $\underline{e} = 0$ to simplify matters but it is important to remember that the principal has also an opportunity cost of entering into the relationship with the agent which plays a symmetrical role to $\underline{u}$.

The very idea that moral hazard is an issue translates into $\underline{e} < e^*$, the fact that the principal would like the agent to perform more than the minimally contractible effort. This also implies that $\underline{\pi} < \pi^*$ (recall that the profit $q(e) - \underline{w}(e)$ is concave with a maximum at e^*).

Verifiable Output

Although effort is not verifiable, it can be the case that the output produced can be contracted upon. Since there is a one-to-one relationship between effort and production, any effort $\hat{e}$ is uniquely associated to the output $\hat{q} \equiv q(\hat{e})$ and to the compensation $\hat{w} \equiv \underline{w}(\hat{e})$. Let us then consider the following contract "produce at least $\hat{q}$ and i'll pay you $\hat{w}+1$€, otherwise nothing". It is readily observed that upon accepting this offer, the agent chooses to perform exactly $\hat{e}$. As he obtains a utility level $u\left(\underline{w}(\hat{e})+1\right) - c(\hat{e}) > u\left(\underline{w}(\hat{e})\right) - c(\hat{e}) = \underline{u}$, he will accept the contract in the first place. Hence the principal can finely tune the contract (reduce the 1€ bonus) to leave no more than $\underline{u}$ to the agent. Choosing $\hat{e} = e^*$ implies that the first-best can be achieved although effort is not contractible.

The previous scheme is however quite sensitive to the precise observation of $q(\hat{e})$ and would fail to work properly if risk was taken into account. A standard *piece rate* scheme can achieve the same outcome; it consist of a base salary $\hat{w}$ and a piece rate factor β so that $w(q) = \hat{w} + \beta q$. Upon signing such a contract, the agent has utility $u\left(\hat{w} + \beta q(e)\right) - c(e)$ and is thus lead to expand the first best effort e^* whenever the principal sets $\beta = 1$ i.e., makes the agent the *residual claimant* of the activity. The base salary is then set so as to satisfy the participation constraint i.e., $u\left(\hat{w} + q(e^*)\right) = \underline{u} + c(e^*) \Leftrightarrow \hat{w} = \underline{w}(e^*) - q(e^*) = -\pi^*$.

Franchising

Up to a slight change in the sequentiality of payments, the previous scheme is identical to franchising or "selling the store". Here, the agent pays the principal π^* up-front (or its expectation if there is uncertainty) and then becomes *residual claimant* of the activity. Although effort may not be contractible, the agent's wage is now $w(e) = -\pi^* + q(e)$ so that his utility is $u\big(q(e) - \pi^*\big) - c(e)$. The optimal effort solves the first-best FOC (20.3) and since $q(e) - \pi^* = \underline{w}(e)$ at e^*, the solution is the first-best effort e^*. The obstacle to such a clever method is *limited liability* i.e., in most cases where moral hazard matters, the agent is poor compared to the principal and cannot pay upfront the potentially large amount π^*.

Bargaining power

It is realistic to assume that in their first encounter it is the principal who makes a "take-it-or-leave-it" offer to the agent and not the reverse. Once the agent has been working for her a while, he has acquired a valuable human capital for the firm and thus may be able to dictate his conditions for the next period.

In that case, the agent will not buy the store but offer the principal the contract (e, w) maximizing his own utility $u(w) - c(e)$ under the constraint that the principal accepts the offer. Turning down the agent, the principal could do the job herself or hire a new agent but in any case she would miss the participation of the old and more experienced agent and she would end up earning a lower profit $\underline{\pi} < \pi^*$; this level $\underline{\pi}$ therefore represents her opportunity cost. The offer of the agent must satisfy $q(e) - w \geq \underline{\pi}$ for it to be accepted by the principal (participation constraint). Like the principal, he likes money, thus proposes the minimally acceptable offer, that saturating her participation constraint.

The agent's utility is thus $u\big(q(e) - \underline{\pi}\big) - c(e)$ and his optimal effort is again efficient.[4@] Under this scheme, the agent reaches a utility level of $u^* > \underline{u}$ since all gains of the economic activity are passed from the principal to him (by duality $\underline{\pi} < \pi^* \Leftrightarrow \underline{u} < u^*$). Hence, by sliding the reservation utility from $\underline{u}$ to u^*, we are able to implement any surplus distribution between the principal and the agent (cf. §2.4.3 on bargaining). The "take-it-or-leave-it" hypothesis is thus best seen as a tool to simplify the analysis without restricting its generality.

Risk and Uncertainty

In most basic moral hazard settings, the agent applies effort on some project or activity which later on yields a result or profit. It is clear that many unpredictable events like the weather or macro-economic shocks interfere with the work of the agent to increase

or decrease the magnitude of his production, to change a winning project into a losing one. Hence the one-to-one relationship between effort and output is lost. This means that incentives based on output force the agent to bear risk: although he worked hard, the resulting performance might be adverse so that pay might be lower than expected. Whenever the agent is risk-averse, the use of incentives becomes more costly to the principal because she must compensate the agent for his risk bearing.

20.2 Managerial Incentives

Before presenting the general model of moral hazard, we develop here a simple model that captures the issues of risk and incentives. We show how contractual terms are optimally distorted towards lower effort to provide the agent with some insurance. The resulting optimal contract is called second-best because the associated profit falls short of the first-best level, the difference is called the agency cost of moral hazard. We then turn to a series of extensions that account for realistic features absent from the base model.

We first derive the optimal incentive scheme as a decreasing function of the uncertainty surrounding the agent and his degree of risk aversion. Next, we show that if the observable performance measure on which incentives can be build is loosely aligned with the principal's real objective then incentives are watered down to impeach the agent to "game" the scheme. Next we analyze succinctly relative performance measure and show that evaluation against peers, although fundamentally a worse instrument than piece rates, does eliminate firm specific risks unlike absolute performance measures; for that reason, "employee of the month" and promotions can become optimal incentive schemes. The next step is job design and whether people ought to work alone or in team, so to say. We show that team work dominates individual accountability when uncertainty is not so much of an issue and technological complexity makes it difficult to trace specific effort from outcomes. Lastly, we deal with implicit incentives or career concerns, the fact that one works harder at the beginning of his professional life to signal high ability and enjoy later on a better position or earnings.

20.2.1 Individual Compensation

To simplify our study of uncertainty we make a number of simplifying assumptions in this section. The equivalent monetary cost of effort is quadratic and after normalizing adequately the measure of a unitary effort, becomes $c(e) = \frac{1}{2}e^2$. The manager has a zero opportunity cost ($\underline{u} = 0$) and constant risk aversion ρ; as we saw in §19.2.2 (eq. 19.2), if he

earns a random normally distributed income $\tilde{x}$, he maximizes $u \equiv \mathbb{E}[\tilde{x}] - \frac{\rho}{2}\mathbb{V}[\tilde{x}]$. We posit a constant productivity γ of labour and an additive noise, blurring the relation between input and output i.e., $q = \gamma e + \tilde{\epsilon}$. We further assume that the noise $\tilde{\epsilon}$ follows a centered normal law with variance σ^2. To conclude, the contractible performance measure is an imperfect signal of effort BUT, in this section, the risk or randomness is independent of effort.

The wage scheme used by the owner consists of piece rate or bonus factor β and a base salary $\underline{w}$, so that total (random) wage is $\tilde{w} = \underline{w} + \beta(\gamma e + \tilde{\epsilon})$ leading to an expected utility of[2]

$$u(e) = \mathbb{E}[\tilde{w}] - \tfrac{1}{2}\rho\mathbb{V}[\tilde{w}] - \tfrac{1}{2}e^2 = \underline{w} + \beta\gamma e - \tfrac{1}{2}e^2 - \tfrac{1}{2}\rho\sigma^2\beta^2\gamma^2 \tag{20.4}$$

Observe that a greater bonus factor β increases the manager's risk exposure; the last term in (20.4) is the corresponding risk premium which is independent of effort due to the assumed additivity in the performance measure. The optimal level of activity maximizes u by equating marginal wage β to marginal cost of effort i.e., $\hat{e}(\beta) = \beta\gamma$. As intuition would suggests, effort increases with the bonus factor and productivity. Plugging the optimal activity into (20.4), the agent's expected utility simplifies into $u(\hat{e}) = \underline{w} - \frac{1}{2}\beta^2\gamma^2(\rho\sigma^2 - 1)$. The agent is willing to accept the wage scheme if this is positive thus the principal sets $\underline{w}(\beta) = \frac{1}{2}\beta^2\gamma^2(\rho\sigma^2 - 1)$ so as to saturates this *participation constraint*.

Summarizing, a bonus β motivates the manager to expand effort $e = \beta\gamma$ at expected cost

$$\hat{w}(e) = \underline{w}(\beta) + \beta\gamma\hat{e}(\beta) = \tfrac{1}{2}(1 + \rho\sigma^2)\beta^2\gamma^2 = \tfrac{1}{2}(1 + \rho\sigma^2)e^2 \tag{20.5}$$

We can now solve for the optimal contract of the firm. Since expected output $Q = \mathbb{E}[q]$ is linear in effort with $Q = \gamma e$, the cost function is $C(Q) = \frac{1+\rho\sigma^2}{2\gamma^2}Q^2$. Profit[5@] $\pi = Q - C(Q)$ is thus maximized for $\hat{Q} = \frac{\gamma^2}{1+\rho\sigma^2}$, yielding a maximum $\hat{\pi} = \frac{\gamma^2}{2(1+\rho\sigma^2)}$.

To relate these findings to the first-best seen in the previous section, recall that the true cost of effort to the manager is $c(e) = \frac{1}{2}e^2 \leq \hat{w}(e)$. There is equality with (20.5) if either $\sigma = 0$ (no risk) or $\rho = 0$ (risk-neutrality). The objective at the first-best is $\pi^* = \gamma^2/2$ so that the agency cost of moral hazard, expressed in percentage, is $\frac{\pi^* - \hat{\pi}}{\pi^*} = \frac{\rho\sigma^2}{1+\rho\sigma^2}$. Notice that the quantity objective might be replaced by a quality one or a reduction of marginal cost. In all cases, the presence of moral hazard generates an additional diseconomy of scale since the marginal cost of reaching the objective is an increasing function of the objective level.

20.2.2 Misaligned Incentives

The ultimate objective of the firm is profit but since no single person makes a definite impact upon it, the yearly profit is a poor signal of employees' past efforts; it is therefore not a good instrument to use in order to motivate hard work. The numerous elements that concur to profit are the quality, sales and cost of products but also human capital acquisition or cooperativeness among employees. Most of these dimensions are too complex to be contracted upon, so that workable objective performance measures only display a small positive correlation with the objective that the firm may desire for a particular employee. We shall see, within the previous model, that explicit incentives are optimally low powered to avoid distorting the agent's effort towards unnecessary goals. More generally, multiple instruments such as yearly cash bonus, promotion or job design enable to span better the fundamental objectives of the firm and avoid misaligned incentives.

Suppose that the valuable output is $q = e + \tilde{\epsilon}$ (assuming unitary productivity) but that the observable contractible performance measure is $x = \theta e + \tilde{\epsilon}$ where θ is an index of the divergence between the principal and the agent objectives. Prior to the relationship, this parameter is unknown to all; we assume variance σ_θ^2 and a unitary mean so that, on average, the performance measure is well designed. Once the agent starts to work for the principal, he becomes knowledgeable about the firm's technology and thus learns the true realized value of θ i.e., whether his effort tends to over or under emphasize the observable performance measure (with respect to its real impact on profits).

Based on our previous model, it is immediate to see that a bonus β based on x leads the informed employee to choose effort $\hat{e} = \beta\theta$ and expect total utility $u(\hat{e}) = \underline{w} - \frac{1}{2}\beta^2(\rho\sigma^2 - \theta^2)$. The base salary guaranteeing ex-ante that the agent is willing to sign is thus $\underline{w} = \mathbb{E}\left[\frac{1}{2}\beta^2(\rho\sigma^2 - \theta^2)\right] = \frac{1}{2}\beta^2(\rho\sigma^2 - \sigma_\theta^2 - 1)$ as $\mathbb{E}[\theta^2] = \sigma_\theta^2 + \mathbb{E}[\theta]^2$ (recall that ex-ante the firm ignores the value of θ). Since expected effort is $\mathbb{E}[e] = \beta$, the expected cost for the principal is then an extension of (20.5):

$$\hat{w}(e) = \underline{w} + \beta\mathbb{E}[x] = \underline{w} + \beta\mathbb{E}[\theta e] = \underline{w} + \beta^2\mathbb{E}\left[\theta^2\right] = \frac{1}{2}\left(1 + \rho\sigma^2 + \sigma_\theta^2\right)e^2 \tag{20.6}$$

Thus, even for a risk neutral agent ($\rho = 0$), the principal's adjusted cost of effort is increased by the agent's ability to exploit his insider information at her expense (a behavior known as "gaming"). Whatever the value of the output, optimal incentives will be deliberately muted and optimal effort will be reduced further (as compared to the previous analysis since marginal cost is now higher).

20.2.3 Rank-Order Tournaments †

Instead of relying on an absolute performance to reward an agent, one can use a relative performance by comparing an agent against a yardstick such as the previous year sales or against peers both within and outside the firm. Rank order refers to the fact that the margin of winning does not affect the level of compensation. In many instances, it is easier to directly compare two outcomes (make an ordinal ranking) rather than assess them against an exogenous scale (make a cardinal ranking). There is thus a metering advantage (lower cost) of relative performance from being an ordinal measure as opposed to piece rate which is a cardinal measure. Relative measures are also more *flexible* in the sense that if the environment changes e.g., technology improves, then all the peers are likely to be affected in the same manner so that incentives remain unaffected quite differently from the case of a piece rate scheme whose absolute bonus become automatically distorted.

Following Lazear and Rosen (1981), we analyze the simple setting where two agents with current salary $\underline{w}$ compete for a higher rank position paying a bonus β. The first agent's wage is then the random $\tilde{w}_1 = \underline{w} + \beta \mathbb{1}_{q_1 > q_2}$ which is a binomial variable. The probability of winning is $\mathbb{E}\big[q_1 > q_2\big] = \mathbb{E}[\tilde{\epsilon}_1 - \tilde{\epsilon}_2 > e_2 - e_1] = H(e_1 - e_2)$ where H is the law of a centered normal variable with variance $2\sigma^2$. The expected wage is thus $\underline{w} + \beta H(e_1 - e_2)$ while its variance is $\beta^2 H(e_1 - e_2)(1 - H(e_1 - e_2))$ which, as a function of the bonus, is less risky than a piece rate scheme. As before, each agent has constant risk aversion ρ, thus maximizes

$$u(e_1) = \underline{w} + \beta H(e_1 - e_2) - 2\rho\beta^2 H(e_1 - e_2)(1 - H(e_1 - e_2)) - \tfrac{1}{2}e_1^2$$

The FOC of optimal effort is $\beta h(e_1 - e_2) = e_1 + h(e_1 - e_2)(1 - 2H(e_1 - e_2))$. Since all the setting is symmetric, so is the equilibrium, hence the FOC reads $e_1^* = e_2^* = \beta h(0)$ as $H(0) = \frac{1}{2}$. Recalling that H is the law of $\tilde{\epsilon}_1 - \tilde{\epsilon}_2$, one can compute $h(0) = \frac{1}{2\sigma\sqrt{\pi}}$. Hence, the greater the noisiness of the performance measure, the lower is the induced effort. To induce high effort, it is thus necessary to set a large bonus and force the agent to support a greater risk. As we shall now see, tournament ends up being more costly.

The salary is set to solve the agent participation constraint $u(e^*) = 0$ i.e., $\underline{w} = -\frac{1}{2}\beta + \frac{1}{2}\rho\beta^2 + \frac{1}{2}(e^*)^2$. The total cost of achieving effort level e per agent is thus $c(e) = \underline{w} + \frac{1}{2}\beta = \frac{1}{2}\rho\frac{e^2}{h(0)^2} + \frac{1}{2}e^2 = \frac{1}{2}(1 + 4\pi\rho\sigma^2)e^2$ as $\frac{1}{h(0)^2} = 4\pi\sigma^2$.

Evaluation against a yardstick $\bar{q}$ is a form of contest quite similar to the rank order tournament yielding a twice lower risk premium.[6@] Comparing the reward schemes is now easy in our setting since we only need looking at the risk premium factor in each cost formula. Since $4\pi \simeq 12.6$, the risk premium is greatest under tournament, then twice lower under yardstick and lowest under piece rate.[7@] As soon as uncertainty matters, the

tournament (relative evaluation) is the worse method of providing incentive, followed by yardstick while the direct evaluation enabled by piece rate is best. Note though that all are inefficient since moral hazard cannot be eliminated as it must be traded against insurance.

However, if the noise blurring the observation of a worker's effort is made of a firm specific component plus an idiosyncratic component i.e., $q_1 = e_1 + \tilde{\eta} + \tilde{\epsilon}_1$ then the variance factor in the piece rate cost is $\sigma = \sigma_\eta^2 + \sigma_\epsilon^2$ while there is no change in the case of a tournament because the firm specific components cancel out. It is now clear that the tournament is optimal whenever the firm specific noise is greater than the idiosyncratic one i.e., $\sigma_\eta^2 > (4\pi - 1)\sigma_\epsilon^2$. Notice that the yardstick evaluation is not immune against firm shocks. This results rationalizes the use of tournaments for managers because their effort relate to firm wide strategies who are particularly sensitive to external shocks. Workers at lower levels of the hierarchy operate in an environment relatively safe from external influences so that piece rate schemes are optimal for them. Lastly, our previous observation that tournaments generate more income variance than piece rate schemes ($\pi\sigma^2 e^2$ vs. $\sigma^2 e^2$), means that "risk averters" prefer piece rate schemes while "risk-lovers" prefer tournaments. This rationalizes the fact that "risk averters" occupy lower levels of the hierarchy and are paid according to their productivity (with a piece rate scheme) while "risk-lovers" choose risky occupations in which few win very large prizes.

20.2.4 Multi-Tasking

Corts (2007) studies team vs individual compensation when agents must perform several tasks (cf. §13.1.2). We consider two functions that apply to two product lines α and β; together they generate four activities. At a given level of employment, say two managers, an assignment scheme must be devised to perform all four activities. Assuming high returns from specialization, each of the two agents performs two tasks.

The effort level in any activity has a cost $c(e) = \frac{1}{2}e^2$ and yield a one-to-one monetary return for the owner i.e., $\pi = e_1 + e_2 + e_3 + e_4$. The performance measure of the two products are $q_\alpha = \gamma_1 e_1 + \gamma_2 e_2 + \epsilon_\alpha$ and $q_\beta = \gamma_3 e_3 + \gamma_4 e_4 + \epsilon_\beta$ where ϵ_α and ϵ_β are independent noises blurring the observations. We cannot normalize the productivities to unity as before unless they are identical, thus we shall later on assume $\gamma_1 = \gamma_3 = 1$ and $\gamma_2 = \gamma_4 = \gamma \geq 1$ where the latter parameter measures the severity of the multi-task problem. As usual, the wage scheme includes a base salary and bonuses related to the observable outputs i.e., $w = \underline{w} + \alpha q_\alpha + \beta q_\beta$. As before, the salary $\underline{w}$ will be tuned to meet the participation constraint. Assuming that the managers are risk averse with (identical) coefficient of absolute risk aversion ρ, their gross expected utility is $u(e) = \underline{w} - \frac{1}{2}\rho\sigma^2\left(\alpha^2 + \beta^2\right) + \alpha(\gamma_1 e_1 +$

$\gamma_2 e_2) + \beta(\gamma_3 e_3 + \gamma_4 e_4)$ which is an immediate extension of (20.4) except for the omission of effort costs.

If an agent works in a team, he must provide the pair (e_1, e_3) or (e_2, e_4)[8@] while if he works alone on a product line, he must provide the pair (e_1, e_2) or (e_3, e_4). In team work, the optimal level of activity e_1 maximizes $\alpha\gamma_1 e_1 - \frac{1}{2}e_1^2$, thus he chooses $e_1 = \alpha\gamma_1$ and by symmetry $e_3 = \beta\gamma_3$. Efficiency commands a unit effort in each activity; this can be achieved by tuning the appropriate bonus factor for that agent, whatever the productivity γ may be in each activity. This is the advantage of team work as we shall see right-away. Since the owner must compensate the agent for the risk burden, her expected cost is $\frac{1}{2}\left(1 + \rho\sigma^2/\gamma_1^2\right) e_1^2$ as in (20.5) up to the productivity factor γ_1. The profit over the first activity is thus

$$\pi_1 = e_1 - \frac{1}{2}\left(1 + \rho\sigma^2/\gamma_1^2\right) e_1^2 = \alpha\gamma_1 - \frac{1}{2}(\rho\sigma^2 + \gamma_1^2)\alpha^2$$

so that the optimal bonus factor is $\alpha = \frac{\gamma_1}{\rho\sigma^2+\gamma_1^2}$ and the maximum profit reduces to $\pi_1 = \frac{1}{2}\frac{\gamma_1^2}{\rho\sigma^2+\gamma_1^2}$. By symmetry for the third activity, $\pi_3 = \frac{1}{2}\frac{\gamma_3^2}{\rho\sigma^2+\gamma_3^2}$. The case of the other agent is identical. Using the normalization of the productivities and summing the four activities, we obtain $\Pi_{\text{team}} = \frac{1}{\rho\sigma^2+1} + \frac{\gamma^2}{\rho\sigma^2+\gamma^2}$.

Let us study now individual line work. The α product manager chooses effort e_1 to maximize $\alpha\gamma_1 e_1 - \frac{1}{2}e_1^2$, hence $e_1 = \alpha\gamma_1$. Observe now that his optimal choice for the other task is $e_2 = \alpha\gamma_2$ because his pay depends only on the α output.[9@] Incentives are thus worsely aligned but there is also less risk exposure since $\beta = 0$ for that the α product manager. The profit over product α is thus

$$\pi_\alpha = e_1 + e_2 - \tfrac{1}{2}(e_1^2 + e_2^2) - \tfrac{1}{2}\rho\sigma^2\alpha^2 = \alpha(\gamma_1 + \gamma_2) - \tfrac{1}{2}(\rho\sigma^2 + \gamma_1^2 + \gamma_2^2)\alpha^2$$

so that the optimal parameter is $\alpha = \frac{\gamma_1+\gamma_2}{\rho\sigma^2+\gamma_1^2+\gamma_2^2}$ and the maximum profit reduces to $\pi_\alpha = \frac{1}{2}\frac{(\gamma_1+\gamma_2)^2}{\rho\sigma^2+\gamma_1^2+\gamma_2^2}$. Using the normalization of the productivities and the perfect symmetry for the other product, we obtain total profit $\Pi_{\text{ind}} = \frac{(1+\gamma)^2}{\rho\sigma^2+1+\gamma^2}$. One can show that[10@] $\Pi_{\text{team}} > \Pi_{\text{ind}}$ if $\gamma > 3 + 4\rho\sigma^2$. We can conclude:

> When the degree of risk aversion is low, or when signals of effort are informative or when the multi-task problem is serious, team work dominates individual accountability.

In terms of firm organization, this suggest reasons when functional rather and divisional structure is preferred (cf. §13.1.4).

20.2.5 Career Concerns

A manager can be motivated by explicit incentives such as performance based bonuses or promotions, but also by the implicit incentives channeled through the labor market. If innate ability, which is highly variable, was easily observed (could be proved), managers would always be paid according to their productivity. Education is one way to signal this personal characteristic because of the differential cost to acquire human capital (cf. §21.1.3 on signaling). However people change jobs several times (on average) during their first decade of professional life, thus the market valuation of their ability comes to dominate that incorporated in the diploma. The fact that the market is able to infer a good estimate of real productivity out of observable performances gives rise to a different kind of *signaling*: people work hard as juniors in order to achieve great performance, signal their worth and ultimately achieve better senior positions. The incentive at work here is implicit because the market does not conscientiously design the inference process revealing abilities and it is also dynamic as future pay is based on past performance. We follow here the seminal contribution of Holmstrom (1982b). On top of the moral hazard seen before (effort is delegated), the fact that ability is a private information to the worker introduces an adverse selection dimension to the problem.

Abilities in the population are measured by a zero mean index so as to capture superior or inferior individual ability. The agent whose ability is θ expands effort e in his current occupation. We keep referring to the market or his future employer as the principal. The observable performance measure is $q = \theta + e + \tilde{\epsilon}$ where $\tilde{\epsilon}$ is a white noise as before. From the point of view of the principal, the ability is a centered normal variable with variance σ_θ^2. Let us denote $\tau \equiv \frac{\sigma_\theta^2}{\sigma^2+\sigma_\theta^2} \leq 1$, a (positive) measure of the informativeness of the signal q upon the ability θ. Upon observing an output q and anticipating an effort level $\hat{e}$, the principal derives a realization of the random variable $\theta + \tilde{\epsilon}$, he thus infers a new estimate of the underlying ability $\mathbb{E}\left[\theta | q, \hat{e}\right] = \tau(q - \hat{e})$ computed using the statistical laws of the involved random variables.

Ex-ante, during the junior part of his working life, the agent with ability θ must choose how much effort to invest in the signaling activity; he knows that his ex-post wage, as a senior manager, will depend on the effort $\hat{e}$ anticipated by the principal and on the noise $\tilde{\epsilon}$ (trough the performance q), his expected wage is thus $\mathbb{E}_\epsilon[w] = \mathbb{E}_\epsilon\left[\mathbb{E}\left[\theta | q, \hat{e}\right]\right] = \tau(\theta + e - \hat{e})$. Given the quadratic cost of effort $\frac{1}{2}e^2$, the optimal effort is $e^* = \tau$ which is increasing in the market's ability to infer productivity out of the observable performance.[11@] At the limit where ability is fully observable i.e., $\tau = 1$, the effort is efficient since welfare is here $q - c(e)$.

Since ability is unknown to the market, there are returns to effort because perfor-

mance influences the ability perception. The agent thus tries to bias the process of inference in his favor. However, in equilibrium the market anticipates the effort level and adjusts the output measure accordingly i.e., no one can fool the market. Yet, the agent is trapped in supplying the equilibrium level that is expected of him, because, as in a rat race, a lower supply of labour will bias the evaluation procedure against him. Including many periods of repeated effort and performance reveals that effort gradually decreases as the end of the professional life approaches because the market inference has been refined to the point of identifying exactly the ability i.e., there is nothing left to signal which means that effort becomes useless.

20.3 State of Nature Approach

In this section, the valuable output is $q = \Phi(e, \theta)$ where θ is a choice of "Nature" such as a macro-economic shock. In the case of automobile insurance, θ can take one of two values, "accident" or "none" and e is the cautiousness taken by the driver (e.g., respect speed limits). The presence of uncertainty does not by itself ruin our hopes to get insurance. If the insurer could install a black box registering every driving decisions, then effort (driving care) would be contractible because it could be verified after the occurrence of an accident; the first-best would be achieved. What is reality for airplanes and trucks is not yet implemented for individual cars so that moral hazard is an issue. Another basic example is when the agent is the manager of a firm. The parameter θ can then reflect the strength of demand for the good he is in charge of developing and producing and e is the amount of time dedicated to this project.[12@] It is quite clear that the overall profit of the firm will depend on both the market demand and the care with which the agent managed the project; the interconnection between the two components is so complex that there is no way to decipher the whole sequence of actions that the agent took to judge whether he worked hard or not. This means that effort is *unverifiable* and since it cannot be inferred from the output, it becomes a free decision for the agent. To repeat ourselves, the asymmetry of information in settings of moral hazard forces the principal to delegate crucial decisions to the agent.

20.3.1 The Second Best program

As we just argued, in the majority of real cases it is too costly to perform an audit after production has occurred to discover how much the agent has worked (and pay him accordingly). Hence, the principal can only propose a wage contingent on the production level, knowing that the agent will choose its effort to maximize his expected utility (over

the distribution of the state of nature). Our task in this section is to characterize the optimal such contract; Figure 20.1 is the game tree describing the relationship between the principal and the agent. The players are the principal **P**, the agent **A** and a special player which is not strategic, nature **N**. The first player to move is the principal who offers a contract out of a large range of possibilities (represented by the cone); it is a rule $w(q)$ stipulating a wage for each level of output. Then the agent either refuses or accepts the contract and immediately after he chooses his effort (in our simple sketch he has only two choices, high and low). Later on Nature decides whether the state of the world is favorable or not which gives us the final output $q = \Phi(e,\theta)$ jointly determined by his effort and the state of nature nature.

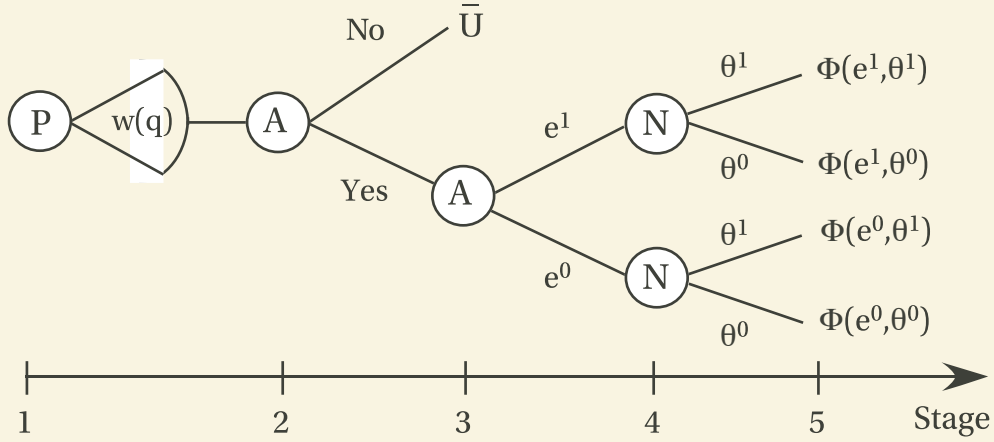


Figure 20.1: Principal-Agent Contractual Relationship

The Principal-Agent game of Figure 20.1 is solved by backward induction, a concept seen in §2.4.2. The agent's ex-post utility (at stage 5) is $u\big(w(q)\big) - c(e)$ where $q = \Phi(e,\theta)$, hence the agent's expected utility at the time where he has to choose his effort (stage 3) is $U(w,e) \equiv \mathbb{E}_\theta\left[u\left(w\left(\Phi(e,\theta)\right)\right)\right] - c(e)$. Being rational, the agent will pick the effort $\hat{e}(w)$ maximizing $v(w,e)$ over the available choices. Moving one step backward (stage 2), the rational agent will accept the contract $w(q)$ if and only if it guarantees him more than its outside option i.e., $U(w,\hat{e}(w)) \geq \underline{u}$. Provided that w is attractive enough to be accepted by the agent, the principal's ex-ante profit (at stage 1) is $\Pi(w) \equiv \mathbb{E}_\theta\left[\Phi(\hat{e}(w),\theta) - w\left(\Phi\left(\hat{e}(w),\theta\right)\right)\right]$. This cumbersome formulation is presented as the following optimization problem:[13@]

$$\max_{w(.)} \mathbb{E}_\theta\left[\Phi(\hat{e},\theta) - w\left(\Phi\left(\hat{e},\theta\right)\right)\right] \text{ s.t.} \begin{cases} U(w,\hat{e}) \geq \underline{u} & (IR) \\ \forall e \neq \hat{e}, U(w,\hat{e}) \geq U(w,e) & (IC) \end{cases} \tag{20.7}$$

where (IC) is the *incentive compatibility* constraint stating that the agent always chooses the effort best for him given the contract $w(.)$ he signed and (IR) is the *individual ratio-*

nality constraint that guarantees participation.

Since this program is very difficult to solve we study a simpler version and assume that the agent manages a risky project yielding either success (value 1€) or failure (value 0€). A contract is now a pair (w_0, β) where β is the bonus for good results (e.g., stock options for managers). The effort positively influences the probability of success but with decreasing return to scale i.e., the probability $q(e)$ is concave increasing. We assume *separability* of effort and income with $u(e, w_0) = u(w_0) - c(e)$ where u is concave increasing (monetary risk aversion) and c is convex increasing.

20.3.2 Resolution

Let us first find out which contracts induce the agent to take some specific effort e. The expected utility of the agent is $(1-q(e))u(w_0) + q(e)u(w_0+\beta) - c(e)$ hence the optimal effort solves $u(w_0+\beta) - u(w_0) = \frac{c'(e)}{q'(e)}$. As $c'' > 0$, $q'' < 0$ and $u' > 0$, there exists $\beta(w_0, e)$ such that $u\big(w_0 + \beta(w_0, e)\big) - u(w_0) = \frac{c'(e)}{q'(e)}$. Independently of risk aversion, the greater the effort wanted by the principal, the greater the bonus he has to give i.e., $\frac{\partial \beta}{\partial e} > 0$. Among the contracts implementing e, the least costly solves

$$\begin{aligned} \underline{u} + c(e) &= (1-q(e))u(w_0) + q(e)u(w_0+\beta) = u(w_0) + q(e)\frac{c'(e)}{q'(e)} \\ \Rightarrow w_0 = \hat{w}(e) &\equiv u^{-1}\left(\underline{u} + c(e) - \frac{q(e)c'(e)}{q'(e)}\right) < \underline{w}(e) = u^{-1}\big(\underline{u} + c(e)\big) \end{aligned} \tag{20.8}$$

where $\hat{w}' > 0$ and $\hat{w}'' > 0$ (composition of two convex functions). The difference with $\underline{w}(.)$ characterizing the participation constraint without moral hazard (cf. eq. (20.1)) is the fraction term. The profit for the principal when implementing $\hat{e}$ is then

$$\begin{aligned} \Pi(e) &= (1-q(e))\,(0 - \hat{w}(e)) + q(e)\big(1 - \hat{w}(e) - \beta(\hat{w}(e), e)\big) \\ &= q(e) - \hat{w}(e) - q(e)\beta\,(\hat{w}(e), e) \end{aligned} \tag{20.9}$$

In the first-best regime where effort is contractible, the profit with a contract $(e, \underline{w}(e))$ is $q(e) - \underline{w}(e)$. The difference in the presence of moral hazard is twofold: one the one hand, there is the costlier bonus $\beta\,(\hat{w}(e), e)$ that must be given to the agent to induce him to perform e while on the other hand, the base salary $\hat{w}$ is cheaper. The last step for the resolution of (20.7) is to choose an optimal $\hat{e}$ to maximize (20.9). The optimal effort solving program (20.7), known as the second best effort, is smaller than the first-best one and we may conclude that the presence of moral hazard forces the principal to distort the optimal contract towards less effort since incentives are now more costly to provide.

If the agent is risk-neutral, the bonus implementing an effort e is independent of the salary, it is $\beta(e) = \frac{c'(e)}{q'(e)}$ while its associated salary is $\hat{w}(e) = \underline{u} + c(e) - \frac{q(e)c'(e)}{q'(e)}$. In that case, the principal's payoff is $\Pi(e) = q(e)(1-\beta) - \hat{w}(e) = q(e) - c(e) - \underline{u}$ which is the first best.

Pay for outputs, not inputs

Mirrlees (1974) shows the following paradox: if there existed an outcome $\hat{q}$ that would occur with zero probability after the first-best effort e^* but would occur with positive probability for every lower effort then the principal would successfully offer the agent the following contract "I pay you $\underline{w}(e^*) + 1€$ but if $\hat{q}$ appears you will be executed". Backward rationality tells us that e^* is optimal under acceptance of the contract and that accepting is optimal in the first place. The purpose of this silly example is to show that rewards should not be linked to output but to input. Hence if a lazy effort increase the occurrence of some output then the contingent wage should be very low to deter laziness. Likewise if the optimal effort yield more often some output then the contingent wage should be very large to encourage this effort.

20.3.3 The Mirrlees Approach †

Holmstrom (1979) provides a powerful yet simple characterization of the optimal contract using the Mirrlees (1974) approach of turning random variables into distributions admitting densities. The production is a random variable $\tilde{q}$ whose law is $H(e,q) \equiv Pr\left(\tilde{q} \leq q \,|\, e\right)$ depends on the effort e previously chosen by the agent. The density is $h(e,q) = \frac{\partial H(e,q)}{\partial q}$; we denote $h_e = \frac{\partial h}{\partial e}$.

From the agent's ex-post utility $u(w) - c(e)$, we deduce the ex-ante utility conditional on the wage scheme ω:

$$U(\omega, e) \equiv \int u\big(\omega(q)\big)\, h(e,q)\, \mathrm{d}q - c(e)$$

Moral hazard means that the agents chooses effort $\hat{e}$ to maximize $U(\omega, e)$ i.e., satisfies the incentive compatibility constraint (IC)

$$c'(e) = \int u\big(\omega(q)\big)\, h_e(e,q)\, \mathrm{d}q \tag{20.10}$$

Furthermore, he will accept the wage scheme ω only if $U(\omega, \hat{e}) \geq \underline{u}$ (IR).

Likewise the ex-post utility of the principal being $\pi(q-w)$, she expects $\Pi(\omega, e) \equiv \int \pi(q - \omega(q)) h(e,q)\, \mathrm{d}q$ when the agent has accepted the scheme ω. Her objective is thus to maximize $\Pi(\omega, e)$ under the above (IC) and (IR) conditions. The Lagragean (with non negative

multipliers λ and μ) is

$$\mathscr{L} = \int \pi(q-\omega(q))h(\hat{e},q)\,\mathrm{d}q + \lambda\left(U(\omega,\hat{e}) - \underline{u}\right) + \mu\left(c'(\hat{e}) - \int u\left(\omega(q)\right)h_e(\hat{e},q)\,\mathrm{d}q\right)$$

and can be maximized point-wise in the variable w i.e., by solving $\frac{\partial \mathscr{L}}{\partial w} = 0$ to obtain

$$-\pi'(q-w) + \lambda u'(w)h(e,q) + \mu u'(w)h_e(e,q) = 0 \Leftrightarrow \frac{\pi'(q-w)}{u'(w)} = \lambda + \mu\frac{h_e}{h}(e,q) \tag{20.11}$$

A first best situation, one without moral hazard, is found by maximizing $\Pi(\omega,e) + \lambda U(\omega,e)$ in ω and e for some multiplier λ (this is equivalent to maximize Π under the (IR) constraint). The optimal risk sharing is found by a point-wise maximization as above; the FOC for every q is

$$\frac{\pi'(q-w)}{u'(w)} = \lambda \tag{20.12}$$

and the solution is an increasing[14@] function $\omega^*(q)$. The efficient effort e^* solves[15@]

$$c'(e) = \int \left(u(\omega(q)) + \frac{\pi(q-\omega(q))}{\lambda}\right)h_e(e,q)\,\mathrm{d}q \tag{20.13}$$

To be able to compare the risk-sharing equations (20.11) and (20.12), Holmstrom (1979) first proves that $\mu > 0$; then using the fact that the RHS $\frac{\pi'(q-w)}{u'(w)}$ is increasing in w, he can deduce that $\omega^*(q) \gtrless \hat{\omega}(q) \Leftrightarrow h_e \gtrless 0$. The meaning of this result is that higher effort is promoted: indeed, at all productions whose probability is increased by effort ($h_e > 0$), the wage increase faster than the first best one while the reverse holds true at all productions whose probability is decreased by effort ($h_e < 0$). If we further assume that the ratio $\frac{h_e}{h}$ is increasing in q then the second best $\omega(q)$ is increasing.[16@] This motivation towards hard work is nevertheless countered by the need to share risk so that in the end the second best effort is inefficiently low. Indeed, comparing (20.10) and (20.13) reveals that $\hat{e} < e^*$ if $\int \pi(q-\omega(q))h_e(e,q)\,\mathrm{d}q > 0$ which happens to be a consequence of $\mu > 0$.[17@]

20.3.4 Automobile Insurance

Since the second best program (20.7) is difficult to solve we present Rasmusen (2006)'s simple application to automobile theft insurance where the analytical solution can be derived. Assume that effort has two levels, care $\bar{e}$ and no-care $\underline{e}$ generating theft probabilities of $\frac{1}{2}$ and $\frac{3}{4}$ respectively. The utility of the car is 12 and the owner is risk averse. An insurance contract $C = (p,\delta)$ is a premium p and a reimbursement δ of damages; the absence of contract is $C_0 = (0,0)$. In the absence of an insurance contract the expected util-

ity under care is $u(C_0, \bar{e}) = \frac{u(12)}{2} - c$ where c is the disutility of taking care (e.g., checking every night that doors are locked). Without care the utility is $u(C_0, \underline{e}) = \frac{u(12)}{4}$. Assuming c not too large, the non-insured car owner will optimally take care.

Observe that a contract can also be labeled $C = (w_{NT}, w_T)$ where $w_{NT} = 12 - p$ and $w_T = \delta - p$ are the income level of the owner in the two possible states of nature (theft and no-theft). No insurance can thus be labeled $C_0 = (12, 0)$. As the probability of theft is 50% under care, the owner is indifferent among C_0 and $C_4 = (0, 12)$. On Figure 20.2 the indifference curve joining C_0 to C_4 passes below $C_1 = (6, 6)$ because the owner is risk-averse.

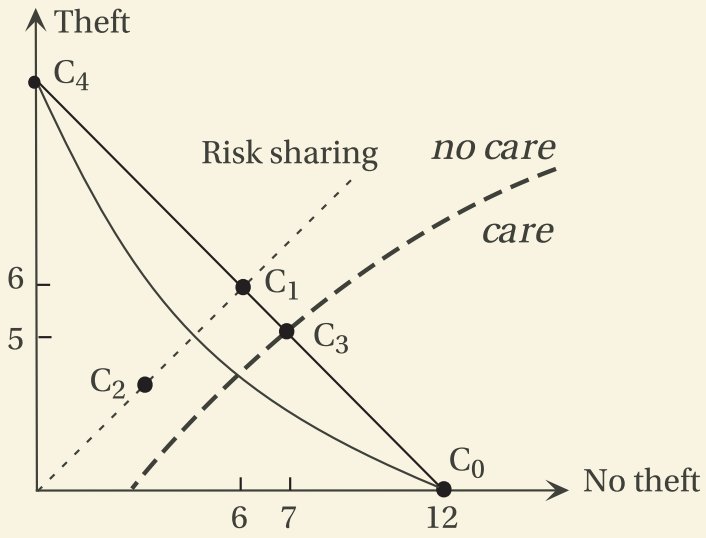


Figure 20.2: Risk Sharing

Competition among risk neutral insurers should drive them to offer the full insurance contract ($p = 6, \delta = 12$) denoted C_1 that leaves an insurer break even (provided that the policyholder takes care). The economic surplus $S = u(6) - \frac{u(12) - u(0)}{2}$ goes to the car owner.

Moral hazard appears immediately because the car owner has no more incentive to take care since $u(C_1, \bar{e}) = u(6) - c < u(C_1, \underline{e}) = u(6)$: why take care if the car will be fully reimbursed whatever happens ? Given this change of behavior, the insurer now expects losses of $6 - \frac{3}{4}12 = -3$ and will raise the premium from 6 to 9 (contract C_2 on Figure 20.2). Under this contract the utility of the owner is $u(C_2, \underline{e}) = u(3)$. If this amount is less than $u(C_0, \bar{e})$ (as shown on Figure 20.2) then the owner won't buy insurance.

Full insurance is thus incompatible with both the insurer *IR* constraint (offering a contract) and the owner *IR* constraint (agreeing to buy it). The optimal contract $C^* =$

(w^*_{NT}, w^*_T) has to give an incentive to the owner to take care of its car i.e., must satisfy

$$\begin{aligned} \frac{u(w_{NT})+u(w_T)}{2} - c &\geq \frac{u(w_{NT})+3u(w_T)}{4} \\ \Leftrightarrow 4c \leq u(w_{NT}) - u(w_T) &= u(12-p) - u(\delta - p) \end{aligned} \tag{IC}$$

i.e., the contract has to be below the bold dashed curve on Figure 20.2 known as the *IC* curve.

Since insurer competition bestow all the surplus on the owner, the equilibrium contract will leave the insurer break even i.e., be the intersection C_3 of the *IC* curve and the insurer *IR* line joining C_0 to C_4. Let us assume to simplify exposition that the optimal reimbursement is $\delta^* = 10$. Then the break even premium is $p^* = 5$ so that the income levels are $w^*_{NT} = 7$ and $w^*_T = 5$. The difference $12 - \delta^* = 2$ is the deductible of the optimal insurance contract. The larger the deductible, the larger the risk supported by the policyholder thus the greater the care he takes of the insured item. Full insurance is rare because 12 stands for the value of finding, buying and bringing home the item while a traditional "full insurance" contract only repays the market value. Hence the owner has always a little incentive to take care of his property.

20.4 Renegotiation and Auditing

In this last section, we treat additional topics relating to the renegotiation of contracts and the game of auditing/cheating between taxpayers and the IRS.

Binary choice

Let us introduce the idea of renegotiation and its aftermath by way of a simple example. An individual earns w by working the regular time but can obtain a bonus β by working extra-hours although this has an opportunity cost $c > \beta$ to him so that his current dominant choice is to perform no extra hours. Let us now introduce into the picture a car of objective (market) value p and subjective value π to the individual with $p < \pi < \beta$ i.e., the car is more valuable than the market price p but lesser than the maximum salary. If $\pi - p > c - \beta + w$ i.e., it is efficient for the individual to own the car, then a bank can lend p to the agent to buy the car with the promise to repay p later on by working extra-hours. The reason why such a contract works is the following: by working hard, the agent makes enough money to repay the loan and enjoy the car, hence his final utility is $\pi - p + \beta - c > w$. If on the contrary he decides to shirk then he won't be able to repay the loan, the bank will seize the car and his final utility will be w. The previous condition is

the incentive compatibility one.

If renegotiation is possible then when the bank faces a defaulter it cannot help but renegotiate. Indeed, the car being already used, it loses much of its market value thus it is better for the bank to accept the lower payment w from a shirker (instead of p) than seizing and reselling for even less. This totally changes the incentives for work since the utility of leisure is know $w+\pi-w=\pi$ which can be greater than $\pi-p+\beta-c$. In that case, the agent optimally chooses to shirk and in response the bank ceases to offer a loan knowing that it will make a loss $p-w$. There is a market failure generated by the inevitability of renegotiation. Banks are therefore forced to be tough (commit to seize no matter what) in order to be able to operate a Pareto improving intermediation service.

Contractual renegotiation can be a problem because it is often unavoidable. If one party into an agency relation cannot engage into a long term relationship then contracts will be short-term and renewed frequently which is like the renegotiation of a long-term contract. This happens for regulators since the prices they impose on firms last no more than the regulatory period and there can be no insurance that the legislator or the government will not change its policy the next time. Likewise, a sovereign state, unless tied by a treaty like the EU one or by membership of the WTO, will easily renege a contract with a firm, whether national or foreign. Beyond these cases of forced renegotiation, we find the voluntary ones: if the parties agree to tear-up the contract and write a new one, no one can stop them.

Continuous choice

Let us use our formal set-up to analyze the consequences of renegotiation. Once the agent has exerted the second best effort $\hat{e}$, it not necessary anymore for the principal to use the costly bonus $\beta\left(\underline{w}(\hat{e}),\hat{e}\right)$. Indeed the bonus was offered to motivate the agent although it was forcing him to support unwanted risk; now that effort has been expanded, the bonus has played its role and should be removed before uncertainty resolves itself in order to fully insulate the agent from risk. Technically, the constant wage $\hat{w}$ solving $\left(1-q(\hat{e})\right)u\left(\underline{w}(\hat{e})\right)+q(\hat{e})\,u\left(\underline{w}(\hat{e})+\beta\right)=u(\hat{w})$ is cheaper than the expected payment $\underline{w}(\tilde{e})+q(\tilde{e})\beta$ stipulated in the original contract (this is so because u is concave).

The principal just got trapped because the agent, anticipating this change to come, has no more incentive to produce the effort $\hat{e}$; he is actually fully conscious that his final payoff will be constant although it appears at first sight to contain a bonus. The rational agent therefore chooses the minimum level $\underline{e}$ instead of $\hat{e}$. This reasoning is true for any effort $e>\underline{e}$ because an incentive contract with a positive bonus $\beta(w_0,\tilde{e})$ is always renegotiated to a nil one $\beta=0$ which precludes (in equilibrium) the agent from performing e.

To resolve this time-inconsistency, the principal must offer a *menu* of contracts contingent on effort that make the agent indifferent between all efforts in $[\underline{e};\hat{e}]$ and, given the equilibrium distribution of effort σ over $[\underline{e};\hat{e}]$, makes the principal indifferent between renegotiating and not. We shall not delve into the computations but it should be clear that the average effort implement under this scheme will fall below the second best one $\hat{e}$; thus we may say that committing to a single contract or committing to never renegotiate is helpful for the principal. In other words it pays "to be true to one's word".

Notice lastly, that if the agent leads the renegotiation[18@] and is rich enough to "buy the store" at price $\Pi(\hat{e})$, then renegotiation is not a problem anymore. Indeed the principal will accept a change in the contract only if the agent offers more than $\Pi(\hat{e})$ which can be achieved only by expanding the second best effort $\hat{e}$!

Debt renegotiation

We show here that the seller of an item has an incentive to finance her activity through debt to dilute the bargaining power of the buyer. We consider in turn the two cases where debt is not renegotiable and renegotiable to show that the latter fosters the strategic purpose of debt.

In most trading situation, the surplus generated by an exchange depends on many factors like future market demand or future price of oil not yet known when the parties devise their trade. From an ex-ante point of view the surplus is a random variable $x \in \mathbb{R}_+$ whose distribution function is $H(x)$ (h denotes the density). In this situation, the trading price has to be set trough bargaining once the surplus is known to the parties. The seller and the maker share it according to their respective bargaining abilities λ and $1-\lambda$.

Whenever the surplus x is inferior to the debt service F, the seller is caught in the debt overhang problem thus she refuses to deal and closes her firm. As a consequence, the value of debt is only $(1-H(F))F$. The value V of the firm for the seller is thus the value of her share $1-\lambda$ of profits $x-F$ when she does not go bankrupt plus the funds raised from the debt emission i.e.,

$$V = \int_F^\infty (1-\lambda)(x-F)\,\mathrm{d}H(x) + (1-H(F))F = (1-\lambda)\int_F^\infty x\,\mathrm{d}H(x) + \lambda(1-H(F))F \tag{20.14}$$

The optimal level of debt F^* solves $-(1-\lambda)h(F)F + \lambda(1-H(F)) - \lambda h(F)F = 0 \Leftrightarrow \lambda = \frac{h(F)F}{1-H(F)}$ and is positive since the RHS is zero for $F=0$.

If debt has been issued by a single bank then the event of bankruptcy is not so sure because the bank might prefer to renegotiate the debt service F down to x. Credibly threatening not to trade, the seller gets all the surplus and use it integrally to repay the bank. The value of debt is now $V_D = \int_F^\infty F\,\mathrm{d}H(x) + \int_0^F x\,\mathrm{d}H(x)$ so that the value of the firm

for the maker becomes

$$\tilde{V} = (1-\lambda)\int_F^{\infty}(x-F)\,\mathrm{d}H(x) + V_D = (1-\lambda)\mathbb{E}[x] + \lambda V_D \tag{20.15}$$

which is increasing function of F, hence the optimal level of debt is maximum. Renegotiation fosters the strategic purpose of debt which is to reduce the bargaining power of the buyer. This is so because the buyer loses all of his bargaining power when the firm is in danger of going bankrupt.

Audit

In all countries the tax authority (Internal Revenue Service in the US), is faced with "suspect" income declarations by tax payers. We can view the relation of a liable citizen (agent A) with the tax authority (principal P) as a game. Two forms can be thought of:
► *Simultaneity*: the tax authority chooses to audit or trust a declaration while the citizen chooses to cheat or reveal its true income.
► *Sequential*: the tax authority chooses to audit a proportion α of all declarations and then, knowing this policy, the citizen chooses to cheat or reveal its true income.

In the first game, the strategy of the tax authority is the probability of auditing σ_P while the strategy of the citizen is the probability of cheating σ_A. Letting t denote the income tax, f the fine paid if caught cheating and c the unit cost of audit, payoffs are

$$\Pi_A(\sigma_P,\sigma_A) = -\sigma_A\sigma_P\left(t+f\right) - (1-\sigma_A)t \tag{20.16}$$

$$\Pi_P(\sigma_P,\sigma_A) = \sigma_P\left(T-c+\sigma_A f\right) + (1-\sigma_P)(1-\sigma_A)t \tag{20.17}$$

There is no pure strategy equilibrium since $\sigma_A = 1$ (cheat) triggers $\sigma_P = 1$ (audit) itself triggering $\sigma_A = 0$ (reveal). In that case there is no point to audit ($\sigma_P = 0$) and therefore the citizen is better off cheating ($\sigma_A = 1$). To solve this conundrum, we have to look for a mixed strategy equilibrium. The audit frequency σ_P^* making the citizen indifferent between revealing and cheating solves

$$\frac{\partial\Pi_A(\sigma_P,\sigma_A)}{\partial\sigma_A} = -\sigma_P\left(t+f\right)+t = 0 \quad \Rightarrow \quad \sigma_P^* = \frac{t}{t+f}$$

and likewise

$$\frac{\partial\Pi_P(\sigma_P,\sigma_A)}{\partial\sigma_P} = \left(t-c+\sigma_A f\right) - (1-\sigma_A)t = 0 \quad \Rightarrow \quad \sigma_A^* = \frac{c}{t+f}$$

is the cheating frequency making the tax authority indifferent between auditing and

trusting. The final payoffs are $\Pi_A^* = -t$ and $\Pi_P^* = t - \frac{tc}{t+f}$ as if the tax authority was always auditing and citizens would therefore always pay their taxes. Still in equilibrium some people cheat and some do not while some are audited and some are not.

In the second game we are in fact assuming that the tax authority can *credibly* announce that it will audit α% of all income declarations. Choosing $\alpha = \frac{t}{t+f} + 1\%$ forces each citizen to reveal because the probability of audit is now strictly superior to σ_P^*. Final payoffs are identical but no cheating takes place anymore thus the following year the tax authority will tend to (secretly) decrease α since it is a costly action but then citizens will anticipate this and start to cheat again !!

Chapter 21

Adverse Selection

The private information held by an agent is crucial to determine the efficient decision in the agency relation. To make sure that this optimal action will be carried out, the principal must elicit this information. As intuition suggests, this can only be achieved by giving up an *information rent* to the agent (with respect to the complete information case). However this mechanically increases the marginal cost of production and creates a distortion or welfare loss. The principal must therefore arbitrate between productive efficiency and information acquisition.

This chapter contains two section. We first model the unraveling and signaling of information in a market and then study various models of screening where firms design multiple contracts that are offered to all their potential clients to take advantage of the induced self-selection.

21.1 Information Unraveling

In this section, the item for sale has a quality only known to its producer (service) or current owner (good). In Akerlof (1970)'s model of partial equilibrium for second-hand goods, a market failure occurs that can even lead to complete market breakdown. When there is a single seller, Levin and Tadelis (2005) show how the firm adapts its internal organization to resolve the information revelation conundrum. In an application to the theory of the firm, it is shown that the partnership structure can dominate the corporation structure if the quality of the service for sale is difficult to observe.

In this section we shall get a better understanding of the law on currencies stating that "bad money drives out good"[1@] or Groucho Marx joking "I wouldn't want to belong to any club that is willing to give me membership" or businessmen saying that "trade is difficult in developing countries".

21.1.1 Market for Used Cars

Akerlof (1970) observes that in many markets, buyers use statistics to judge the quality of prospective purchases. In that case, there is an incentive for sellers to offer poor quality merchandise; indeed the returns for selling good quality items accrue to the entire group whose statistic is affected rather than to the individual seller. As a result of this free riding, there tends to be a reduction in the average quality of goods and also in the size of the market.

A startling example is the large price difference between new cars and those which have just left the showroom. Automotive companies try their best to build perfectly good cars but there are always some copies that turn to be "lemons" (a colloquialism for defective used cars). Since the car is brand new, neither the buyer nor the dealer knows the exact quality of the car; there is symmetric information (or more precisely lack of). Things are different when you look at a used car because the current owner has probably learned whether her car was good or bad, she has acquired superior information by experiencing the car. There is now an asymmetry of information between buyers and sellers. Nevertheless, good cars and bad cars must still sell at the same price since it is impossible for a buyer to tell the difference. Rationally anticipating this fact, buyers will not accept a high price so that owners of good cars are trapped, they cannot sell their car for their real value nor for the price of a new car. Hence they remove their cars from the market so that actual sales are mostly lemons and since some potentials sellers have withdrawn, the total number of transactions tend to be low.

Common Values

We will show in a simple setting how an extreme market failure may occur: if sellers and buyers value cars identically then the market completely breaks down. To see this, consider a number of used cars whose quality or value θ varies in the interval $[\underline{\theta};\bar{\theta}]$. If the price is $\bar{\theta}$, then all owners want to sell so that $S(\bar{\theta}) = 1$, the total mass of cars while if the price is $\underline{\theta}$, none of them find it profitable to sell, hence $S(\underline{\theta}) = 0$. By the same token, demand is maximum for $p = \underline{\theta}$ and nil for $p = \bar{\theta}$. If, demand and supply are price responsive in the intuitive way then the two curves must intersect at some price $p_1 \in]\underline{\theta};\bar{\theta}[$ for some quantity q_1 as shown by the plain lines on Figure 21.1. We shall see that this price cannot be an equilibrium price at which trade can take place.

When all owners offer their car for sale at price $\bar{\theta}$, the average quality of cars for sale is $\theta_1 = \mathbb{E}[\theta]$, the average quality of all cars. when the price goes down to p_1, the supply shrinks to $q_1 = S(p_1) < S(\bar{\theta}) = 1$ and we know for sure that some of the best cars have been pulled out of the market because their owners are the first to be hurt by the

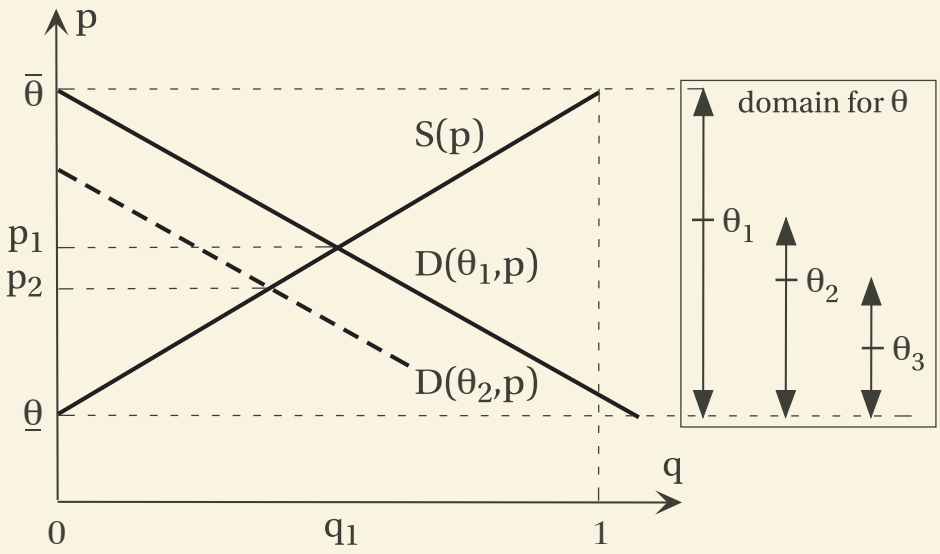


Figure 21.1: Market for Lemons

price decrease. Since potential buyers known that sellers are rational, they understand what's going on and therefore revise their estimate of the average quality of cars for sale down to $\theta_2 = \mathbb{E}[\theta \mid \theta \leq \theta_1] < \theta_1$. As the demand for cars depends also on their quality, the effective demand curve will drop to reflect this updating. It is now clear that the price p_1 generates an excess supply since $D(\theta_2, p_1) < D(\theta_1, p_1) = q_1 = S(p_1)$; it cannot be an equilibrium price. A further price drop down to p_2 will not solve the problem either. Indeed, it generates a lower supply, thus another downward revision of expected quality to some $\theta_3 = \mathbb{E}[q \mid \theta \leq \theta_2] < \theta_2$ by buyers so that supply $S(p_2)$ will still exceed the updated demand $D(\theta_3, p_2)$.

This spiral of descending prices ends at $\underline{\theta}$ where cars of exactly this value (the worst of the market) are sold for that price. If instead of cars we consider the metallic content of coins we obtain the law on currencies. Likewise, Groucho Marx considering himself a man of the street realizes that his joining a posh club would lower the average status of the club to the point where it is not worthwhile anymore to join. As regards general business, the adverse selection problem just mentioned can be circumvented by *certification* or *guarantees*. Private intermediaries like Veritas or TUV Rheinland and public ones likes the national or international certifications authorities certify conformity with security, health rules or ISO standards of products in order to remove part of the uncertainty regarding their quality (cf. §24).[2@] Likewise, guarantees assure buyers that they will not end up with a "lemon". Since fulfilling the guarantee is costly for the seller when he makes products of poor quality, offering a guarantee is also a signal of quality (cf. §11.3) for consumers and competitors. For instance, car dealers often propose used cars with a 6 month guarantee (at a higher price) after revising them extensively.

Differing Values

Although shocking, the "lemons" example does not display a true market failure (there is no efficiency loss) since preferences were assumed identical between buyers and sellers so that ownership did not matter. Assume now that buyers value a car of quality θ at $(1+\lambda)\theta$ and are more numerous than sellers so that the latter capture all gains from trade. The market is in equilibrium at the price p^* if the expected value of car sold for that price is exactly that price i.e., if $(1+\lambda)\mathbb{E}\left[\theta \,|\, \theta \leq p^*\right] = p^*$. Using the uniform distribution over $[\underline{\theta};\bar{\theta}]$, we obtain $p^* = \frac{1+\lambda}{1-\lambda}\underline{\theta}$. The equilibrium is said to be partially *pooling* because all owners of type $\theta \leq p^*$ act in the same way, they sell at price p^*. Likewise, all owners of type $\theta > p^*$ act identically by keeping their cars out of the market. Although the market is active, its outcome is inefficient as buyers value cars more than sellers; in a world of perfect information, all cars would be sold. This will occur here if $p^* = \bar{\theta} \Leftrightarrow \lambda = \frac{\bar{\theta}-\underline{\theta}}{\bar{\theta}+\underline{\theta}}$, that is to say if the differential in intrinsic value between buyers and sellers is large enough to overcome the information asymmetry.

21.1.2 Corporation vs Partnership

In service activities requiring a large human capital such as lawyers, doctors, architects or auditors, firms are rather organized as partnership instead of corporations. Levin and Tadelis (2005) explain this choice as a consequence of the large information asymmetry present in these services, namely that the market (potential clients) observes imperfectly the quality.

Set-up

In the activity under study, people's intrinsic quality θ is drawn randomly in $[0;1]$. The market wage for that activity is an average $w \in]0;1[$ of quality over people active in the entire economy. Due to state regulation, workers must study to earn degrees and pass many exams that enable a potential employer to screen them and thus hire only the most talented people. If the ability of the last person hired is q then the firm size is $1-q$ and its average quality is $\mathbb{E}\left[\theta \,|\, \theta \geq q\right] = \frac{1+q}{2}$. Setting up a firm involves a fixed cost F.

A priori, clients pay for the average quality of the firm but since it is imperfectly observed, there is a probability λ that they pay for an average expect quality q^e. The expected price is thus

$$p = (1-\lambda)\mathbb{E}\left[\theta \,|\, \theta \geq q\right] + \lambda q^e = (1-\lambda)\frac{1+q}{2} + \lambda q^e \tag{21.1}$$

Parameter λ measures the extend of information asymmetry between sellers and buy-

ers. By choosing its size (how many employees), a firm is choosing an average quality which affects the price more or less according to the severity of information asymmetries. We can now study the difference between corporations and partnerships.

Corporation

A corporation (A) maximizes the economic profit where employees cost is evaluated at the market wage, thus $\pi = (1-q)(p-w) - F$. The FOC for optimal size is

$$0 = \frac{\partial \pi}{\partial q} = \frac{(1-\lambda)(1-q)}{2} - (p-w) = w - \left((1-\lambda)q + \lambda q^e\right) \tag{21.2}$$

i.e., the expected ability of newest employee as assessed by the market is the market wage. As $\frac{\partial \pi}{\partial q}$ is decreasing with q, there is a unique solution to (21.2). In equilibrium, the market expectation for average quality is correct, thus $q^e = \frac{1+q}{2}$ so that

$$w = (1-\lambda)q + \lambda\frac{1+q}{2} \Leftrightarrow q_A = \frac{2w-\lambda}{2-\lambda} \leq w \tag{21.3}$$

since $w < 1$ but with equality for $\lambda = 0$. Using the fact that in equilibrium $p - w = \frac{(1-\lambda)(1-q)}{2}$, the equilibrium profit reads

$$\pi_A = \frac{2(1-\lambda)(1-w)^2}{(2-\lambda)^2} - F \tag{21.4}$$

Partnership

A partnership (B) maximizes its per-capita economic profit $u = \frac{\pi}{1-q} = p - w - \frac{F}{1-q}$. Since $\frac{\partial u}{\partial q} \propto (1-q)\frac{\partial \pi}{\partial q} + \pi$, the FOC for optimal size is $\frac{\partial \pi}{\partial q} = \frac{-\pi}{1-q} = -u < 0$. Since $\frac{\partial \pi}{\partial q}$ is decreasing, the marginal ability of a partnership is larger than that of a corporation (solution of (21.2)); hence a partnership always employs less people than a corporation. Using (21.2) and taking out the market wage on both sides, the FOC for an optimal partnership reads $(1-\lambda)\theta + \lambda q^e = p - \frac{F}{1-q}$ i.e., the expected ability of the newest employee as assessed by the market is equal to the average gross profit share. In equilibrium, the market expectation is correct, thus $q^e = \frac{1+q}{2} = p$ (from (21.1)) so that

$$(1-\lambda)q + \lambda\frac{1+q}{2} = \frac{1+q}{2} - \frac{F}{1-q} \Leftrightarrow \frac{(1-\lambda)(1-q)}{2} = \frac{F}{1-q} \Leftrightarrow q_B = 1 - \sqrt{\frac{2F}{1-\lambda}} \tag{21.5}$$

The partnership profit then simplifies to

$$\pi_B = (1-w)\sqrt{\frac{2F}{1-\lambda}} - \frac{2-\lambda}{1-\lambda}F \tag{21.6}$$

Notice that our claim $q_B > q_A$ is strictly equivalent to $\pi_A > 0$. Inspection of (21.4)

reveals that if c or λ is too large, then the corporation is not sustainable and neither is a partnership, a result reminiscent of the lemons problem.

To compare the two organizational structures notice that for $\lambda = 0$, the corporation is efficient because it maximizes the true economic value of the firm i.e., $q_A = w$. The resulting profit $\pi_0 \equiv \frac{(1-w)^2}{2} - F$ is thus the overall maximum and we may conclude that corporation dominates partnership. Now, for the cleverly chosen $\lambda_0 \equiv \frac{\pi_0}{F+\pi_0}$, we have $\frac{2F}{1-\lambda_0} = 2(F+\pi_0) = (1-w)^2$, hence $q_B = w$ by (21.5). This time, it is the partnership that maximizes the true economic value of the firm and is therefore the most efficient organization. By the implicit function theorem, there must exists a threshold $\lambda_1 \in]0;\lambda_0[$ such that $\pi_A = \pi_B$.[3@]

> We conclude that the optimal organizational form is
>
> **Corporation** for low information asymmetries ($\lambda < \lambda_1$).
> **Partnership** for intermediate information asymmetries ($\lambda_1 \leq \lambda \leq \lambda_0$).
> **Neither** for large information asymmetries ($\lambda > \lambda_0$).

21.1.3 Signaling

As in the previous analysis, when an economically relevant personal characteristic is unknown to the market or other trade counterparts, he may want to act so as to *signal* his peculiarity. The concept is introduced by Spence (1973) in the context of education where a worker decides to acquire a costly education to signal his innate high productivity to potential employers.[4@] In the same vein, advertising , on top of being informative (cf. §11.5.2), can also be a signal of high quality. The idea applies also to a central bank whose monetary policy today is a signal of its willingness to accept inflation in the future. Employers then act according to the belief they form about future inflation.

The difficulty with signaling is that the valuable agent faces the threat that a cheap imitator might mimic his attitude and ruin his reputation– and the wage that goes with it–when results reveal the true characteristics. The only way to reveal credibly his true identity is to undertake something so costly that imitators would not dare follow the same path. Yet this commitment or bonding being costly it must be wisely chosen.

Education

It is agreed that at least in the hard sciences, education adds cognitive skills to students. Their market productivity is thus increased and so is the market value of their labor. But education is also a process of socialization where one acquires skills such as the

carrying out of assigned tasks, getting along with others, regularity, punctuality. Lastly, and this the point we pursue here, education is a filter in a world of imperfect information. Employers have a very poor idea of a candidate's productivity although they may know the distribution. By delivering verifiable certificates, schools sort out candidates in small groups whose productivity is easier to assess. The education system thus reveals information about those who go through but also about those who abstain.

Consider a population of workers of either high or low ability in proportion α and $1-\alpha$. Their respective productivities are θ_h and a lower θ_l. If abilities were observable then competition between firms (zero profit condition) would drive differentiated wages to $w_h = \theta_h$ and $w_l = \theta_l$. More plausibly, ability is unobservable. Each worker can then go to university for a duration of q months in order to signal itself to future employers. The disutility of schooling (sic) is $c(q,\theta) = \frac{q}{\theta}$ i.e., both cost and marginal cost are lower for the high ability worker. Notice that education does not increase productivity, it only serves a signal of ability. Outside opportunities are normalized to zero.

Firms compete for workers by offering a wage scheme $w(q)$. In the absence of the educational system the market wage is equal to the average productivity $\mathbb{E}[\theta] = \alpha\theta_h + (1-\alpha)\theta_l$. When a worker shows up with a level of education q, a firm's belief that he is of high ability is α_q. Its rational wage offer is thus $\hat{w}(q) \equiv \alpha_q\theta_h + (1-\alpha_q)\theta_l$.

In a pooling equilibrium all workers choose the same level of education $\tilde{q}$ thus the belief after seeing $\tilde{q}$ remains α so that the equilibrium wage is $\mathbb{E}[\theta]$. The belief after seeing $q \neq \tilde{q}$ can be $\alpha_q = 0$ (if something that should not appear still appears then employers being prudent think that the worker is of the worst type) so that the optimal wage offer is θ_l. Given this behavior, the best alternative to $\tilde{q}$ for workers is to maximize $\theta_l - \frac{q}{\theta}$ i.e., choose no education whatever the innate ability. The proposed equilibrium choice $\tilde{q}$ is thus optimal if $\mathbb{E}[\theta] - \frac{\tilde{q}}{\theta} \geq \theta_l \Leftrightarrow \tilde{q} \leq \alpha\theta(\theta_h - \theta_l)$ for both θ_h and θ_l. The condition is thus $\tilde{q} \leq \alpha\theta_h(\theta_h - \theta_l)$ in order that the *least able* workers accept to go to university. Obviously the efficient equilibrium is $\tilde{q} = 0$ as it saves workers the cost of going to university.

In the separating equilibrium different types of workers chooses different level of education q_h and q_l, thus they are identified and paid at their productivity i.e., $\hat{w}(q_h) = \theta_h$ and $\hat{w}(q_l) = \theta_l$. Beliefs following q_h and q_l are degenerate and for a different q we may set $\alpha_q = 0$. As for optimal schooling it is clear that whatever belief held by employers a low ability type cannot get a wage lesser than θ_l thus he shall either pretend to be a high type by choosing q_h to get $w = \theta_h$ or never go to school to avoid the disutility. Hence $q_l = 0$ and he will not masquerade if $\theta_h - \frac{q_h}{\theta_l} \leq \theta_l \Leftrightarrow q_h \geq q_0 \equiv (\theta_h - \theta_l)\theta_l$. Clearly a high ability type that spends more than q_0 in education cannot be mistaken for a low type. Hence we get a class of equilibria where $q_h \geq q_0$ and $q_l = 0$. It is now clear that other beliefs agree with this outcome; for instance $\alpha_q = 0$ if $q < q_h$ and $\alpha_q = 1$ if $q \geq q_h$. Agents utilities in the

efficient equilibrium where $q_h = q_0$ are $u_l = \theta_l < u_h = \theta_h - \frac{(\theta_h - \theta_l)\theta_l}{\theta_h}$.

There exists equilibria where one type chooses one action while the other randomize between imitation and separation. For instance the high type chooses q_h while the low type imitates with probability λ and separates with $q = 0$ otherwise. The equilibrium belief is thus $\alpha_\lambda = \frac{\alpha}{\alpha+(1-\alpha)\lambda}$ and the associated wage $\hat{w}(q_h) = \alpha_\lambda \theta_h + (1-\alpha_\lambda)\theta_l$. The low type plays a mixed strategy only if he is indifferent between q_h and 0 i.e., $\hat{w}(q_h) - \frac{q_h}{\theta_l} = \theta_l \Leftrightarrow q_h = \bar{q}_\lambda \equiv \frac{\alpha(\theta_h - \theta_l)\theta_l}{\alpha+(1-\alpha)\lambda} \leq q_0$ showing that the efficient separating equilibrium is the limit of the class of mixed equilibrium as λ approaches zero. The off-equilibrium beliefs can be $\alpha_q = 0$ if $q < \bar{q}_\lambda$ and $\alpha_q = \frac{\alpha}{\alpha+(1-\alpha)\lambda}$ if $q > \bar{q}_\lambda$.

Advertising

Kihlstrom and Riordan (1984) show that advertising can be a signal of quality in an oligopoly setting if a high quality producer is patient enough.

The problem faced by a high quality producer is that in the absence of a cheap and reliable certification agency, it is not able to inform correctly potential customers about the quality of its products. For such *experience goods*, the opinion of consumers is either confirmed or revised (upward or downward) according to what they bought and what they anticipated. The following "hit-and-run" strategy immediately comes to mind to take advantage of this delay in the revelation of true quality: package a cheap and low quality product as if it was top-notch, sell it at a high price for one period and make a large profit. Then, one lowers the price since the word has spread regarding what was the real quality inside the gleaming package.

Let π_{ij} denote the profit during one period made by selling quality i when consumer think it is quality j. We obviously have the ranking $\pi_{lh} > \pi_{hh}$ since cost are lower for a lower true quality but also $\pi_{lh} > \pi_{ll}$ since consumers are ready to pay more for (what they think is) a higher quality. The present value of aggregated profits over many periods for the "hit-and-run" strategy is $\pi_{lh} + \frac{1}{r}\pi_{ll}$ where r is the interest rate[5@] and is obviously greater than the profit $\pi_{ll} + \frac{1}{r}\pi_{ll}$ gained by a truthful maker of a low quality product.

To force out the deceptive strategy of "hit-and-run", a high quality producer can simply spend an amount of advertising A larger than the difference $\pi_{lh} - \pi_{ll}$ because it will nullify the benefit of mimicking by a low quality producer. Yet the high quality producer should take care of not spending too much in advertising because he can still change side and switch to the low quality product.[6@] The latter condition reads $\pi_{hh} - A + \frac{1}{r}\pi_{hh} \geq \pi_{ll} + \frac{1}{r}\pi_{ll} \Leftrightarrow A < (\pi_{hh} - \pi_{ll})\frac{1+r}{r}$. A first necessary condition for advertising to be a signal of quality is $\pi_{hh} > \pi_{ll}$ i.e., high quality products must generate more profits than low quality ones in a perfectly informed market. As shown in §11.3, the condition is likely to be satisfied.

Comparing the lower and upper-bound for A we derive a second necessary condition for advertising to be a signal of quality: $r < \frac{\pi_{hh}-\pi_{ll}}{\pi_{lh}-\pi_{hh}}$ i.e., the producer is patient and cares mostly for future profits. The numerator is the truthful profit difference between the qualities while the denominator is the cost difference between the qualities.

21.2 Screening & Self-Selection

The problematic of information acquisition is referred to as screening. For instance, §4.3.3 shows a monopoly (or a firm with market power) trying to extract consumer surplus from its clients (agents). The valuable private information of agents is their willingness to pay for the good sold by the principal. When direct discrimination is unavailable, the firm resorts to indirect discrimination and designs a menu of contracts that is offered to all customers and out of which each type of client picks his preferred option, a procedure called "Self-Selection". This way, the monopolist succeeds to obtain the private information although not completely since more interesting types (high WTP) ends up with an information rent.

This section presents cases of procurement, monopolistic and regulatory screening.[7@] Models are presented in order of difficulty so that each builds heavily on the previous one. In the first case, a principal wants to procure a service from an agent at the lowest possible cost and tries to screen candidates. In the second situation, known as non linear pricing, a firm holding market power discriminate among her clients with a variety of contracts to maximizes profit. The third model studies a regulator trying to elicit cost information from the regulated firm in order to apply an efficient pricing scheme and maximize welfare. We deal with information relative to demand and cost. The fourth model adds a moral hazard dimension to the basic procurement case; the scheme must now modulate the effort incentives on top of screening innate productivities. Lastly, we present the original insurance problem of "adverse selection". When insurers compete perfectly, it is possible to pick up the best clients (aka cream skimming) with a well designed proposal. This may force firms left with risky clients to go out of business so that a market failure occurs.

21.2.1 Procurement

Consider a firm procuring refuse collection to a city or a specialized firm procuring IT services to a manufacturer. According to whether the technology is cutting-edge or obsolete, a quite different outcome is called for. It would be useful for the principal[8@] to know the agent's cost in order to be able to fine tune the volume and price of the service.

Yet, this information is often private knowledge to the agent; the principal thus faces a problem of "information revelation" and must "screen" the agent to uncover it.

Symmetric Information The principal's valuation for output q is $V(q)$ while her WTP for an additional unit is $P(q) = V_m(q)$. The agent's marginal cost θ is either low ($\theta_l = c$) or high ($\theta_h = c+\delta$). This parameter which synthesizes the private information problem here is referred to as the *type* of the agent; in our simple polar example we have the *good* and *bad* types or the *cutting-edge* and *obsolete* technologies.

The objectives are $\pi = V(q) - t$ for the principal and $u = t - \theta q$ for the agent; their sum is the welfare $W(\theta, q) = V(q) - \theta q$. Money is thus simply a means to transfer utility among parties. Ex-ante, before the agent's type is determined, we deal with the expected welfare

$$\mathbb{E}[W] = \alpha \left[V(q_l) - c q_l \right] + (1-\alpha) \left[V(q_h) - (c+\delta) q_h \right] \tag{21.7}$$

Pareto optimality calls for its maximization and can be achieved type-by-type with output $q_\theta^* = P^{-1}(\theta)$, the solution of $V_m = \theta$, for $\theta = l, h$.

If the principal knows the agent's type θ, he can offer him a contract (q, t) with $t \geq \theta q$ in order to meet his participation constraint $u \geq 0$. Since the principal dislikes payments, it is optimal for her to saturate the previous constraint with $t = \theta q$. Profit is now $\pi = W$ and output remains the sole element to be chosen; its optimal level is the efficient one q_θ^*. This so-called "first-best" or ideal scheme is discriminatory since different contracts are proposed to different types of agents. Note the similarity with the basic moral hazard case of §20.1.

Asymmetric Information When the principal ignores the firm's type, she can offer an "attractive" contract guaranteeing full participation i.e., satisfying the participation constraint of any type. Because types are ordered, the condition $\{\forall \theta, t \geq \theta q\}$ boils down to $t \geq \theta_h q$ i.e., the problem is bring the bad type on board. As before, it is optimal to saturate the latter so that the principal's objective becomes $\pi = V(q) - \theta_h q$. This is as if the agent was, for sure, of the worst type, in which case the optimal output is q_h^*. This crude solution completely forsakes the fact that, on average, the agent is much more efficient and that welfare (or principal's profit) could be vastly improved.[9@]

The adequate way to manage the diversity of possible technologies (types) is to offer a menu of contracts and let each type of agent picks his most preferred one. For instance, the principal could offer simultaneously the ideal contracts $\gamma_l^* = (q_l^*, t_l^*)$ and (q_h^*, t_h^*), hoping that each type of agent will pick the one intended for him. The bad type will stick to γ_h^* (over γ_l^*) since by construction we have $0 = t_h^* - \theta_h q_h^* > t_l^* - \theta_h q_l^* = -q_l^*(\theta_h - \theta_l)$; this is because γ_l^* requires a heavy work load paid at less than the actual cost. Sadly, the good

type will also prefer γ_h^* as $0 = t_l^* - \theta_l q_l^* < t_h^* - \theta_l q_h^* = q_h^*(\theta_h - \theta_l)$; this time, the workload is light but paid handsomely because from the point of view of the good type, he stands to make a profit equal to the marginal cost difference on every unit of output.

The question then is whether we can we design a complex screening scheme that would lead to discriminating behavior so as to uncover the agent's type and have him perform in a relatively efficient manner ? The answer is twofold. Yes, we can discriminate (or screen) using only as many contracts as there are types provided they satisfy some incentive constraints. Yet, in the process, the principal has to give a rent to the good type and accept that the bad type under performs wrt. the ideal perfect information situation.

Revelation Principle

We tackle here the issue of complexity with the revelation principle. In order to motivate the agent to reveal his type (private information), the principal can offer a menu of contracts; for instance a red, a green, a yellow and a blue contracts. If type l picks the green one while type h picks the red one then it must be the case that type l prefers the green over the red while type h ranks them inversely. It is then obvious that the blue and yellow contracts were unnecessary in the first place. It is even useless to name contracts with colors; we can directly name them "low cost" and "high cost" (or use labels l and h). As shown by Myerson (1979), this intuition carries on to much more mathematically advanced models of asymmetric information.

In our simple two types settings, the revelation principle allows us to restrict attention to pairs of contracts that are *self-selecting* in the sense that each type of agent prefers the contract designated for him over any other. The conditions that can be deduced from these observations are called the *incentive compatibility* constraints (IC). Lastly, we must not forget the participation constraint i.e., the contract designed for a specific type of agent must leave him with a non negative profit for otherwise he would decline it. This yield a set of *individual rationality* constraints (IR).

Second Best Solution

Given a pair of contracts (q_h, t_h) and (q_l, t_l), the principal's objective is to maximize expected profit

$$\mathbb{E}[\Pi] = \alpha \left[V(q_l) - t_l \right] + (1 - \alpha) \left[V(q_h) - t_h \right] \tag{21.8}$$

under constraints

$$(IC)\begin{cases} t_l-\theta_l q_l \geq t_h-\theta_l q_h & ① \\ t_h-\theta_h q_h \geq t_l-\theta_h q_l & ② \end{cases} \quad \text{and} \quad (IR)\begin{cases} t_l \geq \theta_l q_l & ③ \\ t_h \geq \theta_h q_h & ④ \end{cases} \tag{21.9}$$

The (IC) constraints simplify into $\theta_l(q_l-q_h) \leq t_l-t_h \leq \theta_h(q_l-q_h) \Rightarrow 0 \leq (\theta_h-\theta_l)(q_l-q_h) \Rightarrow q_l \geq q_h$ since $\theta_h > \theta_l$. We also derive $t_l \geq t_h$. Observe now that ④ and ① imply ③ since $t_l-\theta_l q_l \geq t_h-\theta_l q_h > t_h-\theta_h q_h \geq 0$. Neglecting ② for the moment,[10@] the optimum satisfies $t_h = \theta_h q_h = (c+\delta)q_h$ (no rent for the high cost agent) and $t_l = t_h+\theta_l(q_l-q_h) = cq_l+\delta q_h$ (minimum rent for the low cost agent). We may then rewrite the expected profit as

$$\mathbb{E}[\Pi] = \alpha\left[V(q_l)-cq_l\right]+(1-\alpha)\left[V(q_h)-(c+\delta)q_h\right]-\alpha\delta q_h \tag{21.10}$$

We see that wrt. the welfare objective of (21.7), an added cost appears; it is the information rent left to the efficient type to motivate him to reveal his private information (here, his ability to perform at low cost). The optimum in the incomplete information setting is called "second best" because the informational rent generates an allocative distortion as we now demonstrate.

The FOCs for maximizing (21.10) are $P_l(q_l) = c$ and $P_h = c+\frac{\delta}{1-\alpha} > c+\delta$.

The first result is called "no distortion at the top" because the low cost firm (good type) is ordered to produce the efficient quantity. The high cost firm (bad type), on the other hand, is instructed to reduce output (and perform inefficiently) in order to lower the rent left to the good type. This distortion is meant to hurt a good type pretending to be a bad type, while not hurting the bad type as much. Thus, the principal is trading-off the efficient behavior of the low type with the information rent of the high type. Lastly, we check that ②, the incentive constraint for the low type, is satisfied at the candidate optimum as it reads $0 \geq t_l-\theta_h q_l \Leftrightarrow q_l \geq q_h$.

21.2.2 Non Linear Pricing

Non linear pricing by a monopolists involves selling to consumers of unknown type while procurement involves purchasing goods or services from providers of unknown type. The optimization procedures are thus dual one of another.

Indirect Price Discrimination

A firm with marginal cost c sells a good or service under a package (q,t) where q is consumption and t total price (fee). If potential clients form an homogeneous population sharing WTP $P(.)$, the optimal discrimination scheme is to offer the efficient quantity q^*

solving $P(q) = c$ for total fee $t^* = V(q^*)$ where $V(q) \equiv \int_0^q P(x)\,\mathrm{d}x$ is the gross utility enjoyed from consuming q units (cf. §4.1.3).

Consider now the differentiated "home" and "pro" market segments whereby the latter agree to pay premium δ over the former i.e., we have WTPs $P_l(.) = P(.)$ and $P_h(.) = P(.) + \delta$ (and gross utility functions V_l and V_h).[11@] Notice that contrary to the previous section, the "good" type is h. We may interpret the premium $\delta > 0$ as a measure of the information asymmetry affecting the monopolist who cannot distinguish among the two types.

By the revelation principle, the monopoly need only design two contracts (q_h, t_h) and (q_l, t_l) satisfying self-selection. Letting $u_i \equiv V_i(q_i) - t_i$ for $i = l, h$, the per-type profit is $\pi_i = t_i - cq_i = V_i(q_i) - cq_i - u_i$ (note that $\pi_i + u_i$ is welfare). The individual rationality constraints (IR) are $u_i \geq 0$ while the incentive compatibility constraints (IC) are $u_l \geq V_l(q_h) - t_h = u_h + \delta q_l$ and $u_h \geq V_h(q_l) - t_l = u_l - \delta q_h$ which simplify into

$$\delta q_l \leq u_h - u_l \leq \delta q_h \tag{21.11}$$

from which we deduce $q_h \geq q_l$ (in a manner quite similar to (21.9)).[12@] The information rent of the "pro" is again the lower bound δq_l in (21.11); it is the net surplus he can secure by grabbing the proposal designed for a "homer" i.e., by pretending to be of a different type.

As in the procurement case, we come to the conclusion that the firm leaves no rent to a "homer" (i.e., sets $u_l = 0$) and leaves the minimum information rent to a "pro" (i.e., sets $u_h = \delta q_l$). The expected profit is thus very much like (21.10) with

$$\mathbb{E}[\Pi] = \alpha\pi_h + (1-\alpha)\pi_l = \alpha\left[V_h(q_h) - cq_h\right] + (1-\alpha)\left[V_l(q_l) - cq_l\right] - \alpha\delta q_l \tag{21.12}$$

The FOCs are $c = P_h = P(.) + \delta$, leading to choose the efficient quantity $\hat{q}_h = q_h^*$, and $P_l = P(.) = c + \frac{\alpha\delta}{1-\alpha}$ leading to choose a reduced quantity $\hat{q}_l < q_l^*$.[13@] The ratio $\frac{\alpha\delta}{1-\alpha}$ may be interpreted as the marginal cost of information acquisition.

Observe that if α and δ are large then $P_l(0) < \frac{\alpha\delta}{1-\alpha}$ might become true. In that case, it is better to exclude "homers" by setting $q_l = 0$. The monopolist does this to avoid paying the information rent to "pros"; recall indeed that if the proposal is unappealing to "homers", the firm can extract all the consumer surplus from "pros" because she faces a unique segment of homogeneous people.

Generalization

In the most simple model, the average WTP for the good is $P(q)$ so that average TWTP is $U(q) \equiv \int_0^q P(x)\,\mathrm{d}x$. The heterogeneity among customers is captured by a random variable $\tilde{\theta}$ of law H and unitary mean with the following meaning: a consumer of type θ has WTP $\theta P(.)$. Taking into account the marginal cost of service c, the gross welfare associated to type θ is $w(\theta, q) = \theta U(q) - cq$. Net welfare is $W(\theta) \equiv w(\theta, q_\theta^*)$ where the efficient quantity q_θ^* solves $\theta P(q) = c$. If the monopolist could perfectly discriminate, he would sell type θ the amount q_θ^* against total payment $t_\theta^* = \theta U(q_\theta^*)$ or price at marginal cost together with a subscription $f_\theta^* = t_\theta^* - cq_\theta^*$.

When discrimination is impractical, the monopolist sets a variable unit price $p(q)$ or equivalently a tariff $T(q) \equiv \int_0^q p(x)\,\mathrm{d}x$. When faced with this tariff, a consumer of type θ derives a net surplus or utility $u\big(\theta, q\big) = \theta U(q) - T(q)$, thus chooses or demands the optimal level $q(\theta)$ satisfying the FOC[14@]

$$\theta U'(q) = T'(q) \Leftrightarrow \theta P(q) = p(q) \tag{21.13}$$

Since the WTP is decreasing, a larger θ moves the LHS upwards and thus forces an intersection with the $p(.)$ curve further on the right (whatever its shape may be); this means that $q(\theta)$ is increasing.[15@]

If the firm wants type θ to consume some particular level $\hat{q}$, it is enough to set $p(\hat{q}) = \theta P(\hat{q})$. This shows that the firm is able to induce any consumption path across types by tuning her tariff adequately. Hence, the search for an optimal tariff $T(q)$ is a quest for an optimal path of consumption $q(\theta)$. We already know that the firm cannot fully extract the consumer surplus and must leave each type a rent above his opportunity cost. Indeed, the indirect utility or equilibrium surplus is $v(\theta) \equiv u(\theta, q(\theta))$ and satisfies $v' = U$ by the envelope theorem. Integrating, we obtain a cumulative formula with $v(\theta) = v(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} U(q(x))\,\mathrm{d}x$.

Setting w.l.o.g. the opportunity cost at zero, the participation constraint is $u\big(\theta, q(\theta)\big) \geq 0$; it must hold for all types but it is in fact enough to make sure it holds for the lowest one $\underline{\theta}$ since $u\big(\theta, q(\underline{\theta})\big) \geq \big(\theta - \underline{\theta}\big) U\big(q(\underline{\theta})\big) \geq 0$ i.e., higher types can always use the scheme designed for the lowest type and get more than their opportunity cost. Now, it is obvious that the firm will adjust her scheme so as to leave no unnecessary rent to the lowest type i.e., $v(\underline{\theta}) = 0$ holds at the optimum. Ex-ante, the rent paid to the entire population is

$$\mathbb{E}\big[v(\tilde{\theta})\big] = \int v(\theta) h(\theta)\,\mathrm{d}\theta = \int \int_{\underline{\theta}}^{\theta} h(\theta) U(q(x))\,\mathrm{d}x\,\mathrm{d}\theta = \int (1 - H(\theta)) U(q(\theta))\,\mathrm{d}\theta$$

using integration by part and assuming the distribution H has a density h. Finally,

$$\mathbb{E}\left[v(\tilde{\theta})\right] = \int \frac{1-H(\theta)}{h(\theta)} U(q(\theta))\,\mathrm{d}H(\theta) = \mathbb{E}\left[\rho\left(\tilde{\theta}, q(\tilde{\theta})\right)\right] \tag{21.14}$$

where $\rho(\theta, q) \equiv \frac{1-H(\theta)}{h(\theta)} U(q)$. The usefulness of this expression is to disentangle the exogenous type parameter from the endogenous quantity choice of the consumer.

By construction, profit is welfare minus utility i.e., $T(q) - cq = w(\theta, q) - u(\theta, q)$. Taking into account the fact that each type of consumer demands a particular amount $q(\theta)$ of the good, the firm earns $\pi(\theta) = \theta U\left(q(\theta)\right) - cq(\theta) - v(\theta)$ per-capita. Using (21.14), the ex-ante profit simplifies into $\Pi = \mathbb{E}\left[\pi(\tilde{\theta})\right] = \mathbb{E}\left[\varphi\left(\tilde{\theta}, q(\tilde{\theta})\right)\right]$ where $\varphi(\theta, q) \equiv \theta U(q) - cq - \rho(\theta, q)$. It is now obvious that the maximization of the ex-ante profit requires an ex-post maximization for each distinct type of customer: the optimal consumption $q(\theta)$ is that maximizing $\varphi(\theta, q)$. Using (21.13), the FOC is

$$\left(\theta U'(q) - c\right) h(\theta) = (1 - H(\theta))\, U'(q) \Leftrightarrow \frac{p-c}{p} = \frac{1-H(\theta)}{\theta h(\theta)} \tag{21.15}$$

The rightward version is a Lerner index (cf. eq. (3.4)) showing how the efficiency condition is distorted away by the monopolist. Note however the absence of distortion (at the top) for the highest type who also consumes most. Conversely, the distortion is greatest for the lowest type. The leftward version of (21.15) shows that the monopolist willing-fully creates a deadweight-loss to reduce the information rent left to high type customers. Indeed, the LHS is the gain from marginally extending sales to type θ times the population of this particular type while the RHS is the infra-marginal loss of allowing a greater rent to all better types.

For most statistical distributions, the hazard rate $\frac{h(\theta)}{1-H(\theta)}$ is increasing, so that the RHS of (21.15) is decreasing with type i.e., the price solution is decreasing. Now, rewriting (21.13) as $P(q) = p(q)/\theta$, we see that the RHS decreases with type. Since WTP is decreasing, it must be the case that the optimal output is increasing as required to be the optimal choice of the consumer.[16@] We may thus conclude:

21.2.3 Public Firm Regulation

As explained in §17.2, the ideal (second best) regulation for a public service is a per unit price close to marginal cost together with a subscription large enough to cover fixed cost. Choosing these two parameters is not an easy task for the regulator because she can only use the cost report handed over by the firm's manager who might therefore have an incentive to overstate true cost. This possibility is a real issue given his better knowledge

of the production technology. Combined with the fact that public funds involve a loss of efficiency (cf. §17.1.2), the optimal regulatory policy will be distorted as if there was an additional marginal cost of eliciting the private information of the firm.

Another issue affecting regulation is the cyclicality of demand of many public services. The existence of peaks and off-peak periods warrants a flexible price policy from an efficiency point of view. For reasons better explained in §25, inter-temporal discrimination is often forbidden and leads to excess capacity and large deadweight losses. In a few cases such as the pricing of highways operated under franchise, regulators are more and more willing to adapt their regulatory framework to account for this reality. Yet they face the problem of eliciting the demand from the operator in order to set the price at the efficient level. This is what we explore in the first part before delving into the more involved case of unknown cost.

Private Information on Demand

To assess the variability of demand, we simply assume that the market size parameter a in the usual demand formulation $D(p) = a - bp$ is a random variable observed only by the regulated firm, thus unknown to the regulator. The firm has to build a capacity of service k (at marginal cost δ) and can produce up to k units at marginal cost c.

Efficient Pricing The efficient pricing rule is found by following a simple procedure. Try first to price at marginal cost; if the resulting demand can be met ($D(c) \leq k$) this is it; if on the contrary there is potential congestion ($D(c) > k$), then raise the price until excess demand vanishes ($D(p_k) = k$) i.e., in our example, set $p_k = \frac{a-k}{b}$. This is the most efficient manner to ration consumers i.e., that which minimizes the welfare loss generated by the limited service capacity (cf. §25.3.2 on congestion). The threshold market size such that marginal cost pricing does not generate congestion is $a_k \equiv k + bc$. We thus obtain $p_k(a) \equiv \min\{c, \frac{a-k}{b}\} = \begin{cases} c & \text{if } a < a_k \\ \frac{a-k}{b} & \text{if } a \geq a_k \end{cases}$.

The fact that the efficient price is a rule depending on some piece of information unknown to the regulator poses a problem for he has to trust the firm to reveal it correctly. However this is not going to happen because the firm has an incentive to overstate the true demand. Indeed, if the firm claims that the demand parameter is some $\hat{a}$ greater than the true level a, then the regulator will allow the price $p_k(\hat{a})$; it then remains to adjust $\hat{a}$ so as to equate $p_k(\hat{a})$ with the monopoly price.[17@]

Fortunately, Riordan (1984) displays a simple contract restoring the incentives of the firm toward the common good: the firm is allowed to set the price p, has to meet demand up to capacity and receives subsidy $s_k(p) \equiv \delta k - (p - c)k$, contingent on her announced

price.[18@] The scheme involves a fixed subsidy covering the capacity cost and a variable tax component to extract any margin made over sales. The firm's profit is now $\pi(p) = (p-c)\min\{D(p),k\} - \delta k + s_k(p) \leq 0$ by the very definition of $s_k(.)$. Hence, the most that she can expect is to earn zero (extraordinary) profit. We now show that $p_k(.)$ is a rule that precisely enables to achieves this maximum. Indeed, the firm earns:

$$\pi(p_k) = \begin{cases} (c-c)D - \delta k + (c+\delta-c)k & = 0 \quad \text{if } a < a_k \\ (p_k-c)k - \delta k + (c+\delta-p_k)k & = 0 \quad \text{if } a \geq a_k \end{cases}$$

> We conclude that the optimal regulatory mechanism is akin to a two-part tariff whereby both the subsidy and the unit price respond to shifting demand conditions.

Efficient Capacity The efficient capacity is found by equating the marginal cost δ of capacity expansion to its marginal value which is the wedge between the WTP of the first rationed consumer (i.e., the price) and the marginal cost of service c. Since the price varies with the intensity of demand, the efficient capacity k^* solves $\mathbb{E}\left[p_k(.)\right] = c + \delta$.

When the regulator offers the price contingent subsidy $s_k(.)$, the firm prices efficiently and thus reveals the state of demand. The regulator is thus able to compute the expectation $\mathbb{E}\left[p_k(.)\right]$ to ascertain whether the actual capacity is too small or too large.

Once the firm has been instructed to invest efficiently at the level k^*, the subsidy mechanism becomes budget balanced; indeed, the expected subsidy is $\mathbb{E}\left[s_k(.)\right] = k\left(c+\delta-\mathbb{E}\left[p_k\right.\right.$ since the firm is motivated to price efficiently. This expression is zero precisely when the efficient capacity k^* is build. This is a very convenient property since deficit or surplus are always source of costly haggling among stake-holders in the regulation process.

Private Information on Costs

When the regulated firm knows better than the regulator its ability to perform the job, we have a situation similar to procurement. The only difference is the principal's objective: a regulator cares for both consumer surplus and firm profit but is reluctant to tax the rest of the economy to finance the firm. The marginal cost of public λ funds alluded to in §17.1 will introduce a slight modification of the optimal second-best regulatory policy.[19@] We follow Laffont and Tirole (1986) and Baron and Myerson (1982).

The procurement model for variable output seen in §21.2.1 is easily adapted. The regulated firm's profit is $\pi = t - \theta q$ where the marginal cost is $\theta_l = c$ with probability α or $\theta_h = c + \delta$ with complementary probability. Raw consumer surplus is $U(q)$ (with WTP $P(.)$) while a contract is an output-transfer pair (q,t). The regulator's objective is welfare

net of the cost of public funds i.e.,

$$W = U(q) - \theta q - \lambda t = U(q) - (1+\lambda)\theta q - \lambda\pi \quad \propto V(q) - \theta q - \tau\pi \tag{21.16}$$

with $V(q) \equiv \frac{U(q)}{1+\lambda}$ and $\tau \equiv \frac{\lambda}{1+\lambda}$.

The (IC) and (IR) constraints are identical to (21.9) leading to an information rent $\tau\delta q_h$ once transfers are minimized. Expected welfare is thus

$$\mathbb{E}[W] = \alpha\left[V(q_l) - cq_l\right] + (1-\alpha)\left[V(q_h) - (c+\delta)q_h\right] - \alpha\tau\delta q_h \tag{21.17}$$

which is identical to (21.10) except for the multiplier τ in the last term.

The FOC for the low cost firm is $V_m = c \Leftrightarrow P(.) = (1+\lambda)c$ i.e., she produces the efficient quantity (*no distortion at the top*). The FOC for the high cost firm is $P(.) = (1+\lambda)(c+\delta) + \frac{\alpha\lambda\delta}{1-\alpha}$ i.e., equalizes marginal benefit with a *virtual* marginal cost summing the true marginal cost of production and the cost of eliciting the firm's private information.[20@] Although the regulator does not care per-se for wealth distribution, he tries to avoid distortionary tax collection and is thus eager to minimize the information rent left to the low cost firm. Consequently, the dearer firm is instructed to produce less than what would be efficient to avoid imitation by a cheap firm $\left(\hat{q}_h < q_h^*\right)$.

21.2.4 Procurement and Moral Hazard

In many instances of procurement, the principal (e.g., government, firm) contracts the agent (e.g., builder, maker, consultant) to produce a single and idiosyncratic item such as a bridge, an industrial design or a firm re-organization. In that case, the principal would like the agent to specialize over that task and invest to develop specific skills that would reduce the overall cost the project.

As before, the agent's basic technology can be cutting-edge or obsolete, an information known only to him. The novelty here is the possibility to meliorate the technology (reduce cost) by investing into assets or capital specific to the delegated task (cf. §14.2). Denote β the cost saving undertaken be the agent under a regime of residual claimancy (cf. §20.1).[21@] This investment, however, is not contractible as it involves personal and unobservable skills and actions. This element of moral hazard (cf. §20) generates an adverse selection problem.

Worst Case: CP and FP

Under a Cost-Plus Contract (CP), the principal repays the agent all his expenses (related to the task). The latter seeing that any improvement effort falls on his shoulders

will undertake none (cf. §20.1.2). The principal thus reimburses the full cost θ and on expectation pays $\bar{\theta} \equiv \mathbb{E}[\theta]$.

Under a fixed price (FP) contract, the principal offers the fixed price t for completion of the item. The agent, whatever his type, then becomes residual claimant of any technology melioration thus invest optimally and saves β. His payoff is then $t - \theta + \beta$. He will accept the initial offer only if $t \geq \theta - \beta$. If the principal insists on getting the item no-matter-what, she has to set a high price to attract the high cost firm (bad type) i.e., $t \geq \theta_h - \beta$ (with equality at the optimum). In that case, the low cost firm (good type) enjoys a large rent.

This "no risk FP" contract is better than CP if $\theta_h - \beta \leq \bar{\theta} \Leftrightarrow \beta \geq \alpha\delta$ i.e., when the gain from optimized technology is large, so that providing good incentives is important. Conversely, if the information asymmetry is large (high δ) then CP is better to avoid leaving such a large rent to the low cost agent.

Middle Case: a simple option

A frequently observed contract is a CP with an option to switch to FP. Obviously, if the fixed price t is low, no one ever picks the option so we are back to the pure CP case. Likewise, if t is large, everyone picks the FP option and we are back to the "no risk FP" case. For intermediate t, only the low cost agent picks the FP option and implements the maximal cost saving β. Among these, $t = \theta_l - \beta$ is optimal as it minimizes the payment. The expected cost for the principal is thus $\alpha(\theta_l - \beta) + (1-\alpha)\theta_h = \bar{\theta} - \alpha\beta$ which clearly dominates the CP contract. The practical value of this simple option contract is that it can be computed without any knowledge of the agent's ability to improve his technology i.e., without resorting to a screening mechanism.

Best Case: no information asymmetry

To develop fully the model we follow footnote 26.3 and denote e a marginal cost reduction and $d(e)$ the financial cost of achieving it. The principal's objective is to minimize the expected cost

$$\mathbb{E}[C] = \alpha\left[\theta_l - e_l + d(e_l)\right] + (1-\alpha)\left[\theta_h - e_h + d(e_h)\right] \tag{21.18}$$

where e_i is the level of cost savings implemented by the agent (which depends on the contract he agreed to).

If the principal could observe the agent's type, he could taylor a FP contract for each type $i = l, h$ with $t_i^* = \theta_i - \beta$. This offer would be accepted and followed by implementation of the efficient cost savings so that realized cost would be $\theta_i - \beta$. The principal would thus get the item done for an expected cost of $\mathbb{E}[C] = \bar{\theta} - \beta$ i.e., the maximal improvement over

the CP base case. Equivalently, the principal offers the agent to bring cost down to $\theta_i - \beta$ and pays him exactly that amount.

In the realistic case where the principal does not observe the agent's type, she has to devise a revelation mechanism. We already know from the previous models that naively offering the efficient FP contracts t_h^* and t_l^* fails because both types prefer the greater payment t_h^* i.e., the outcome is the same as with to the "no risk FP" contract.

Second Best: complex option

Our basic instrument is the target contract (c, t) whereby the principal agrees to pay the agent $t + c$ if the cost $\theta - e$ is brought down to c. When type θ accepts such an offer, he must implement a minimum effort saving $e = \theta - c$ to get the payment.[22@] His profit (rent) is then $\pi(\theta, c, t) = t - d(\theta - c)$ since procurement expenses are paid by the principal.

The principal offers contracts (c_l, t_l) and (c_h, t_h). We denote $e_i = \theta_i - c_i$ for $i = l, h$ and $\pi_i = \pi(\theta_i, c_i, t_i)$. The (IC) and (IR) constraints are then $\pi_i \geq \pi(\theta_i, c_j, t_j)$ and $\pi_i \geq 0$ for $i = l, h$ and $j \neq i$. With target contracts, if the low cost agent picks contract (c_h, t_h), he will have to exert effort $\theta_l - c_h = e_h - \delta < e_h$ because he is intrinsically more efficient than the high cost agent. On the contrary, if the high cost agent picks (c_l, t_l), he will be forced to exert effort $\theta_h - c_l = e_l + \delta > e_l$ since it is more difficult for him to pretend to be a low cost agent.

The principal's objective is to minimize the expected cost of producing the item

$$\mathbb{E}[C] = \mathbb{E}[c_i + t_i] = \mathbb{E}[\theta_i - e_i + d(e_i) + \pi_i] \tag{21.19}$$

under the (IC) and (IR) constraints

$$\begin{cases} t_l - d(e_l) \geq t_h - d(e_h - \delta) & (IC_l) \\ t_h - d(e_h) \geq t_l - d(e_l + \delta) & (IC_h) \end{cases} \quad \text{and} \quad \begin{cases} t_l \geq d(e_l) & (IR_l) \\ t_l \geq d(e_l) & (IR_h) \end{cases} \tag{21.20}$$

As before, the (IC) conditions simplify to $d(e_l + \delta) - d(e_h) \geq t_l - t_h \geq d(e_l) - d(e_h - \delta) \equiv v$ which is the information rent that the low cost agent receives to reveal his type. Minimizing rents by setting $\pi_h = 0$ and $\pi_l = v$, the objective becomes

$$\mathbb{E}[C] = \alpha[\theta_l - e_l + d(e_l)] + (1 - \alpha)[\theta_h - e_h + d(e_h)] + \alpha[d(e_h) - d(e_h - \delta)] \tag{21.21}$$

which is the complete information formula (21.18) plus the cost of information acquisition, the informational rent paid to the low cost firm (good type). The FOCs are

$$d'(e_l) = 1 \quad \text{and} \quad d'(e_h) = 1 - \frac{\alpha}{1 - \alpha}\left(d'(e_h) - d'(e_h - \delta)\right) \tag{21.22}$$

leading to optimal choices $e_l = e^*$ and $e_h < e^*$ i.e., no distortion at the top and an optimal reduction of effort at the bottom to reduce the information rent. Lastly, we need to check that (IC_h) is satisfied at the candidate optimum. The condition reads $\pi_h = 0 \geq t_l - d(e_l + \delta) \Leftrightarrow d(e_h) - d(e_h - \delta) \leq d(e_l + \delta) - d(e_l)$ which is true because we assumed d convex and we have $e_h < e_l$ at the optimum.

This complex scheme improves over the simple CP-FP option but not so much as we now proceed to show.

Extension to many types †

To compare the second-best optimal self-selecting menu of contracts with the simple CP-FP option, Rogerson (2003) assumes that the agent's type is uniformly distributed between θ_l and θ_h. We also take $d(e) = \frac{e^2}{4\beta}$ in order that $e^* = 2\beta$ and that β is indeed the maximum of $e - d(e)$.

A CP-FP option at t leads all types below $\hat{\theta} = t + \beta$ to pick the FP option and all the types above to stick to CP. The contract is thus parametrized by t or $\hat{\theta}$. The expected cost is then $C(\hat{\theta}) = \int_{\theta_l}^{\hat{\theta}} (\hat{\theta} - \beta) \, \mathrm{d}F(x) + \int_{\hat{\theta}}^{\theta_h} x \, \mathrm{d}F(x)$ and its derivative is proportional to $f(\hat{\theta})(\hat{\theta} - \beta - \hat{\theta}) + F(\hat{\theta})$. With the uniform distribution, the equation becomes $\frac{\hat{\theta} - \theta_l}{\delta} = \frac{\beta}{\delta}$ and the optimal cutoff level is thus $\hat{\theta} = \min\{\theta_h, \theta_l + \beta\} = \theta_l + \min\{\delta, \beta\}$. The optimal optional FP is then $t^* = \theta_l + \min\{0, \delta - \beta\}$. The expected cost is $\hat{C} = \bar{\theta} - \min\left\{\beta - \frac{\delta}{2}, \frac{\beta^2}{2\delta}\right\}$. The percentage improvement of the simple option over the basic CP in terms of the maximum cost saving k is then $\hat{\rho} \equiv \frac{\bar{\theta} - \hat{C}}{\beta} = \min\left\{1 - \frac{\delta}{2\beta}, \frac{\beta}{2\delta}\right\}$.

Under the optimal second-best contract, the principal offers a menu $(c(\theta), t(\theta))$ for $\theta \in [\theta_l, \theta_h]$. We let $e(\theta) = \theta - c(\theta)$ and $\pi(\theta) = t(\theta) - d(e(\theta))$. The IC condition is $\dot{\pi}(\theta) = -\dot{d}(e(\theta))$ while the expected cost is $\mathbb{E}[\theta - e(\theta) + d(e(\theta)) + \pi(\theta)]$. The FOC for θ is then $\dot{d}(\theta) = 1 - \frac{F(\theta)}{f(\theta)} \ddot{d}(\theta)$. In our uniform distribution example, it reads $\frac{e}{2\beta} = 1 - \frac{\theta - \theta_l}{2\beta}$ leading to $e(\theta) = \min\{0, 2\beta + \theta_l - \theta\}$. Finally Rogerson (2003) shows that $\rho^* = \min\left\{1 - \frac{\delta}{2\beta} + \frac{\delta^2}{12\beta^2}, \frac{2\beta}{3\delta}\right\}$.

The ratio of $\hat{\rho}$ to ρ^* is always greater than $\frac{3}{4}$ and equal to this value for $\delta \geq 2\beta$. Hence, a simple CP contract with a fixed price option computed without knowledge of the agent's preferences is within 75% of the second-best efficiency frontier.

Accounting Regulation

Laffont and Tirole (1986) studies the previous regulation problem with the slight improvement that realized cost are retrievable from the accounting statements. This allows her to instruct the firm to produce efficiently, conditional on the realized cost. The regulator's problem now is that she is unable to disentangle the role of effort e from the underlying marginal cost θ in the final unit cost $c = \theta - e$. A low observed cost occurs

only when luck and hard work are combined together while a high cost is a sure indication of low effort (and back luck). The identification problem lies with an intermediate observation that can be due to luck if $\theta=\theta_l$ or hard work if $\theta=\theta_h$.

The adverse selection formulation follows §21.2.3 and adds a moral hazard issue as in §21.2.4. Welfare is thus a combination of (21.16) and (21.19) with $W=V(q)-(\theta-e)q-d(e)-\lambda t$. Since profit is $\pi=t-(\theta-e)q-d(e)$, the expected welfare

$$\begin{array}{rcrl}\mathbb{E}[W] &=& \alpha & \left[V(q_l)-(1+\lambda)\left((\theta_l-e_l)q_l+d(e_l)\right)-\lambda\pi_l\right]\\ &+& (1-\alpha) & \left[V(q_h)-(1+\lambda)\left((\theta_h-e_h)q_h+d(e_h)\right)-\lambda\pi_h\right]\end{array} \tag{21.23}$$

is to be maximized under the (IC) and (IR) constraints (21.20). As in the previous model, the rents are minimized so that (21.23) becomes

$$\begin{array}{rcrl}\mathbb{E}[W] &=& \alpha & \left[V(q_l)-(1+\lambda)\left((\theta_l-e_l)q_l+d(e_l)\right)-\lambda\left(d(e_h)-d(e_h-\delta)\right)\right]\\ &+& (1-\alpha) & \left[V(q_h)-(1+\lambda)\left((\theta_h-e_h)q_h+d(e_h)\right)\right]\end{array} \tag{21.24}$$

Observe, as claimed initially, that the FOC for quantity is $P=(1+\lambda)(\theta-e)$ i.e., production is efficient conditional on the achieved level of marginal cost. As for moral hazard, the effort of the cheap firm is the efficient level e_l^* since it solves $d'(e_l)=q_l$ but the dear firm's FOC is $d'(e_h)=q_h-\frac{\lambda\alpha}{(1+\lambda)(1-\alpha)}\left(d'(e_h)-d'(e_h-\delta)\right)$ so that $\hat{e}_h<e_h^*$. There is an indirect effect on production i.e., $\hat{q}_h<q_h^*$, since a lesser effort raises the marginal cost while for the low cost firm, we have $\hat{e}_l=e_l^*$ and thus $\hat{q}_l=q_l^*$.

Continuous types † To simplify we consider the procurement case where just one unit of the good is needed ($q=1$). The random variable θ follows the distribution H. Since profit is $\pi=t-(\theta-e)-d(e)$, the first best effort e^* solves $d'=1$. The regulator can either mandate effort e^* or requires to bring cost down to $c^*=\theta-e^*$, the transfer is then adjusted to cover all costs i.e., $t=\theta-e^*-d(e^*)$.

Under incomplete information, although c is observed, θ is not anymore so that the efficient effort cannot be verified. The regulator can propose the firm a payment against a cost level. By the revelation principle, he will offer only schemes $(t(.),c(.))$ inducing truthful revelation of types. Profit upon revealing $\tilde{\theta}$ but being type θ is $\Phi(\theta,\tilde{\theta})=t(\tilde{\theta})-c(\tilde{\theta})-d(\theta-c(\tilde{\theta}))$. The incentive compatibility constraint is $\frac{\partial\Phi}{\partial\tilde{\theta}}=0$ so that $\pi(\theta)\equiv\Phi(\theta,\theta)$ satisfies $\dot{\pi}=-\dot{d}(e(\theta))$ where $e(\theta)\equiv\theta-c(\theta)$. We may write $\pi(\theta)=\pi(\bar{\theta})+\int_\theta^{\bar{\theta}}\dot{d}(e(x))\mathrm{d}H(x)$ and since there is no need to leave excessive rent to the firm, we have $\pi(\bar{\theta})=0$.

As the value of the procured object is fixed, the regulator objective is to minimize the expected cost $\mathbb{E}[(1+\lambda)(\theta-e+d(e))+\lambda\pi]$ over IC schemes. If he increases the effort of type θ, whose density is $h(\theta)$, he reduces cost by $(1+\lambda)(\dot{d}-1)$ but the IC constraint forces him

to increase the rent left to all better types by $\ddot{d}(e)$ and there are $H(\theta)$ of these. The level of effort for type θ is optimal when $h(\theta)(1+\lambda)(\dot{d}-1) = H(\theta)\lambda\ddot{d}(e) \Leftrightarrow \dot{d} = 1 - \frac{\lambda H(\theta)}{(1+\lambda)h(\theta)}\ddot{d}(e)$. Observe that only the most efficient firm picks the efficient effort (no distortion at the top) since $H(.) = 0$ at the bottom of the distribution. For worse firms, effort is sub-optimal but their information rent is lesser.

Variable Quantity † Assume total cost is $C(\theta, q, e) = (\theta - e)q + d(e) + F$ where θ is the base marginal cost, F the fixed cost, e is the effort or investment of the management into cost reductions and d the opportunity cost of effort or disutility (assumed increasing convex). The profit is $\Pi = qP(q) - C(\theta, q, e) + t$ and the optimal effort for the manager is $\hat{e}(q)$ solving $\frac{\partial C}{\partial e} = d' - q = 0$. Total welfare is

$$W(\theta, q, e) = V(q) - C(\theta, q, e) - \lambda t = V(q) - (1+\lambda)C(\theta, q, e) + \lambda qP(q) - \lambda\Pi$$

so that the efficient effort is also $\hat{e}(q)$; hence it is efficient to leave the firm elicit effort. Observe now that

$$\begin{aligned} W\left(\theta, q, \hat{e}(q)\right) &= V(q) - (1+\lambda)\left[(\theta - \hat{e}(q))q + F + d(\hat{e}(q))\right] + \lambda qP(q) - \lambda\Pi \\ &= W_D(q) - (1+\lambda)\theta q - \lambda\Pi \end{aligned}$$

where $W_D(q) \equiv V(q) + (1+\lambda)\left[\hat{e}(q)q - d(\hat{e}(q)) - F\right] + \lambda qP(q)$ is the so-called "social valuation" of the production that the regulator can compute provided he knows d. (cf. 2.1 of Laffont and Tirole (1993) for public goods). The efficient production, called the Ramsey solution, solves $(1+\lambda)\theta = W_D' = (1+\lambda)\left(P(q) + \hat{e}(q)\right) + \lambda qP'(q)$

$$\Leftrightarrow P + \frac{\lambda}{1+\lambda}qP' = \theta - \hat{e}(q) = c \Leftrightarrow L = \frac{p-c}{p} = \frac{\lambda}{1+\lambda}\frac{1}{\epsilon}$$

where L is the Lerner index (cf. eq. (3.4)) and ϵ is the price elasticity of the demand. If there is no tax distortion ($\lambda = 0$), we fall back on marginal pricing.

21.2.5 Insurance Cream Skimming †

We present here a simplified version of Rothschild and Stiglitz (1976)' analysis of insurance markets. Consider insurance against economic losses c due to an unexpected event like a fire that forces a firm to stop production or an accident stopping an individual from working. The underlying risk for each client that seeks insurance is different, thus the insurer will try to gather a maximum amount of information to assess each riskiness. Yet some differences among clients remain unobservable. There are *safe* customers whose

accident probability is θ_l and *unsafe* ones whose accident probability is a greater θ_h. A policy consists of a premium p and a reimbursement d in case of damages. The expected utility of a type θ customer is

$$U(q,c) = \theta u\big(d-c-p\big) + (1-\theta)u(-p) \tag{21.25}$$

where u is increasing concave since customers are all risk averse (otherwise they wouldn't seek insurance). The per capita profit of a risk neutral insurer is $\pi = p - \theta d$.

Pareto efficiency commands to eliminate risk by setting $d = c$. This efficient risk sharing outcome is achieved by competitive insurance markets insofar as risk classes are correctly identified by insurers; indeed, free entry and perfect competition drive economic profit to zero i.e., to actuarially fair premiums $p = \theta c$.

Cream Skimming

As soon as the different types of agents cannot be distinguished, the insurance market breaks down; more precisely, there cannot exist a pooling equilibrium where the same contract $\gamma = (p,d)$ is bought by all types of agents. The underlying reason is that the average insurance cost, $\mathbb{E}[\theta]$, stands between the real values θ_l and θ_h meaning that safe customers generate profits that can be used to cover the losses generated by unsafe customers.

Cream skimming is the process whereby a competing insurer succeeds to attract profitable safe customers and woo away the unprofitable unsafe ones. The trick is to offer a deductible increase ($\Delta d < 0$) i.e., force customers to assume more risk, together with a premium reduction ($\Delta p < 0$) that compensate safe risk customers only. To prove this claim, we compute

$$\begin{aligned}
\Delta U(\theta) &= \theta u'\big(d-c-p\big)\Delta d - \big[\theta u'\big(d-c-p\big) + (1-\theta)u'(-p)\big]\Delta p \\
&\propto \Delta d - \Delta p - \frac{1-\theta}{\theta}\frac{u'(-p)}{u'\big(d-c-p\big)}\Delta p \\
\Rightarrow \frac{\Delta U(\theta)}{\Delta d} &\propto 1 - \frac{\Delta p}{\Delta d}\left[1 + \frac{1-\theta}{\theta}\frac{u'(-p)}{u'\big(d-c-p\big)}\right]
\end{aligned} \tag{21.26}$$

and observe that $\Delta U(\theta) < 0$ for a small ratio $\frac{\Delta p}{\Delta d}$ while it is positive for a large ratio $\frac{\Delta p}{\Delta d}$. Hence we may adjust $\frac{\Delta p}{\Delta d}$ so that $\Delta U(\theta_l) = 0$ holds. Now, observe that the last term in the bracket decreases with the risk index θ, thus $\Delta U(\theta_h) < 0$ i.e., the conditions for cream skimming hold true. The change in the contract leaves a low risk indifferent while repelling a high risk.

For the original γ to be a candidate equilibrium, it must not generate losses, thus

$p - \mathbb{E}[\theta]d \geq 0$ so that $p - \theta_l d > 0$. We can now choose Δd small enough to guarantee that the insurer offering the altered contract earns the positive payoff $p + \Delta p - \theta_l(d + \Delta d) = p - \theta_l d + \left(\frac{\Delta p}{\Delta d} - \theta_l\right)\Delta d > 0$ (whatever the sign of $\frac{\Delta p}{\Delta d} - \theta_l$ since Δd is small). We have thus shown that cream-skimming is at work with this alteration of the original pooling contract that simultaneously attract low risks, repel high risks and generate profits.

Separation of Risks

The equilibrium is thus separating i.e., insurers offers menus of contracts and different risks elicit different contracts that identifies them afterwards. Since in our simple setting there are only two classes of risk or types, insurers need only offer two contracts $\gamma_l = (p_l, d_l)$ and $\gamma_h = (p_h, d_h)$ that will be picked up by the safe and unsafe customers respectively (cf. revelation principle seen in §21.2.1). Technically those incentive compatibility (IC) conditions are $U(\theta_l, \gamma_l) \geq U(\theta_l, \gamma_h)$ and $U(\theta_h, \gamma_h) \geq U(\theta_h, \gamma_l)$. Furthermore competition among insurers guarantees zero per-capita profits i.e., $p_l = \theta_l d_l$ and $p_h = \theta_h d_h$.

The unsafe type IC condition $U(\theta_h, \gamma_h) \geq U(\theta_h, \gamma_l)$ reads

$$\begin{array}{rcccc} f(d_h) & \equiv & \theta_h u((1-\theta_h)d_h - c) & + & (1-\theta_h)u(-\theta_h d_h) \\ \geq \; g(d_l) & \equiv & \theta_h u((1-\theta_l)d_l - c) & + & (1-\theta_h)u(-\theta_l d_l) \end{array} \tag{21.27}$$

As u is concave, $f' \propto u'(-p_h - c + d_h) - u'(-p_h) > 0$ thus (21.27) is more easily satisfied by increasing d_h. This means that the Bertrand type competition among insurers[23@] will drive d_h to c i.e., unsafe types receive full insurance and obtain a utility $f(c) = u(-\theta_h c)$. Likewise competition tends to increase d_l but (21.27) is violated at $d_l = c$ since everybody prefers the low premium p_l when both contracts give full insurance. Therefore the safe risk will have to support some risk ($d_l < c$). Since $\theta_l < \theta_h \Rightarrow \theta_h(1-\theta_l) > (1-\theta_h)\theta_l$, we have

$$\begin{array}{rcl} g'(d_l) & = & \theta_h(1-\theta_l)u'((1-\theta_l)d_l - c) - (1-\theta_h)\theta_l u'(-\theta_l d_l) \\ & > & (1-\theta_h)\theta_l\left[u'\left((1-\theta_l)d_l - \bar{d}\right) - u'(-\theta_l d_l)\right] \quad > 0 \end{array}$$

Having proved that g is increasing, there exists $d^* < c$ such that $g(d^*) = u(-\theta_h c)$ so that (21.27) is satisfied for any reimbursement $d_l \leq d^*$. The benchmark reimbursement d^* leaves the unsafe risk indifferent between revelation and imitation. Once again competition among insurers drives d_l to its upper limit d^*, so that $\gamma_l = (\theta_l d^*, d^*)$. Lastly, we must check the incentive condition for the safe type: $U(\theta_l, \gamma_h) \leq U(\theta_l, \gamma_l) \Leftrightarrow$

$$\begin{array}{rcccc} u(-\theta_h c) = g(d^*) & = & \theta_h u((1-\theta_l)d^* - c) & + & (1-\theta_h)u(-\theta_l d^*) \\ & \leq & \theta_l u((1-\theta_l)d^* - c) & + & (1-\theta_l)u(-\theta_l d^*) \end{array}$$

is true since the weight on the larger term $u(-\theta_l d^*)$ is increased. increases.

Equilibrium

What we have characterized is the optimal pair of separating contracts. An immediate observation is that a cream-skimming contract is a non-optimal separating contract thus it is not a candidate equilibrium. To know whether $\gamma_l = (\theta_l d^*, d^*)$ and $\gamma_h = (\theta_h c, c)$ form an equilibrium we must check that there does not exists a pooling contract $\hat{\gamma}$ preferred by both types of risk.

Observe indeed that an insurer could offer $\hat{\gamma} = (\hat{p}; c)$ where the premium is computed to make the safe risk indifferent i.e., $U(\theta_l, \gamma_l) = u(-\hat{p})$. Then the IC condition for the safe type tells us that unsafe risks would have a strict benefit in switching from γ_h to $\hat{\gamma}$. The expected profit is then $\hat{p} - (\lambda\theta_l + (1-\lambda)\theta_h)\, c$ and will be positive if the proportion λ of safe types in the population is larger than $\frac{\theta_h - \hat{p}/c}{\theta_h - \theta_l}$. The reason behind the existence of this profitable deviation is that the proportion of unsafe risk being small, it is not worthwhile for insurance companies to seek separation of types because it forces the large majority of safe risk to support costly risk; hence any insurer can offer a Pareto improving trade (an insurance service) to all consumers even if this means losing money on the few unsafe risks that are around.

So, when λ is large, a pooling contract $\hat{\gamma}$ can successfully attack the separating (γ_l, γ_h) but it is itself attacked by a cream skimming contract which is not a candidate separating equilibrium. This circularity proves that there is no equilibrium. Still, the model of insurance competition should be completed because the reaction of an insurer whose customers are stolen by a competitor matters. When attacked by a cream skimming contract $\bar{\gamma}$, the "old" pooling contract $\hat{\gamma}$ is withdrawn because the zero profit condition $p - (\alpha\theta_l + (1-\alpha)\theta_h)\, d = 0$ yield losses once the safe types are gone. But then the cream skimming contract $\bar{\gamma}$ has to serve all types and becomes itself vulnerable to a cream skimming attack. The same reasoning applies in case of a pooling attack over a separating contract.

To solve this inconsistency, Wilson (1977) proposes to redesign the competition between insurers as a stage game where:

- Insurers simultaneously offer "old" contracts (pooling or separating).
- Insurers simultaneously offer "new" contracts.
- Insurers simultaneously withdraw "old" contracts if they wish to.
- Customers can sign any contract.

In this game, the equilibrium is the separating one characterized earlier for small λ; otherwise it is the pooling contract $\hat{\gamma}$ giving the highest level of utility to safe types.

Indeed if λ is large, a cream skimming attack against $\hat{\gamma}$ needs $\lambda = \theta_l < \frac{\Delta p}{\Delta d} < \beta = \theta_h$ hence generates a profit of $(\lambda\theta_l + (1-\lambda)\theta_h)\,\Delta d - \Delta p < \lambda(\theta_l - \theta_h)\Delta d < 0$.

Chapter 22

Auctions

Auctions are very competitive trading mechanisms, so much so that the textbook examples of "perfectly competitive" markets are often the colorful markets for fish, cattle, wine or flowers which all use an auction to allocate commodities. An auction is an organized contest (cf. §7) that is cheap to set-up and enable the prompt selling of almost any item. It also has the crucial ability to extract the precious information that economic agents might possess regarding the item for sale; this explains the appearance of this chapter in the Part on asymmetric information. Modern references are Klemperer (2003) and Milgrom (2004).

The chapter is organized as follows: we first shed light upon the origin and main uses of auctions with an emphasis on the assignment of natural monopolies. We then compare the main auctions before inquiring into the optimal auction (for the seller) and efficient one (for society).

22.1 Purpose of Auctions

22.1.1 Origins

Auctions are a very useful and old exchange mechanism. There are records of auctions for slaves around 1900 BC in Assyria (Irak), for virgin brides and slaves around 500 BC in Babylon. In BC Rome, auctions were commonly used to sell real estate, slaves or goods as follows: an auctioneer sets a low starting price and waits for participants to signal a higher price; they can call out openly or nod in which case it is the auctioneer who sets the new (higher) standing price. When no one dares to bid above the standing price, the bidder who made it gets the item and pays the seller his bid. As astounding as it may sound, the entire Roman empire was once sold in an auction!

This auction form, which we call Roman to distinguish it from other forms,[1@] has been used in Europe, the Middle East and Asia continuously over a thousand year.[2@]

Nowadays, the real life auction closest to the open ascending model prevalent in the literature is a variant of the *Roman* auction used in Japanese fish markets.

In the XXth century, governments have used auctions to sell treasury bills, foreign exchange, mineral rights, oil fields, assets of firms to be privatized, public land or properties and more recently air waves for television or telephone. In the private markets, houses, cars, agricultural production, livestock, art and antiques are commonly sold by auction. Even more recent are the internet auctions to sell used items and the business to business (B2B) procurement auctions whereby firms compete to sell or buy at bargain price their inputs or outputs.

22.1.2 The case for auctioning

Perfect Competition

An auction is the practical trade mechanism that comes closest to perfect competition. Standardized goods such as financial assets (stocks, options, derivatives), grain or minerals are traded in exchanges (trading posts, clearinghouses) where anonymous buyers meet anonymous sellers in a double auction (cf. §22.2.1).[3@] The remarkable property of such markets is to extract the private information of participants. Since nobody has market power in these large markets, the optimal pricing strategy is to be "price-taker". As a consequence, individual demand equates (market) price and (personal) marginal willingness to pay; symmetrically, individual supply equates price and marginal cost. The bidding behavior of all participants therefore reveal perfectly all their economically relevant information and it is aggregated in the equilibrium price. We obtain the first welfare theorem according to which efficiency is reached in a competitive market.

Revelation Mechanism

It would seem that the previous mechanism fails to work properly to sell a unique item such as a painting, a house, a mineral field or a mobile phone license. Indeed, being so exceptional or unique, the item does not have a well known market value so that both sellers and buyers are unsure of how much they should ask or bid.

Organizing an auction is an answer to this problem. The owner needs to advertise the event to attract the largest possible number of participants because each potential buyer will bring his own small piece of information relative to the item for sale. The seller can then hope that during the auctioning process these information bits will be revealed by the participants through their bidding behavior (like in a competitive market). This way, the seller can sell the item to the person who values it most, thereby maximizing her

revenue. Similarly, whenever an economic agent wants to buy or procure a service or a very specific item like a museum or a railroad line, he can use a procurement auction to attract many potential contractors (sellers of the service) and award the production of the item to the least demanding candidate in order to minimize spending.

Natural Monopoly

Government intervene natural monopoly markets (cf. chap. 17) and often deliver one or a few licenses to operate. Ideally, this should be done with a view to maximize efficiency i.e., look for the firm best able to provide high quality at low cost.[4@] In the past, the allocation process was often a "beauty contest" where contenders would propose a detailed plan of activities and the government would select that best fitting its needs (cf. Table 7.2). Such a scheme presents two obvious problems. Firstly, the promises included in the plan are hard to check and enforce, thus hard to believe in the first place which ultimately means that few elements can really be used to decide between offers. Secondly, this selection process is open to wasteful lobbying and corruption; it is widely believed that such methods have seriously limited entry, challenging and innovation (cf. §10 & §12), an issue better known as regulatory capture.

The Chicago school, may be inspired by ancient Greek politics, proposed an alternative method immune to rent-seeking: random allocation. Once some lucky person gets the license, all interested firms will try to buy it back from him.[5@] The ensuing competition should guarantee efficiency since the license will go to the firm with the highest WTP. The underlying strong assumption here is the absence of transaction costs i.e., the negotiation between parties is neither time nor lawyers consuming.

This option was tried in the US during the 1980s when the FCC organized lotteries to allocate radio and TV waves; it lasted until 1993 and lead to a severe fragmentation and to very costly negotiations between telecommunication firms and all sort of arbitrageurs. Experience (and the comparison with Europe) has revealed the existence of large transaction costs. It seems quite obvious that the lucky arbitrageur who wins a license and the telecommunication firm willing to buy it back have private information; none of them knows clearly how much the license is worth for the other. The arbitrageur knows nothing about telecommunications and the firm ignores whether the arbitrageur already got an offer from a competitor. As we formally show in §22.3.3, this asymmetry of information makes bargaining inefficient in the sense that the license is some times resold to a firm that does not value it most.

This failed experiment reinforce the case for devising a wise allocation mechanism that extracts the relevant information from the participants and identifies directly the highest value for the license. Economists are inclined to believe that auctions have fared

quite well in this respect as we shall argue in the theory section.

Collusion

Bidders to an auction have an incentive to collude (form an illicit cartel) in order to get the item at better conditions than if they were competing one against the other. In the case of auctions among private economic agents the presence or absence of collusion is a matter of surplus distribution among buyers and sellers but when the public power is involved, reducing its surplus is akin to increasing public spending which causes inefficient distortions in the economy. Antitrust authorities therefore actively fight collusion in procurement auctions for public works; the large number of prosecutions and judgments against groups of firms is proof that collusion is widespread in auctions.

McAfee and McMillan (1992) show that whenever the members of a cartel are unable to make monetary transfers (because of anti-trust surveillance) the best they can do to manipulate an auction is to bid the same amount, a feature commonly observed in procurement auctions. If they are able to redistribute the rents of collusion among themselves, then they behave as a monopsonist in the original auction and resell the item among themselves in a private auction.

Robinson (1985) argues convincingly that collusion is easier in a second-price auction (open or sealed) than in a first-price one. Indeed, for an item of known value 50€, it is enough for the designated winner to bid a high value like 100€ and for all other colluders to bid a low 7€ ; this way no one will ever cheat. Indeed, with a bid like 101€, the cheater would end up paying 100€ and lose 50€ while if he bids 30€, he doesn't win the item but hurts the winner who will retaliate in subsequent auctions. Such a deception is much more difficult to play in a first-price auction because the designated winner ought to bid only slightly more than his mates, say 8€; but then anyone can outbid him at 9€ and make a profit of 41€ .

22.2 Comparing Auctions

In our theoretical presentation, the original owner of the item for sale (she) stands also as the auctioneer who organizes and runs the auction. We shall first consider her encounter with a single potential buyer (he), before generalizing to several.

22.2.1 Typology

Auctions can be classified according to their rules. A one-sided auction sees a unique seller or buyer proposing an item to several bidders. Two-sided or double auctions put

buyers and sellers in contact through the auctioneer. Bids can be secret (sealed) or public (open). This last case opens the possibility of competition among bidders by alternating bids. We present the most frequent ones and indicate in parenthesis the familiar name of each.

- *Open ascending* (Roman): in the Japanese electronic version, the price starts low and goes up with the clock time. Bidders push a button to enter the auction and release it to leave (no reentry permitted) so that the winner is the last holding bidder; he pays the standing price.[6@] Examples: art auction houses.
- *Open descending* (Dutch): the price starts from a very high level and falls with the clock time until someone says mine (pushes a button); the winning bidder gets the item and pays the standing price. Examples: Flower sales in the Netherlands (cf. case study by Kambil and van Heck (1996)) or the Google IPO.
- *First-price sealed* (Sealed): Bids submitted secretly in written form to the auctioneer who sorts them and award the item to the highest bidder who pays his bid. Examples: procurement contest or airwaves licenses.
- *Second-price sealed* (Vickrey (1961)): same as before except that the winner pays the second-highest bid. Example: stamps sales by mail in the US during the late XIXth.
- *All-pay*: every bidder pays his bid and the highest bidder receives the object. Examples: any lottery (sweepstakes) or equivalently any match opposing two teams; each participant's bid is the effort he produces to win the match.
- *War of Attrition*: bidders put repetitively an equal amount (of money, time or effort) on the table until all but one drop out. The item goes to the last standing bidder (cf. §7.4).[7@]
- *Double Auction*: participants submit supply and demand bids, then the auctioneer sort demand bids in descending order and supply bids in ascending order. Matching the two curves yields an equilibrium price at which all feasible transactions are executed.

Drawing on the characteristics of the double auction, Walras (1874) offers a theoretical description of a competitive market: the auctioneer (called Walrasian to distinguish from real ones) cries out a price and waits for demands and supplies. He then lowers the price if supply exceeds demand and raises it otherwise. This "tâtonnement" process eventually converges to an equilibrium price. Today most electronic markets use an automated Walrasian auctioneer that offers an equilibrium price continuously equating demand and supply.

The other crucial element in the study of auctions which justifies its presence in this Part is the information setting, more precisely what the seller and the potential

buyers know regarding the item for sale. In the case of *private values*, each bidder has a privately known willingness to pay for the item, called the value and no one else knows it. This applies for a fixed quantity of raw standard material or a license for a mobile telephone network because the willingness to pay of a potential buyer is determined by the technology where the item will be used as an input. In the *common value* situation, the item for sale has a unique market value which is all that matters for buyers but no one knows it exactly e.g., an oil field. Lastly we can *mix* the previous settings, letting the item having an intrinsic objective value but also a subjective one for each bidder e.g., a painting which you might appreciate on top of knowing that it also has a market value for resale.

Before delving into theory to compare the 4 *standard* one-sided auctions, let us recall that the obvious advantage of open cry auctions is the speed at which large quantities can be sold; this is particularly crucial for fresh food or flowers and explain why Roman and Dutch auctions have been used for centuries.

22.2.2 Standard Auctions

Some formal comparisons can be made among the 4 *standard* one-sided auctions. In all the sealed auctions a strategy is simply a price (to be written down the porposal) while in the open auctions, a strategy is a stopping rule, a limit price that can depend on what the player has observed while the standing price was evolving.

First-Price Auctions: Dutch and Sealed

In the *Dutch* auction no information is revealed until some says mine triggering the end of the auction, thus a strategy is a single figure just like in a sealed auction. Given that in the *Dutch* and *Sealed* auctions, the winner pays his bid, these two auctions are strategically equivalent; this is true whatever the information context. Hence we might see the *Dutch* auction as open first price. What distinguishes them in the real life relates mostly to collusive behavior among bidders and transaction costs, the speed at which it is conducted and can be repeated.

Second-Price Auctions: Roman and Vickrey

When values are private, the *Roman* and the *Vickrey* auctions are also equivalent albeit in a weaker sense. In the *Roman* auction, if values are private and statistically independent then a player learns nothing from the fact that some people are dropping out as the price goes up. In fact, it is a dominant strategy (optimal whatever do the other bidders)

for him to stay until the public price reaches his own value because dropping out before is to lose an opportunity to get the item cheap while staying too long is taking the risk of buying dear.

In a *Vickrey* auction, a strategy is a price $\hat{v}$ (equal or different from the true value v). The optimal strategy for a bidder, say #1, is to play his personal value v whatever the behavior of other bidders and whatever the information structure. Indeed, letting u be the highest bid made by someone else, say #2, we see in Table 22.1 that telling the truth is identical or better than understating and also identical or better than overstating. The $\star$ cells indicate when lying reduces the payoff of bidder #1. In both cases, the winner is the bidder who values most the item but he pays only the second highest valuation, thus the *Roman* auction could be relabeled "open second price" since the *Vickrey* is the sealed-bid second price auction.

bidder #1 / bidder #2	$u < \underline{v}$	$\underline{v} < u < v$	$v < u < \bar{v}$	$\bar{v} < u$
understate $\hat{v} = \underline{v} < v$	win, $v - u$	lose, 0 $\star$	lose, 0	lose, 0
truth $\hat{v} = v$	win, $v - u$	win, $v - u$	lose, 0	lose, 0
overstate $\hat{v} = \bar{v} > v$	win, $v - u$	win, $v - u$	win, $v - u$ $\star$	lose, 0

Table 22.1: Gain for bidder #1

Affiliated information

The private information of bidders is rarely independent, rather it is often affiliated in the sense that when one bidder is optimistic i.e., receive information stating that the item for sale is very valuable, it is more likely that other bidders' are also optimistic. In this context and assuming risk-neutral bidders, whose signals are drawn from symmetric distributions, and whose value functions are symmetric functions of the signals, Milgrom and Weber (1982) show that the following ranking in terms of seller's revenue: *Roman* $\gg$ *Vickrey* $\gg$ first-price auctions (*Dutch* or *Sealed*). This theory is backed by the prominent use of the *Roman* auction through history and settings.

The intuition behind this classification is that the surplus of the winning bidder is due to her private information. Hence to maximize revenue, the seller looks for an auction able to extract the winner's information. When information is affiliated, this effect will be stronger the more the price paid depends on others' information. The standard auction where the price most depends on all bidders' information is the *Roman* auction; indeed, the winner has seen everybody else drop out and has been able to infer much of these events so that his own winning bid reveals a maximum amount of private information. In the *Vickrey* auction, a similar but weaker phenomenon takes place because the price depends only on the second-highest bid. Lastly, in the first-price auction (*Dutch* or

Sealed), a player's bid incorporates no information in addition to his own.

By the same token, if the seller has some private information, he should release it to augment the information at the disposal of bidders since this will motivate them to bid higher. The general principle stating that expected revenue is raised by linking the winner's payment to information that is affiliated with the winner's information, is known as the Linkage Principle. In practice, sellers of art or exploration rights pay independent experts to assess and reveal the likely value of the item for sale.

22.3 Optimal Auctions

22.3.1 Revenue Equivalence

Vickrey (1961)'s revenue equivalence theorem states that if a fixed number of identical, risk-neutral bidders, who each want a single unit, have independent information, and bid independently, then all 4 *standard* auctions yield the same expected revenue to the seller.

Revelation Principle

We follow Myerson (1981) to prove this important result. The first step is to demonstrate Myerson (1979)'s revelation principle: When a person dressed in red participates in the auction (or any other selling mechanism with very complex rules) taking a series of actions, the seller can record them on a sheet and title it "red strategy"; she can do likewise with all the participants dressed in blue, green or any other color. The fact that the red dressed man plays the red strategy in equilibrium of the auction game means that, for him, it dominates the green strategy; likewise the green strategy dominates the red one for the green dressed man. Since the preferences of participants are captured by their WTP for the item, any two people with the same WTP will behave identically. Thus, the seller will record (possibly) different strategies only for people with different WTPs. This imply that we can drop the color labeling system and use instead one based on the WTP of participants.

The second step is to use this principle to characterize a participant's payoff. A bidder with WTP v (his private information) will get the item with a probability $\varphi(v)$ that depends on the rules of the auction, the behavior of other participants and obviously his own optimal strategy.[8@] Likewise, he expects to pay some amount[9@] $t(v)$, so that his expected surplus is $u(v) = v\varphi(v) - t(v)$. In equilibrium, it does not pay for bidder v to act as if he was $\hat{v}$ i.e., use the optimal strategy of that person so as to win the item with probability $\varphi(\hat{v})$ and pay an expected $t(\hat{v})$. Analytically, this reads:[10@]

$u(v) \geq v\varphi(\hat{v}) - t(\hat{v}) = u(\hat{v}) + (v - \hat{v})\varphi(\hat{v})$. Since it won't pay either for $\hat{v}$ to act as if he was v, we obtain

$$\varphi(\hat{v}) \geq \frac{u(\hat{v}) - u(v)}{\hat{v} - v} \geq \varphi(v) \tag{22.1}$$

out which we deduce two results. Firstly,

$$\hat{v} \geq v \Rightarrow (\hat{v} - v)\left(\varphi(\hat{v}) - \varphi(v)\right) \geq 0 \Rightarrow \varphi(\hat{v}) \geq \varphi(v) \tag{22.2}$$

By playing optimally in an auction, a bidder with higher WTP guarantees himself a greater probability of winning the item i.e., $\varphi' \geq 0$.

The second result is obtained by letting $\hat{v}$ converge towards v; at the limit $u'(v) = \varphi(v)$ must be true. Integrating this equality, we can write

$$u(v) = u(0) + \int_0^v \varphi(x)\,\mathrm{d}x \tag{22.3}$$

The interpretation goes as follows: given that the bidder follows an optimal (equilibrium) bidding strategy, his final surplus depends only on the probability of winning (and his payoff in the worst case). Notice that because φ is increasing, u is convex meaning that additional WTP becomes more and more valuable to win the item at a good price and earn a maximum surplus.

Whenever two auctions generate for all bidders the same probabilities of winning and the same surplus in the worst case then they generate the same surplus function $u(.)$, the same payment function $t(.)$ and finally the same revenue for the seller; this is the general form of the revenue equivalence theorem. In the 4 *standard* auctions, the item goes always to the bidder who has the highest value thus the probability of winning is the same because it depends only on the statistical distribution of values. Furthermore, a bidder drawing the lowest value $\underline{v}$ cannot win (with positive probability) thus her surplus is always 0 so that the theorem applies yielding our initial claim.

Optimal Bidding

As we already saw, the optimal bidding strategy in the Vickrey auction with independent values is one's own value, a behavior known as truth telling or truthful revelation (of the private information). If values are private and independent then the same is true in the Roman auction.

The revenue equivalence enables to compute the optimal bid in first-price auctions because expected payments are identical to those of second price auctions. Now, the expected payment is the product of the winning probability by the expected price, and we know that the winning probability is the same in all 4 *standard* auctions if values are

private and independent. In that case, the bid $b_i(v)$ in the first-price auction is equal to the expected price $\mathbb{E}[\hat{v}|\hat{v} < v]$ in the second price auction with $\hat{v}$ being the second highest bid. Hence, my optimal bid in a first-price auction is the expected second highest value conditional on being lesser than my own.

Let us compute this for player $\#n$. We first look for the distribution of the random variable u, the highest bid of the other $n-1$ players. Since all values are drawn form the distribution H with density $h = H'$, the probability that player #1 bids in $[\hat{v}; \hat{v} + d\hat{v}]$ is $h(\hat{v})$ while the probability that the remaining $n-2$ players bid less than $\hat{v}$ is $H(\hat{v})^{n-2}$; since identities do not matter the density of $\hat{v}$ is $g(\hat{v}) = (n-1)h(\hat{v})H(\hat{v})^{n-2}$. Using integration by parts below, we obtain the optimal bidding strategy as

$$b(v) = \mathbb{E}[\hat{v}|\hat{v} < v] = \frac{\int_0^v x g(x)\,\mathrm{d}x}{\int_0^v g(x)\,\mathrm{d}x} = v - \int_0^v \left(\frac{H(x)}{H(v)}\right)^{n-1} \mathrm{d}x \tag{22.4}$$

which, as expected, is lesser than the value to reflect the fact that players want to make a profit in these first-price auctions and are aware of the possibility to win the item without bidding their private value.

To give a practical use to formula (22.4), let us assume that each buyer's value is uniformly distributed over $[0; \bar{v}]$ i.e., $H(v) = v/\bar{v}$, then $b(v) = v - \int_0^v \left(\frac{x}{v}\right)^{n-1} \mathrm{d}x = v - \left[\frac{x^n}{nv^{n-1}}\right]_0^v = v - v/n$. Notice that the optimal bid does not depend on the top value but only on the number of contenders. We see here why it is interesting to attract many participants in an auction: it motivates all of them to bid more aggressively which raises the seller's revenue.

A phenomenon well known among practitioners is the *winner's curse* according to which the winner of an auction has paid more than the value of the item. Whenever some people receive optimistic information while others receive pessimistic information, values are not private anymore, they become correlated (common value model). In that case, the winning bid, being made by the player who received the highest signal, is overtly optimistic and frequently overshoots the true value of the item. In theory, player take this phenomenon into account when computing their optimal bid. The key to compute my optimal bid is to forget about the item's expected value (based on my private information) and concentrate on its value when my bid is the highest, hence taking into account that other did not dare bid so high.

22.3.2 Optimal Selling Mechanism

Intuition

Bulow and Roberts (1989) introduces this topic with the follow example: Wallace would

be ready to pay some price between 0 and 10 while Gromit would pay some price between 10 and 30 for a painting. In a Roman auction, Gromit systematically outbids Wallace and pays exactly his value so that on average the seller gains 5. Setting a starting price of 10 would raise that payoff to 10 because Wallace would never bid and Gromit would wisely bid his minimum WTP. An even better idea is to set a starting price of 15. True, not even Gromit would participate when his value is less than 15 (probability $\frac{1}{4}$) but the average revenue would jump to $\frac{3}{4}15 = 11,25$.[11@] The question is then to identify criteria that can help us build rules of an optimal auction.

Observing that the seller's pricing problem is quite similar to that of a standard monopolist, one can assimilate the distribution of bidders' values to a demand curve and compute a marginal revenue curve. An optimal (revenue maximizing) auction is then one where the item goes to the bidder whose marginal revenue is greatest; this includes the seller herself if no marginal revenue surpass her reservation value v_0. This scheme is implemented by a *modified second-price auction* where each participant bids a value (openly or secretly), then the seller then computes the marginal revenues and gives the item to the bidder with the greatest marginal revenue. This auction is incentive compatible in the sense that participants truthfully reveal their value.

The previous analysis of an optimal auction (for the seller) makes the search for an efficient auction (socially optimal) quite simple. We only need to give the item to the bidder with the highest valuation; this includes the seller herself if no bidder value surpass v_0.

To achieve this, we can use the Vickrey auction with reserve price v_0 since we already proved that bidders reveal their true WTP in that auction. By the revenue equivalence theorem, the 4 *standard* auctions with reserve price v_0 are efficient. This result looks as a promising allocation mechanism but a word of caution is necessary when dealing with the auctioning of governmental licenses. As we explain in §12.2.3, an incumbent operator (already owning a license) has more to lose from entry than the challenger can hope to win after entry, thus the incumbent will strategically raise his bid for a second license to outbid the challenger in order to block his entry. This could lead to an inefficient outcome if the challenger had a better technology or simply because a monopoly is maintained, instead of evolving towards a more competitive duopoly market structure.

To compare the previous optimal selling mechanism and the efficient one, we observe that since the marginal revenues of the potential buyers are all increasing functions, the ranking of values yields the same ranking of marginal revenues so that the object goes to the same person whenever it is sold. The only time where efficiency fails is when the monopolist computed reserve price p^M is greater than his true valuation v_0 because in that case she keeps the item even though someone may have a higher valuation.

Discriminating Monopoly Analogy

We now develop the previous intuition to uncover the optimal selling mechanism. If the seller was an omniscient druidess, she would be able to rank the WTP of participants $v_1 \geq v_2 \geq ... \geq v_n$ and ask bidder #1 to pay v_1 in a "take-or-leave-it" manner i.e., act like a perfectly discriminating monopoly (cf. §4.2.1), rip a maximum revenue and perform an efficiency enhancing trade since the item would go to the economic agent best able to use it (unless her own valuation v_0 is greater so that efficiency commands her to keep the object). In real life situations, the WTP of participants is a private information to each of them. Maximizing revenue is therefore akin to maximizing information revelation. We adopt Bulow and Roberts (1989)'s heuristic analogy of auctions with monopoly price discrimination as treated in §3.3.1.

The seller's dilemma is almost identical to that of a monopoly facing a market demand because facing one buyer whose WTP you ignore is formally identical to facing a large population of indistinguishable buyers.[12@] Indeed, a typical demand curve $D(.)$ like that displayed on Figure 2.4 can be seen as a series of $n = D(0)$ people ready to buy a single unit of the good, ranked by their WTP. At the price p, monopoly sales are the fraction $D(p)/D(0)$ of the market size i.e., there is a proportion $H(p) \equiv 1 - D(p)/D(0)$ of buyers whose WTP is lesser than p. An alternative way to read this is the following: if a potential buyer is picked at random from the population and offered the item for the price p, he will accept with probability $D(p)/D(0) = 1 - H(p)$.

Now, an auctioneer does not pick a buyer, rather a buyer presents himself at the auction and although she ignores how much he would be ready to pay for the item, she knows his population of origin. The latter is completely described by the distribution function $H(.)$ which we assume known to the auctioneer, just like we always assume that a monopolist knows the demand function $D(.)$ without knowing what it is made of.

Marginal Revenue †

Consider the encounter between the seller of the item and a potential buyer drawn from a population whose statistical distribution is H. When the seller offers the item for the price p, the probability of a greater buyer WTP is $1 - H(p)$ which is also the probability of a sale; this amount plays the role of quantity q in the standard monopoly analysis as illustrated on the left panel of Figure 22.1. From an ex-ante point of view, the expected revenue is $\hat{R}(p) \equiv p \times \big(1 - H(p)\big)$ which can be written $R(q) = qH^{-1}(1-q)$ using the $p \to q$ change of variable. We can then compute the marginal revenue $R_m(q) = H^{-1}(1-q) - \frac{q}{H'(1-q)}$

and using once again the change of variable q into p, we derive

$$\hat{R}_m(p) \equiv R_m\big(1-H(p)\big) = p - \frac{1-H(p)}{h(p)} < p \tag{22.5}$$

which is plotted[13@] on the right panel of Figure 22.1 for the uniform distribution over $[0;\bar{v}]$. We now use the fact that $R(q) = \int_0^q R_m(x)\,\mathrm{d}x$ to express

$$\hat{R}(p) \equiv p\big(1-H(p)\big) = R\big(1-H(p)\big) = \int_p^{\bar{v}} \hat{R}_m(v)\,\mathrm{d}H(v) \Rightarrow p = \mathbb{E}\big[\hat{R}_m(\tilde{v})|\tilde{v}>p\big] \tag{22.6}$$

meaning that the price is the expected marginal revenue conditional on the bidder's value being larger than this price.

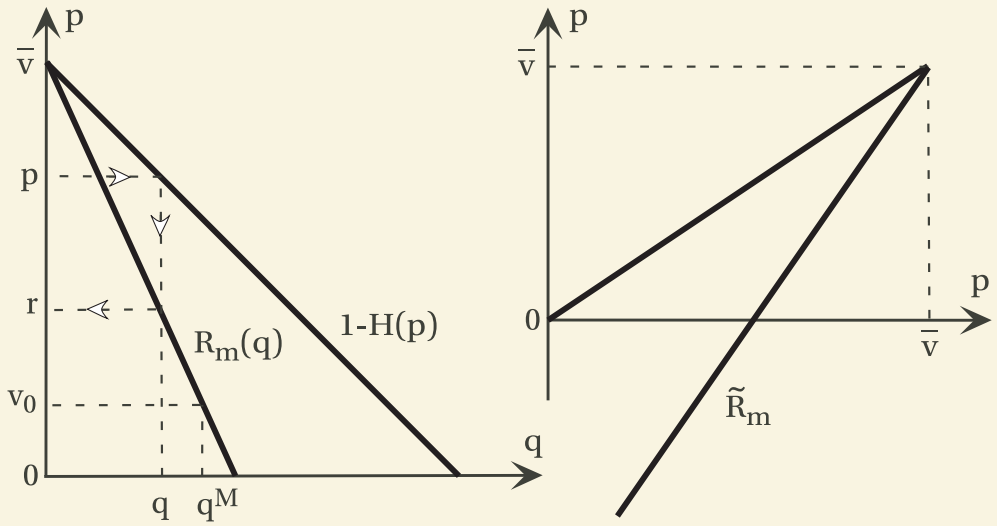


Figure 22.1: Virtual Utility

Optimal Bilateral Sale †

We devise here the optimal mechanism to sell one item to someone whose WTP is drawn form the distribution H. The seller's opportunity cost is v_0. The expected surplus of the potential buyer, $\mathbb{E}[u]$, is computed from (22.3):

$$\begin{aligned}
\mathbb{E}[u] &= \int_0^{\bar{v}} u(v)\,\mathrm{d}H(v) = u(0) + \int_0^{\bar{v}} h(v)\,\mathrm{d}v \int_0^{v} \varphi(x)\,\mathrm{d}x \\
&= u(0) + \int_0^{\bar{v}} \varphi(x)\,\mathrm{d}x \int_0^{x} h(v)\,\mathrm{d}v = u(0) + \int_0^{\bar{v}} \varphi(v)\,(1-H(v))\,\mathrm{d}v \\
&= u(0) + \int_0^{\bar{v}} \varphi(v)\,\tfrac{1-H(v)}{h(v)}\,\mathrm{d}H(v)
\end{aligned} \tag{22.7}$$

by an exchange of integration order.

A bidder with WTP v is willing to take part in the auction if $u(v) \geq 0$ and this constraint will be always satisfied if it is satisfied for the lowest possible WTP i.e., the only participation constraint that matters is $u(0) \geq 0$. Since revenue is $t(v) = v\varphi(v) - u(v)$, we can use (22.7) to check that the seller earns on expectation:

$$\mathbb{E}[t] = \int_0^{\bar{v}} \left(v - \frac{1 - H(v)}{h(v)} \right) \varphi(v) \, \mathrm{d}H(v) - u(0) = \mathbb{E}\left[\varphi(\tilde{v}) \hat{R}_m(\tilde{v}) \right] - u(0) \tag{22.8}$$

Taking into account her opportunity cost v_0 of renouncing to the item, her (producer) surplus[14@] is

$$W_S = \mathbb{E}\left[\varphi(\tilde{v}) \hat{R}_m(\tilde{v}) + (1 - \varphi(\tilde{v})) v_0 \right] - u(0) = v_0 - u(0) + \mathbb{E}\left[\varphi(\tilde{v}) \left(\hat{R}_m(\tilde{v}) - v_0 \right) \right] \tag{22.9}$$

so that its maximization leads her to sell the item i.e., set $\varphi(v) = 1$ only for values v such that $\hat{R}_m(v) \geq v_0$. Letting q^M solve $R_m(q) = v_0$ and p^M solve $1 - H(p) = q^M$, we check on the left panel on Figure 22.1 that this optimal rule generates sales to a proportion q^M of potential buyers which is exactly the standard monopoly quantity given that the opportunity cost v_0 is the marginal cost of "producing" the item. Furthermore, the rule satisfies the condition $\varphi' \geq 0$ since φ is nil over $[0; p^M]$ and unitary over $[p^M; \bar{v}]$.

To implement this outcome it is enough to behave as a standard monopoly i.e., ask a price $p^M \equiv H^{-1}(1 - q^M)$ for the item or to set-up the following auction: the buyer bids a WTP $\hat{v}$ and is allowed to buy the item at the price p^M if the marginal revenue corresponding to his offer, $\hat{R}_m(\hat{v})$, is greater than v_0. It is a simple exercise using the left panel of Figure 22.1 to check that a buyer with WTP v has no interest to lie i.e., he will truthfully reveal $\hat{v} = v$ (the proof for the general case is provided in the next paragraph).

When comparing the two previous mechanisms, the second looks dumb since it adds a bidding stage that seems irrelevant. That is correct in the present setting but once there are several bidders, it becomes a useful device to force a maximum revelation of information.

Optimal Auction †

We derive here the optimal auction for the seller and the efficient one which are close.

Consider n independent but not necessarily identical bidders i.e., bidder #i's value v_i has statistical distribution is H_i with density h_i. We count the seller as a dummy bidder #0 whose value distribution is entirely concentrated at v_0 and set $t_0 = u_0 = 0$.

When participating in the auction, each bidder will bid optimally so that the previous results will apply. We denote $\varphi_i(\mathbf{v})$ the probability that bidder #i wins the object when

the vector of values is $\mathbf{v}$. Thanks to the dummy bidder trick, the seller's revenue is also her producer surplus

$$W_S = \mathbb{E}\left[\sum_{i\geq 0} \varphi_i(\mathbf{v})\hat{R}_{m,i}(v_i)\right] - \sum_{i\geq 0} u_i(0) \tag{22.10}$$

where $\hat{R}_{m,i}$ is computed as in (22.5) but using the distribution function H_i.

An optimal (revenue maximizing) auction is now easy to identify: for every vector $\mathbf{v}$, the item should go to the bidder whose marginal revenue $\hat{R}_{m,i}(v_i)$ is greatest; this includes the seller herself if no marginal revenue surpass v_0.[15@] It remains to build an auction where this outcome is implemented for each possible combination of private values.

Consider the *modified second-price auction* where each participant bids a value v_i (openly or secretly), the seller then computes the marginal revenues $r_i \equiv \hat{R}_{m,i}(v_i)$ for $i \geq 1$ and $r_0 \equiv v_0$. Assume for simplicity that after ranking these marginal revenues we have $r_1 \geq r_2 \geq \ldots \geq r_n$; let then q_2 be defined by $\hat{R}_{m,1}(r_2) = q_2$. The object is awarded to bidder #1 at the price $p_2 = H_1^{-1}(1-q_2)$ as shown on Figure 22.2. Notice that since the price is always positive, a bidder with minimum WTP never wins it, thus derives a nil utility so that $u_i(0) = 0$.

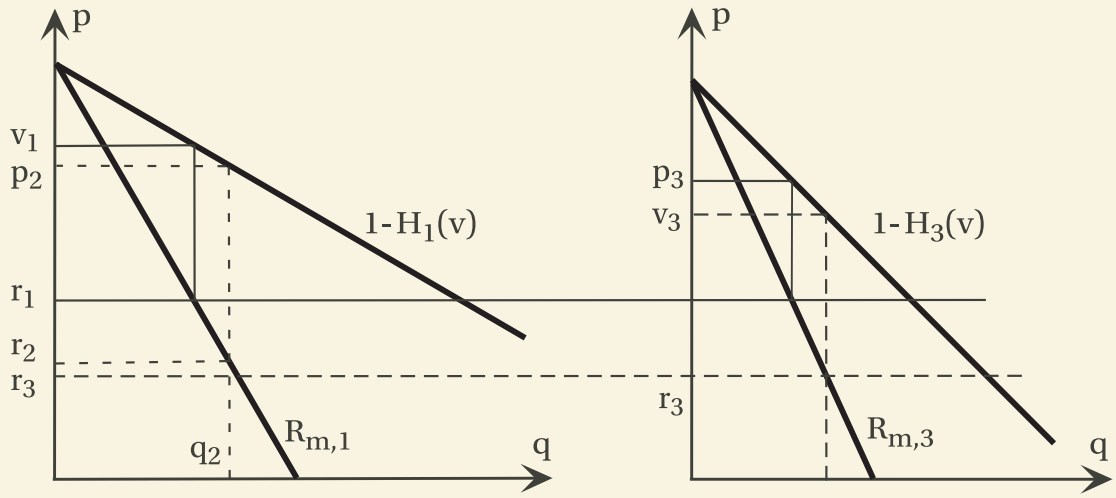


Figure 22.2: Optimal Auction Design

We claim that everybody has an incentive to announce truthfully his value in this auction. There are two cases to consider; either you are a winner, say bidder #1 with the truthful bid v_1 or a strategic bid. In the first case, the price paid is a function of r_2 not r_1. As can be seen on Figure 22.2, overstating v_1 by announcing $\hat{v}_1 > v_1$ has no effect whatsoever. Understating v_1 by announcing $\hat{v}_1 < v_1$ will have no effect either meanwhile $\hat{v}_1 > p_2$ but if the understatement is too strong, the item will go to someone else since the computed marginal revenue $\hat{r}_1$ is now inferior to r_2; this represents a loss since telling

the truth leaves a net surplus $v_1 - p_2 > 0$. Now, for a bidder, say #3, who does not win when truthfully bidding v_3, there is nothing to do. Indeed, the only way to win the object is to outbid the current winner (#1) with some bid $\hat{v}_3$ that has to be greater than p_3 the price computed using H_3 and the winning marginal revenue r_1 which itself is larger than the true valuation v_3 as can be seen on the right panel of Figure 22.2.[16@]

When the auctioneer has absolutely no discriminating information regarding the participants, the distribution functions are identical. This implies that the highest value translate into the highest marginal revenue, thus our optimal auction becomes the Vickrey auction with an optimal reserve price. By the revenue equivalence theorem, the 4 *standard* auctions are optimal. This result is correct for private as well as common value settings.

An efficient auction is one that maximizes welfare $W = \mathbb{E}\left[\sum_{i\geq 0} v_i \varphi_i(\mathbf{v})\right]$. Thus, it is enough to award the object to the highest WTP, including the seller as a dummy bidder with value v_0.

22.3.3 Bilateral Trade under Uncertainty †

Informed Trade

When two people are interested into trading an item, it is enough that the value b for the potential buyer be larger than the value s to the current owner (and potential seller); they might agree on the price $\frac{b+s}{2}$ if they have equal bargaining abilities (cf. §2.4.3).

This simple analysis underlies the case for "free trade" or (economic) "liberalism": whenever something in the economy is owned by someone, say a man, unable to make good use of it then someone else, say a woman, will offer to buy the item. The rationale behind this offer is that she thinks she's able to create more market value from the item, hence she is ready to pledge enough money to buy it because later on she will be able to recoup the investment. The immediate conclusion is that initial ownership does not matter, only *ability* does.[17@]

This story runs into a difficulty as soon as the valuations b and s are not public knowledge. If each agent only knows his or her own valuation then bargaining over the exchange price is more difficult and can sometimes fail to implement the efficient outcome i.e., the item might be exchanged when it should not be and conversely, it might fail to be traded when although it would have been desirable.

Information Revelation

The Vickrey (1961)-Clarke (1971)-Groves (1973) mechanism (VCG) provides a solution to this problem and more general ones; it involves a broker and the two agents. Each agent announces his value, $\hat{b}$ for the buyer and $\hat{s}$ for the seller. Trade occurs if $\hat{b} > \hat{s}$ in which case the buyer has to pay $\hat{s}$ to the broker who pays $\hat{b}$ to the seller. Exactly as in the case shown in Table 22.1, no one has an interest to lie because the payment does not depend on one's own announcement and lying can only bring inefficiency in trade. The item is therefore traded exactly when it is efficient to do so (iff $b > s$); yet the broker exactly loses $b - s$, the gain from trade. Hence, efficiency can be reached but it requires some prior funding by the parties to make the broker willing to play his part.

To see if there is a way around this deficit issue, we assume the individual values b and s are drawn from a continuous distributions F_b and F_s with support $[\underline{v}; \bar{v}]$ for $v = b, s$. Both parties know the statistical information regarding their partner' value but do not observe its actual realization. If $\underline{b} \geq \bar{s}$ then it is always efficient to sell the item and partners know perfectly this fact: although the buyer ignores the exact value v, he knows that $b > v$ for sure and likewise, the seller who ignores the true b knows that $b > v$. Somehow, the asymmetry of information does not matter; a balanced and efficient mechanism is a sales contract stipulating any fixed price $p \in [\bar{s}; \underline{b}]$. Since $b \geq p$ and $p \geq s$ for sure, it is always in the interest of both the seller and the buyer to sign this contract, whatever their own valuation and without worrying for the actual value of the partner. The price will be negotiated ex-ante using the expected values of b and s; for instance, the fair division price is $p = \frac{\mathbb{E}[s+b]}{2}$.

Problematic Trade

The more realistic but more problematic case is that where ownership swapping is not always the efficient decision ($\underline{b} < \bar{s}$). Chatterjee and Samuelson (1982) (CS) consider the following simple exchange mechanism whereby the seller asks a price p_s, the buyer offers a price p_b and exchange takes place for the price $\frac{p_s+p_b}{2}$ if and only if $p_s \leq p_b$.

These authors show (cf. proof) that the optimal strategies are to announce the non-truthful prices $p_b^*(b) = \frac{2}{3}b + \frac{\bar{s}+3\underline{b}}{12}$ and $p_s^*(s) = \frac{2}{3}s + \frac{3\bar{s}+\underline{b}}{12}$. In equilibrium, trade occurs at price $\frac{p_s+p_b}{2} = \frac{2(s+b)+\bar{s}+\underline{b}}{6}$, if and only if $p_b \leq p_s \Leftrightarrow b - s \geq \frac{\bar{s}-\underline{b}}{4}$ which is inefficiently rare as soon as $\underline{b} < \bar{s}$; the fact that trade might be inefficient (socially undesirable) distorts the optimal bidding strategies away from truthful revelation which is the only way to be always efficient in trade.

Let us now analyze the optimal bidding strategies. The seller can ask a very high price and make sure he keeps the item, thus he can guarantee himself his private val-

uation s. One would think that asking more than the valuation ($p_s(s) \geq s$) is an even better strategy, since trade takes place only if p_b is larger so that the actual price $\frac{p_s+p_b}{2}$ is mechanically greater than the valuation. The optimal strategy does not follow this too simple intuition; indeed, $\underline{b} < \bar{s}$, which is true, implies $p_s^*(\bar{s}) < \bar{s}$ i.e., the seller strangely understates his valuation when the latter is very large. The reason is that trade will actually take place only if $b > \bar{s} + \frac{\bar{s}-\underline{b}}{4}$ and in that case the price will be greater than $\frac{11\bar{s}+\underline{b}}{12}$ i.e., for some unfrequent values of b, the seller will make a loss but this will be more than compensated by the increased frequency of profitable trades. The same observations apply in a completely symmetric manner to the buyer. The crux of the problem here is whether trade is efficient or not is to be discovered by the parties, hence

> The optimal strategies must not only concentrate on making a gain from trade but also on making trade happen.

Limiting Efficiency Losses

Generalizing this study, Myerson and Satterthwaite (1983) (MS) show that a bilateral trading mechanism cannot be at the same time efficient in trade, balanced in budget and guarantee the participation of both agents i.e., there is no way to eliminate the deficit issue in the VCG mechanism without introducing some inefficiency in trade (cf. proof). The intuition uses our previous results. Suppose the broker tries to devise a trading mechanism that is always efficient and the least costly to him. The efficiency will be satisfied only if both the seller and the buyer reveal their private information to the broker. The latter then tries to buy cheap on one side and sell dear on the other. Provided that the broker has learned s, the buyer should get the item as soon as his own value is greater. As a consequence, the price paid is exactly s for otherwise some socially beneficial trade opportunities would be lost. In that case, the buyer receives all gains from trade. By an exact symmetry, the seller will also receive all gains from trade from the broker. The latter can do not better than loose the expected gains from trade.

MS further demonstrate that the CS trade mechanism maximizes the gains of trade among budget-balanced mechanism guaranteeing participation of both agents (cf. proof). The previous result was akin to show that a natural monopoly forced to price at marginal cost would incur losses while this new result is akin to show that average cost pricing is the least damaging deviation from efficiency that avoids losses. As noticed by Bulow and Roberts (1989), the latter is a Ramsey-Boiteux problem.

To fix ideas, assume $\underline{b} = \underline{s} = 0$, $\bar{b} = \bar{s} = 1$ and that values are uniformly distributed. In the first-best world (efficient trade with a sponsor broker), the gains of trade are $\int_0^1 \int_0^b (b-s)\,\mathrm{d}s\,\mathrm{d}b = \frac{1}{6}$. In the second best world where trade occurs only if $b - s > 1/4$, the

gains of trade are only $\int_0^1 \int_0^{b-1/4} (b-s)\,\mathrm{d}s\,\mathrm{d}b = \frac{13}{96}$, a loss of $\frac{\frac{1}{6}-\frac{13}{96}}{\frac{1}{6}} \simeq 19\%$.

Chapter 23

Entrepreneurship

According to Schumpeter (1942), capitalism is a dynamic evolutionary process coming from within the economic system. It does not develop by adapting to exogenous changes but by mutating in a discontinuous fashion, succumbing to revolutions which displace old equilibria and structures to create radically new ones. For instance, the factory wiped out the workshop, the car superseded the horse and buggy, and the corporation overthrew the proprietorship. This process of creative destruction is the essence of capitalism.

The key actor in this enduring vision is the *entrepreneur* whose characteristic trait is *innovation*, the ability to see external change as an opportunity, not a threat and therefore to do new things or perform old ones differently.[1@] This novelty often stems from invention or research but not necessarily given that its driving force is economic profit rather than intellectual satisfaction. The entrepreneur needs control over the means of production to "get things done"; in that respect, ownership is helpful but hardly necessary. The innovation can be the introduction of a new item (good or service), a new technology to produce an old item, a new commercial strategy (new market, new source of supply), a new internal organization or a new market structure such as monopolization.

The reason why entrepreneurship is so important for government policy is because technical change and innovation explains much of the steady growth in advanced economies since the industrial revolution.[2@] Most of the academic interest towards entrepreneurship is found in the finance and business literature although many important contributions use the standard toolbox of information economics.

In this chapter, we propose a self encompassing introduction to moral hazard and adverse selection within a single framework, the bilateral agency relationship between an entrepreneur (agent) and an investor (principal). The former, aka the agent, is innovative but penniless. He is thus forced to seek external capital from a wealthy but clueless financier to start up his project. The first section will look at the inefficiencies created by equity finance while the next one concentrates on debt. Our findings apply word for word to just about any firm inso far as managers can be assumed to perfectly represent

the owners. The last section is thus devoted to managerial incentives.

23.1 Agency Cost of Equity Finance

In this section, we show that outside equity can be a source of moral hazard and adverse selection. The first segment sets the entrepreneur's playing field and relates the financial perspective to the economics one. We then enter the heart of the matter to show the sharing of future profits through the emission of equity demotivates the entrepreneur and rationally lead him to under-invest. In the next segment, we explain why going to the equity market always conveys bad news about the quality of the firm. As a consequence, less money is raised and more generally, good firms are underpriced; this is an instance of adverse selection. Lastly, we study how the retention of a large fraction of the equity by a risk averse entrepreneur may act as a signal of quality to counter the previous "bad news" effect.

23.1.1 Background

Finance vs. Economics

The neoclassical theory of the firm used in microeconomics focuses on decisions directly relevant to market competition and tend to treat profit as a function of production and/or price. The emphasis is thus more on revenue than cost; there is also a neglect of the fact that costs are disbursed before revenues accrue, thereby creating a need for liquidity. Corporate finance, as its name indicates, aims at solving this later problem for firms (corporations) while keeping in mind that the ultimate objective remains the maximization of profits. This amounts to invert priorities and focus on expenditure (aka. investments), treating revenue (aka. cash-flow) as a consequence. Corporate finance therefore studies what projects to undertake and how to finance them. In this paragraph we show that those views are dual or the two facets of the same objective, profit maximization.

From a financial point of view, a firm's technology is an ability to implement projects i.e., perform investments into assets and later receive returns. We therefore identify a firm with a list of projects opportunities $(c_j, r_j)_{j\leq n}$ where c_j is the investment cost of project[3@] $\#j$ and r_j its rate of return (i.e., the return is $c_j(1+r_j)$). As illustrated on Figure 23.1, one can rank the potential projects by decreasing rate of return. The total cost and total revenue of the first i best projects are respectively $k_i \equiv \sum_{j\leq i} c_j$ and $R_i \equiv \sum_{j\leq i} c_j(1+r_j)$ which satisfy $1+r_i = \frac{R_i - R_{i-1}}{k_i - k_{i-1}}$.

If we pass to the continuum by considering many small size projects, then R becomes a function of k and this revenue–cost relation $R(k)$ has derivative $1+r(k)$. If the risk-free

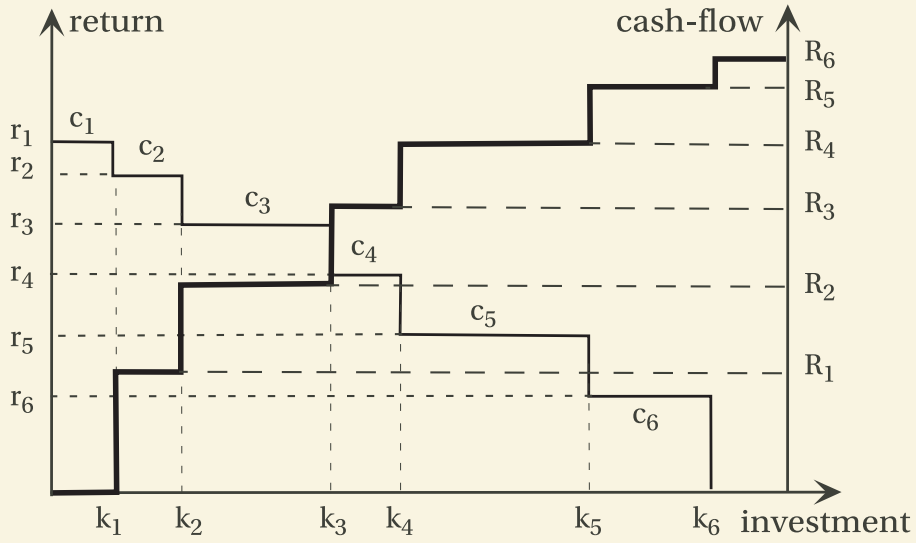


Figure 23.1: Investment Projects and Profitability

interest rate is r_0 then the profit function expressed in net present value[4@] is

$$\pi(k) = \frac{R(k)}{1+r_0} - k \tag{23.1}$$

so that the optimal level of investment k^* maximizing profit solves $r(k) = r_0$. A comparative static exercise, useful later on, is to find out how k^* adjusts to changes in r_0. If the risk-free rate is initially r_4 (cf. Fig. 23.1), then the optimal investment is k_4. If the risk-free rate drops to r_5, the optimal investment increases up to k_5 while if the rate goes up to r_3 then investment recesses down to k_3.

It is now easy to relate the financial vision of profits to the economic one. Denoting p the market price of the commodity produced by the firm, the ratio of the net present value (NPV) of revenue to price $q \equiv \frac{R(k)}{p(1+r_0)}$ can be interpreted as production. Inverting this relation yields the neoclassical cost function $k = C(q)$ so that the profit is now

$$\pi(q) = pq - C(q) \tag{23.2}$$

The optimal production is that solving $p = C_m(q) = \frac{p(1+r_0)}{R_m(k)} = \frac{p(1+r_0)}{1+r(k)}$ which, not surprisingly, leads to solve $r(k) = r_0$. Since $r(k)$ is a decreasing function of k, it is everywhere lesser than the average return, thus at the optimum, the firm's average return is strictly greater than the risk-free one r_0 which means that the firm is making extraordinary profits. Such a situation characterizes a short-term equilibrium since in the long run, free entry enables competitors to claim some of these "high returns" projects and therefore reduces the overall profitability of the firm.

Efficient Finance

We study the contractual relationship of an entrepreneur (she) and an investor (he). The former has a limited personal wealth and owns an unalienable human capital summarized by a cash generating technology as discussed above. By its very nature, the entrepreneur's knowledge cannot be sold so that the investment decision can only be taken by her.

To simplify notations, we denote R the present value of future cash-flow so that the project NPV is simply $\pi = R(k) - k$ (as if the risk-free rate r_0 was nil in (23.1)). As explained above, the technology has decreasing returns to scale ($R_m > 0, \searrow$) so that the efficient investment maximizing the NPV is k^* solving the first order condition $R_m = 1$ i.e., one should equate the productive value of 1€ to its opportunity cost evaluated at market conditions.[5@]

If the entrepreneur was rich enough, she would be able to afford the efficient investment out of her initial wealth and efficiency would be reached (one speaks of a "first-best" situation). Nonetheless, the realistic case is precisely the opposite one and to simplify further we simply assume a zero initial wealth. The two basic ways to raise capital are *debt* where you promise repayment to a lender with an interest or emission of new *equity* where you agree to share future profits with someone else. The corresponding financial instruments are bonds and shares. The fundamental difference between them is the seniority of debt over equity in case of bankruptcy (inability to meet financial obligations).

23.1.2 Incentives to Under-invest

Jensen and Meckling (1976) show that recurring to *outside equity* has an agency cost because it leaves the entrepreneur with partial residual claimancy while the cost of putting time, effort or personal wealth into the firm remains the same. Thus, she is left with partial incentives toward investment. Indeed, if at the margin, 10€ invested by the entrepreneur into the firm yields in return 15€ then it ought to be invested; now if the entrepreneur gets only 50% of it because she sold half of the firm to an outsider, she shall rationally forgo this investment. In a nutshell, "there is no way to make more than one person the residual claimant of an economic activity"; for that reason, incentives toward effort or investment are diminished for most members and under-investment occurs (cf. also Holmstrom (1982a) on moral hazard in teams).

To prove this result, we use Figure 23.2 where the NPV of the project is plotted as a function of total investment k. We denote $k_0 = k^*$ the efficient investment and $\pi_0 \equiv R(k_0) - k_0$ the maximum NPV. To finance her project the entrepreneur sells α% of the project's future cash-flow to an investor who in return pledges an amount F. Even if these

shares have voting power, the technology is uniquely controlled by the entrepreneur so that she remains the only one able to decide on the investment level. Her profit[6@] is $u_\alpha(k) \equiv (1-\alpha)R(k) - k + F$ for any investment satisfying the financing constraint $k \leq F$. Neglecting this constraint for the moment (as if F was large), profit maximization leads her to solve $R_m = \frac{1}{1-\alpha} > 1$. As a result, she under-invests at a level $k_\alpha \leq k_0$ (recall the comparative static exercise of Figure 23.1). As can be grasped on Figure 23.2, the final NPV of the project, $\pi_\alpha \equiv R(k_\alpha) - k_\alpha$ is lesser than the maximum π_0 obtained with the efficient investment and furthermore, the larger the share sold to the outsider, the worse the resulting inefficiency.[7@] The intuition has already been given: since the entrepreneur has to share the profits of her firm with the investor, the marginal benefit from investing her personal wealth is lowered to $(1-\alpha)R_m$ while the opportunity cost of money remains 1, thus the entrepreneur is demotivated and lead to under-invest. This distortion can be called an agency cost because the entrepreneur serves as an agent for the outside investor.

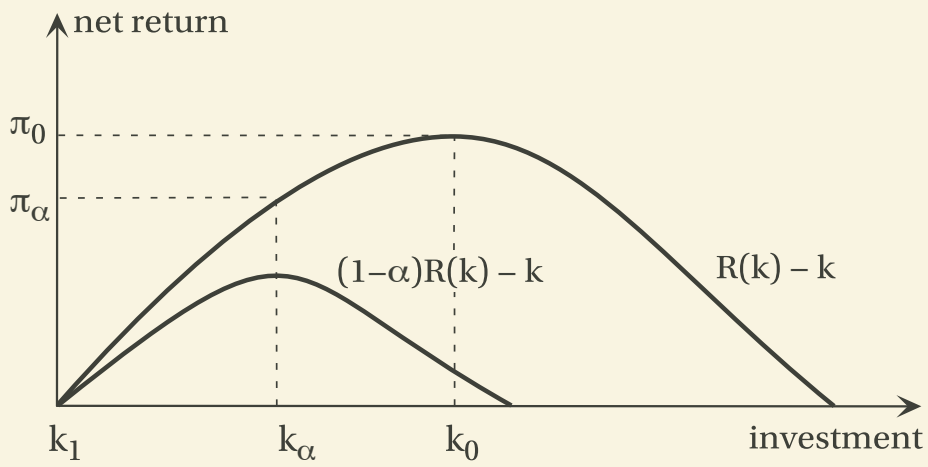


Figure 23.2: Equity Financing

To close the study we only need to take care of the investor; he pledges F only if α is large enough but this leads to severe under-investment and a serious reduction of the cash flow generated by the project, so much that it might not cover the initial funding. Formally, the participation constraint for the investor is $F \leq \alpha R(k_\alpha)$; it involves not just any value of k but k_α because the investor anticipates the future choice that will be undertaken upon his acceptance of α shares. Combining this condition with the financing constraint $k_\alpha \leq F$, we obtain a necessary condition

$$k_\alpha \leq \alpha R(k_\alpha) \quad \Leftrightarrow \quad \alpha \geq \varphi(\alpha) \equiv \frac{k_\alpha}{R(k_\alpha)}$$

i.e., the offered share must be large enough. This constraint is obviously violated at $\alpha = 0$

since the entrepreneur is seeking funds but it is satisfied for $\alpha = 1$ because a minimal investment is still worthwhile.[8@] Observing now that the ratio φ is decreasing[9@] with α, the constraint is satisfied with equality for some $0 < \bar{\alpha} < 1$: this is the minimal share the entrepreneur must relinquish to get the capital necessary to undertake the investment that will then be optimal given the shares she kept. Yet, she won't offer a greater portion because she realizes that the cake to be shared with the investor, the NPV π_α, is decreasing with α and there is no way for her to increase her absolute share of it.[10@]

Notice that in equilibrium, investors just break even, thus it is the entrepreneur who bears all the (agency) cost of his own inefficient behavior due to her inability to commit today to perform in a given way tomorrow. If she could credibly commit to the efficient investment k_0, investors knowing they will get a share of $R(k_0)$ would accept to pay $F = \alpha R(k_0)$ for α% of the business. It would then remain to choose $\alpha = \frac{k_0}{R(k_0)}$ to satisfy the financing constraint. The crux of the problem lies obviously in the credibility of the announcement "I will invest k_0", especially if it incorporates subjective elements such as human capital (the part of her working time really devoted to make the project a success).

23.1.3 Equity Underpricing

Myers and Majluf (1984) deal with the known difficulty of firms to raise capital when they have private information about their current profitability. Formally, this is a problem of adverse selection similar to Akerlof (1970)'s lemons: if a firm issues equity to raise capital for an investment, it accepts to share future cash-flow with the new shareholders, a cash flow that comes from the new investment but also from the current assets. If the latter are very valuable, they will yield a lot of cash, so it not a good idea to share with foreigners; it is better to fund the new project with debt and emit zero additional equity. Potential investors understanding this logic, rightly suspect that an equity issue is a rather bad signal regarding the value of current assets; if the latter are not worth much, investors won't pay much for the firm's new equity either. If, contrary to this belief, the firm has really valuable assets, its equity will be underpriced, a phenomenon frequently observed in financial markets.

To see formally this phenomenon suppose that a firm owns some capital assets that will yield tomorrow a cash-flow $\tilde{x} > 0$ that is random from the point of view of investors; that is to say, the market faces a large population of firms whose future cash-flow $\tilde{x}$ has a statistical distribution H. We study the consequence on the quotation of the current shares of going to the equity market to finance a new project whose lumpy cost is k and whose positive return rate is r.[11@]

The current true value of the firm is $\tilde{x}$, the cash-flow of current assets while its current market valuation is $\mathbb{E}[\tilde{x}]$, the expectation made by investors according to the distribution H. If the firm renounces to the investment when it was publicly known it had the opportunity to undertake it (event no), the value of the firm remains $\tilde{x}$ but the market value of equity becomes $V_{no} = \mathbb{E}[\tilde{x}|\text{no}]$ because investors update their beliefs given the information just revealed by the "not going" decision. Likewise, if the investment is announced together with the emission of new equity (event go), the market value of actual equity changes to $V_{\text{go}} = \mathbb{E}[\tilde{x}|\text{go}] + rk$. The true value of the firm is now $\tilde{x} + (1+r)k$ but the value of old equity is only a fraction $\frac{V_{\text{go}}}{V_{\text{go}}+k}$ of it since it was diluted by the new emission that raised the amount k needed to undertake the new project.

Comparing the two decisions and their payoffs, we see that the original owners decide to launch the new project only if the value they will retain tomorrow is greater than today's i.e., the go event is characterized by

$$\tilde{x} < \frac{V_{\text{go}}}{V_{\text{go}}+k}(\tilde{x} + (1+r)k) \quad \Leftrightarrow \quad \tilde{x} < (1+r)V_{\text{go}} \tag{23.3}$$

i.e., current assets are not very valuable. As intuition told us, it is not a good idea to sell valuable assets (large realized x) and indeed, the firm will rationally pass the opportunity to launch the new project. Since the no event is characterized by $\tilde{x} > (1+r)V_{\text{go}}$, we can check[12@] that

$$V_{\text{no}} = \mathbb{E}\left[\tilde{x}|\tilde{x} > (1+r)V_{\text{go}}\right] > (1+r)V_{\text{go}} > V_{\text{go}} \tag{23.4}$$

i.e., the decision to go is "bad news" and is chastised by the market. This explains why stock price falls when a firm announces a new emission of equity.

Beyond the under-pricing issue which is a consequence of the revelation of private information by the firm to the capital market, we have an efficiency problem: a positive NPV project will not always be implemented because the no event has a positive probability. If the firm could sell its old assets it would credibly reveal to the market the value of $\tilde{x}$; then it would not suffer under-pricing and efficiency would be restored since the new investment would always occur. Indeed, the fair value of the actual shares would be $V_{\text{go}} = \mathbb{E}\left[\tilde{x}|\text{go}\bigcup\tilde{x}\right] + rk = \tilde{x} + rk$ since the market information contains the decision to go and the true value of the firm's current assets. Now, the condition $\tilde{x} < (1+r)V_{\text{go}}$ is always true for $r > 0$ i.e., every new project of positive NPV is undertaken. The value of the firm after the "go" decision would increase from $\tilde{x}$ to $\tilde{x} + rk$ because this time, the market rewards the decision to go on with the new project.

One could nevertheless wonder why the decision to issue equity could not be the "good" signal that the firm has encountered a new project of positive NPV? The reason has to do with the possibility that the rate of return r is itself uncertain. The firm will

never undertake an investment with $r < 0$ because the capital would be better invested in riskless bonds. Yet for $r = 0$, the firm will go on whenever $\tilde{x} < V_{\text{go}}$, so that the equity issue could still be bad news.

The conclusion generally drawn from the previous model is called the "pecking order" theory stating that a firm should finance its projects internally before recurring to debt; equity should be used as a last resort only.

23.1.4 Signaling Quality

We leave aside the value of current assets and inquire the pricing of new assets (new projects). When the market is unable to distinguish the good projects from bad ones, the equity of good firms end up being underpriced which in turn might lead these firms to under-invest. Leland and Pyle (1977), in a model replicating Spence (1973) to the current setting, show that the entrepreneur can alleviate this information problem by retaining a large fraction of the firm's equity to convince the market of its intrinsic quality. This behavior leans on the well known fact that the founder is human, thus she is risk averse and prefers to replace the risky cash flow of her firm by the certain return of an equity sell. Now, if the entrepreneur keeps most of her equity, it must be true that the return is high enough to compensate her for the risk she is bearing. Lenders therefore value more the few shares put for sale.

To demonstrate this claim, we consider a population of entrepreneurs displaying constant absolute risk aversion and facing normally distributed risk. As shown in §19.2.2, each maximizes $\mathbb{E}[\tilde{w}] - \frac{1}{2}\rho\mathbb{V}[\tilde{w}]$ where $\mathbb{E}$ denotes expectation and $\mathbb{V}$ the variance of the random income $\tilde{w}$. Each entrepreneur owns a technology of two possible types, good (g) or bad (b), characterized by a random future cash flow $\tilde{x}_i$ for $i = g, b$; the distributions have distinct expectations $\mu_g = \mu > \mu_b = 0$ (w.l.o.g.) but the same variance σ^2; we can thus w.l.o.g. scale this parameter in order to eliminate the constant multiplier $\frac{1}{2}\rho$ from further formulas. After selling $(1-\alpha)\%$ of its equity for an amount v_i an entrepreneur of type $i = g, b$ will have income $\tilde{w}_i = \alpha\tilde{x}_i + v_i$ with $\mathbb{E}[\tilde{w}_i] = \alpha\mu_i + v_i$ and $\mathbb{V}[\tilde{w}_i] = \alpha^2\sigma^2$, thus her expected utility is $u_i(\alpha) = \alpha\mu_i + v_i - \alpha^2\sigma^2$.

If there were no asymmetry of information, the market would always distinguish the two technologies and being risk neutral, it would pay $v_i = \mu_i$. It is then obvious that each risk averse entrepreneur would be better off selling all her equity to the market in order to maximize u_i; this would result in the optimal risk-sharing allocation. In the more realistic situation where the market cannot distinguish the two technologies, there are two possible equilibria called "pooling" and "separating". If the two types of entrepreneurs behave in a similar fashion, they leave the investors in the fog, the latter

therefore put a single price for any equity and there is pooling of the types. If the two types of entrepreneurs behave in a very distinct manner, they might succeed to convince investors that one behavior is typical of a good firm and the other of the bad firm; there is separation of the types.

In the first case, the single price v is an average between the value of a good and bad firm, thus a good firm is underpriced (and a bad one overpriced). In the second case, the market is able to price each type of equity at its real value; if we denote α_i the share kept by an i-entrepreneur, we have $v_i = \mu_i$ for $i = g, b$. Compared to pooling, separation is desired by a good firm but feared by a bad one so that the former will try to force it and the latter will try to avoid it. Wise people say that "talk is cheap" which means that the only credible behavior available to an entrepreneur to signal the quality of her firm passes through the decision to sell more or less of her equity. The incentive condition to avoid that a bad firm pretends to be a good one is that her utility when lying (i.e., keeping α_g as if she were a good firm) is lesser than when telling the truth (i.e., keeping α_b):

$$u_b(\alpha_g) \leq u_b(\alpha_b) \quad \Leftrightarrow \quad (1-\alpha_g)\mu \leq \sigma^2(\alpha_g^2 - \alpha_b^2) \tag{23.5}$$

If the bad firm cannot micmick the good one then it is better off selling all of her equity to eliminate risk, hence we need only consider (23.5) with $\alpha_b = 0$ so that the condition becomes after some algebraic manipulations

$$\alpha_g \geq \underline{\alpha} \equiv \frac{-\mu + \sqrt{\mu(\mu + 4\sigma^2)}}{2\sigma^2} \tag{23.6}$$

Conversely a good firm should not be tempted to mimic a bad one by selling all of its equity i.e.,

$$u_g(0) \leq u_g(\alpha_g) \quad \Leftrightarrow \quad \alpha_g \leq \bar{\alpha} \equiv \sqrt{\frac{\mu}{\sigma^2}} \tag{23.7}$$

Algebraic manipulations[13@] enable to show that $\underline{\alpha} < \bar{\alpha}$, hence both conditions are compatible for $\alpha \in [\underline{\alpha}; \bar{\alpha}]$. Now given that risk-sharing is the motive for selling equity in the first place, a good firm will sell the maximum and retain no more than $\underline{\alpha}$% of its equity to signal credibly its quality to market investors (this also proves that only bad firms are tempted to mimic good ones).

Given our convention, the parameter μ measures the extent of asymmetry of information and it is not difficult to check that $\underline{\alpha}$ is increasing with μ; hence, the stronger the information asymmetry, the greater the risk the entrepreneur has to bear (by retaining more shares). When applied to the real world, our findings predict that in fast growing industries (e.g., IT services) which display large asymmetries of information, manager/founders of firms retain more equity than managers of firms in mature sectors

(e.g., traditional industry) where there is little asymmetry of information.

23.2 Agency Cost of Debt Finance

We now switch instrument and concentrate on debt. If capital is raised from debt with face value d to be repaid at the interest rate r, the entrepreneur's profit is $R(d) - (1+r)d$ and the FOC of maximization is $R_m = 1 + r$. Assuming a perfect financial market (no transaction cost) the price of money is the same for borrowing and lending i.e., $r = 0$ given our previous convention regarding the PV of R. In that case, the incentives to invest are adequate since the optimal debt choice is $d = k^*$, the efficient investment.

In the absence of uncertainty (debt is riskless), *lump sum* finance such as debt is efficient in the sense that the incentives of the entrepreneur to invest are not distorted (with respect to the case where she does not need external capital). This section will precisely introduce realistic features such as uncertainty or premium for borrowers to test the robustness of this efficiency result. Asset substitution generates excessive investments on the part of the entrepreneur while debt overhang works in the opposite direction ; we also present a model blending the two effects to understand better under which conditions one is more likely to take place. Next, we look at credit rationing, an adverse selection phenomenon induced by debt finance.

23.2.1 Asset Substitution and Free Cash Flow

Intuition

Jensen and Meckling (1976) is again the seminal article that first called the attention on the *asset substitution* induced by debt (aka. the over-investment effect of debt). In the presence of *limited liability* and future uncertainty, the entrepreneur is indifferent with respect to the size of bankruptcy and only cares for the happy times where she can pay her debt and keep all remaining profits. This leads her to undertake too risky and even unprofitable investments.

To get the idea suppose that as a firm's manager you borrow 60 to finance a new project. You can either invest in a very secure idea yielding 100 and make a sure profit of 40, or, take a gamble with a risky project that either yield 180 or 0. Under the risky alternative you either win 120 or 0 since you are protected by limited liability, hence on average you make 60 which turns you into a risk lover. Notice that you selected the inefficient project since its average return is only $90 < 100$. This phenomenon is not limited to finance. In sports like football where matches have fixed duration, it is frequent to see the lagging team take increasingly more risk in a desperate attempt to

catch up, the result being either an unlikely victory or a more presumable overwhelming defeat. The rationale is quite obvious whenever the ultimate goal is to achieve victory:[14@] given the current score and the current field strategy, defeat is almost certain, thus it cannot worsen to change the strategy into a more aggressive one that will increase the probability of winning although it also increase the probability of losing big (but who cares since the team would have been eliminated anyway).[15@]

Model

To see the asset substitution formally, we modify the previously used DRS technology so as to account for market price uncertainty; the profit is $\pi = \tilde{p}R(k) - k$ where the market price $\tilde{p}$ is random with a distribution function H and mean $\mathbb{E}[\tilde{p}] = 1$ (so as to maintain our initial convention). The expected NPV is

$$\pi^*(k) = \int_0^{+\infty} \big(pR(k) - k\big)\, \mathrm{d}H(p) = \mathbb{E}[\tilde{p}]R(k) - k = R(k) - k$$

so that the optimal investment is again the efficient level k^* solving $R_m(k) = 1$.

We consider a risk-neutral entrepreneur who has neither personal wealth nor is able to emit new equity to finance her project; she is thus forced to finance it entirely with debt. Following our convention, the risk-free interest rate is $r_0 = 0$. We shall show that the uncertainty over the market price generates an agency cost in the sense that the entrepreneur's optimal decision is now to over-invest (with respect to the first best decision k^* that a wealthy entrepreneur would pick).

As we explained in the previous examples, whenever the realized cash-flow $\tilde{p}R(k)$ does not cover the debt obligation k, the entrepreneur defaults and walks out with a zero profit but no penalty since she is protected by limited liability. The cut-off level below which defaults takes place is $\hat{p} \equiv \frac{k}{R(k)}$. At the time where she must decide on her investment, she realizes that only the good times matter i.e., her expected profit is

$$\hat{\pi} = \int_{\hat{p}}^{+\infty} \big(pR(k) - k\big)\, \mathrm{d}H(p) = \big(R(k)\mathbb{E}\left[\tilde{p} | \tilde{p} \geq \hat{p}\right] - k\big)\big(1 - H(\hat{p})\big) \tag{23.8}$$

We see that, with respect to a self-financed entrepreneur, she earns more because although $\hat{\pi}$ and π^* are the integral of the same surplus, the former goes over positive surpluses only. To ascertain the incentives towards investment, we look at the FOC of profit maximization:

$$0 \quad = \quad \frac{\partial \hat{\pi}}{\partial k} = \int_{\hat{p}}^{+\infty} \big(pR_m(k) - 1\big)\, \mathrm{d}H(p) + \underbrace{\big(\hat{p}R(k) - k\big)}_{=0} h(\hat{p})$$

$$= \left(R_m(k)\mathbb{E}\left[p|p \geq \hat{p}\right] - 1\right)\left(1 - H(\hat{p})\right) \tag{23.9}$$

$$\Leftrightarrow R_m(k) = \frac{1}{\mathbb{E}[\tilde{p}|\tilde{p} \leq \hat{p}]} \geq \frac{1}{\mathbb{E}[\tilde{p}]} = 1 \tag{23.10}$$

i.e., the debt-financed entrepreneur *over-invests* (recall the comparative static exercise of Figure 23.1).[16@]

This behavior generates default with a positive probability so that lenders do not recoup their loan on expectation; they will have to ask for a risk premium ($r > 0$) whose effect is to induce under-investment (cf. §23.1.1 and next paragraph). Nevertheless, this (second) indirect effect remains dominated by the (first) direct one as long as the interest rate is determined by the participation constraint of lenders requiring they do not make a loss on expectation (cf. proof). Over-investment in this simple model has a multiplier effect on the rate of return of the project, it thus acts as a gamble (risk-taking) with respect to the efficient level.

Free Cash Flow

Jensen (1986) identifies another source of moral hazard, arguing the financial resources at the disposal of managers, the so-called free cash flow leads them towards over-investment Indeed, if managers have the power to decide on the use of cash flow, they are likely to waste it in projects with low return (negative NPV), into acquisitions to derive more power or worse, into perks (e.g., build a pool at home with the firm's money). Hence everything should be done to limit the quasi-rents at the disposal of managers in order to reduce their uncontrolled waste and force them to incur the monitoring of the capital markets when they must raise new capital to finance investments. This problem is more important in sectors that generate large cash flows but have low growth prospects like regulated monopolies. According to Jensen (1986), it would be therefore desirable to force managers to pay large dividends to shareholders whenever they have free cash flow; but since promises are not credible (recall that "talk is cheap") one way of achieving this objective is to emit debt in exchange of a repurchasing of shares. This way, a former shareholder becomes a debt-holder and has the right to take the firm into bankruptcy court if it defaults on debt service. This swap of securities makes debt a credible substitute for dividends and reduces the agency costs of free cash flow (inefficient over-investment). Empirical evidence support this view since share prices tend to respond positively to debt-equity swaps indicating that investors interpret these decisions as value enhancing.

23.2.2 Debt Overhang

Option to Invest

A radically different point of view with respect to the role of debt in agency situations is taken by Myers (1977) who claims that an entrepreneur faces a *debt overhang* problem because at any point in time, investments are optional choices to be undertaken. So, whenever the net return of a project is lesser than the outstanding debt service, it is better for the entrepreneur to drop it and go bankrupt since she is protected by limited liability; somehow, she free rides on the investor. The overall effect is under-investment since some valuable projects are not implemented.

To explain formally this phenomenon, we add a fixed cost $\tilde{c}$ whose level is not yet known by the entrepreneur at the time where she decides on the investment; its distribution function is H.[17@] If the entrepreneur gets financed exclusively by debt at the interest rate r (recall that the risk-free rate is zero under our convention) her random profit is $\tilde{\pi} = R(k) - \tilde{c} - (1+r)k$ so that she is forced into bankruptcy whenever $\tilde{c} > \hat{c} \equiv R(k) - (1+r)k$ i.e., whenever the fixed cost realization is large. The cut-off $\hat{c}$ must be positive for otherwise the technology would be useless (at the current interest rate), but this also means that default occurs with positive probability. Hence, the expected repayment to lenders is strictly lesser than $(1+r)k$, so that r must be positive to compensate their initial investment of k. The entrepreneur's expected profit being

$$\mathbb{E}[\tilde{\pi}] = \int_0^{\hat{c}} (R(k) - (1+r)k - c)\, \mathrm{d}H(c) \tag{23.11}$$

we have

$$\frac{\partial \mathbb{E}[\tilde{\pi}]}{\partial k} = \frac{R_m(k) - (1+r)}{H(\hat{c})} \tag{23.12}$$

thus the optimal investment solves $R_m = 1 + r$ and since R_m is decreasing and $r > 0$, it involves under-investment.

To understand the difference with the previous over-investment result, notice that unlike a multiplicative shock like price, an additive shock like fixed cost has no effect on the marginal productivity, thus does not distort incentives. However, both generate default which is risk from the point of view of lenders; this means that they demand a positive risk premium whose (indirect) effect is to dampen incentives to invest.

Empirically, firms in mature sectors (e.g., heavy industries) have free cash flow and little investment opportunities so that they tend to invest in negative NPV projects. The reverse holds for firms in high growth sectors (e.g., IT services) who lack funds but not ideas and cannot implement as many as efficiency would command. To conclude, over-investment should be more severe in mature industries while under-investment should

be more severe in high growth industries.

A rejoinder

Loosely speaking our previous models conclude for over-investment occurs if the risk is multiplicative (e.g., market price uncertainty) and under-investment if the risk is additive (e.g., fixed cost uncertainty). The following adaptation[18@] draws a clearer frontier among the two effects by blending together the original arguments of Myers (1977) and Jensen and Meckling (1976): over-investment is likely to happen if the firm's realized cash-flow is greater than expected (free cash flow) while under-investment takes place in the reverse situation (debt overhang).

Assume that the entrepreneur uses long-term debt to finance a project developing in two stages, with a promise to repay d at the end. Ex-ante, she raises funds to run her normal business and to perform R&D with a view to develop a better technology. At the interim period, the current assets yield a cash-flow x, a part of which k can be used to invest into the new technology R which is now available; the efficient investment k^* solving $R_m(k) = 1$ can be undertaken at the interim period.

What creates a moral hazard problem at the interim stage (when k is chosen) is the market price uncertainty that plagues every period: from the ex-ante point of view, the interim cash flow x is random, hence there is no way to tune the ex-ante investment so as to obtain exactly $x = k^*$ which is necessary in order to re-invest optimally into the new technology. Now, at the interim period, the entrepreneur disposes of a realized cash flow x and must decide how much to invest in her new technology knowing that the ex-post market price $\tilde{p}$ will also be random (with unit mean).

If the realized interim cash flow is $x < k^*$, then the entrepreneur is forced to put up money to invest at the desired level.[19@] Since she faces the debt overhang problem, she won't put the missing $k^* - x$ but less, so there is under-investment. If, on the contrary, the interim cash flow is a generous $x > k^*$, then the entrepreneur has *free cash-flow* and takes advantage of the existence of debt to gamble over the uncertainty of the final return; she undertakes excessive investment.

To derive precisely these results let us study the optimal investment of the entrepreneur in two polar cases. By investing $k \leq x$ in the *high cash-flow* scenario, her final wealth is $v_h = \max\{0, \tilde{p}R(k) - d + x - k\}$. By investing $k > x$ (putting up $k - x$ out of her pocket) in the *low cash-flow* scenario, her final wealth is $v_l = \max\{0, \tilde{p}R(k) - d\} + x - k$.

Let p_h and p_l be the solutions of $\tilde{p}R(k) + x - k = d$ and $\tilde{p}R(k) = d$. The respective interim

expected wealth are thus

$$\begin{aligned}\mathbb{E}[v_h] &= \int_{p_h}^{\infty} \big(pR(k) - d + x - k\big)\,\mathrm{d}H(p) \quad \text{if } k \le x \\ \mathbb{E}[v_l] &= x - k + \int_{p_l}^{\infty} \big(pR(k) - d\big)\,\mathrm{d}H(p) \quad \text{if } k > x\end{aligned}$$

The optimal investments in each case are k_h and k_l solving

$$\begin{aligned}R_m(k_h) &= \frac{1-H(p_h)}{\int_{p_h}^{\infty} p\,\mathrm{d}H(p)} = \frac{1}{\mathbb{E}[\tilde{p}|\tilde{p}\ge p_h]} \quad \text{if } k \le x \\ R_m(k_l) &= \frac{1}{\int_{p_l}^{\infty} p\,\mathrm{d}H(p)} \quad \text{if } k > x\end{aligned}$$

Basic probability tell us that

$$\int_{p_l}^{\infty} pdH(p) < \int_{0}^{\infty} pdH(p) = 1 = \mathbb{E}[\tilde{p}] < \mathbb{E}[\tilde{p}|\tilde{p} \ge p_h]$$

so that since R_m is decreasing, we have the ranking $k_l < k^* < k_h$.

A bit of logic is required to conclude. If the cash flow realization is large ($x > k_h$) then the optimum investment is $k_h > k^*$, there is thus over-investment (yet never by more than $k_h - k^*$). Conversely, if the cash flow is low ($x < k_l$) then the optimum investment is $k_l < k^*$, there is under-investment (yet never by more than $k^* - k_l$). Finally, when the cash flow is intermediate, it is entirely invested.[20@] There is thus either a moderate under or over investment depending on where x falls with respect to k^*.

23.2.3 Debt as the Optimal Security

Intuition

We tackle here the problem of optimal security design i.e., finding the most efficient contract an entrepreneur and an investor might sign. Recall for instance that equity leads to under-investment by the entrepreneur either in physical or human capital because she is forced to share part of the future profits with the investor. Under the reasonable assumption of limited liability for the entrepreneur, Innes (1990) shows that debt is optimal to promote effort in situations of potential moral hazard.

The result builds on two simple observations. When the cash-flow is low, the investor gets to keep all of it so that the entrepreneur's motivation towards effort is minimal. On the contrary, when the cash-flow is large enough to cover the debt obligation, the entrepreneur is the *residual claimant* of any increase in cash flow, thus is optimally motivated towards effort. Obviously, efficiency corresponds to zero debt so as to make sure the entrepreneur is always the residual claimant. The need for external finance will nec-

essarily introduce a distortion in the sense that for some values of the cash-flow realization, the entrepreneur will receive only a fraction of this cash-flow; as a consequence, she will have incentives to under-invest, thereby generating an inefficient outcome, deemed a “second best”.

Our original question boils down to decide where to put these distortions. It turns out that debt is optimal because it concentrates the distortion on low levels of cash-flow which represent small prizes, thus small disincentives while it makes the entrepreneur residual claimant for large cash-flow which represent large prizes giving large incentives toward effort. On Figure 23.3, we show four reimbursement rules: γ_d is the debt rule corresponding to loan d (bold curve), γ_α is the equity rule corresponding to the sale of a share α of future profits, γ is the strange rule where there is no repayment until cash-flow reaches a minimum, then the repayment increases faster than cash-flow until it achieves a maximum; lastly $\hat{\gamma}$ is a weird repayment rule that does not make any economic sense but which is nevertheless imaginable. Clearly, the debt rule is the closest to the diagonal for low cash-flows and then the farthest away for large cash-flows; this means that when compared to another rule, the debt rule is firstly above then permanently below.

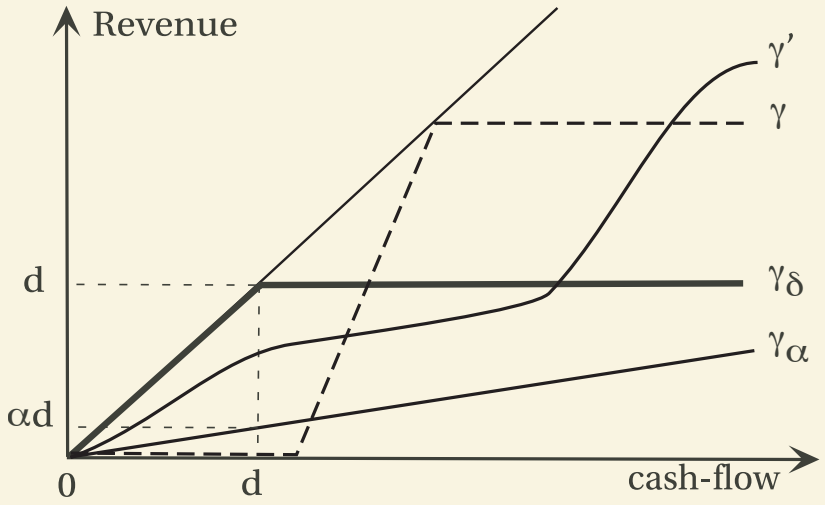


Figure 23.3: Repayment Rules

Model

To prove formally this result, consider a penniless entrepreneur investing k into a project. The future cash-flow is a random variable $\tilde{x}$ such that $H(k,x) = Pr(\tilde{x} \leq x \,|\, k)$, the probability to observe a result lesser than x depends on the investment k. For any function f,we denote $\mathbb{E}\left[f(\tilde{x}) \,|\, k\right] = \int f(x)\,\mathrm{d}H(k,x)$.

An investor pledges k in exchange for a repayment rule γ function of the future cash-flow. Popular rules shown on Figure 23.3 are debt with $\gamma_d(x) = \min\{x, d\}$ or equity with $\gamma_\alpha(x) = \alpha x$. The restrictions for an admissible rule are $\gamma(x) \leq x$ because the entrepreneur is protected by limited liability (she has no collateral to pledge to her creditor) and $\gamma' \geq 0$. To understand this later property, imagine that γ is decreasing at some level of cash-flow x; the creditor could then artificially reduce the available cash flow down to x' by calling a costly external audit and get a contractual repayment $\gamma(x') > \gamma(x)$. The actual repayment rule would, at most, be flat but never decreasing. The investor's expected repayment is $\mathbb{E}\left[\gamma(\tilde{x}) \,|\, k\right]$ so that the entrepreneur's expected profit is $\pi(\gamma, k) = \mathbb{E}\left[\tilde{x} - \gamma(\tilde{x}) \,|\, k\right] - k$.

Assume that the agreed rule γ is optimal and is not a debt rule. The entrepreneur elicits the investment k maximizing $\pi(\gamma, k)$. i.e., solving $R_m(k) = 1 + \frac{\partial \mathbb{E}[\gamma(\tilde{x}) \,|\, k]}{\partial k}$. Since $\mathbb{E}\left[\gamma_d \,|\, k\right]$ is increasing with leverage d, there exists a unique debt level d such that $\mathbb{E}\left[\gamma \,|\, k\right] = \mathbb{E}\left[\gamma_d \,|\, k\right]$. By construction of a debt rule, the difference $f(x) \equiv \gamma(x) - \gamma_d(x)$ is negative for small x, zero for some y value and then becomes positive for $x > y$ (check on the right panel on Figure 23.3); it displays the single crossing property (SCP).

We now assume that a higher effort induces a change in cash-flow distribution satisfying (MLRP)[21@] so that we can conclude (after Milgrom (1981)) that the expectation of the random variable $f(\tilde{x})$ is increasing with the investment k. In other words, its derivative with respect to k is positive which given the definition of f reads

$$\frac{\partial \mathbb{E}\left[\gamma \,|\, k\right]}{\partial k} > \frac{\partial \mathbb{E}\left[\gamma_d \,|\, k\right]}{\partial k} \tag{23.13}$$

Property (23.13) of the cash-flow distribution means that a small additional investment increases the expected debt repayment less than under any other rule. Thus, this small additional investment increases the entrepreneur's expected profit faster under debt finance than under any other rule i.e.,

$$\frac{\partial \pi\left(\gamma_d, k\right)}{\partial k} > \frac{\partial \pi\left(\gamma, k\right)}{\partial k} \tag{23.14}$$

By optimality of k under the rule γ, the RHS of (23.14) is nil, meaning that under the new debt rule γ_d, the entrepreneur can increase her expected profit by choosing an investment $\hat{k} > k$ which is welcomed by the investor since $\mathbb{E}\left[\gamma_d \,|\, \hat{k}\right] > \mathbb{E}\left[\gamma_d \,|\, k\right] = \mathbb{E}\left[\gamma \,|\, k\right]$.[22@] This fact ends the proof that the original rule γ was not optimal.

Risk Aversion

If we wish to take into account that the entrepreneur is most often risk-averse (relative to the investor), Matthews (2001) shows that debt optimality continues to hold if the

original contract can be renegotiated by the entrepreneur after she invested but before the cash flow is realized. Consider a non debt candidate optimum γ that is renegotiated towards a contract ϕ_γ. The objective of the entrepreneur at that point is to eliminate risk which would require $\phi(x) = x - cte$ but this would violate the investor limited liability, thus the entrepreneur would choose something akin to the inverse of debt with $\phi(x) = \max\{0, x - cte\}$ tuning the constant so as to generate acceptation by the investor. The problem then is that this new final contract generates poor incentives to invest (it's the opposite of debt!) and therefore cannot raise a lot of money from the investor. A mid-point will have to be struck to preserve investment incentives with respect to risk sharing.

Since the entrepreneur final payoff is $\pi(\phi_\gamma, k) = \mathbb{E}\left[u(\tilde{x} - \phi_\gamma(\tilde{x}) - k) \,|\, k\right]$ for some concave function u, the investor expects the investment k^* to maximize $\pi(\phi_\gamma, k)$; he will thus accepts ϕ_γ only if $\mathbb{E}\left[\phi_\gamma \,|\, k^*\right] \geq \mathbb{E}\left[\gamma \,|\, k^*\right]$. If now the initial contract is changed for a debt one γ_d such that $\mathbb{E}\left[\gamma_d \,|\, k^*\right] = \mathbb{E}\left[\gamma \,|\, k^*\right]$ then the entrepreneur can still invest k^* and offer ϕ_γ in renegotiation which proves she can't lose from the change towards debt. We still need to show that the investor won't fear a change of investment that is bad for him, thus making the initial offer uninteresting.

Let ϕ_d and k_d be optimal after γ_d i.e., k_d maximize $\pi(\phi_d, k)$ and $\mathbb{E}\left[\phi_d \,|\, k_d\right] \geq \mathbb{E}\left[\gamma_d \,|\, k_d\right]$. If, on the one hand, $\pi(\phi_d, k_d) = \pi(\phi_\gamma, k^*)$ then (ϕ_γ, k^*) is optimal after γ_d i.e., the change to debt does not change the final contract nor investment, thus the investor does not lose. If, on the other hand, the change to debt is profitable for the entrepreneur then she invests more[23@] which is unilaterally good for the investor.

23.2.4 Credit rationing

Credit rationing as well as unemployment are important preoccupations for macroeconomists but also for micro-economists since these long lasting disequilibrium phenomena fail to be explain by the classical general equilibrium theory. Inspired by Akerlof (1970)'s lemons model of the used car market (cf. §21.1.1), Stiglitz and Weiss (1981) show how asymmetries of information among borrowers and lenders can create a market imperfection that limits the volume of credit and endogenously generate credit rationing.

In a credit market, like any other market, demand is decreasing in the nominal interest rate while supply is increasing with the effective interest rate (return on unit loan). In a perfectly competitive market where asymmetries of information are absent, the two kinds of rate are identical. The novelty is to demonstrate how uncertainty about the quality of lenders can make a difference between the *nominal* and *realized* interest rates. More precisely, it might be the case that the *realized* interest rate reaches a maximum such that the corresponding (larger) *nominal* rate generates excess demand for

credit. There is credit rationing because no lender agrees to loan more as he knows this would only lower his effective return.

The story behind this phenomenon builds on the well known positive correlation between leverage and default. Understanding this relationship, lenders tend to believe that firms asking much credit are signaling a high probability of default. This in turn leads them to ask for a large risk premium. The problem with this attitude is that it may affect the average quality of applicants (adverse selection) and their behavior once financed (moral hazard). As we shall demonstrate afterwards, raising the nominal rate does not reduce the demand for credit in an even manner because the safest entrepreneurs drop out so that the pool of remaining applicants are of lower intrinsic quality and worse still, they take more risk than ever. These negative effects diminish the effective interest rate, so much that it might be the case that they outweigh the original nominal increase.

The adverse selection effect is quite similar to the asset substitution effect: riskier projects are more profitable on expectation whenever the entrepreneur uses debt and is protected by limited liability. To see this formally, consider two projects looking for the same funding k with random cash-flows $\tilde{x}$ and $\tilde{y}$, that have the same expected value i.e., $\mathbb{E}\left[\tilde{y}\right] = \mathbb{E}[\tilde{x}]$. Since the repayment of debt $R(\tilde{x}) = \min\{(1+r)k, \tilde{x}\}$ is linear then constant, it is concave, thus the entrepreneur's profit, $\pi(\tilde{x}) = \tilde{x} - R(\tilde{x})$, is convex. Rothschild and Stiglitz (1970) introduce a statistical notion of riskiness such that if $\tilde{y}$ is more risky than $\tilde{x}$, then $\mathbb{E}\left[\pi(\tilde{y})\right] \geq \mathbb{E}[\pi(\tilde{x})]$.[24@] When a lender increases the nominal rate r, firm profits are lowered, hence also fall on expectation so that the less risky project $\tilde{x}$ drops out i.e., cease to demand credit. This means that the average riskiness of applicants increases (only $\tilde{y}$ remains) and therefore the average expected repayment drops (at constant interest rate).[25@]

If there are only two classes of risk, safe and risky people, there is a cut-off nominal rate, say $\bar{r}$ where safe people drop out; at that point the average repayment falls much more than the gain generated by the nominal rate increase. In other words, the *realized* interest rate reaches a maximum at $\bar{r}$. The optimum for lenders is therefore to offer the highest nominal rate that guarantee participation from all entrepreneurs i.e., $\bar{r}$. This also means that the supply of funds has reached a maximum because supply is an increasing function of the effective rate. Yet this optimal rate proposed by lenders generates a demand for funds larger than the supply so that credit rationing occurs.

Regarding moral hazard, notice that after a nominal rate increase, the structure of profits becomes more convex,[26@] thus motivate entrepreneurs to gamble and choose more risky projects thereby worsening the expected repayment to lenders and the *realized* return on their loans. Indeed, we have $\mathbb{E}[\pi(\tilde{x})] = \int_{(1+r)d}^{\infty}(x-(1+r)d)\,\mathrm{d}H_1(x)$ where H_1 is the

distribution function of the random variable $\tilde{x}$. By the definition of the cut-off where the entrepreneur defaults,

$$\frac{\partial \mathbb{E}[\pi(\tilde{x})]}{\partial r} = -\left(1 - H_1((1+r)d)\right) d < 0$$

hence

$$\frac{\partial \mathbb{E}[\pi(\tilde{x})]}{\partial r} < \frac{\partial \mathbb{E}[\pi(\tilde{y})]}{\partial r} \Leftrightarrow H_1((1+r)d) < H_2((1+r)d)$$

where H_2 is the distribution function of the random variable $\tilde{y}$. If, at the initial interest rate, the firm was indifferent between two projects $\tilde{x}$ and $\tilde{y}$ ($\pi(\tilde{x}) = \pi(\tilde{y})$) then after the increase of r, the project with a higher probability of default $\tilde{y}$ is preferred to the safer one (profit was reduced for both but less for $\tilde{y}$).

This Stiglitz and Weiss (1981) result is however highly sensitive to the nature of the uncertainty. If the screening process of lenders identifies returns but leaves doubts regarding the probability of success, de Meza and Webb (1987), show that too many projects are undertaken.[27@] The reason is quite simple: when the return in case of success R is known the entrepreneur's profit is increasing with p, thus the projects that ask for funding are those above the participation threshold, contrary to what happens in the Stiglitz and Weiss (1981) case. The reverse equilibrium sub-optimality then follows.

23.3 Managerial Incentives

Success in business is rarely immediate which means that an entrepreneur needs time to build experience and try different options before she can hope to hit the jackpot, so to say. We saw previously that equity finance had a demotivating effect but debt financing is not perfect either. Indeed, since cash-flow is likely to be low during the first years there is a serious possibility of defaulting on the debt obligations. Now, going bankrupt is a very dark prospect for the entrepreneur. Not only does she loses the prestige of her position but above all she loses all the human effort she invested in the firm; in other words, the human capital she amassed is complementary to the physical assets of the firm.

A similar albeit weaker argument holds for the manager of a dispersedly owned firm: failing to maximize profits in the presence of debt increases the probability of bankruptcy, thus the probability of losing one's job and the perquisites associated with it.[28@]

The first two works presented below use this observation to show that debt can alleviate agency problems of adverse selection and moral hazard. In the third part we present the efficiency wage theory explaining how asymmetries of information force firm to pay higher than necessary wages (and also generate unemployment).

23.3.1 Debt as a Signal of Profitability

Ross (1977) in an early application of Spence (1973)'s signaling theory, shows how the manager of a firm can use debt to signal the profitability of her firm (cf. §21.1.3).

In the absence of uncertainty and under complete information regarding profitability, the value of a firm is correctly assessed by the market and the financial structure (debt or equity) does not matter.[29@] If now the profitability or future cash-flow is a private information of managers, the market will price each firm at an average which means that a profitable firm will be underpriced. Most attempts to signal a high profitability to the market will be mimicked by lower quality firms. In that situation, debt can help a manager. To see this, imagine there are only two firms, one "good" and one "bad" with certain future cash flows x and $y < x$. The "good" manager needs to emit debt $d \in]y;x]$ to avoid bankruptcy but make sure that an imitator would surely go bust. Then it must be the case that bankruptcy is a concern for her to tell the market that her choices reveal her desire to avoid bankruptcy. As we previously argued, this is naturally verified for an entrepreneur. As for a manager, she must tie her own remuneration in a significant way to the final cash flow of her firm to convince the market that she cares to avoid bankruptcy. Her remuneration contract might incorporate a share of the bankruptcy costs.

The formal model uses a continuum of types τ uniformly distributed over $[a;b]$. The future cash-flow of a type τ firm is a random variable $\tilde{x}_\tau$ uniformly distributed in $[0;2\tau]$. Given a debt level d, the terminal value of a τ-firm is $\tilde{x}_\tau$ if she successfully repays her debt (case $\tilde{x}_\tau \geq d$) and $\tilde{x}_\tau - L$ otherwise, L being the cost of reorganization generated by defaulting on the debt obligation. The expected final value is thus computed as

$$V_1^\tau(d) = \int_0^d (\tilde{x}_\tau - L)\,\mathrm{d}x + \int_d^{2\tau} \tilde{x}_\tau\,\mathrm{d}x = \int_0^{2\tau} \tilde{x}_\tau\,\mathrm{d}x - L\int_0^d \mathrm{d}x = \tau - Ld/2\tau$$

assuming $d < 2\tau$ (debt lesser than maximum cash-flow).

Let us consider the remuneration $w = \gamma_0 V_0 + \gamma_1 V_1$. In equilibrium the choice d_τ of a type-τ firm must be optimal for the manager, hence $\frac{\partial w}{\partial d} = 0 \Leftrightarrow \gamma_0 V_0' + \gamma_1 V_1' = 0 \Leftrightarrow V_0'(d_\tau) = \frac{\gamma_1 L}{2\gamma_0 \tau}$. Since in equilibrium types are revealed we have $V_0(d_\tau) = \tau$ hence $V_0'(d_\tau) d_\tau' = 1$. Plugging into the previous equation we obtain $d_\tau' = \frac{2\gamma_0\tau}{\gamma_1 L}$, thus $d_\tau = \frac{\gamma_0\tau^2}{\gamma_1 L} + c$ where c is the integration constant. Taking into account the fact that the worst type will not emit any debt to eliminate the risk of bankruptcy, we derive $d_\tau = \frac{\gamma_0(\tau^2 - a^2)}{\gamma_1 L}$. Lastly we must check that our initial assumption $d < 2\tau$ is satisfied which requires eliciting γ_1 large in front of γ_0 i.e., the manager's remuneration must strongly depend on the future where bankruptcy might happen to credibly transmit information to the market as regard the type of her firm.

23.3.2 Debt as a Signal of Obedience

A *moral hazard* issue tantalizing financiers is the possibility that the manager of a firm follows the pursuit of happiness rather than the pursuit of benefits. As we explained in the introduction, it is reasonable to assume that bankruptcy is costly for a manager-entrepreneur; this fact leads Grossman and Hart (1982) to argue that issuing debt is a pre-commitment or bonding behavior aimed at convincing investors that the firm will be managed to maximize profits, so as to avoid bankruptcy. Hence, a high leverage could *signal* the good prospects of the firm and be the guarantee that investments will be carried on at the efficient level.

To check formally this claim, we consider first the manager of a firm that is a distinct from the owners. The firm (aka. the owner) raises an amount F of funds by selling a mix of equity and a debt obligation d. The moral hazard issue here is the fact that it is the manager who allocates k into physical capital and the remnant $F-k$ into her human capital which we interpret as a private benefit since it is an unalienable asset that cannot be taken away by the owners.[30@] What disciplines the manager is the fact that she will enjoy her private benefits only if she remains at the head of the firm.

The project's future cash flow is $R(k)+\tilde{x}$ where $\tilde{x}$ is a random shock of zero mean[31@] so that the expected NPV of profits is $R(k)-k$. Given the uncertainty regarding the future cash flow, the firm will go bankrupt whenever $R(k)+\tilde{x}<d$, hence the expected utility of the manager is the average of her perks $F-k$ over the states of nature where she enjoys them i.e.,

$$\mathbb{E}\left[F-k|R(k)+\tilde{x}\geq d\right]=(F-k)\left(1-H\left(d-R(k)\right)\right)$$

The optimal investment $\hat{k}$ solves the FOC

$$(F-k)R_m(k)h\left(d-R(k)\right)=1-H\left(d-R(k)\right)\Leftrightarrow\frac{1}{(F-k)R_m(k)}=\phi\left(d-R(k)\right)$$

where $\phi(x)\equiv\frac{h(x)}{1-H(x)}$ is the hazard rate of the distribution function H.[32@] In equilibrium of the capital market, investors anticipate the choice $\hat{k}$, thus value the firm's securities F as the present value $R(\hat{k})$ of the project; the first order condition determining $\hat{k}$ then becomes $\frac{1}{(R(k)-k)R_m(k)}=\phi\left(d-R(k)\right)$ and we observe on Figure 23.4 that an increase in debt from d_1 to d_2 moves the RHS up, hence the optimal investment increases from k_1 to k_2; we can therefore conclude that the leverage chosen by the entrepreneur is a commitment to invest mostly into the firm's future value and not into her personal satisfaction.

Adding a greater degree of realism is possible without changing the qualitative nature of the result. Firstly, the utility for the manager need not be linear in the perks she keeps for herself, it can be $u(F-k)$ for a concave increasing utility function. Then

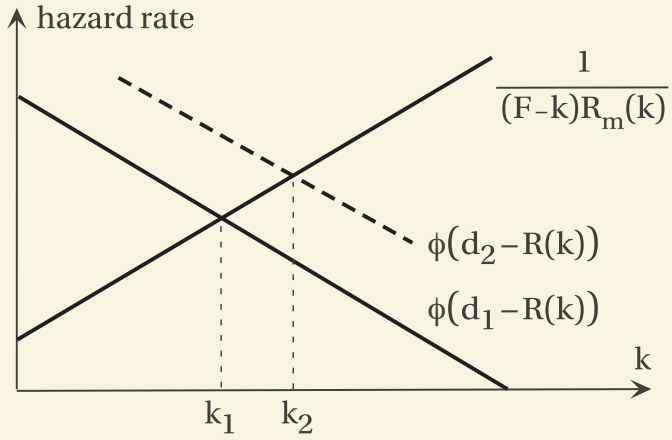


Figure 23.4: Bonding Behavior of Debt

the FOC becomes $\frac{u'(F-k)}{u(F-k)R_m(k)} = \phi(d - R(k))$ but since the ratio $\frac{u'(x)}{u(x)}$ is a decreasing function of x, our previous conclusions remain unchanged. Then we are able to consider the case of an entrepreneur instead of a manager; in that case the utility she maximizes is $\mathbb{E}[u(F-k) + R(k) + \tilde{x} - d | R(k) + \tilde{x} \geq d]$ so that the FOC becomes $\frac{u'(F-k) - R_m(k)}{u(F-k)R_m(k)} = \phi(d - R(k))$ and once again, the qualitative effect of raising the leverage remains a commitment to later choose a greater investment.

23.3.3 Efficiency Wage

Intuition

The persistence of unemployment in competitive economies has long been a puzzle that standard neoclassical theory could not solve. Indeed, in the many economies around the world that do not guarantee a minimum wage, why can't the market wage descend low enough (with subsistence as a lower bound) to stimulate demand and thereby provide a job to every person who seeks one? Shapiro and Stiglitz (1984) offer an innovative explanation based on incentives and asymmetry of information.

To understand their theory it is useful to work by contradiction. If there existed an equilibrium with full employment, a fired worker would be instantaneously rehired at the same equilibrium market wage. Thus, in terms of utility, she would barely notice the change which means, adopting the firm's point of view, that there is no way to penalize a worker caught shirking (not working as hard as stipulated in her contract). Thus, to motivate hard work, firms are forced to pay above-market wages in order that the loss of one's job be painful. But, as always in economy, if one firm finds it attractive to pay above-market wages, then all firms will do the same. This means that the market wage increases and exactly matches the wage paid by each firm, hence work incentives

have been destroyed once again. However, wages being above their natural equilibrium values, labor has become a more expensive input (lower productivity) and its aggregate demand falls generating *unemployment*. Now, the consequence of being fired (for being caught shirking) is more dreadful than before because the individual will have to wait before finding a new job and will have to live poorly in the meantime.

We have thus seen that the wage is not only the price of the labor input equating demand and supply but also a device to provide work incentives inside the firm. As intuition suggests, when one instrument is used to solve two problems, some inefficiencies are bound to appear.

Model

To check this claim, consider identical risk neutral workers whose utility is $w - q$ where w is the wage and q the effort they exert (0 if jobless). The common rate of discount is r. For simplicity, the efficient effort is a fixed level $q > 0$. At each period, a person can be unemployed (type u) or enjoy a job (type j) in which case he will either work obediently (o) or shirk (s). Jobless people receive a benefit b that can be provided by the State or family. There is an exogenous probability ρ of losing one's job (e.g., the firm goes bankrupt) and a probability λ of being caught by the monitoring technology if shirking.

The value of being employed and unemployed are denoted V_j and V_u; the present value of remaining in a given situation forever is $V \times \sum_{k\geq 1} \frac{1}{(1+r)^k} = rV$. For an employed "shirker" with present value V_j^s we also have

$$V_j^s = w - (\rho + \lambda)(V_j^s - V_u) \tag{23.15}$$

i.e., what she expects is equal to her current wage minus the expected loss of utility relative to the possible job termination that might occur at the end of the period (if caught shirking). For an "obedient" worker, the effort must be accounted while the probability of losing one's job is lower, thus the equation is

$$rV_j^o = w - q - \rho(V_j^o - V_u) \tag{23.16}$$

Solving the two equations so obtained we derive

$$V_j^s = \frac{w + (\rho + \lambda)V_u}{\rho + \lambda + r} \quad \text{and} \quad V_j^o = \frac{w - q + \rho V_u}{\rho + r} \tag{23.17}$$

Shirking won't take place only if $V_j^o > V_j^s \Leftrightarrow w > \hat{w} \equiv rV_u + \frac{(\rho+\lambda+r)q}{\lambda}$. We can now assume that employers will pay exactly $\hat{w}$ to avoid shirking by their employees.

For a jobless person, there is a unique equation since he does not have to decide if he shall work or shirk. Letting α denote the job acquisition rate, the value of being unemployed satisfies

$$rV_u = b + \alpha(V_j - V_u) \tag{23.18}$$

Since in equilibrium, workers do not shirk we have $V_j = V_j^o$; (23.17) and (23.18) form a system whose solution is

$$V_j = \frac{(\hat{w} - q)(\alpha + r) + b\rho}{\alpha + \rho + r} \quad \text{and} \quad V_u = \frac{(\hat{w} - q)\alpha + b(\rho + r)}{\alpha + \rho + r} \tag{23.19}$$

Replacing the latter formula into the definition of $\hat{w}$ further yields $\hat{w} = b + q + \frac{(\alpha+\rho+r)q}{\lambda}$. In terms of comparative statics the critical wage must be larger when work is more painful (larger q), when the unemployment benefit b is larger, when the probability of being caught shirking λ is lower, when the interest rate r is larger (preference for the present) and when the economy is more unstable (larger ρ).

Lastly, using the total worker population N and the employed population L, we can derive α since in equilibrium, the flows out of unemployment equal the flows in so that $\alpha(N - L) = \rho L \Rightarrow \alpha = \frac{\rho L}{N-L}$ and $\hat{w} = b + q + \frac{q}{\lambda}\left(\frac{\rho N}{N-L} + r\right)$. As we can see on Figure 23.5, this relation draws a frontier $L = S(w)$ between combinations of employment and wages that induce shirking or work inside firms. We can interpret this curve as a labor supply function.

.

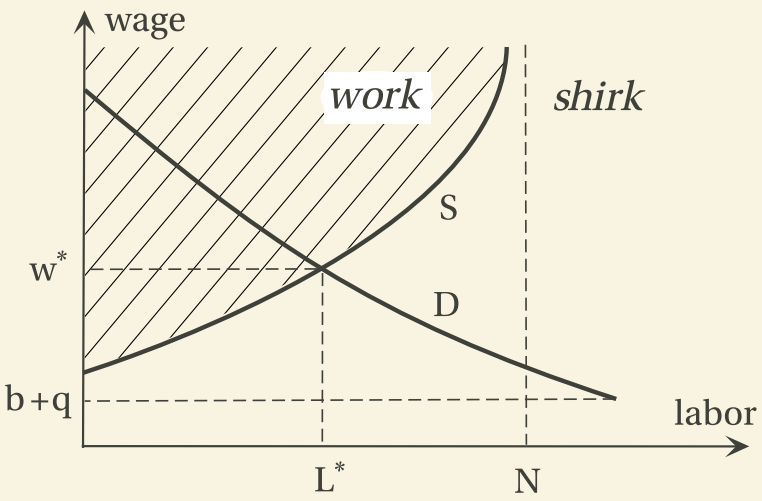


Figure 23.5: Efficiency Wage

To derive the equilibrium on the labour market we have to consider the behavior of firms; we only need to assume that the marginal productivity of labour is decreasing

with employment to obtain a downward slopping demand curve D. Also the productivity of the fixed effort q must be large enough to warrant full employment at the complete information equilibrium i.e., $D(b+q) > N$ as represented on Figure 23.5. Given the moral hazard issue of shirking, firms are forced to add the incentive constraint $L \leq S(w)$ to the traditional technological constraint $L \leq D(w)$ i.e., they limit their demand schedule to the upper part of the D curve so that the equilibrium is a pair (w^*, L^*) generating unemployment.

Some interesting observations are:

- All the above comparative statics regarding conditions that increase the critical wage turn out to increase unemployment.
- Jobless workers would accept a lower wage but since they are unable to commit not to shirk, firms prefer to let them out.
- During a recessions the labor demand moves down which lowers wages but the probability of shirking rises (larger ρ) so that unemployment is doubtlessly increased.
- The equilibrium is inefficient because firms employ too few (they equate w^* not e to their productivity) and because each of them causes a negative externality on others given that V_u is increased for other firms when one of them hires more people.

If the workers are the communist owners of firms then the optimal level of employment maximizes workers utility $(w-q)L + b(N-L)$ under the "no shirking" constraint $w \geq \hat{w}$ and the industry participation constraint $wL \leq \Phi(L)$. Given that $w \geq \hat{w} \Rightarrow w > b+q$, the social objective turns out to be the maximization of employment under the two previous constraints i.e., the optimum solves $\hat{w}(L) = \Phi(L)/L$. For most industries, there are diseconomies of scale so that the average productivity is greater than the marginal one and leads to a greater employment than at the market equilibrium. If workers and owners are different economic agents then the only change in the above Pareto program is $wL \leq \Phi(L) + \theta$ where $\theta \geq 0$ is a non negative parameter. As a result, optimal employment is reduced but it still remains above the market equilibrium level. Hence we may conclude that asymmetric information and costly monitoring generate a market failure creating too much unemployment. The government is thus warranted to intervene to reduce unemployment by using (not too distorting) taxes.

Part I

Network Industries

Chapter 24

Standards and Components

Listening to music, taking photos, watching a video, using a computer are occupations involving bundled goods because each of these activities uses a media for storage, an interpreter for treatment and a human-interface to enjoy. There are many other everyday life examples of *components* that are assembled to make up a final good e.g., a house, an airplane, a meal, a car or an electronic device.

When two components can be successfully combined they are said to be *compatible* or to follow the same *standard*. More generally, a standard is a convention, a specification, a protocol, or an understanding that allows successful interaction between humans, between machines or between humans and machines. Agreement on a standard[1@] is desirable for society as it enables:

- interchangeability in consumption (CD, DVD, tape, disk)
- interchangeability in production (car parts, electronic components)
- ease of communication (telephone, keyboard, units, english language)

Most final goods are sold as pre-assembled bundles because it is much cheaper to have the bundling realized by an industrial producer than by the end-user.[2@] However, the consumer's desire to mix and match himself the components has lead to increased connectivity, a concept we could define as compatibility made user-friendly. For instance, loudspeakers can be connected to any hi-fi system or the SIM card sold by a mobile network operator can be inserted in any mobile handset. More connectivity means more markets and therefore more competition.

Bundling is also present in services such as energy or telecommunications. Indeed, these are made of several components displaying complementarity as well as substitutability for the consumer. Due to network externalities (cf. §25), their supply displays economies of scope and scale so that firms tend to be active in the market for the final service rather than the market for a single component.[3@]

In this chapter, we first look at the market power that a firm can derive from the complementarity of the various goods or services it sells. Next, we study the incentive

for firms to agree or disagree on a standard i.e., whether to make their components connectable for the end-user. We also look at the pattern of adoption for a standard or a new technology. In the following section, we look at the critical mass issue and discuss some myths related to over-stretched conclusions derived from some simple models of network externalities. Detailed examples explain why history and dynamics should never be lost of sight when dealing with standards. Our penultimate section analyzes two sided platforms which are intermediary networks trying to connect end-users and service providers. Finally, we conclude this chapter with social interaction, the effects of conformity and vanity on our everyday purchasing behavior; we show that the traditional conclusions of demand theory can be altered by network phenomena.

24.1 Components: Tie and Bind

Before considering competition among firms on standard setting, we investigate how one firm with market power, for instance a monopoly, can use tie and/or bind as a price discrimination device to increase profits (cf. also §4.1.3). Needless to say, bundling may also be triggered by technical complementarity, cost savings and network effects. An early reference discussing this strategy is Burstein (1960).

24.1.1 Tying

Tying (aka tied selling or pure bundling) is the commercial practice that conditions the sale of product B on the purchase of another product A. A frequent example is the installed operative system (OS) on a computer; requiring an OS free hard-drive does not entitle the buyer to a rebate which means that hardware and software are tied. The practice although common has nevertheless been judged anti-competitive in the EU and the US in a number of cases among which:

Hilti a producer of fastening systems used in the building industry, was abusing its dominant position by supplying cartridge strips only when purchased with the necessary complement of nails.

Tetra Pak required its customers to use only the Tetra Pak's cartons with the Tetra Pak's filing machines and moreover, to obtain those cartons only from the supplier itself.

IBM required the purchasers of its mainframe computers in the 1970s to buy exclusively its tabulating cards.

Kodak tied the sales of part for machines to the sales of repair services for these machines.

Microsoft tied the "Internet Explorer" browser to the "Windows" operating system.

In all the previous examples, a firm sells a unique durable good A (protected by patents) so that the firm holds market power. However, this durable good consumes a perishable good or service that can be delivered competitively. In the absence of tying, the monopoly can freely fix p_A but is obliged by the competitive fringe offering good B to set $p_B = c$, the marginal cost of delivering good B. This situation is quite similar to the monopolist using a two-part tariff; when consumers are homogeneous, the unit price is optimally set at the marginal cost to generate maximum demand and thus maximum surplus which is then siphoned through the subscription. The latter plays the role of p_A while the unit price plays the role of p_B.

Now, in the more realistic case where consumers are heterogeneous, we already saw in §4.3 on quantity discrimination that the optimal pricing scheme involves increasing the unit price and decreasing the subscription so as to rip higher margin on all the consumers eager to consume good B, the consumable, in large quantities. Translated to the present situation, the firm would like to bind the purchase of the two items in order to be able to set $p_B > c$ which is the absence of tying is impossible due to the competition of consumable makers.

24.1.2 Bundling

A commercial practice more flexible than tying is mixed *bundling* whereby each good can be purchased either as a separate item or as part of a bundle.

Imagine that potential consumers belong to two groups $i = 1,2$ of equal size whose willingness to pay are w_A^i and w_B^i for $i = 1,2$. If individual characteristics were observable and discrimination was legal, the optimal prices would be the reservations prices. In the more realistic case where individuals cannot be distinguished, the seller must charge $p_A = min\{w_A^1, w_A^2\}$ if he wants to include everyone and similarly for good B. Bundling is almost always superior because it averages the differences among consumer groups; indeed the optimal bundle price is $p_{AB} = min\{w_A^1 + w_B^1, w_A^2 + w_B^2\} > p_A + p_B$ as soon as there is some asymmetry in the consumers preferences. Examples of famous bundles are the Microsoft Office suite which bundles programs for typing texts, calculating with spreadsheets, maintaining databases and making presentations.

Another case where bundling can appear is when the two goods show complementarity in use, an issue already seen in §3.2.3 on multiple products. Consider for instance the demand for the bundle $D(p) = 1 - p$ where p is the price of the bundle. If the two components are sold separately at prices p_A and p_B then the bundle price perceived by consumers is $p = p_A + p_B$ so that firms A and B receive demands $q_A = 1 - p_A - p_B$ and

$q_B = 1 - p_A - p_B$. The inverse demands are $p_A = 1 - p_B - q_A$ and $p_B = 1 - p_A - q_B$; assuming zero cost, the optimal prices (found from the optimal quantities) are then $p_A = \frac{1-p_B}{2}$ and $p_B = \frac{1-p_A}{2}$. The equilibrium is thus $p_A = p_B = \frac{1}{3}$ leading to industry profits of $\frac{2}{9}$.

If the two firms integrate, then the usual monopoly price is $\frac{1}{2}$ leading to the greater profit of $\frac{1}{4}$ which could rationalize the fact that firms making complementary goods tend to merge. Noticing that sales increases, the merger is efficiency enhancing, contrary to intuition. The reason for this paradox is that the independent firms fail to account for the externality their own price impose on the demand for the other component; when competing one against the other they unduly restrict trade much like in the double marginalization problem (cf. §14.1.3).

Adams and Yellen (1976) illustrate this reasoning with the help of Figure 24.1 by first abstracting from cost issues. The distribution of WTPs for the two goods are shown on the two axes and are assumed to be independently drawn so that a point represent a consumer. Treating each good separately, the firm finds two profit maximizing prices p_A^* and p_B^* which divide the space into four zones. This is illustrated on the left pane. People whose WTP pair lie in **(i)** buy neither A nor B, those in **(ii)** buy B only, those in **(iii)** buy A only while those in **(iv)** buy both products. When applying pure bundling or tying, the firm finds an optimal price p_{AB}^* and only those consumers whose WTP pair satisfies $w_A + w_B > p_{AB}^*$ will buy. Hence zones ($\boldsymbol{\alpha}$) and ($\boldsymbol{\beta}$) are separated by a line of slope -1 as displayed on the central pane of Figure 24.1. Lastly, when the firm combines the two previous approaches, consumers are sorted again into four groups with the novelty that group **(iv)** whose members buy the bundle is made of zones ($\boldsymbol{\alpha}$), ($\boldsymbol{\beta}$), ($\boldsymbol{\gamma}$) and ($\boldsymbol{\delta}$) where people would, respectively, buy none of the separate goods but the bundle, only good B, only good A and finally those who under any circumstance end up with the two goods. If production cost are zero then mixed bundling with $p_{AB} < p_A^* + p_B^*$ is a superior strategy since the α people now buy the bundle instead of nothing.[4@]

In the general case where values are not independently drown, bundling is attractive when values are negatively correlated. If some people display a low WTP for good A and a high one for B, they might buy the bundle instead of solely B. When this happens, the firm is making a loss over the good A which would not happen under pure component pricing.

More generally, McAfee et al. (1989) show that mixed bundling is profitable relative to separate sales for the simple reason that offering more (cleverly designed) options to consumers cannot hurt wrt. optimal separate prices. Indeed, if we initially set the bundle price to be the sum of the optimal separate prices, it must be the case that one can increase profit with a slight change of bundle price. We already saw that with independently distributed values, it is optimal to discount the bundle price so as to generate

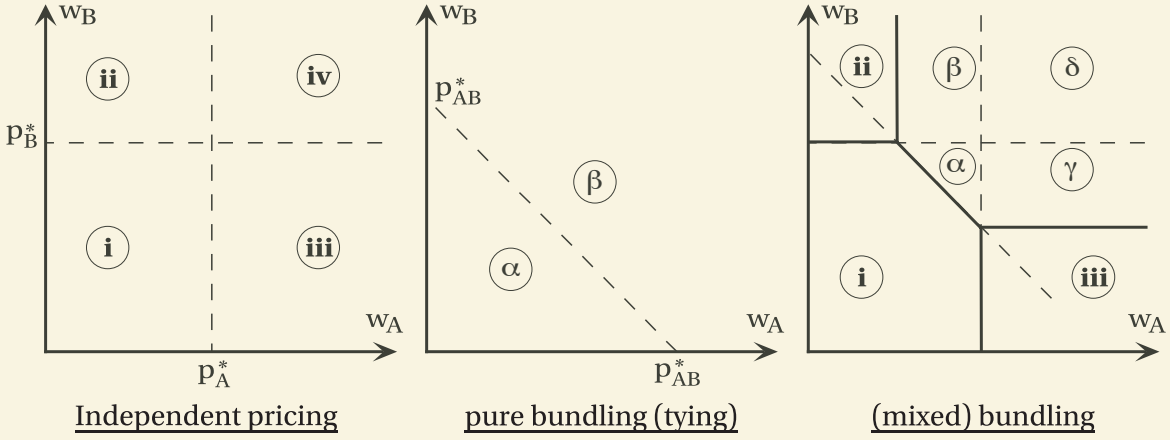


Figure 24.1: The advantage of Bundling over pure pricing

additional sales (which is welfare improving). The more problematic case is when it is optimal to raise the bundle price because then the firm must monitor sales to keep buyers from assembling the bundle by themselves (i.e., buying the independent components).

Mixed bundling might be unfeasible because of antitrust pressure (cf. case of dearer bundle wrt. separate items), technological reasons or transaction cost.[5@] Hence, the firm must decide whether to sell two good or one bundle good. Fang and Norman (2006) show that under likely assumptions regarding the distribution of WTPs, bundling reduces variance which is good for the firm when a good should be sold with high probability (either because costs are low or because valuations tend to be high). Conversely, the reduction of taste dispersion is detrimental when the goods have only a thin market (either because the costs are high or because valuations tend to be low). Thus a low marginal cost or high mean valuation make bundling more likely to be optimal.

24.2 Creation of a Standard

The International Standard Organization (ISO) as its name makes clear, organizes the activities of many research committees among which the Moving Pictures Expert Group (MPEG) and the Joint Photographic Experts Group (JPEG). In other instances, the standard is developed privately as in the cases of CD and DVD which were the result of lengthy negotiations among industrials (cf. history of the CD-audio standard). In many instances, a single private initiative succeeds to grab a dominant market share and then becomes a "de facto" world standard. However, as the *high-definition DVD* story illustrates, it sometimes happen that several private initiatives are developed competitively, offering incompatible standards. This situation is extremely inefficient as it retards the development of the consumer market until one of the contender is defeated.

We shall consider in turn the competition among two substitutable candidates for achieving the status of standard, then we look at the effect of complementarity among the two products and finally at the issue of compatibility among the two products.

24.2.1 Substitutes

Two firms Apple (A) and Microsoft (B) endorse a different standard (Macintosh and Windows) for the operative system of their computers. The payoffs of their simultaneous interaction on the consumer market are:

$A \setminus B$	Mac	Bin
Mac	x, y	s, r
Bin	r, s	y, x

If $\min\{x, y\} > \max\{r, s\}$ then Mac for both or Bin for both are Nash equilibria while if $\max\{x, y\} < \min\{r, s\}$ then asymmetric choices are Nash equilibria including Apple producing a Bin-compatible product and Microsoft producing a Mac-compatible one.

This example illustrates the fact that if profits are high for compatible lines then firms are lead to cooperate on any standard. At the opposite if firms benefit from being the sole supplier of a given standard then cooperation does not appear. The question we ask is under which condition the payoff of the above reduced matrix present one configuration or the other. The answer can be found only by digging into consumer choice as we already saw in Part E on differentiation but this time taking into account the effects of compatibility and networks.

24.2.2 Complements

Bundled goods are made of complementary components such as a computer operative system working with software, a game console using games cartridges or a movie player reading movies tapes. Let us study the competition among firms A and B who produce consoles running compatible games which come in respective numbers N_A and N_B.

For some historical reason (not modeled here), consumers have a more or less marked preference x for a brand which is distributed uniformly over the $[0;1]$ interval. The surplus of consumer x is $U_x^A = x\sqrt{N_A}$ when adopting system A and $U_x^B = (1-x)\sqrt{N_B}$ when adopting system B (complementary services have decreasing return to scale).

The indifferent consumer (cf. Fig. 11.1) is $\tilde{x}$ such that $U_{\tilde{x}}^A = U_{\tilde{x}}^B$; his location determines the respective market shares or consoles and it is trivial to compute $\frac{D_B}{D_A} = \frac{1-\tilde{x}}{\tilde{x}} = \sqrt{\frac{N_B}{N_A}}$. Hence if new games become available only for a brand, that brand's market share

increases. It is also straightforward to see the effect of prices when the games supply is proportional to the expenditure on the console i.e., when $N_A = D_A(Y - p_A)$ where Y is the consumer leisure budget.[6@] We obtain $\frac{1-\tilde{x}}{\tilde{x}} = \frac{Y-p_B}{Y-p_A} \Leftrightarrow D_A = \frac{Y-p_A}{2Y-p_A-p_B}$ so that an increase in price p_A reduces the market share D_A and even more strongly the supply N_A.

We have here a direct and simple explanation of the recurrent and fierce price battle occurring in the markets for PC's and game consoles.

24.2.3 Compatibility

We study oligopolistic competition for the establishment of a standard like mobile vs. fixed phone or laptop vs. desktop computer.

Let us assume that consumers suffer a disutility c of not consuming their preferred model, either A or B. Otherwise they value the customer base x, so the surplus of a type i consumer from product j is $U_i = x_j - c1_{j\neq i}$. If the intrinsic attachment to a particular brand is not too large ($c < 1$) then both brands can become the de-facto standard i.e., achieve 100% market share. Indeed if brand A is the standard, no B customer can find it profitable to switch (alone). Otherwise, when $c > 1$, it is a dominant strategy to adopt one's preferred brand whatever its actual market share; hence both brand achieve a positive market share.

To analyze an equilibrium with coexistence of different standards we have to consider the share a of A-lovers. It is easy to see that $D_A = a$ and $D_B = 1 - a$ is an equilibrium if the switching cost is large enough: $c > 1 - 2\max\{a, 1-a\}$. Indeed, when A-lovers consume A, none of them wants to change to B if $a > 1 - a - c$. In the remaining cases, the preferred brand of the majority imposes itself as the standard.

Defining social welfare as $U_A D_A + U_B D_B$, it is clear that the efficient standard is the one preferred by the majority but the two-standards outcome may dominate both. As its welfare is $a^2 + b^2$ while that associated to standard A is $a + b(1-c)$, incompatibility dominates A if $c > 2a$. Hence incompatibility is the best outcome if $c > 2\max\{a, 1-a\}$ i.e., when switching cost is high and/or types are evenly distributed.

Arguments of coordination through repeated interaction[7@] guarantee that clients will coordinate on the correct standard when both are Nash equilibria; thus no market failure occurs. There is a market failure if incompatibility is the Nash outcome and it is an inefficient one i.e., if $\frac{1-c}{2} < \max\{a, 1-a\} < \frac{c}{2}$. To conclude, we can say that according to the parameters values there are inefficient equilibria, efficient ones, displaying standardization or incompatibility.

Network effects have also an influence on preemption strategies when for instance each customer's valuation of a product grows with the number of others adopting the

product. In this case, an incumbent can profit from aggressive pricing to prevent entry, because the present losses are recouped later in the form of large profits derived from the larger base of captive customers. This is especially true if the prevention of entry encourages standardization on the incumbent's product and thereby lessens subsequent risks of entry (think of free software and M$).

In the late 1980s MS bundled its software "Word" and "Excel" at a price slightly in excess of its leading product "Word" alone to gain the critical mass against "Lotus" and "Quattro". Then in the 1990s MS did the same with the Powerpoint program but this time in a preemptive manner to prevent the entry of another presentation program.

24.2.4 Adoption of a Standard

Intuition

When looking at the speed of adoption of a given technology, whether a standard or an innovation, the percentage of users plotted against time always displays an S shape. Take-off is very slow because consumers fear embarking on a dead-end so that early adopters are risk loving people. Later on, there is moment where adoption accelerates because everybody is "jumping on the bandwagon" being eventually convinced by the usefulness of the standard and the fact that the more people use it the more valuable it is. The last phase of slower adoption signals that the product has reached its potential. As intuition suggests, the absence of coordination delays adoption because the optimal individual behavior has a free rider flavor: "it's better to wait until others switch, before doing it myself"; excess inertia could thus appear (cf. §2.4.4).

Farrell and Saloner (1985), using the famous backward induction reasoning, show this won't happen in a complete information world: if everybody has already switched then it is dominant strategy to do so for the last consumer. Then the penultimate consumer, conditional on the previous ones having switched, will also find it a dominant strategy to switch given that its own move will be followed by a last switch. This way everybody sequentially switches and the new standard is readily adopted.

Obviously, this line of proof breaks down under incomplete information because no one can be sure that its own move will be followed by more switches. These authors provide us with a model of adoption of a new technology in a world where firms know precisely how much they would pay to adopt e.g., they know how the switch will be received by their clients. What they ignore is how much would their challengers pay, that is to say if they are more or less impatient to switch.

Model †

Each firm's willingness to adopt the new technology depends on a private information parameter θ. From the point of view of challengers, θ is uniformly drawn from $[0;1]$. Since there are 2 firms able to make 2 choices each, we are left with 4 situations. Assume then that a firm's profit as a function of her own parameter θ is

$$B^{old}_{alone}(\theta) = -1, \quad B^{old}_{both}(\theta) = 0, \quad B^{new}_{alone}(\theta) = 8\theta - 4, \quad B^{new}_{both}(\theta) = 8\theta - 3$$

as shown on Figure 24.2.

To inquire how fast firms adopt the new technology we use a two periods model of adoption among two firms. Assuming irreversibility, the available strategies are:

- s_1: never switch
- s_2: wait and switch if the other did
- s_3: switch immediately
- s_4: wait and switch if the other did not
- s_5: wait and always switch

The important properties of these benefit functions are the existence of a positive network effect for each technology ($B^{old}_{alone} < B^{old}_{both}$ and $B^{new}_{alone} < B^{new}_{both}$), people with low θ's will play s_1 no-matter what ($B^{new}_{both}(0) = -3 < -1 = B^{old}_{alone}(0)$) and people with high θ's will play s_3 ($B^{new}_{alone}(1) = 4 > 0 = B^{old}_{both}(1)$). Note that s_4 and s_5 are dominated by s_1 and s_2 respectively.

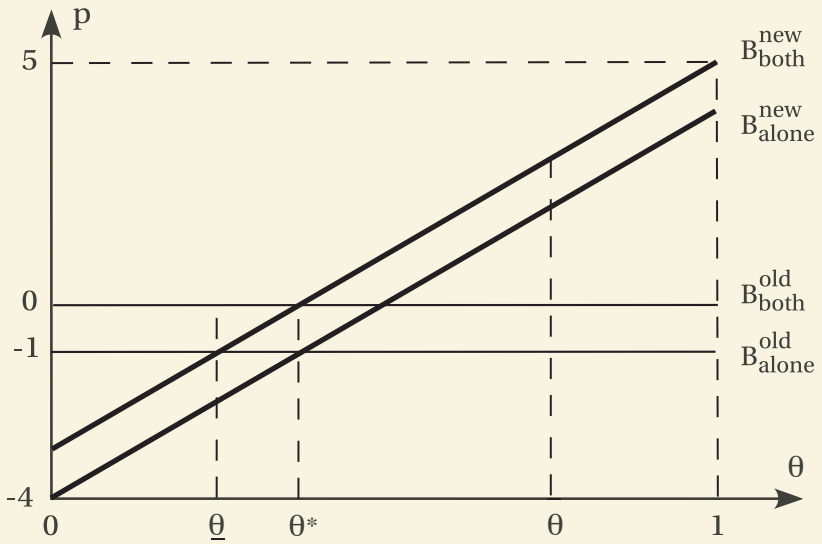


Figure 24.2: The Bandwagon Effect

We show that there is a Bayesian equilibrium where firms with an intermediate willingness to change decide to "jump on the bandwagon",[8@] that is to say to adopt the new technology once the other has adopted (s_2). Technically the whole equilibrium strategy is

$$\sigma(\theta) = \begin{cases} s_1 \text{ if } \theta < \underline{\theta} \\ s_2 \text{ otherwise} \\ s_3 \text{ if } \theta > \bar{\theta} \end{cases}$$

where the thresholds satisfy $\underline{\theta} < \bar{\theta}$. We now compute the payoff associated to s_1, s_2 and s_3 when the opponent follows σ. We have:

$$\begin{aligned} \Pi_1(\theta) &= \underline{\theta} B^{old}_{both}(\theta) + (1-\underline{\theta}) B^{old}_{alone}(\theta) = -(1-\underline{\theta}) \\ \Pi_2(\theta) &= \underline{\theta} B^{old}_{both}(\theta) + (1-\underline{\theta}) B^{new}_{both}(\theta) = (1-\underline{\theta})(8\theta - 3) \\ \Pi_3(\theta) &= \bar{\theta} B^{new}_{alone}(\theta) + (1-\bar{\theta}) B^{new}_{both}(\theta) = 8\theta - 3 - \bar{\theta} \end{aligned}$$

The threshold $\underline{\theta}$ is part of an equilibrium only if the firm with type $\underline{\theta}$ is indifferent between strategies s_1 and s_2 i.e., $\Pi_1(\underline{\theta}) = \Pi_2(\underline{\theta}) \Leftrightarrow \underline{\theta} = \frac{1}{4}$. Likewise $\bar{\theta}$ is part of an equilibrium only if the firm with type $\bar{\theta}$ is indifferent between strategies s_3 and s_2 i.e., $\Pi_2(\bar{\theta}) = \Pi_3(\bar{\theta}) \Leftrightarrow \bar{\theta} = \frac{3}{4}$. We can conclude that

> Equilibrium behavior leads to excessive inertia because firms fear to adopt alone and thereby lose the benefit of the network externality.

Observe indeed that it is efficient to have both firm adopting whenever

$$B^{new}_{both}(\theta_1) + B^{new}_{both}(\theta_2) > 2B^{old}_{both}(\theta) \Leftrightarrow \frac{\theta_1 + \theta_2}{2} > \theta^* = \frac{3}{8}$$

(cf. Fig. 24.2) while adoption is delayed by at least one of the firm if either $\theta_1$or θ_2 is lesser than $\bar{\theta}$.

It can be shown that

> Equilibrium behavior leads to excessive inertia because firms fear to adopt alone and thereby lose the benefit of the network externality.

Farrell and Saloner (1985) project the result of their model to the political arena. On a controversial question, staunch supporters or opponents always commit themselves without waiting to see whether their stance become the popular view. Professional (wiser ?) politicians often wait awhile to test the political waters, declaring themselves to be "for" the measure if the bandwagon begins to roll and "against" otherwise.

24.2.5 On Deregulation

In the US, the deregulation of the telephone sector took two steps. Firstly the ATT monopoly was broken into one long-distance company and several regional local operators; secondly the long-distance market was opened to competition. Economides (1999) points at a neglected network effect: the total quality of a phone call is the minimum of sound quality on the two local lines and the long-distance one. This author claims that breaking a monopoly telephone company into local and long-distance companies can reduce the overall quality of the service because the service is made up of complementary elements. Although the new firms will coordinate their quality levels, the double marginalization (cf. §14.1.3) lowers individual profit margins and thus lowers quality choices.

Consider the market for good AB with components A and B respectively produced by firm A and B. The final quality of good AB is $q \equiv \min\{q_A, q_B\}$ where q_A and q_B are the qualities chosen by firms ex-ante. The utility of consumers is $u(x) = xq - p$ where p is the price of good AB and the type x is uniformly distributed in $[0;1]$. Assume zero marginal cost of production and total fixed cost $\phi_A(q_A) + \phi_B(q_B)$ where $\phi_i(q_i) = c_i \frac{q_i^2}{2}$ for $i = A, B$. In the present situation there are no increasing returns to scale, hence no potential benefit to integrate the two firms.

Consider first the behavior of the integrated monopolist. Since the sales for price p are $1 - \frac{p}{q}$, the optimal price is $p^M = \frac{q}{2}$ so that the optimal final quality q maximizes $\Pi^M(q) = \frac{q}{4} - \phi_A(q_A) - \phi_B(q_B)$. The optimal qualities are thus $q_A = q_B = q^M \equiv \frac{1}{4(c_A + c_B)}$ and the total profit is $\Pi^M = \frac{1}{32(c_A + c_B)}$.

After deregulation firm A and B sell their component at prices p_A and p_B. Sales of both goods are identical and equal to $1 - \frac{p_A + p_B}{\min\{q_A, q_B\}}$. Assuming that prices are chosen simultaneously we obtain a variant of the Hotelling model. The best replies are therefore $p_A = \frac{q - p_B}{2}$ and $p_B = \frac{q - p_A}{2}$ leading to the equilibrium, $p_A^D = p_B^D = \frac{q}{3}$ and most notably to the final deregulated price $p^D = \frac{2q}{3} > p^M$ i.e., a lower market coverage. We obtain here the inefficiency of double marginalization identified by Cournot (cf. §14.1.3). The remaining question is wether the consequence on quality choices is also negative.

Ex-ante, the best reply of a firm to its competitor cannot be a higher quality since it does not result in any advantage, only costs. The profit of firm A being $\frac{\min\{q_A, q_B\}}{9} - \phi_A(q_A)$, the best reply to q_B is $\min\left\{\frac{1}{9c_A}, q_B\right\}$. The marginal benefit factor of quality has dropped from $\frac{1}{4}$ for an integrated monopoly to $\frac{1}{9}$, thereby generating a lesser investment into quality. This problem is also know as *hold-up* (cf. §14.2) because the necessity to bundle A's component to B's puts firm A at the mercy of firm B's opportunism (it does indeed appear at the price stage). Firm A feeling that some of the rents of a higher quality are hold-up by firm B is lead to under invest.

The choice of qualities has a unique equilibrium

$$q_A^D = q_B^D = \min\left\{\frac{1}{9c_A}, \frac{1}{9c_B}\right\} < \frac{1}{4(c_A + c_B)}.$$

The difference is minimized at $\frac{1}{72c}$ when $c_A = c_B = c$. In this symmetric case, total profits are $\Pi_A^D + \Pi_B^D = \frac{1}{81c} < \frac{1}{64c} = \Pi^M$. The consumer surpluses are respectively $S^M = \frac{1}{64c}$ and the much lower $S^D = \frac{1}{162c}$.

Observe that the utility function $\min\{q_A, q_B\}$ is the limit of a constant elasticity of substitution (CES) function as the elasticity of substitution tends to zero. Hence the previous results hold for goods whose components are highly complementary.

24.2.6 Network and Competition

Since the interaction between a consumer and its service provider lasts several periods, we may start to care for the size of our network in the future because we anticipate a greater level of service satisfaction. The general impact of this network effect is to intensify competition because consumers are more valuable inso far as the network effect turns them into prisoners of the firm i.e., they are unlikely to leave a large network to a smaller one.

We build on Hotelling's model of competition seen in §5.2.2. The specific addition is that people display an additional WTP μ per fraction of the entire population present in their network. The utility from staying with firm A in the second period is $u_x(p_A, q_A) = v + \mu q_A - tx - p_A$ where q_A is the first period market share of firm A; a symmetric formula holds for firm B. We assume full market coverage, so that $q_A + q_B = 1$. Working out the indifferent consumer yields demand $D_A = \frac{p_B - p_A + t + \mu_A}{2t}$ where $\mu_A = \mu(2q_A - 1)$. The equilibrium is then $\bar{p}_A = t + \frac{\mu_A}{3}$, as in the standard asymmetric setting and profit is $\frac{1}{2t}\bar{p}_A^2$, after simplifications.

The full profit over both period, as a function of first period price p_A is then $\pi_A = p_A q_A + \frac{1}{2t}\bar{p}_A^2$. The FOC is $0 = q_A - \frac{1}{2t}\left(p_A + \frac{2\mu}{3}(1 + \frac{\mu_A}{3t})\right)$ which simplifies into $\frac{1}{2} = \frac{1}{2t}(p + \frac{2\mu}{3})$ at the symmetric equilibrium where $q_A = \frac{1}{2}$. The equilibrium price is then $p^* = t - \frac{2\mu}{3}$ which is smaller than in the standard case due to the exacerbated competition caused by the value of catching valuable consumer first.

24.3 Critical Mass

Positive network externalities create increasing returns to scale for firms which often give rise to extreme market structure where a single standard remains. It is thus impor-

tant for each contender to be the first one to reach the critical mass, if not to dominate, at least to survive.

24.3.1 Pros and Cons of joining a Standard

In the late 1990s two technical hardware specifications for the DVD were competing to establish themselves as a standard, the single layer and the double layer technologies. A few years later, once this was settled, there was another battle among two technical software specifications for rewritable DVD, the so-called "DVD+R" and "DVD+RW". Most manufacturers of consumer electronics and computers choose to support only one standard. In 2004, the battle rages over the next generation DVD between the so-called "blu-ray" and "hd-dvd" formats.

The decision to join a given alliance is guided by three effects, private cost, social cost and competition. On the one hand one wishes to adhere to the standard best suited to one's own technology but on the other hand it is important to embark on the largest alliance because it is more likely to produce learning gains and positive feedback. Yet being more numerous to share the same standard is a guarantee for tougher market competition in the future. It is therefore not trivial to assess the right decision. The following model tries to give hints.

Assume that the n potential adopter firms are differentiated by their innate preference $\theta_i \in [0;1]$ among the two standards called A and B. The network effect is captured as follows, let n_A and n_B be the number of member of each alliance then the marginal cost is $c_i = c - n_A(1-\theta_i)$ for a member i of alliance A and $c_j = c - n_B\theta_j$ for a member j of alliance B.

The decision to leave one alliance to join another affects one's marginal cost in two ways, directly through the innate proximity to standards and indirectly through the network effect. Recall now the n firm Oligopoly Cournot model studied in §5.1.3 and in particular eq. 5.14 for asymmetric constant marginal cost c_i. We found that the equilibrium individual sales were

$$q_i^* = \frac{a - b\left(nc_i - \sum_{j\neq i} c_j\right)}{n+1}$$

Since the equilibrium profit of firm i is proportional to $(q_i^*)^2$, what matters for a member of alliance A is to minimize $nc_i - \sum_{l\neq i} c_l$ or equivalently to maximize

$$n_A(1-\theta_i) - \frac{n_A\sum_{k\in A}^{k\neq i}(1-\theta_k) + n_B\sum_{m\in B}\theta_m}{n}$$

In this formula, $n_A(1-\theta_i)$ is the product of the private technology parameter by the size of one's actual network while the remaining term is the negative competition effect induced by the network affiliation of other firms.

Likewise a member j of alliance B maximizes

$$n_B\theta_j - \frac{n_A\sum_{k\in A}(1-\theta_k) + n_B\sum_{m\in B}^{k\neq j}\theta_m}{n}$$

hence the decision for i to leave alliance A to join B is profitable if

$$\begin{aligned} & nn_A(1-\theta_i) - n_A\textstyle\sum_{k\in A,k\neq i}(1-\theta_k) - n_B\sum_{m\in B}\theta_m \\ < \quad & n(n_B+1)\theta_i - (n_A-1)\textstyle\sum_{k\in A,k\neq i}(1-\theta_k) - (n_B+1)\sum_{m\in B,m\neq i}\theta_m \end{aligned}$$

or

$$nn_A(1-\theta_i) - n(n-n_A+1)\theta_i < \sum_{k\in A,k\neq i}(1-\theta_k) - \sum_{m\in B,m\neq i}\theta_m = n_A - 1 - n\bar{\theta} + \theta_i$$

where $\bar{\theta} \equiv \frac{1}{n}\sum_l \theta_l$ is the mean of the θ distribution. Hence, rearranging, firm i remains in alliance A if $\theta_i \leq \frac{n_A(n-1)+1+n\bar{\theta}}{n^2+n+1}$. It is quite immediate to see that low θ's all belong to alliance A while large θ's belong to alliance B hence the last member of alliance A is the largest integer i satisfying

$$\theta_i \leq \frac{i(n-1)+1+n\bar{\theta}}{n^2+n+1}$$

which shows that the dispersion of all firms along the axis of innate preference among standards determines the final size of alliances. If for example the θ's are evenly distributed with $\theta_i = \frac{i}{n+1}$ then the size of alliance A is $\frac{n+1}{2}$ which is approximately half of the industry.

24.3.2 Consumption Externality

We adapt Economides and Himmelberg (1995)[9@] to show how the presence of a network externality in consumption changes the usual conclusions of the neoclassical theory.

Introduction

The most striking feature is that there is no small size market in equilibrium, a critical mass of consumer must subscribe to the service for its existence to be sustainable. Furthermore this property is mostly independent of the market structure. We shall see that the network externality introduces discontinuities or phenomena that might recall chaos theory in the sense that a small change of an underlying parameter triggers a strong and far reaching reaction.

We model the telephone service using bits of models seen for horizontal and vertical differentiation. The willingness to pay for a phone call is Sx where S is the total quality of the service and $x \in [0;1]$ captures the heterogeneity of consumers. Total quality is made up of an intrinsic component $s \geq 0$ and a network component, the expected subscribers base b. For a price p, the last subscriber will be $\tilde{x}$ such that $(s+b)\tilde{x} = p$ so that the aggregate demand is $q = 1 - \tilde{x}$. We can therefore express the price as a function $P(q,b) = (s+b)(1-q)$ of both the actual demand n and the expected one b with $P_q = -s-b$ (traditional negative price effect) and $P_b = 1-q$ (positive network effect).

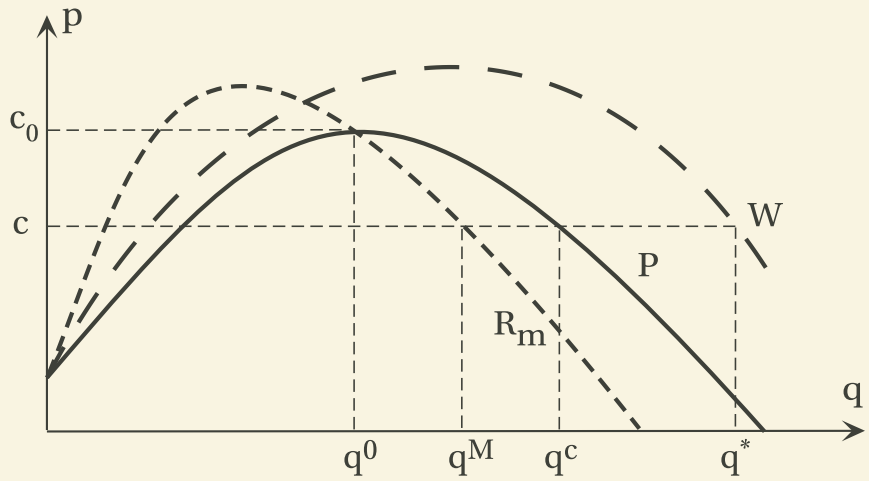


Figure 24.3: Network Effect

For normal goods, demand always slopes downward but for network goods, the willingness to pay for the last unit increases as the number expected to be sold increases. If expected sales rises with actual sales, then the willingness to pay for the last unit may increase with the number of units sold.To see this we use the equilibrium condition stating that expectations are to be fulfilled ($b = q$); the demand q thus solves $P(q,q) = p$ hence $D(p) = \frac{1}{2}\left(1 - s + \sqrt{s(s+2)+1-4p}\right)$ as plotted on Figure 24.3 (the smaller root is unstable: a small increase in customers raises the utility of a few more people who then subscribe and convince even more people to do so). As can be observed from the figure, the market is either void (large price) or reach at least the critical mass q^0 that maximizes $P(q,q)$.[10@] The impossibility to get a small market coverage stems from the vicious circle according to which many consumers are not interested in subscribing because the installed base is too small, and the installed base is too small because an insufficiently small number of consumers have subscribed.

Equilibrium Demand

In a perfectly competitive market the equilibrium is characterized by $p = c$, hence $q^{pc} = D(c)$.

Turning to efficiency we observe that if b units are sold, the net consumer surplus of unit $q \leq b$ is $P(q,b) - c$ thus the total surplus of consuming b units is $W(b) = \int_0^b \big(P(q,b) - c\big)\,\mathrm{d}q$. For the same reason as before, the efficient subscriber base is the largest root[11@] q^* of the equation $W' = 0$ and since $W'(b) = P - c + \int_0^b P_b(q,b)\,\mathrm{d}q > P - c$ we deduce that $q^* > q^{pc}$ i.e., the first theorem of the welfare does not hold because competing firms do not account for the positive externality in their profit maximization.

We now inquire how the exercise of market power is affected by the presence of the network externality. As seen on Figure 24.3, the marginal revenue $P + q\frac{dP}{dq} = P + qP_q + qP_b$ is greater than P for small sales, equal for sales of exactly q^0 and lesser when the demand has the traditional shape. The optimal price is nevertheless found without ambiguity as the largest root [12@] n^M of $P + q\frac{dP}{dq} = c$ because the smaller root corresponds to a deficit ($p < c$). By the same token we see that if $c > c^0$, the monopoly does not supply this market because it would lead to losses; otherwise it supplies $q^M < q^{pc}$ by quoting $P(q^M, q^M)$.

Telecommunication

Phone, SMS, email are obvious examples of services with a high positive network externality. For each of them the market can be divided among high and low valuation consumers (e.g., firms and households)

Consider an equal number n of both types; the individual surplus is $s_i B - p$ where B is the consumer base, p the price of the service and s_i the willingness to pay ($s_h > s_l$). Assume $2s_l < s_h$, then

- if $p > s_h n$, then nobody connects.
- if $p \in [2s_l n; s_h n]$, then only firms get connected because the network effect is too small (even if all households connect the price remains too large).
- if $p < 2S_l n$, everybody connects to the service.

The ratio p/s_h can therefore be interpreted as the *critical mass* of firms that need be convinced of the service values in order not to be deceived later on. If the ratio s_h/s_l is large then a non regulated monopoly would choose not to serve households; this may explain why PTT is either a public service or a regulated activity in most countries.

24.3.3 Big Push

In this section, pecuniary externalities as opposed to purely technological ones[13@] are shown to matter for economic development.

Context Ever since the industrial revolution, advanced economies have followed a similar developmental path whereby most sectors industrialize simultaneously i.e., massive investments into large scale production take place within a decade. Such a move is generally unprofitable for a single firm operating in a backward economy because fixed cost are too high and/or the output market is too small given the low wealth level. However, if all sectors develop at the same time, typically in large cities, fixed cost items become cheaper because of closeness. Furthermore, industrial salaries increase (due to scarcity or union pressure) giving rise to a middle class that can afford the new products. This demand pull, in turn, rationalizes for each firm the decision to invest into large scale production. It thus appear that the same economy can reach equilibrium with or without industrialization because the latter requires coordinated move to become profitable. Murphy et al. (1989) formalize this "big push" theory of industrialization.

Model The supply side of the economy features an obsolete technology (e.g., agriculture, handicraft) transforming labour into output on a one-to-one basis (after normalizing units) and a modern technology (e.g., industry, services) with marginal cost $c < 1$ (i.e., labour productivity $\frac{1}{c} > 1$) but only if the fixed cost F (measured in labour units) is sunk. The choice for entrepreneurs is thus between constant and increasing returns to scale. Potential Bertrand competition among modern entrepreneurs leaves room only for a monopoly.[14@] At the other extreme, the obsolete technology can be operated at any scale by anyone so that this sector forms a competitive fringe pricing at marginal cost. Taking the labour unit as money, wage is unitary. An obsolete firm earns zero profit while a modern firm producing q earns $\pi = q(p-c) - F$. An entrepreneur is willing to create a modern business if he can get enough demand to cover his fixed cost out of his producer surplus; formally $\pi > 0 \Leftrightarrow q > \bar{q} \equiv \frac{F}{p-c}$, the minimum efficient scale.

There is a continuum (size L) of consumers/savers who can either work for the modern industry (inelastic supply) or operate the obsolete technology. Their aggregate income m is total labor earnings L (at unitary wage) plus aggregate firm profits (over the two technologies). The driver of the model is the assumption of a unit continuum of product varieties upon which consumers hold identical Cobb-Douglas preferences. When charging $p < 1$, the industrial monopolist of a given variety has demand $\frac{m}{p}$ hence variable profit $m(1 - \frac{c}{p})$ so that he wishes to increase his price. However, the competitive fringe forces him to adopt the unitary price. Hence, whether a sector (variety) industrializes or

not, the equilibrium price is unitary. This in turn means that the aggregate income m is equally shared among the varieties i.e., each receives demand $q = m$.[15@] If a fraction λ of entrepreneurs develops the modern technology, aggregate income is $m = L + \lambda\pi$. Developing profit as a function of output, we obtain $q = m = \frac{L-\lambda F}{1-\lambda(1-c)}$ and the entry condition becomes independent of λ as it reads $q > \bar{q} \Leftrightarrow \frac{L-\lambda F}{1-\lambda(1-c)} > \frac{F}{1-c} \Leftrightarrow F < (1-c)L$.

This means that the (unique) equilibrium is either full industrialization or none. The outcome thus depends on the fixed cost F or the market size (proxied by L) as well as whether the innovation is drastic (low c).

Multiple equilibria Multiple equilibria appear only if the industry wage is exogenously maintained above the rural one at $w > 1$ (as we show hereafter). This premium may be a compensation for disutility of work, moving to town or the result of union pressure made possible by the life in the city. Alternatively, once everyone lives in the city, the fixed cost of establishing a modern firm is diminished because all required services are closer. At the other extreme, innovating in the country side is difficult (high fixed cost) because specialized talents do not reside on site but have to be brought in.

No development remains be an equilibrium if $(1-c)L < wF$ i.e., if the fixed labour requirement, paid at city wages, is too expensive. Imagine now that everyone adopts the innovation, then the entire labour force gets paid w as if there was no city premium ($w = 1$); this is akin to lowering the fixed cost. This opens the door for unanimous adoption to be an equilibrium. More precisely, if everyone works at a modern firm, output is $q = \frac{L-F}{c}$, thus profit is $\pi = \frac{L-F}{c}(1-cw) - wF = \frac{L-F}{c} - wL = \frac{(1-cw)L-F}{c}$. Multiple equilibria occur for $\frac{1-c}{w} < \frac{F}{L} < 1-cw$ which is possible as soon as $1 < w < \frac{1+\sqrt{1-4c+4c^2}}{2c}$.

Since two qualitatively different equilibria are possible, there is a role for government to try to force the better one by building expectations or jump start industrialization in all sectors aka the “big push”.

Infrastructure The previous intellectual construct can be applied to infrastructure building which has been recognized as a key ingredient to economic development.

The economy enjoys two IRS sectors of similar size, say service with a low fixed cost and industry with a high fixed cost so that firms in the first sector generate twice more surplus. Assume then that aggregate surpluses of 4 and 2 (say bn$) are generated only if an infrastructure costing 5 is build. Industrialization of the service sector alone and appropriation of its surplus via discriminatory charging for the use of the infrastructure is not enough for financing ($4 < 5$) although construction would be efficient since total surplus is $6 > 4$. If both sectors industrialize but the infrastructure operator cannot price discriminate among them (for whatever reason), it will raise only[16@] $2+2 = 4$, go bankrupt

and interrupt service; it may even anticipate this outcome and refuse to build (especially if he is risk averse). At a further round of anticipation, industrialists from both sectors may decide not to industrialize. There is thus a role for the country's business association or a large bank or even the government to try to design a financing scheme for the infrastructure that guarantees its viability. Otherwise, the country is kept in a so-called "poverty trap".[17@] Note finally the futility of building a magnificent harbor if local businesses cannot engage into exports. Thus, authorities must also provide cheap finance for potential users of the infrastructure to help (convince) them to switch to the modern technology. This last issue is the complementarity of the inputs and services needed to allow an industry to develop.

Banks The history of continental Europe saw large banks playing a coordinating role in financing industrialization. Da Rin and Hellmann (2002) rationalize this fact by showing how a bank with large market power may force the big push even when all actors hold pessimistic beliefs about the wave of industrialization.

Assume to simplify that the profitability of a single firm is an increasing function π of the share of firms having industrialized and that $\pi(0) < 1 + r < \pi(1)$ holds where r is the marginal cost of money (rate of refinance on international money markets) for the competitive banking fringe i.e., if no one industrializes it is not possible to finance the first one but if all do then all upgrades can be financed. There exists a large bank with a cost advantage (refinancing rate $r_0 < r$) able to force unanimous adoption by offering extremely attractive financing conditions to early adopters. If there are n firms (ordered according to some accounting screening), firm $\#i$ is offered debt at rate $\min\{r, r_i\}$ where $1 + r_i = \pi\left(\frac{i-1}{n}\right)$ (the competitive fringe must be kept at bay). If the belief is that only the first $i-1$ firms adopt then they will indeed adopt since their loan is profitable but crucially, firm i will also adopt since her rate is designed so as to make her indifferent (or happy if her rate is r). This proves that only 100% adoption is a coherent (equilibrium) belief. Now, this peculiar strategy is profitable for the large bank if she can recover the losses from early adopters out of the profits she makes on later ones where her margin is $r - r_0$. The feasibility condition is thus dependent on this margin and her protected market share.

We have here an instance where an oligopolistic structure is dynamically efficient because it alone can avoid the pessimism that would keep the economy in a poverty trap. The key is that big banks accept losses early on because they can recoup them later on thanks to their maintained market power.

24.3.4 Rent-Seeking

In line with the previous model, Murphy et al. (1993) show that the potential existence of rent-seeking (cf. §7.2.1) creates indeterminateness i.e., an economy may reach either an efficient or inefficient equilibrium. If history has brought the economy in the bad equilibrium, a revolution rather than an evolution is necessary to break out and reach the good equilibrium.

Imagine that individuals can engage into agriculture (A), business (B) or rent-seeking (R) with potential returns α, β, γ. Compared to activity A, B is less local, more export oriented, more intensive in capital (human and physical), more risky, operates at a greater scale and is potentially more profitable i.e., $\beta > \alpha$. However, all these characteristics make B an easier target for rent-seekers (compared to A). To simplify, we assume that only B can be plundered by an amount $\gamma < \beta$. Private rent-seeking will be theft or fraud while public rent-seeking will be corruption for permits and licenses that business needs. In line with the paradox of exploitation (cf. §7.2.5), rent-seekers are assumed to secure a greater income than producers of basic staples, thus $\gamma > \alpha$.

Let n_i for $i \in \{A, B, R\}$ denotes the share of the population in each occupation. When $n_R = 0$, the population structure settles at (n_B^*, n_A^*), a distribution obeying complex cultural and historical relationships not modeled here. Rich businessmen and poorer peasants earn respectively β and α per capita. As peasants start to switch to rent-seeking, each earns the greater γ. Given a distribution of the population among activities, the total income to activity B is $\beta n_B - \gamma n_R$, thus the average income is $\pi_B = \beta - x\gamma$ where $x \equiv \frac{n_R}{n_B}$ is the ratio of rent-seekers to businessmen (our main variable of analysis). Meanwhile $\pi_B > \alpha$, more peasants flock into rent-seeking but for $x \geq \hat{x} \equiv \frac{\beta - \alpha}{\gamma}$, each businessman can threaten to revert to agriculture if his rent-seeker does not leave him at least α (so to speak). This means that the pie available to rent-seekers is no more than $(\beta - \alpha) n_B$, so that their per capita income is $\pi_R = \frac{\beta - \alpha}{x}$, a decreasing hyperbole. Let $\bar{x} \equiv \frac{\beta - \alpha}{\alpha} > \hat{x}$ be the solution of $\pi_R = \alpha$. Figure 24.4 graphs these per capita profits as a function of x; intersections between π_B and π_R occur at $\underline{x}$ and $\bar{x}$.

There are two equilibria where no one desires to change occupation. One involves no rent-seeking. Although each peasant would like to become a rent-seeker, it is cheap to deter such a behavior when the phenomenon is rare. This is an efficient equilibrium where wealth creation is maximum. The second and inefficient equilibrium is $\bar{x}$ where there is so much rent-seeking that everybody earns the low secure income α. Lastly, $\underline{x}$ is not an equilibrium because there is already a significant proportion of rent-seekers and any small increase creates a larger wedge in returns that triggers more entry into the rent-seeking activity.

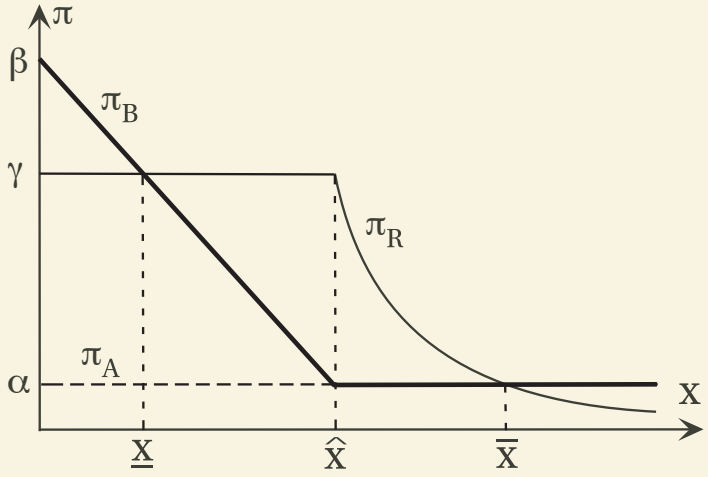


Figure 24.4: Multiple Rent-seeking Equilibria

24.3.5 Myths

Rule of the Road

In every country, the rule of the road compels driving on the same side of all roads. Contrary to popular belief, this choice is not a caprice of history but an optimal response to changing technological conditions. The driving factor is that most people are right handed and thus use their right arm to apply force or precision. The evolution of the devices they had to operate decided which side of the road was more convenient for travel.

As for walking, most people appear to have a natural tendency to keep to the right. During the middle ages, transportation used animals lead by the right hand of a walking person who was then on the left side of the cart. The driver would keep to the right of the road in order to position himself towards the center and avoid collisions with incoming traffic.[18@]

Later on, wagons and passenger coaches were driven by a single man holding a whip in his right hand and sitting or standing at the back of the cart; he would sit on the right side to avoid hurting passengers with the whip. Having changed side of the cart, he needed to change side of the road and travel on the left in order to remain close to the center.

At the same time, large distance transportation developed in the US and continental Europe and it made sense to use large wagons and stagecoach pulled by many horses where the driver would ride on a rear horse. Since the driver keeps the whip in the right hand, he must ride a left horse to use it properly and is thus lead to travel on the right side to remain close to the center of the road. This new custom was enforced

on turnpikes (toll roads) and quickly became widespread in the entire US. The story for continental Europe follows a different kind of rationale. It is believed that Napoleon enforced walking on the right side for his soldiers because the widely used musket (riffle) was better aimed towards the left (as horsemen do with their lance in jousting).[19@] This new walking custom, agreeing with France's custom for wagons, was then imposed onto conquered countries i.e., most of continental Europe.

Lately, some (surprisingly?) English authors have claimed that driving a car on the left side of the road provides with a greater level of security due to some asymmetries in the brain. Even if this advantage was noticeable, it would become increasingly irrelevant as more traffic takes place on protected highways. This illustrates the fact that the current equilibrium may not follow what the latest science might call for but the ups-and-downs of historical adaptations to seemingly unrelated technological changes.

Video Cassette Recorder

As recalled by Spulber (2002), the battle between Sony and JVC overs VCRs in the 1980s was a competition for market dominance where both tried to impose its standard (*Beta* and *VHS* respectively). This story is often presented as a path-dependent market failure where the bad standard, *VHS*, won (quite inexplicably) over the good one, *Beta*. The true story is a bit different.

VCRs was invented by the US firm Ampex in 1956 and was marketed to professionals during the 1960s. Ampex licenced its patents to several japanese firms, among them Sony and Matsushita, in order to penetrate this market but all VCR models were incompatible. In 1970 no less than five firms tried repeatedly to launch a VCR for home use. All failed but Matsushita learn that the playing time of its VCR was deemed too short by consumers. In 1975, Sony launched its *Betamax* with 1 hour playing time using compact tapes. In 1976, JVC, a subsidiary of Matsushita, introduced a similar machine with bigger tapes that allowed 2 hour playing time but less possibilities for editing, a key feature in the professional market. When Sony launched the *Beta II* with a 2 hours recording capability for the US market in 1977, it took only 6 weeks for JVC to launch a 4 hour recording VCR. The latter feature proved crucial as *VHS* quickly outsold *Beta*.

During the 1980s, both competitors made alliances to increase the market share of their standard in a very close timing. They both improved their technology and lowered their prices but the *VHS* sales advantage never vanished so that the installed base of *VHS* consumers grew much bigger than the *Beta*. This fight repeated in every single market with the same outcome, *VHS* outselling *Beta*. Still *Beta*, by permitting editing, remained the de-facto standard of the professional broadcasting industry. This may be why it had a supposedly better picture quality but this was never confirmed, if not

contradicted, by consumers reports. With respect to the alleged market failure we may conclude that there was no bad standard but two standards adapted to two markets, the professional and the individual ones.

If the payoffs for compatibility had been higher then both would have agreed on one of the standard but this may have generated a high welfare loss. Indeed, the loser of the battle, Sony, has never lost money with *Beta* thanks to its leadership of the professional market; this has permitted and encouraged Sony to enter into a Bertrand competition against JVC. Both products have improved and their prices have fallen at a pace not observed for a protected good like Polaroid instantaneous photography.

Microsoft

Without judging upon the alleged monopolistic behavior of this company in the Internet era, the success of its spreadsheet and word processor deserve to be told. Before proceeding we must make a detour to the Apple world to understand Liebowitz and Margolis (1999)'s claim: Microsoft gambled on the graphical user interface (GUI) pioneered by Apple, spending time and money to learn well this technology, to later on launch the software that made its fortune.

Back in 1984, Apple introduced the "Macintosh" operating system (OS); although nicer and easier to use than text-based OSes like DOS it was also more expensive (more costly components like memory and video display), exclusive (no licensing to competitors) and much slower since the graphical interface used a significant part of the processor resources (all processors were and still are roughly equally powerful). Business users, by far the largest market, love to crunch numbers and have always been repelled by the slowness of the "Mac" OS. This is the key explanation put forward by analysts to justify that this product never caught a large market share. At the same time, the graphical capabilities of the "Mac" are at the source of its success in graphical intensive industries (a similar dichotomy exists in the *Beta* vs. *VHS* case).

We can now get back to Microsoft who like its competitors of the 1980s produced software not only for its own best selling "DOS" but also for the Macintosh. The "Excel" spreadsheet was introduced for the Mac (1985) before being ported to the "Windows" OS (1987). Likewise "MS Word" was launched for "DOS", then for the "Mac" and later for "Windows".

The spreadsheet leader in the 1980s was "Lotus 1-2-3" who it self had outperformed the first historical leader in just one year thanks to its technological advances. In 1987, Borland launched "Quattro" and MS ported "Excel" to the IBM PC world with the then flaw that it needed the "Windows" OS to operate. The newer products were systematically judged better than the leader, "Excel" being ranked first most of the times although

it initially suffered from a sluggish responsiveness. Both market shares stayed between 10 and 20% from 1988 until 1990 while Lotus managed to retain more than 50%. Then with the improvement of hardware and the "Windows" OS, the superiority of "Excel" became clear given that Lotus never successfully ported "1-2-3" to "Windows" (according to the professional reviewers). The economic results followed since MS's market share steadily gain to reach 70% by 1995.

Liebowitz and Margolis (1999) reason that Lotus's technological failure to match "Excel" capabilities is due to its immaturity with GUIs, more specifically because they never dare spend a $ into the Macintosh world. The latter attitude makes sense from a short term point of view since the Apple market was small thus not very profitable. This is why one may think of the Microsoft association with the Apple world as an investment for the future.[20@] In fact Lotus followed a similar path by investing into the development of a revolutionary spreadsheet for a very advanced OS, "NeXT". The fate of Lotus came from the complete failure of this OS to break up in the mass market, in due time. In a twist of history, NeXT happens to be the ancestor of the "Mac OS X" system that made the fortune of Apple.

The market for word processors was fairly competitive back in 1986 with "WordStar" and "WordPerfect" on top. The latter subsequently gained market share to reach 50% by 1990 and was recognized of highest quality in its "DOS" version by readers of computer magazines; however "WordPerfect" did a poor transition to "Windows" as attested by worse reviews. On the contrary Microsoft succeeded to increase its already good ratings for "DOS-Word" to top-of-the-class for its "WIN-Word". Adding the fact that "WordPerfect" came into the "Windows" world 2 years after its Microsoft competitor, there is no surprise to see that the MS product gained a steady 10% of market share every year from its introduction in 1990 until 1998.

The analysis of prices is also a good indicator of the competition that took place in these two markets in the 10 years period 1988-1997. For spreadsheets, the price of "1-2-3" dropped from 300$ to 20$, that of "Excel" dropped from 250$ to 60$ although they were quite similar most of the time. "Quattro" was a better bargain, starting at 50$, reaching 100$ before dropping to 20$. As for word processors, prices started around 160$ and slightly rose until the introduction of "Windows 3" in 1990; thereafter a strong decline occurred. Microsoft priced "WIN-Word" like "Excel" from 250$ at its beginning in 1990 down to 60$ in 1997. Competitors that were once more expensive ended up at a bargain price. The success of Microsoft resided in the *bundling* of its two star products in the "Office" suite sold at an attractive price (less than the combination of independent software whether from MS or not).

We cannot conclude this chapter without a word on the famous QWERTY contro-

versy. This keyboard was developed to prevent the hammers on early mechanical typewriters from jamming; the letters configuration has carried over into the present day even though the initial reason for its arrangement no longer applies. It has been alleged by many economists that the newer Dvorak Keyboard was superior and so that the lock-in of millions of users into the old standard was a clear example of inefficient path dependency. Liebowitz and Margolis (1990) demonstrate the fallacy in full detail; one detail deserves special mention: the only evidence of the Dvorak superiority are studies undertaken by the inventor Dvorak himself!

24.4 Two sided platforms

To be successful, a platform like a game console must offer many games to attract gamers but it also has to be popular among gamers to attract game developers; the owner must therefore address the celebrated "chicken-and-egg" problem to get "both sides of the market on board". Rochet and Tirole (2003) argue that most markets with network externalities display such a two-sided feature and study the business model of the platform owner, that is how he/she courts each side while making money overall.

24.4.1 Game consoles

Rochet and Tirole (2003) report several failures of highly innovative platforms who provided a better game experience to users but lacked good games to play with; the first generation XBox disappointing results are generally attributed to the limited number of available games. Strangely there are no reverse result.

In this sector, the console developer such as Atari, Nintendo, Sega, Sony or Microsoft charges game creators a two-part tariff. Firstly the development kit that enables testing a game on the console is sold for a fixed fee. Later on, royalties are paid to the console manufacturer over the sales of games. The platform itself is often sold with a loss to end-users, it is called a loss-leader.[21@] For the launch of its XBox, Microsoft tried to appeal to game creators by offering an inexpensive development kit running on (cheap) PCs to free game developers of working on (expensive) workstations.

Similar to game consoles, softwares designed to display web content (e.g., browser, media player, text viewer) are free of charge for consumers while the servers that generate the content are sold to creative firms. Likewise most of the media make money on advertising and subsidize the price of the end-user product be it newspapers or TV.

24.4.2 Portable Devices

Interestingly, operating systems (OS) for the personal computer (PC) and handheld devices have adopted the opposite business model probably motivated by the success of Microsoft in applying it. An OS developer derives profits from consumers through the royalties hardware makers pay to get their machines running the OS (or directly as in the case of Palm and Apple). The loss for the OS developer comes from the obligation of making crucial code freely available years in advance to foster the development of applications running on the OS and make the package attractive to potential clients.

Apple has successfully build a two sided platform for portable music. Late 2001, this company launched the iPod, a portable music player packing thousands of songs into a tiny and nicely designed shell. Despite being much more expensive than its competitors, this product succeeded to conquer a large market share by the end of 2002 thanks to its offer of useful features like compatibility with "MS windows", synchronized agenda (typical of digital organizers) and above all a very easy management of songs thanks to the integration with the iTunes software. The next move by Apple in May 2003 was to turn this software into an internet music store, bursting an instant success crowned by Time Magazine as "invention of the year". The reason behind the success is that the flexible Digital Rights Management (DRM) system negotiated with the major record labels enables iPod users to enjoy their legally bought music on the move. What older online music stores were missing was the integration of the software with the hardware that provides users with greater satisfaction, thus greater willingness to pay.

Yet, Apple has just recognized that iTunes Music Store was a loss-leader due to the high level of royalties paid for DRM. There are two possible explanations; on the one hand, the service might become profitable in the future for the possible presence of scale economies and on the other hand, the software activity (iTunes) might just be a boost to the profitable hardware sales (iPod). Sales figures as well as the market value of Apple from late 2005 seem to point at the latter explanation.

24.5 Social interaction

Veblen (1899) and more recently Leibenstein (1955) and Becker (1974) have argued forcefully that consumption decisions are made to satisfy both material and social needs: the pleasure derived by consuming is affected by the consumption choice of other consumers. Vanity and Conformity (aka) are two complementary explanations for such needs.

In the same vain, Schelling (1969) analyzes location choices and shows that small perturbations such as the arrival of people of different characteristics can trigger moves

that propagate and radically change the entire distribution up to the point where complete segregation results (neighborhoods only contain homogeneous people).

24.5.1 Status seeking

Conformity and Vanity

To maintain cohesion and defend itself against aggressions every society has formed habits and customs together with the censure of nonconformist attitudes. As a result, individuals imitate each other because they feel the desire to avoid social ostracism. In this case, social interaction leads to *conformity*. Products like garments or beverages are very conformist especially in the population that most need social recognition, the youth. Another consequence of social conformity is the "winner takes it all" effect observed in entertainments, sports and the arts where the number one in sales or audience gets much more than its followers because he (she) is the object of a collective passion.

On the other hand, human beings have individualistic values and look for prestige or social recognition. If they can afford it, they try to signal their idiosyncrasies by their consumption of positional goods that are better observable than their income (and not subject to taxation). Examples of luxury goods are perfumes, sport cars, “haute couture” or membership of selective clubs. Social interaction now leads to *vanity*. Veblen held the view that the lower classes were not out to overthrow the upper class as Marx believed but, rather, strived to climb up to it. It is quite straightforward to see that the desire to surpass the neighbor or least to keep up with the average leads to excessive consumption.[22@] Social status can be so important relative to intrinsic utility that individuals are rationally lead to adopt conformism as a standard of behavior, despite heterogeneous underlying preferences.

Impact on firm behavior

We study Grilo et al. (2001)’s adaptation of Hotelling’s spatial model explaining firm’ strategic pricing behavior when consumer preferences exhibit either conformity or vanity.

As in §11.1.2, we consider consumers uniformly distributed within the $[0;1]$ segment. Stores A and B are located at distance x_A and x_B respectively; these figures can be both larger than 1 meaning that the farthest away may sells a product of inferior quality. The consumer with characteristic x, who buys from store $i = A, B$ derives utility

$$u_i(x, n_i) = \bar{p} - p_i - t\,(x - x_i)^2 + \alpha n_i$$

where αn_i is the externality affecting store i with market share n_i. Vanity is present for $\alpha < 0$ and conformity if $\alpha > 0$. The model yields interesting result only in the presence of an asymmetry among the two firms. Let us then assume that firm A has a locational advantage with $x_A + x_B > 1$.

The equilibrium is characterized by the indifferent consumer $\hat{x}$ for which $u_A(\hat{x}, n_A) = u_B(\hat{x}, n_B)$ and two additional market clearing equations $n_A = \hat{x}n$ and $n_B = (1-\hat{x})n$ stating that the expectation n_i for store i corresponds to its actual demand, the product of market share by market size. Solving this system, we obtain

$$\begin{aligned} p_A + t(x - x_A)^2 + \alpha x n &= p_B + t(x - x_B)^2 + \alpha(1-x)n \\ \Rightarrow 2tx(x_B - x_A) - 2\alpha x n &= p_B - p_A + t\left(x_B^2 - x_A^2\right) + \alpha x n \\ \Rightarrow \qquad \hat{x} &= \frac{p_B - p_A + t\left(x_B^2 - x_A^2\right) - \alpha n}{2\left(t(x_B - x_A) - \alpha n\right)} \end{aligned}$$

$$D_A = \hat{x} \propto p_B - p_A + t\left(x_B^2 - x_A^2\right) - \alpha n \tag{24.1}$$

$$D_B = 1 - \hat{x} \propto p_A - p_B - t(x_B - x_A)(2 - x_A - x_B) - \alpha n \tag{24.2}$$

To find out the price equilibrium, observe when there is no network effect i.e., $\alpha = 0$, the threshold $\hat{x}$ is identical to $\tilde{x}$ derived in eq. (11.2) of §11.1.2. Up to a multiplicative constant, the denominator, demand is $D_A = \tilde{x} - \alpha n$, thus when maximizing the profit $\Pi_A = (p_A - c)(\tilde{x} - \alpha n)$, we obtain the best reply $\tilde{p}_A(p_B) - \alpha n/2$ where $\tilde{p}_A$ is the no-network best reply function (eq. 11.3). It is as if the marginal cost c was inflated by αn. The equilibrium is thus the standard Hotelling solution (11.5-11.8) minus the externality computed for the entire market.[23@]

We can compute the equilibrium market share with

$$\hat{x} = \frac{t(x_B - x_A)(2 + x_A + x_B) - 3\alpha n}{6\left(t(x_B - x_A) - \alpha n\right)}$$

It is readily seen that if stores are symmetrically located with $x_A + x_B = 1$ then $\hat{x} = 1/2$ and is therefore unresponsive to the strength of the externality or the market size. If store A has a location advantage then $\hat{x} > 1/2$ increases with αn.

> Vanity yields higher market prices while weak conformity generates lower prices. As the market expands or as the externality grows stronger, the firm with a location advantage loses market share under vanity but gains under weak conformity.

24.5.2 Rationing

Leisure Services

No one likes to dine, drink, dance, listen, view or patronize any other social activity in an empty room; the larger the audience, we more we appreciate the act. Among reasons explaining this behavior, we get confirmation that our choice was good if others do the same and we appreciate to be seen or being able to claim presence at a "successful" social event. Social events therefore incorporate a positive network externality.

Because of this positive externality, we tend to express demand at times where we believe others are likely to do the same, hence strong demand gets concentrated at specific moments. Patterns of modern life also tend to amplify this phenomena since the leisure time gets limited by the standardized work schedule. We shall thus speak of a short lasting peak demand and a long lasting off-peak demand. Examples of peak periods are noon and evening for restaurants, week-ends for cultural events (movie, theater, concert, sport) and high season for holiday trips.

Another characteristic of social events, albeit on the supply side, is the limited capacity to serve the clientele since attendance is roughly limited by the numbers of available seats. Allowing a greater number to attend generates congestion which is strongly resented by most participants; this is one of the main issue addressed in the next chapter. In a free market economy where prices tend to adjust to equate demand and supply, one would expect suppliers of social events to raise their prices at peak time so as to force a demand reduction down to their capacity (since this is profitable). Yet, prices tend to be sticky which often results in queuing or overcrowding.[24@]

A technical reason is that demand varies a lot between peak and off-peak times, thus prices would have to be constantly adjusted which is extremely costly to do in all the places where prices are displayed on boards and menus. Furthermore, in order to discover the market clearing price firms would have to perform auctions. This is what is actually taking place to resolve overbooking in planes:[25@] when there are more passengers waiting at the gate than available seats in the plane, an auction is conducted where a prize constantly growing is offered until enough people accept to board in a later plane or redeem their travel.

A cultural reason is that consumers of social events believe they deserve stable prices so that any attempt to implement time dependent pricing is seen as gouging and backfires into boycott.

Resale

Oddly enough, there are many events like concerts or sports where there exists a very active second hand market that no consumer resents. These markets are operated by brokers (legal entities) and scalpers (dodgy people on the streets) who are pure arbitrageurs; they buy tickets at regular price in advance to resale them close to event time at the price that clear the market (since at that point their expense is a sunk cost). Brokers and scalpers are local intermediaries who have an informational advantage over the promoter regarding the evolution of local demand as the time of the event approaches (i.e., whether the event will be "in" or "out"). As shown in Courty (2003), this enables them to track the market clearing price and out-perform the promoter if he were to compete with them in the late sales market. Two consequences emerge: first, there is a profitable late sales market which explain why arbitrageurs appear in the first place, second, the promoter is bound to lose the competition against them which explain why arbitrageurs remain.

Because promoters lose profitable late sales to resellers, they often lobby their lawmakers for legal protection through prohibition or limitation of ticket resale. Another route against arbitrage is to personalize the emission as done by airlines although this is resented by consumers who see it as an undue introspection into their private matters and creates a value damaging inflexibility (you can't give away your ticket to a friend in case you won't be able to attend).

Arbitrage is also reinforced by the underpricing of early tickets. Promoters do so for two reasons. The first motive is to generate the positive demand externality alluded to before. The second motive is risk management. As we saw in §19.3 on pricing under risk, a risky promoter tries to avoid loss making situations, here the case where tickets are overpriced and sell poorly.

Because prices are sticky, there are some episodes of rationing at peak time i.e., why we have to queue at restaurants, disco clubs, theaters, cinemas or stadiums.

Goods

Producers of toys, microprocessors, game consoles and other fashionable electronic appliances frequently resort to rationing when they launch a novel item whose qualities generates strong demand. Either a quantity known to be lesser than demand is brought to the market to force a "buyer frenzy" or sale is time limited to force "early purchase".[26@]

Raising the price when demand grows large is the efficient way to clear the market because it guarantees that the available units go to those who value them most. This adaptive behavior amount to price discriminate among peak and off-peak demand (cf.

§4 on price discrimination). By sticking to its original price and rationing part of her demand, a firm foregoes an immediate profit; some benefit must therefore accrue from the commitment to perform rationing and not raising prices.

According to Gilbert and Klemperer (2000), this gain is identified by looking at the consequences of rationing. Those who suffer from it are the high-WTP individuals facing the prospect of being rationed when they would be assured to consume if price discrimination was used.[27@] At the same time, low-WTP individuals that would never consume under price discrimination are sometimes lucky enough to get served under the rationing scheme. This last observation is at the root of rationing: the firm wants to give surplus to low-WTP individuals in times of peak demand to motivate them to become clients because this has a sunk cost.

Consider indeed the decision to build a computer around a microprocessor, to buy a new game console or get a culture. The future client must incur an irreversible cost to learn how to use the good or service he will later buy. He will sunk that money only if he perceives a high enough future surplus from his consumption of the good or service. Due to the nature of the good or service under consideration this client, like any other, expresses revolving high and low demands. If the firm price discriminates, someone with low intrinsic interest in the item will be barred from consuming at times of high demand, he will therefore derive surplus from consuming at times of low demand.

If the firm wants to lure that person into becoming her customer, she must compensate his sunk cost by lowering the off-peak price to generate enough surplus. This is a costly move because the discount is enjoyed by the entire customer base. Rationing can improve because it gives this marginal low-WTP individual the opportunity to consume at times where he most enjoys the item i.e., to derive a large surplus. The firm can then raise the off-peak price and avoid the aforementioned costly rebate.

Model: consumer $i \in \{H, L\}$ has a valuation v_i for the item with probability r_i and no interest for it otherwise. W.l.o.g., $v_H > v_L$; we then assume that $r_H > r_L$ i.e., high-WTP clients are more likely to be observed.

24.5.3 Group behavior

The previous models were micro foundations of how individual choices are influenced by their environment beyond the intrinsic qualities of the goods and services they consume.

Segregation

We present here Schelling (1969)'s model. Population is divided into n_i incumbents (citizens of the local state) and n_e entrants (immigrants) who share a preference for the city's

central neighborhood. The dichotomy could apply to group membership for men/women, toddlers/teens, classic/vanguard, black/whites, gay/straight, christian/muslim or any other cultural trait.

The next ingredient is to consider the heterogeneous acceptance of diversity i.e., each person subjectively tolerates a maximum mix of types (for the area); upon overshooting that limit, he/she moves to another neighborhood, presumably, where his/her type dominates. Let assume that the tolerance index is uniformly distributed in each subpopulation between 0 (outright racists) and λ (ultra-progressives).

In each population of size n, let us order people by decreasing tolerance and relabel each by his rank from 1 to n. The number of people with tolerance greater than t is $x = \frac{\lambda - t}{\lambda} n$, thus $t(x) = \lambda\,(1 - x/n)$ is the tolerance of individual x. If the most tolerant people in numbers x_i and x_e live in the neighborhood, the marginal incumbent is going to stay if his tolerance $t_i(x_i)$ is greater than the mix ratio x_e/x_i. If he stays, so do all other more tolerant incumbents. The stability condition is thus $x_e \le \lambda\left(x_i - x_i^2/n_i\right)$. By symmetry the stability condition for entrants is $x_i \le \lambda\left(x_e - x_e^2/n_e\right)$. Taken with equality, these two conditions give rise to tolerance curves.

The left panel of Figure 24.5 depicts them for equally sized populations ($n_i = n_e$) and limited overall forbearance ($\lambda = 2$) which means that the average person tolerates a one-to-one mix i.e., would agree to a 50/50 sharing of the neighborhood. Consider a pair (x_i, x_e), if it lies below the incumbent curve, the neighborhood is attractive to additional incumbents while if it lies above, the neighborhood is repulsive and forces some incumbents to move out. The same applies with respect to the entrant's tolerance curve; we draw arrows to indicate in each zone how the current numbers of incumbents and entrants changes when marginal people find out whether the neighborhood is attractive or repulsive to their group.

An equilibrium is found at the intersection of the curves (diamond) where the tolerant halves of each population live together but this outcome is unstable since any small deviation triggers a process of adaptation that ends with complete segregation at either dot on the axis. Whether the neighborhood ends up filled by incumbents or entrants depends on the respective speeds of change in each group and on the historical distribution.

Hope to see a mixed outcome is not lost however. The right panel of Figure 24.5 uses a greater approval for diversity ($\lambda = 5$) with the effect that three mixing outcomes are candidate equilibrium, one of which (middle star), is stable. Yet if the incumbent's population is twice the entrant's then the entrant curve shrinks down and the stable mixed outcome disappears, segregation follows again.

When the overall degree of approval for diversity is large enough, a mixed outcome can become a stable equilibrium, otherwise complete segregation takes place and is determined by historical conditions.

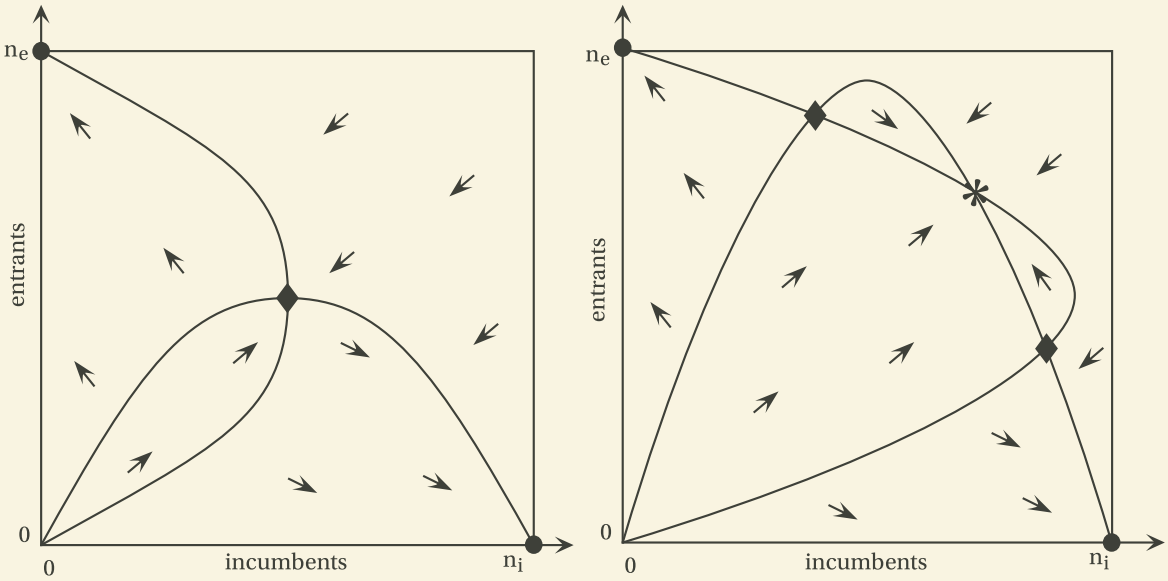


Figure 24.5: Location equilibrium

Rational rationing

Becker (1991) reports having observed that two nearby restaurants offering similar services get vastly different demands. One is packed and has a long waiting queue while the other one is almost empty. This pattern can persist for a long time and strikingly the successful owner does not raise his price as standard microeconomic theory would predict. Alike phenomenon occur for the price of best-sellers books or top sport events. Eating out satisfies consumer's hunger but also their social needs (being "in"). In that case, the social status of the restaurant is its popularity.

On Figure 24.6 we represent a traditional demand function, the marginal revenue function and the optimal price p^* together with the corresponding sales q^* that are lesser than the restaurant's capacity k. What the traditional demand theory neglects is that the individual demand depends positively on the expected attendance α at the restaurant.

The effective demand of the restaurant is a function $D(p,\alpha)$ of price p and expected attendance α. In equilibrium, attendance is equal to demand thus $D(p,\alpha)=\alpha$ which can be inverted into $p=P(\alpha)$. Differentiating, we obtain $P'D_p+D_\alpha=1\Rightarrow P'=\frac{1-D_\alpha}{D_p}$ which is positive if the externality is strong enough i.e., when $D_\alpha>1$ (recall that $D_p<0$ remains

true). The price can thus increase with the expected occupation but such a phenomenon cannot last forever so that the network effect D_α will fall back if the expected occupation is too large. The price function is then bell-shaped and reaches a maximum at p^* inducing a demand α^*.

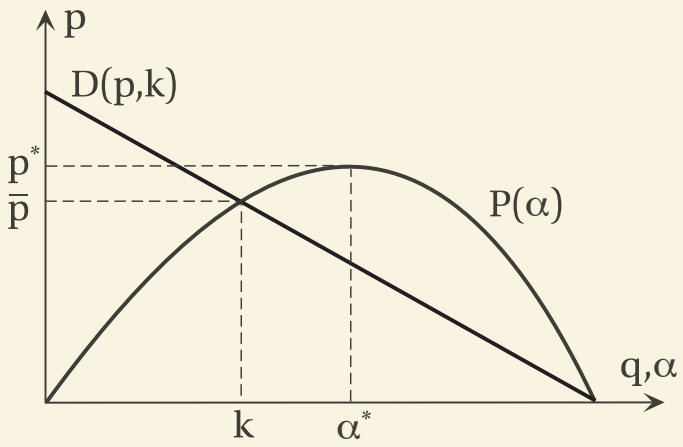


Figure 24.6: The Restaurant Puzzle

Let $\bar{p}$ be the market clearing price i.e., solving $D(\bar{p},k)=k$. The maximum price p^* generates queuing if $\alpha^*>k \Leftrightarrow P'>0 \Leftrightarrow D_\alpha(\bar{p},k)>1 \Leftrightarrow D(\bar{p},k+1)>D(\bar{p},k)+1$ which means that, at the market clearing price, the demand increases by more than one table each time a new table is added to the restaurant.

Assuming that this condition holds, the maximum price p^* is optimal. A lower price entails at best the same effective demand but a lower margin, it is thus a dominated choice. Any higher price generates a zero effective demand because potential customers revise their expectations downward and start to leave the queue in front of the restaurant (technically, for $p>p^*$ there is no α such that $D(p,\alpha)=\alpha$).

24.5.4 Social Cohesion

Introduction

In the business environments where cooperation is key to produce efficiently, the models of labor discipline tell us that a rent has to be left to employees to induce obedient behavior (cf. §20 on moral hazard). This view probably fits well the rather anonymous work relationship observed in advanced economies but stands in starck contrast with the business practices of ethnic communities whose members accept low wages for long periods without creating negative side effects on labor productivity or threatening their cohesion. Chinese, Jewish or Italian communities among others have always displayed

this communitarism in the many countries where they migrated. A simple model of social network externalities can shade light on the economic forces behind their social organization (beyond the obvious cultural reasons).

The idea is that paying a negative rent to workers enables the community firms to generate a greater cash flow so that new investments can be financed internally and rapidly in opposition to the lenghtier and dearer use of external funds. To sustain this apparently exploitative social organization, a carrot and a stick are necessary; the former is a siginificative probability of inheritating the business for people who have no initial rights on it while the second is ostracism for those who betray the social order.

Model †

Let us denote δ the discount factor, ω the wage inside a community firm, T the number of employees in a community firm and ρ the periodic return of the entrepreneur; the present value of being a community entrepreneur is thus $V_e = \frac{\rho - T\omega}{1-\delta}$. An obedient member of the community has a positive probability π of inheriting the business i.e., become entrepreneur later on.[28@] Free-riding on the community while being a member is possible given the delegation level and yields an opportunistic wage α greater than the outside wage $\bar{\omega}$; yet, since information flows perfectly in the community, the cheater is identified and fired which means his outside opportunities are reduced by an ostracism factor ϵ. We assume $\rho > \alpha > \omega$; the present value of cheating (c) is

$$V_c = \alpha + \delta V_o = \alpha + \delta \frac{1-\epsilon}{1-\delta}\bar{\omega} \tag{24.3}$$

while the present value of behaving honestly (h) is

$$V_h = \omega + \delta(1-\pi)V_h + \delta\pi V_e \tag{24.4}$$

thus

$$V_h = \frac{\omega + \delta\pi V_e}{1-\delta(1-\pi)} = \frac{\delta\pi\rho + (1-\delta-\delta\pi T)\omega}{(1-\delta)(1-\delta(1-\pi))}$$

is an increasing function of insider wage ω, thus

$$V_h \geq V_c \Leftrightarrow \omega \geq \underline{\omega} \equiv \frac{(1-\delta)(1-\delta(1-\pi))\left(\alpha + \delta\frac{1-\epsilon}{1-\delta}\bar{\omega}\right) - \delta\pi\rho}{1-\delta-\delta\pi T}$$

The minimum wage $\underline{\omega}$ necessary to sustain patient behaviour within the community increases with α, $\bar{\omega}$ and T but decreases with ϵ, ρ and π (for the latercondition, $V_e > V_c$ is

assumed). Observe that in absence of network effects ($\pi = \epsilon = 0$), the inside wage has to be greater than the outside one since it reduces to $\underline{\omega} = (1-\delta)\, V_c = (1-\delta)\,\alpha + \delta\bar{\omega} > \bar{\omega}$.

The insider wage $w > \underline{\omega}$ will have the appearance of exploitation if it is lesser than the outside wage $\bar{\omega}$; this is possible only if the business is profitable enough:

$$\begin{aligned} \underline{\omega} < \bar{\omega} \quad &\Leftrightarrow \quad \bar{\omega}\left[\delta\left(1-\delta(1-\pi)\right)(1-\epsilon) - 1 + \delta + \delta\pi T\right] < \delta\pi\rho - \alpha\,(1-\delta)\,(1-\delta(1-\pi)) \\ &\Leftrightarrow \quad \rho > \underline{\rho} \equiv \frac{\lambda\bar{\omega} + \alpha\,(1-\delta)\,(1-\delta(1-\pi))}{\delta\pi} \end{aligned}$$

where $\lambda \equiv \delta\,(1-\delta(1-\pi))\,(1-\epsilon) - 1 + \delta + \delta\pi T$ can be positive or negative if the ostracism factor is large enough.

Summarizing, we have shown in a simple agency relation that the expected value of being diligent increases with both insider wage and profitability of the business. To sustain diligent behavior within the firm over cheating and quitting, insider wage has to be large enough. However, if profitability is high, the insider wage may still appear exploitative i.e., be less than the outside wage.

Chapter 25

Network Congestion

Service industries based on physical networks like the transportation of people and merchandise (road, rail, air, water), the transportation of energy (electricity, gas, oil) and communications (radio, TV, internet, telephone) display scale economies which give rise to a technological positive network externality: once available and consumed by some people, additional persons can enjoy them without increasing much the cost of provision.[1@] In this context, natural monopolies are often the efficient market structure and regulation is called for to avoid abuse of market power (cf. §17). On top of this, society views these as public services (cf. §16.1.1) which leads to choose pricing rules arbitrating between efficiency and fairness (preferential treatment for some classes of consumers).

This chapter addresses the issue of *optimal network size*. Indeed, a regulated price close to marginal cost is presumably low, thus likely to generate a strong demand. At some point however, the physical network becomes saturated by excessive use, thereby generating a negative network externality called *congestion*. We shall firstly delve into the origins of congestion and explain why the law of demand fails to eliminate it. We then look at the intuitive response consisting in increasing the network size only to expose its hidden costs. Then, we present *peak load pricing*, the major theory addressing the problem together with a list of cases and their remedies. Whereas most of the chapter uses electricity as the underlying good to illustrate our subject, the last section turns to road congestion to broaden our point of view.

25.1 Roots

In this section, we take a close look at seasonality because this phenomenon is at the root of congestion in networks. We also introduce the notion of arbitrage as a force tending to mitigate seasonal congestion.

25.1.1 Seasonality

Seasonality or cyclicality in economic activities originates in nature and customs; the main historic forms are

- *Daily*: farming takes place from dawn to sunset.
- *Weekly*: Sunday for Christians, is devoted to worshiping the Lord and rest.
- *Yearly*: the first people who could leave for vacations used to do it during the summer to enjoy the better weather.
- *Weather*: the wide changes in exterior conditions (e.g., heat, cold, humidity) drive our energy use, mostly destined at countering their effects.

Many of these patterns have passed into regulation or norms to organize society's activity during day, week and year in order to achieve a better coordination both at work and at home. This collectivism makes herding a dominant strategy since everything is organized for those who follow the rules. For instance,

- Night buses being fewer, it is more lengthy to go to one's night job.
- Shopping on Sunday is difficult since most shops are closed.
- The August slowdown–when most firms are closed for vacations–makes all economic activities dearer since the degree of competition is reduced.
- Labour productivity is significantly lower under extreme temperatures.

Breaking society's norm and choosing a path of action different from the majority is an expression of one's freedom in front of an alternative;[2@] practically speaking it means to *arbitrate* or decide. Since this choice is not without consequences, we may say that there is an *opportunity cost of arbitrage*. We shall explain promptly the similarity with the well known concept of financial arbitrage. Table 25.1 summarizes our approach to cyclicality and the human reactions to it.

Actor	Phenomenon	Attitude
Nature	Weather	Natural Seasonality
Society	Customs	Societal Seasonality
Individual	Freedom	Arbitrage

Table 25.1: Seasonality and Arbitrage

25.1.2 Congestion

Seasonality of demand is rarely a problem for the supply of goods because they can be cheaply stored in advance. For instance, new cars or new electronic devices are produced during a few months and stocked in advance of the first day of sale; this way the manufacturer can respond to demand if the product is a success and avoid to enrage rationed "would be" clients.[3@] When it comes to services, the instantaneous supply is limited by the currently installed capacity measured either by available employees (labour) or machines (capital). In labour intensive activities like retail or personal services (e.g., haircutting or catering), it is relatively cheap to meet demand peaks by recruiting more part-time employees. Transportation activities, on the contrary, are capital intensive. In these industries, the technological obsolescence and the necessary maintenance of current equipment makes it very costly to increase the installed capacity to the point that any peak of demand can be met. Hence these transportation systems are likely to suffer congestion during a small period of time, the peak, and extensive under-utilization during the remaining time, the off-peak period. Table 25.2 summarizes our classification.

Commodity	*Technology*	*Congestion*
Goods (storage)	Indifferent	Light
Services (no storage)	Labour intensive	Mild
Services (no storage)	Capital intensive	Strong

Table 25.2: Impact of Congestion

25.1.3 Arbitrage

Almost a century ago, Pigou (1920) observed that taking the London subway at 8h in the turmoil was unpleasant while traveling on the same route at 15h could be enjoyable. Although the fare is the same, the quality of service is worse in the morning as if the journey was more expensive. There is usually an *opportunity cost of congestion* for the user either through lost time or worse quality of service i.e., he would agree to pay more to avoid the crowd at the crowded time.

Since it is physically impossible to avoid the crowd at the peak time, one has to travel off-peak to avoid the crowd i.e., to arbitrate. However, switching from the peak to the off-peak period to travel may be difficult or, stated in economical parlance, expensive; there is thus an *opportunity cost of arbitrage*.

These two opportunity costs are key to understand the relation between arbitrage and peak load pricing. If the cost of congestion is greater than the cost of arbitrage then it is worthwhile to switch consumption from peak to off-peak period. Yet, customs

are strongly binding for most people which means that arbitrage is extremely costly for them; as a result few are naturally inclined to perform the direct *arbitrage*.

If firms could serve whatever demand addressed to them, they would not mind this situation and would keep charging a single price for peak and off-peak service. Understanding that their installed capacity is usually lesser than the peak demand, firms try to optimize the returns on their valuable assets. The trick is to retain those clients willing to pay more for the on-peak service and to retain exactly as many as the installed capacity enables to serve. The other clients who will be rationed (with respect to their first choice) are then compensated in order to foster their use the off-peak service and to lose their patronage altogether. The emergence of two different prices for the two periods is *peak load pricing*.

Practically, the available capacity is auctioned to the users most eager to use it by increasing the wedge between the peak price p_h (high demand) and the off-peak one p_l (low demand) which, under traditional pricing, is nil. A consumer will arbitrate if his net utility of the service is greater during the off-peak period i.e., when

$$\begin{array}{ccccccccc} \text{utility} & - & \text{cost of} & - & \text{peak} & < & \text{utility} & - & \text{cost of} & - & \text{off-peak} \\ \text{peak} & & \text{congestion} & & \text{price} & & \text{off-peak} & & \text{arbitrage} & & \text{price} \end{array}$$

$$\Leftrightarrow \quad \text{Arbitrage if} \quad \left\{ \begin{array}{c} \text{cost of} \\ \text{congestion} \end{array} \right. + \left. \begin{array}{c} \text{off-peak} \\ \text{discount} \end{array} \right\} > \begin{array}{c} \text{cost of} \\ \text{arbitrage} \end{array} \tag{25.1}$$

For many services, arbitrage does naturally take place and, as we shall explain hereafter, should be encouraged using peak load pricing to smooth out congestion. This reduces the inefficiency or hidden cost thereby generated at society's level by the very existence of congestion. For instance, some people negotiate with their firm a shift of their workday by several hours to avoid traffic jams, other leave for vacation out of the season (and go to tropical countries where the weather is still fair) while some move their residence nearby their working place or look for a job nearer to their home. In many sectors and industries, auctions are used to sort out potential clients by their willingness to pay. Congestion in (regulated) transportation has become such a big problem that seasonal discrimination which was previously prohibited on grounds of fairness is now being introduced in more and more places like highways and city centers.

For material goods, both opportunity costs in (25.1) tend to be very low. Indeed, there is almost no congestion cost for goods because firms can continuously adjust supply to demand (storage and transport are cheap). Next, the opportunity cost of arbitrage sums another low storage cost (cupboard or freezer) to the opportunity cost of money if the good is bought in advance (to arbitrate). These observation means that arbitrage is

solely governed by the off-peak discount or, as buyers tend to see it, the peak overrating. Therefore, firms are prevented to take advantage of the higher peak demand and this is why luxury items like “foie gras”, champagne or perfumes which are mostly bought around Christmas time do not suffer wide price variations during the year (cf. §4 on price discrimination). One of the few exception is fresh sea food because freshness is a non storable attribute.

25.2 Capacity Expansion

In this section, we first show that congestion mitigation by capacity expansion has a hidden cost, then we explain how road network effects can mare completely the supposedly obvious benefit of capacity expansion. We conclude with an application to the internet and communications.

25.2.1 Hidden Cost

Seasonality

The customs and regulations that create the seasonality of demand generate an obvious hidden cost. Take for instance the economic activity in a major city; on working days, 10 million people use the public transportation system (TPS) to commute to work while on week-ends nobody does. This makes 50 million trips a week, thus if these could be spread over the whole week, daily trips would fall to $50/7 \simeq 7$ million/day thereby generating a 30% saving for the transportation system either through less investment and less time lost in traffic jams. Obviously, this reallocation of working time over the whole week would necessitate a revolutionary change of mind and organization in that society. This example nevertheless demonstrates that every seasonality generates a hidden capacity cost; it occurs as soon as production of the service necessitates a costly physical equipment and demand for using the equipment displays seasonality. The ideal, cost minimizing solution, would be achieved if usage could be made constant over time.

Price Discrimination

Consider an airline facing strong demand during the summer for an holiday route. The investment into a new unit of capacity, an additional plane, will be launched only if the cost is lesser than the expected return, the latter being the product of a price by a quantity. The price is the highest one that a typical vacationeer would agree to pay while the quantity is the expected yearly number of holy-days. It is clear that if, for cultural

reasons, strong demand is restricted to a few days, it won't pay for the firm to increase capacity; thus there will be congestion.

Fitting the available capacity to demand is called rationing and there are many ways to apply it. Yet, the most profitable one is to increase the price up to the moment where demand shrinks to exactly the capacity. Tickets on a plane that serves only a holiday route during the summer are thus dearer than tickets on the regular all-year-long service on the same route. This discriminatory pricing behavior thus enables the firm to maximize revenue on the holiday route and will induce it to increase capacity on that route. Congestion will be reduced but never eliminated because each new unit serves additional clients with ever decreasing WTP, hence fetches lower and lower revenues (up to the point where the cost of the unit would cease to be covered). Underlying discriminatory pricing is the fact that the additional planes put for service on the holiday route enable the provision of a new service, holiday travel; this service is therefore priced differently from the original one.

Nevertheless, such discriminating schemes are prohibited by regulators for services to the general public such as energy, water, buses or highways. Acting on ground of fairness and efficiency, unit rates are generally set at a low level so that the service receives an extremely large demand in peak periods. Since rationing is almost invariably viewed as a sign of underdevelopment (or a remembrance of war times), it is a politically untenable situation. Legislators and the government therefore thus make it an obligation for the demand peak to be served by recurring to capacity building. The additional cost is then passed rather uniformly to consumers either in a lump sum way with an increase of the subscription or through a rate increase. Society thus ends-up paying a high bill for the service in order to indulge itself the pleasure of taking communion all together.

Under the no-discrimination rule imposed by politicians to regulators, nothing is done to moderate the naturally large demand occurring at some specific and unfrequent moments. Peak load pricing aims, on the contrary, at computing (and applying) the efficient scheme of price discrimination that equate marginal benefit to marginal cost. This is way an optimally sized capacity is build.

25.2.2 Application to Communications

We apply to internet, "the highway of information", a classic analysis from the economics of public goods. Browsing speed on the internet is sometime slow because the resource being almost free we tend to overuse it. Indeed, we only need to pay a small fee[4@] per hour to surf anywhere on the web so that when a great event takes place, millions of users try to access the pages giving realtime information (images, sound or video) on the

event; as a result, the server gets congested and the users have to wait a long time before receiving the response to their request. This excess demand is caused by the zero price of the service and generates a great loss of utility; it is therefore more efficient to set a positive price to the service even if its cost is zero in order to reduce the waiting time of the vast majority of users.

Let us analyze this claim in a simple model. Consider potential users $i = 1$ to n, each characterized by the flow q_i of information he transmits through the network (per hour). Denote $Q \equiv \sum_{i \leq n} q_i$ be the total amount transmitted and K the capacity of the network. We assume that each user displays decreasing returns to consumption and suffers from congestion in proportion of the delay he suffers which is measured by the ratio Q/K; we thus posit $u_i(q_i) = \sqrt{q_i} - pq_i - cQ/K$ where $c > 0$. The maximization of utility leads to the FOC, $\frac{1}{2\sqrt{q_i}} = p + \frac{c}{K}$ (since $Q = q_i + q_{-i}$). If internet is free ($p = 0$), the optimal individual consumption is $q^0 = K^2/4c^2$ leading to an aggregate of $Q^0 = nK^2/4c^2$. It is readily observed that total usage Q^0 increases with capacity K but decreases with the delay factor c. We notice that the congestion index Q/K is, in equilibrium, proportional to capacity by a factor $n/4c^2$. A 10% increase in capacity will trigger such a large increase in demand that the congestion index will increase by the same 10%. This is a well known result according to which

> An increase in transportation capacity ends-up creating even more congestion.

The socially optimal use of the network maximizes the total utility minus the cost of providing the service which, for simplicity, we assume to be zero (it is quite small compared to the fixed cost of setting up the network). The objective is thus $\sum_{i \leq n} u_i = n\left(\sqrt{q} - c\frac{nq}{K}\right)$. At the social optimum, the price is set equal to the marginal cost, here zero, hence the FOC is $\frac{n}{2\sqrt{q}} = \frac{cn^2}{K}$ and its solution is $q^* = \frac{K^2}{4n^2c^2} = q^0/n^2$, leading to total usage $Q^* = Q^0/n^2$. We may thus conclude that free internet is overused by a factor n^2 independently of the opportunity cost of delayed surfing c. Maximum welfare, equal here to consumer surplus, being $W_D^* = \frac{4K}{c}$, society's willingness to invest into new capacity (internet server) is $\frac{4}{c}$. It is now natural to search for a price that decentralizes the optimum; we look for p^* such that the individual FOC is satisfied at the efficient consumption q^* i.e.,

$$\frac{1}{2\sqrt{q^*}} = p^* + \frac{c}{K} \Leftrightarrow p^* = \frac{1}{2\sqrt{q^*}} - \frac{c}{K} = \frac{(n-1)c}{K}$$

This optimal price increases with the number of potential users when the capacity is fixed. Using this optimal price, an increase in capacity enables a price reduction but still increases congestion since Q^*/K is proportional to K.

If access to the Internet event is managed by a monopoly, most often the owner of

the broadcasting rights, he shall set p to maximize $\Pi(p) = npd(p)$ where $d(p) = \frac{K^2}{4(pK+c)^2}$, thus profit is proportional to $\frac{p}{(pK+c)^2}$ and maximum for $p^M = \frac{c}{K}$ which leads to total traffic $Q^M = \frac{nK^2}{16c^2} = \frac{Q^0}{4}$. The monopoly is aware that a high price reduces demand (as usual) but it is also cognizant that this indirectly reduces congestion, thus increases willingness to pay and generates a positive feedback on demand. This is why, in our simple example, the monopoly price rises more, leading to a monopoly demand that is only a quarter of the competitive demand ($p = C_m = 0$ is the competitive price) to be compared to one half in the absence of externalities. Yet, the monopoly production is still inefficiently large because the congestion externality is mostly ignored.[5@]

Observe also that since the monopoly profit is $\Pi^M = \frac{nK}{16c}$, his willingness to invest into new capacity is $\frac{n}{16c}$; this level is proportional to the number of potential users and inversely proportional to their delay cost (which confirms basic intuition) but above all it is greater than the social willingness to invest which shall lead to over-investment because, in our simple model, users value capacity and are ready to pay for it.

Although appealing, this model should not be over-stretched. No general conclusion can be drawn here because the changes with respect to traditional local equilibrium theory "à la Marshall" utterly depend on the linearity, concavity or convexity of the perceived cost of congestion about which we cannot say anything beforehand.

25.2.3 Application to Air Transport

Air traffic grows in Europe but delay growths faster. According to Eurocontrol, movements rose 4% per year in both 2004 and 2005 to reach more than 9 millions (per year) but the number of delayed flights rose to 40% and then 43%. Likewise the percentage of flights delayed by more than fifteen minutes rose to 18% and then 20%. The average delay per movement rose to 11 minutes increasing by about 8% every year which is about twice the increase in activity.

Legacy airlines used to connect directly all served cities, leading to the so-called fully connected (FC) shape of their networks. Today, the Hub and Spoke (HS)[6@] network has become the standard shape; it was originally developed by FedEx for rapid freight delivery and by Delta Air Lines for passenger transportation. Brueckner (2004) offers an explanation for the dominance of this new organization based on the optimal behavior of a monopolist operating a triangle network where the top vertex can optionally serve as a Hub to connect indirectly the two lower vertices. The HS structure dominates the FC one if fuel cost and/or waiting time are low enough.

We saw in §3.3.2 on quality choice by a monopolist that the profit over a route is $\pi = qP(q,s) - C(s,q)$ where $P(q,s) = \frac{a-q}{b} - \frac{\gamma}{s}$ and $C(q,s) = \theta s + cq$. Under FC, the cost of

a single route is $C(q_1, s)$ where q_1 is the demand for that route. Since all three routes receive the same demand, total cost are $3C(q_1, s)$. Revenues are three times those of one route, thus profits under FC are $\pi^{FC} = 3\pi(q_1, s)$. With an HS network, the cost of a single route is $C(q_1 + q_2, s)$ where q_1 is the demand for direct flights while q_2 is the demand for connected flights. Total cost, using only 2 routes for 3 cities, are thus $2C(q_1 + q_2, s)$. The treatment of revenues for the HS firm is more complex than for the FC one. Revenues over the two direct routes are twice those of one route but revenues for the connected route are lesser because airplane time is twice and there is some waiting time at the hub. To model this fact, we assume that the WTP of consumers is diminish by a constant factor μ. Profits are thus

$$\pi^{HS} = 2q_1 P(q_1, s) + q_2(P(q_2, s) - \mu) - 2c(q_1 + q_2) - 2\theta s$$

The FOCs for quantity and quality in the FC network are

$$a - bc - \frac{\gamma}{s} = 2q_1 \quad \text{and} \quad q_1 = \frac{\theta}{\gamma} s^2 \tag{25.2}$$

while the three FOCs for the HS network include the RHS of (25.2) together with

$$a - \mu - 2bc - \frac{\gamma}{s} = 2q_2 \quad \text{and} \quad q_1 + \frac{1}{2} q_2 = \frac{\theta}{\gamma} s^2 \tag{25.3}$$

The system (25.2) has a unique economically relevant solution (q_1^{FC}, s^{FC}). Since the RHS of (25.2) applies in both situations, if we set initially $s^{HS} = s^{FS}$, then we must have $q_1^{HS} = q_1^{FC}$. Now, assuming that $2q_1^{FC} > \mu + bc$, the LHS of (25.3) yields a positive q_2^{HS}. Plugging this information into the RHS of (25.3) and comparing to the RHS of (25.2), we see that $s^{HS} > s^{FS}$ must hold. Looking back into LHS of (25.2), we get $q_1^{HS} > q_1^{FC}$. The increase in s^{HS} also increases q_2^{HS}. Since both HS quantities are greater than before, the solution of the RHS of (25.3) is an even greater s^{HS}. This process converges because the changes are smaller and smaller. We have thus shown that adopting the HS structure involves higher flight frequency ($s^{HS} > s^{FS}$) and larger traffic on direct routes ($q_1^{HS} > q_1^{FC}$).

From the RHS of (25.2) applied to FC and HS, we get $\frac{q_1^{HS}}{q_1^{FC}} = \left(\frac{s^{HS}}{s^{FC}}\right)^2 > \left(\frac{s^{HS}}{s^{FC}}\right)$ since $s^{HS} > s^{FS}$, thus $k^{HS} = \frac{q_1^{HS}}{s^{HS}} > \frac{q_1^{FC}}{s^{FC}} = k^{FC}$. Since the k ratio is a proxy to plane size, the HS structure also involves bigger planes. Lastly, Brueckner (2004) shows that the HS network is conducive of greater profits i.e., more likely to be the optimal shape, if the parameters μ and c are small.

25.3 Peak Load Pricing

As compared to capacity expansion, peak load pricing is an alternative and more savvy way to equate demand and supply; it fosters consumer arbitrage and thereby increases the off-peak demand while decreasing the peak one; for that reason one also speak of "peak shaving" or "valley filling". We illustrate in details its functioning using the case of electricity.

25.3.1 Variability and Seasonality

Figure 25.1 displays the instantaneous demand for electricity in France, Italy and Spain over the day of greatest demand in 2003/2004.[7@] The graphs clearly show many variations in demand during the day; they result from the living habits of households and from the legal pattern of working hours in businesses. There is however an important difference, the small peaks at 22h and 2h in France are absent from the other countries; they are due to the extensive use of peak load pricing schemes by french industrial consumers.

Non cyclical events such as a solar eclipse or a football match can also provoke large demand variations. On 11 August 1999, many people in the UK stopped their normal activities to watch the solar eclipse, the electricity demand fell sharply during 20 minutes, as if 4 million consumers had suddenly disappeared. At the end of the solar eclipse, demand rose again back to normal in only 20 minutes. Likewise, when watching TV, we stop using other appliances and therefore reduce our electricity consumption. This last until the commercial break when everybody turn on the lights and goes to the kitchen to prepare coffee or tea. The resulting sharp increase in electricity demand is called a TV pick-up.

Figure 25.2 displays the French demand over the year 2003 in two ways; the first graph simply records the weekly consumption and displays it according to calendar date. To understand its shape, it is enough to recall that electricity demand is inversely related to outdoor temperature and that France is dominated by Atlantic weather conditions.[8@] The bottom graph, called the *load duration curve*, displays the same information except that levels of hourly demand are ranked according to decreasing demand. It is readily observed that to meet the maximum demand of that year, which is 83 GW, the utility has to maintain available a larger capacity, here 91 GW. The latter is itself lesser than the installed one, 116 GW, because some units are necessarily out of line for maintenance.[9@]

To see the hidden cost of excessive capacity, imagine that industrial demand can be coordinated by a central planner so as to become perfectly correlated to the opposite of the residential demand; this means that at any hour of the year, total demand is equal

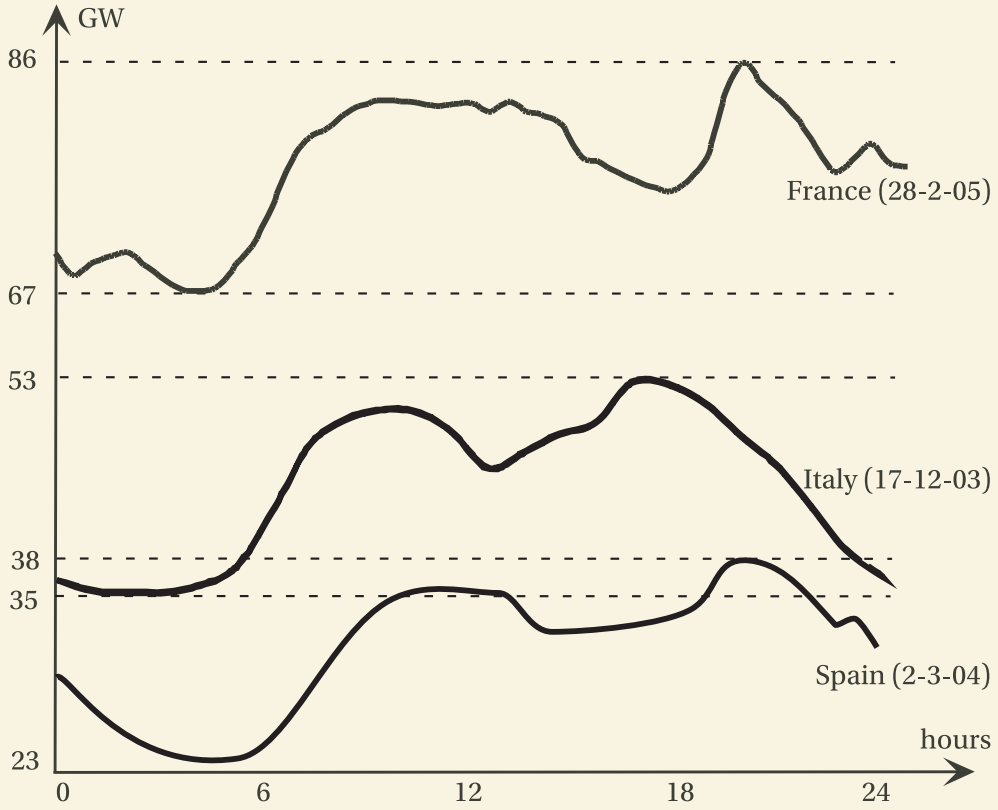


Figure 25.1: Variability of Electricity Consumption

to the average 53 GW. In suh a situation, the installed capacity could be halved to 58 GW without compromising security of supply (the fact that a unit can go down unexpectedly). Once we know that the cost of capacity is more than one half of the total cost of electricity, the electricity bill for the country would be reduced by at least a quarter.[10@] That may be enough to negotiate with industrial users a smoother use and change this dream into reality.

Not everything is lost in the current situation; thanks to the existence of interconnections with neighboring countries, the excess of available capacity over internal demand can be used for exports.[11@] By widening the network, the variability of demand is reduced since cold days in the north with heavy use of heating are fair ones in the south and hot days in the south with heavy use of air-conditioners are fair ones in the north; hence the need for extra capacity is reduced if the generated electricity can be transported from north to south and vice versa to meet demand wherever it is located. The European integration of national electricity markets shall yield a hidden benefit by reducing the hidden cost of excess capacity. France provides a clear example; this country is at the moment the largest European exporter of electricity and is a net exporter at almost every hour of the year. However, during a few very cold hours on 1/3/05, net

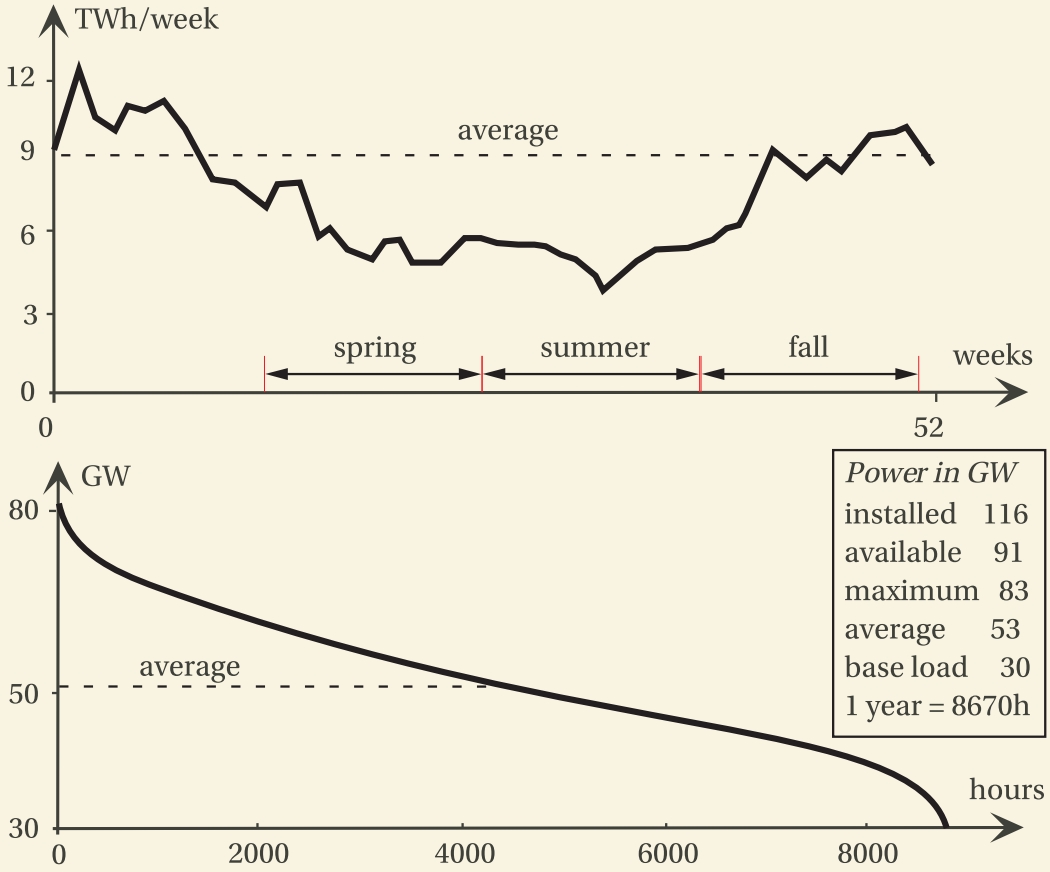


Figure 25.2: Load Curves for Electricity Demand

imports came from Germany and Spain to save the utility from having to start its most expensive generation units.

25.3.2 Optimal Peak Load Pricing and Capacity Choice

Problematic

Peak load pricing is an old commercial practice akin to price discrimination that was neglected by academics until Boiteux (1949) made clear its relation to long term efficiency in the sector where congestion is most pressing: electricity.[12@]

Indeed, congestion in an electrical network is simply unsustainable; either the operator rations some consumers or the whole system fails i.e., a black-out occurs. Both black-outs and rationing have and continue to have a very negative political backlash; governments therefore always take care to instruct regulators to put reliability as the number one requirement of the integrated utility. Since these events are to be avoided at all costs, the utility is lead to build a very large transport capacity together with a very

large generation capacity to be sure that any peak of demand anywhere in the service area can be securely met.

Analysis

When the demand for a service is constant or changes at a sufficiently slow pace to enable capacity adjustments, the relevant time frame for cost is the long-term (LT). Capacity is therefore set at the efficient production level equating willingness to pay and marginal LT-cost, the latter determining the price. The left panel on Figure 25.3 shows this for two demands labeled low (l) and high (h).

If there are two periods of low and high demand that alternate faster than the time needed to adjust capacity, each period will necessarily have to be treated from the short-term (ST) point of view i.e., the relevant cost function is the ST one. To analyze the setting of prices we shall assume that there is *no arbitrage* possibility so that the demand in each period remains independent of the price set for the other period. In each period, efficiency commands to price at the marginal ST-cost.

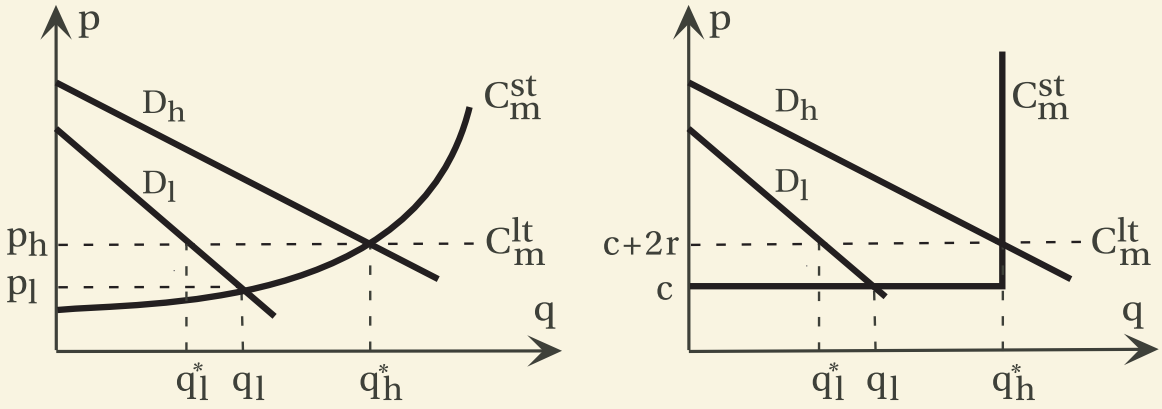


Figure 25.3: Flexible and Rigid Intermittent Demands

We know from §2.1.3 that the ST and LT marginal cost coincide for just one quantity and since the ST function is more convex than the LT one, it is optimal to set the capacity equal to the peak demand (this is quite clear for the rigid technology on the right panel of Figure 25.3). Therefore, the peak quantity is the LT-efficient one, q^*_h, while the off-peak quantity is q_l; the latter is only ST-efficient and greater than the LT-efficient level q^*_l. Indeed, since the installed capacity is greater than the off-peak demand (when priced at the ST-efficient level), production is artificially cheap.

Interpretation

We can now interpret the ST marginal cost at off-peak demand p_l as a unit cost c of service for one period while the difference with the marginal cost at peak demand $p_h - p_l$ is seen as the unit cost of capacity expansion; we denote it $2r$ in order to interpret r as a per period cost of capacity expansion. A cycle sums the peak and off-peak periods; our notation makes clear that peak users are alone to finance the capacity since they pay c for their energy, r for the extra capacity they need and another r for a capacity that is shared with off-peak users. As a matter of example, consider a business hotel for which the week-end is the off-peak period. The unit cost of providing a room for one night is the labour cost of cleaning the room everyday while the cost of making one more room available (capacity expansion) is the capital cost of construction plus the labour cost of employees that are present during the whole week (cycle). The same dichotomy applies for airplanes.

For transportation activities such as highways or trains that are still regulated in some countries, governments have generally been wary of discriminatory pricing and have tended to impose a minimum fairness in the design of tariffs. Uniform pricing is in fact the only way to share the capacity cost $2rq_h$ in proportion of group consumption. Indeed, one easily checks that the unique solution to the system

$$r_l q_l = \frac{q_l}{q_l + q_h} 2rq_h \quad \text{and} \quad r_h q_h = \frac{q_h}{q_l + q_h} 2rq_h$$

is $r_l = r_h = \frac{q_h}{q_l+q_h} 2r > r$ since the uniform price guarantees that final demands satisfy $q_h > q_l$. Contrariwise to the efficient discriminating solution, uniform pricing is unfair with off-peak users since they pay a price for capacity greater than if peak demand did not exist.

Examples

Modern regulators now take a closer look at the hidden costs of capacity expansion and accept some degree of peak load pricing to eliminate unnecessary investments. The case of road transportation is treated in the next section. For water networks, true peak load pricing would require real time metering which would be extremely costly to implement. Instead, the rate structure takes advantage of the near simultaneity of individual loads to overcharge heavy users using an increasing marginal price. This is why in most countries where potable water is not plentiful, efficient discrimination is applied with increasing block tariffs (cubic meters become more and more expensive).

One of the first peak load pricing scheme is probably Boiteux (1957)'s "green tariff"

that was implemented the following year by EDF in France. Industrial users were offered a discounted price for most of the time in exchange of paying a higher price during the winter for peak hours.[13@] Today, in the retail market, beyond the basic contract offering a constant price of 107 €/MWh, EDF proposes a "time-of-use" (TOU) contract where the night time price is reduced to 64 €/MWh. This tariff tends to smooth out consumption across hours within any given day.[14@]

The previous characterization of efficient peak load pricing was based on the inverse demand curves (willingness to pay). A competitive firm would obviously price peaks in exactly the same way while a monopoly or any firm with market power would set higher prices although following the same approach i.e., equates off-peak marginal revenue to cost of service and peak marginal revenue to cost of service plus the total cost of capacity.

Lastly, we must comment the *reversal* problem. If the peak demand priced at LT marginal cost happens to be lesser than the current capacity then we have excess capacity so that none should be replaced meanwhile obsolescence naturally reduces the working capacity. Likewise, if the off-peak demand priced at the (expensive) LT marginal cost is still greater than the current capacity there is an urgent need to build more capacity. Finally, there is the case where peak load prices inverts the original ranking of demands; in that case the peak–off-peak wedge must be reduced (while maintaining capacity financing) until the two demands become equal.

25.3.3 Demand Side Management

Demand Side Management is the commercial name for direct strategies of peak shaving that are an alternative to peak load pricing. The problem with the latter is that it leaves the client take his own consumption decision. If the law of large number can be applied then occasional strange behaviors are averaged out; however, if the cost of arbitrage suddenly increases then the peak price will be too low to motivate arbitrage.[15@] An example is electricity prices during summer days; usually clients avoid using much their air-conditioner because of the high cost but if there is a heat wave they will surrender and switch the apparatus to the maximum, at least for a few hours. Since everybody in the area will do the same, the law of large number won't apply and demand will explode which is precisely the worst situation for the utility. To limit load (i.e., quantity), a direct negotiation with large clients can be carried out:

- *peak clipping*: an interruptible contract whereby the client, mostly industries, agrees to be cut-off on short notice (for a limited time and in exchange of a bargain price).
- *load shifting*: a packages of prices and quantities whereby the client agrees in advance to limit load during the known periods of peak demand.

Modern methods of consumption smoothing involve a more active participation of consumers thanks to the progress of telecommunications (IT revolution) which enables to notify clients either of a price increase or of a rationing (to which they have agreed). EDF's large clients for which interruptible contracts can be proposed account for one third of total electricity output (more than household share); hundreds have contracted one of the above formula.

A different scheme of DSM in use at EDF since 1996 is the "tempo" contract; this peak pricing strongly motivates the client to reduce consumption on some crucial days to compensate for the load increase of more inelastic clients. The 365 days of the year are distributed into 3 classes as follows: 300 days blue (cheap), 43 white (regular) and 22 red (expensive). EDF later decides every day whether the next day falls in any of the 3 categories with some limits and up to the exhaustion of the available days in each category.[16@] Table 25.3 shows the various unit prices associated with this flexible contract together with some EDF cost and prices for various fuels; one sees how the prices in the flexible contract match the market prices of the technologies used to meet peaks of demand.

category	blue N	blue D	white N	white D	red N	red D
# hours	2400	4800	344	668	132	396
€/MWh	44	55	91	107	168	470
1997 data						
Marg. Fuel	Nuke	Nat. Gas	Nat. Gas	Gas Turb.	Gas Turb.	Oil Turb.
Av. Cost	41	40	40	98	98	152
Price	32	43	72	85	136	363

D: day time (6h-22h) N: night time (22h-6h)

Table 25.3: Demand Side Management

The overall effect of the DSM scheme is to smooth out consumption across days within the week. Thanks to these peak load pricing schemes, the demand addressed to EDF is less variable so that maximal available capacity can be reduced while keeping the same margins of security of supply.

25.3.4 Nodal Pricing in Electricity †

Purely competitive trading states that the price of Port wine in England is that in Portugal increased by transportation cost. Transportation in a network is however limited by its size characteristics e.g., the number of lanes on a highway, the section of a pipeline, the voltage of an electricity line, the number of optic fibers in an internet backbone. We will show in a simple model of deregulation the nature of this problem.

Consider two countries like France (#1) and Italy (#2) that regularly trade electricity. Deregulation has taken place so that there is a competitive market in each country where generators and distribution companies bid demand and supply. The excess demand function for region i is $D_i(p_i)$, the aggregate demand minus the aggregate supply. It is a decreasing function thus, for a net import load q_i, the equation $D_i(p_i) = q_i$ has a unique solution $p_i = P_i(q_i)$ measuring the value (shadow price) of an additional unit of imports at the current level q_i. If region i was disconnected from the other one, local production and consumption would equalize and the price would be $p_i^o = P_i(0)$ as shown on Figure 25.4. Now, when net imports reach the level q_i, the net surplus of electricity generation and consumption, called the *regional benefit* is $W_i(p_i, q_i) = \int_0^{q_i} (P_i(x) - p_i)\,\mathrm{d}x$.

In each region there is a transmission operator whose mission is to maximize $W_i(p_i, q_i)$, the regional benefit. Suppose France (#1) is a net exporter ($q_1 < 0$) while Italy (#2) is a net importer ($q_2 > 0$). Since trade between the two regions is performed at the same price p and $q_1 + q_2 = 0$ in the absence of transmission losses, the social objective is to maximize $W_1(p, -q) + W_2(p, q) = \int_0^q P_2(x)\,\mathrm{d}x - \int_{-q}^0 P_1(x)\,\mathrm{d}x$. The efficient cross-border trade q^* solves $P_2(q) = P_1(-q)$ which defines p^* as illustrated on Figure 25.4. This outcome can be easily obtained by opening national spot markets to firms of both countries; this way arbitrage opportunities will guarantee price convergence in France and Italy toward p^*. The welfare so obtained is a maximum W^* also called first-best.

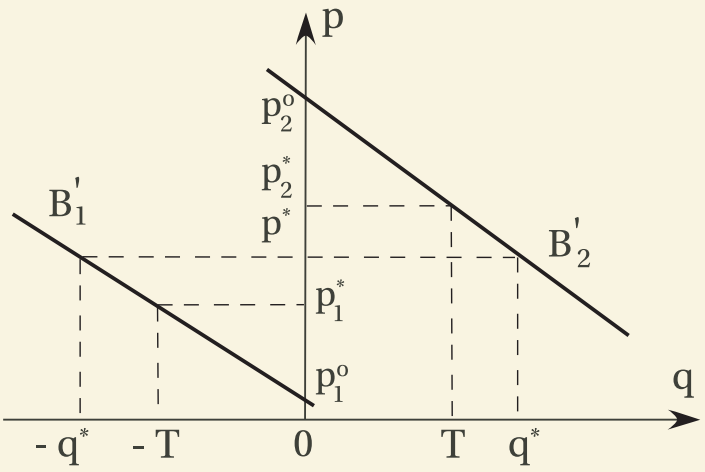


Figure 25.4: Electricity Trade

The transmission line bridging the two countries has a thermal limit T that should not be exceeded without risking a complete failure; it therefore determines a maximum transport capacity. If the market equilibrium level of trade q^* is larger than this limit, the transmission system operators must step in to resolve the imbalance by forcing a maximum trade of T units. Yet they still wish to maximize welfare i.e., they look for a

rationing that reduces minimally welfare (from the first best level W^*). The problem is that no single price can generate the desired reduction $q^* - T$ in demand for transmission. Indeed, if there are exactly T units scheduled for transmission then the willingness to import in Italy is $p_2^* \equiv P_2(T)$ which is greater than the willingness to export in france, $p_1^* \equiv P_1(-T)$ as can be seen on Figure 25.4. As a result, there are still profitable trading opportunities and traders will ask for more transmission service.

Applying a rationing scheme like "first–in–first–served" is not a good idea because it is not necessarily those who value most the transmission service that will come first to reserve it. Neither is it efficient to curtail all trades proportionally by a factor $1 - T/q^*$ in order to force the entire demand q^* down to the capacity T. It would be more efficient to curtail entirely trades with low WTP and allow trades with high WTP. How then do we modify the market rules to implement the efficient rationing? Two methods can be used. The first one is *nodal pricing* whereby different prices are quoted at different locations. This scheme is used with great success in Scandinavia, a market formed by Denmark, Finland, Norway and Sweden. A company called NordPool is granted by the respective governments the exclusive right to buy and sell electricity over the borders of the participating countries.

Applying this idea to our example amounts to set the initial prices in the French and Italian markets at p_1^o and p_2^o, the respective autarkic levels. Then the transmission monopoly starts quoting a larger price in France (standing as a buyer of electricity) and a lower one in Italy (standing as a seller); he will continue until the point where the transmission link is fully used but not congested. The price difference between the two regions $p_1^* - p_2^*$ expresses the opportunity cost of using the link, hence electricity is priced at its "true" marginal cost in each region. At the equilibrium, the transporter buys T units at price p_1^* in France and sells them at price p_2^* in Italy thereby making a windfall profit.

The second method amounts to tax trades that cross the border. Ideally the transmission operators compute the efficient Pigouvian tax $t^* = p_1^* - p_2^*$ in order that arbitrage makes the full price equal on both sides of the Alps. In both cases, the transmission monopoly obtains a rent that has to be redistributed to market players or invested to improve transmission capacity and reliability.

Regarding fairness and price discrimination, the model's result tells us that consumers of energy located far away from production centers (or import harbors) ought to pay more for it since they generate a higher cost of effectively providing the service (transportation cost). In this sense, cities subsidy the countryside whenever a uniform statewide consumer price is imposed by the regulator.

25.4 Road Congestion

25.4.1 Origin

Stated in few words, road congestion occurs in the urban areas of all countries because geographical space is limited while urban population keeps growing. What appeared half a century ago in the largest cities of the most advanced economies has spread over the world and is starting to be thoroughly studied now that government have come to realize the magnitude of the associated costs (cf. Textbook on Transportation Economics).

The root of the problem is the organization of production in modern economies which requires not only task specialization but also geographical specialization. We thus observe the creation of a central business district (CBD) in each city, generally at the center or at least within a pre-existing urban area; this centric localization means that extending the CBD is costly because transaction cost to displace neighbors are significant. For that reason, the CBD tends to grow vertically with the appearance of skyscrapers. On the other hand, residential areas in cities can grow easily (or cheaply) by extending towards the outskirts.[17@]

Now, the growth of population generates a growth in commuting while economic growth enables more people to afford a car and use it for commuting (instead of using public transportation). These two effects compound to create an important growth of traffic on the arteries linking residential suburbs to the CBD. In most cases, urban planning has failed to anticipate this evolution and as a result, road capacity is systematically inferior to the peak demand when commuters drive to and from the CBD. The resulting congestion is recurrent and is refereed to as "bottleneck". Other forms of (non recurring) congestion includes accidents, special events or works.

Although the issue of road congestion now ranks high on the agenda of public decision makers, data has only recently begun to be collected in order to enable well informed choices. For the US, Schrank and Lomax (2005) show that mobility problems in large urban areas have increased at a relatively constant rate during the last two decades. Congestion affects more cities (not only the largest ones), more roads in each city, the time span where congestion can appear is greater (not limited anymore to morning and evening rush hours) and the duration of congestion events is greater. This can be checked in Table 25.4 where data for the 85 largest US urban areas are gathered.[18@] Travels, both private and public, have increased at a faster rate than population growth, thereby generating more congestion in terms of extension, intensity and duration which altogether contribute to a staggering increase of the total cost; the latter which only sums fuel and time wasted thereby ignoring all negative externalities already amounts to one percent of the income of the involved population.

Year	1982	2003	Yearly Δ
Annual bn of vehicle-miles (roads)	387	781	3.4%
Annual bn of person-miles (public transport)	23	43	3.0%
Proportion of roads affected	35%	65%	3.0%
Congested traffic during peak time	33%	67%	3.4%
Daily hours of congestion	4.5	7	2.1%
Delay in hours/year/peak traveler	16	47	5.3%
Average delay over free-flow	12%	37%	5.5%
Reliability delay over free-flow	60%	100%	2.5%
Cost in 2003 bn$	12	63	8.2%
Cost in proportion of income	.4%	1%	4.5%

Table 25.4: Road congestion in the US

25.4.2 Solutions

Since the major cause of road congestion is the inadequacy of supply capacity to meet demand, aka bottlenecks,[19@] the first and obvious answer is to extend transportation systems, both privates ones (road) and public ones (subway, train). Yet the pace of new constructions is systematically inferior to demand growth because of transaction costs for expropriation and the price of land nearby the CBD (cf. §16.1.2). This applies for roads as well as public transportation systems.

A symmetric solution to the aforementioned, advocated by Vickrey (1963), is to curtail demand using either prices or quantities (rationing). The difficulty with applying peak load pricing to road usage at the local level stems from two factors. Firstly, most urban ways are free, their cost being paid by taxes; citizens thus feel the differential pricing scheme as a tax in disguise. Secondly, the benefits are hard to pinpoint since they consist of an average reduction of lost time; indeed, not all the variations in traffic are eliminated, only smoothed. Finally, this efficiency enhancing scheme is believed to be un-equitable (socially regressive) because the poorest users are frequently those with a high cost of arbitrage; this means that their pattern of behavior is not affected by the scheme, they only end up paying for what was previously free.

For the case of road traffic, some cities like Paris have adopted a deliberate policy of rationing the supply below the level afforded by the physical capacity. Car use is made costlier and more time consuming, by limiting parking space, reducing the number of streets available for transit[20@] and generally speaking increasing car related taxes. Such a strategy has mostly cost because traffic worsens and few benefits in terms of better functioning of alternative transport modes (bus, subway, bicycle) as shown by Prud'homme et al. (2006) in their study of the Paris strategy.[21@]

The alternative strategy based on prices is to charge for access to the conflicting zones

during peak hours; this is one occurrence of peak load pricing adopted by many cities starting with Singapore in 1975. In the London case, TfL (2006) reports an 18% traffic reduction in the charging zone leading to a 30% reduction of the congestion delay; at the same time, bus use increased as well as quality of service. Lastly, some positive externalities were notices in the form of lesser road traffic accidents and 12% less emissions of pollutants. Yet TfL (2008) reports that the substitution of car lanes into walking or biking lanes and the important works on the streets have severely curtailed the road capacity for cars so that congestion has returned to the 2002 level. Computations found with the introduction of congestion charging, an elasticity of $-.45$. Yet, when the 60% raise took place a few years later, elasticity dropped to $-.2$. While the initial scheme seemed to deter some drivers from accessing central london by car, those who paid seemed ready to endure any increase to continue to enjoy the privilege. The report states that "the extension of the charging zone to the west has shown none or little improvement in all the dimensions previously commented indicating that a simple and not overtly expensive price instrument cannot solve for all the traffic problems of a city".

Lastly, we may look at the cost of implementing the congestion charging system which uses an extensive network of cameras. Prud'homme and Bocarejo (2005) find it to be quite large and argue that the net benefit to society is not necessarily positive but will surely become so once the aforementioned cost is reduced by learning and experience.

25.4.3 Expansion

Road Pricing

Hau (1992) recants the origin of the debate on road pricing. Pigou (1920) observes that the open-access regime of many natural resources (land, sea) or infrastructures (road) leads to an inefficient over-exploitation. He uses the example of two roads, a fast but narrow one (A) and a wide but slow one (B). As traffic grows on A, travel time increases up the point where the advantage over route B vanishes. A toll equal to marginal cost minus average cost of a trip would bring traffic on route A to the efficient level i.e., the negative traffic externality would be internalized. Knight (1924) agrees with the existence of an externality but sees public intervention warranted only for public ownership. If road property rights are well delineated and competitive pressure is present, a private owner derives a Ricardian rent and faces the same incentives as the social planner when seeking to maximize it; he thus sets the same toll.[22@]

Mohring and Harwitz (1962) study the management of a congestion prone road as follows: the perceived unit cost of a trip is $C(q,K) = v \times \tau(q/K)$ where v is time value, q total output (e.g., flow on road), K capacity and τ the convex cost of delay or congestion.

The capacity cost (including building AND maintaining)[23@] is $\Phi(K)$ (e.g., rK). Demand arises from the elastic WTP of potential users $P(q)$ and equilibrium settles when $P(q) = C(q,K) + p$ where p is service price. Welfare is $W(q,K) = U(q) - qC(q,K) - \Phi(K)$ where $U = \int_0^q P(x)\,\mathrm{d}x$ is the gross surplus from road use. Its maximization involves a congestion term since the output FOC is $P = C + qC_q \Leftrightarrow p = qC_q$ i.e., a Pigouvian tax is required to edge the congestion term. Its effect is to depress demand down to the point where WTP is equal to social cost. This is symmetrical to the monopoly increasing its price to edge out the profit loss due to competitive demand (to be improved, ref needed). Toll revenue is then $R = qp = q^2C_q$. The FOC for efficient capacity is $\Phi_m = -qC_K \Rightarrow K\Phi_m = -qKC_K = R$ because C is a function of the ratio $\frac{q}{K}$ (ie.., satisfies $qC_q = -KC_K$). In the end, the ratio of revenue to cost of the asset is $\frac{R}{\Phi} = \kappa$, the elasticity of capital cost Φ. Thus, toll revenue exactly covers cost if and only if the road construction and maintenance technology displays constant returns to scale.

A monopolist owner (or manager) would maximize profit $\pi = pq - \Phi(K)$ which amounts here to replace WTP P by the smaller marginal revenue $P + qP'$ in the quantity FOC and maintain the same capacity FOC. Thus a larger price would be asked in order to sell a higher quality service at a premium. Capacity in turn would likely be smaller. As noted in footnote 26.3 above, if there exists an alternative road at fixed toll (zero if open access) and an unlimited demand for usage, then the monopolist is faced with an infinitely elastic demand and forced to behave as a social planner. Yet, such an extreme assumption is unlikely to hold.

It is been shown empirically for highways, the existence of mild scale economies (e.g., doubling lanes), meaning that κ is close but lesser than unity. Hence, first-best efficient pricing which is a positive toll to limit congestion would still run a deficit. As with most utilities, a second best pricing is required to allow the asset to stand on its own bottom.[24@] The conclusion is reversed in urban areas because the cost of land tends to increase exponentially (it is a very scarce resource) and interchanges (bridges,...) are more numerous. Hence first-best pricing would generate a surplus.

Paradoxes

Downs (1962) formalizes Pigou (1920) and Knight (1924)'s verbal argument to produce well known capacity expansion paradoxes bearing their names.

City residents commute by car to the CBD using a network consisting of two roads, one narrow and fast (A), the alternative being wide and slow (B). The demand for commuting service is inelastically set to unity. Travel time on route $i = A, B$ depends on the share of total traffic d_i going through this route. We assume that the wide road is never congested so that $t_B(d_B) = \bar{t}$. Free flow travel time on road A is $t_A(d_A) = \underline{t} < \bar{t}$ meanwhile the

density remains below capacity i.e., $d_A \leq k$. For higher densitities, travel time increases linearly with $t_A(d_A) = \underline{t} + \alpha d_A / k$. The left panel of Figure 25.5 uses $d = d_A = 1 - d_B$ on the horizontal axis to display t_A and t_B simultaneously. We consider increasing capacity levels k_1, k_2, k_3 to illustrate the effects of increasing the capacity of the congestion-prone road.

The equilibrium distribution of travelers is characterized by the no-arbitrage condition: $t_A = t_B = \hat{t}$ and is shown with a circle on the Figure. It is readily seen that an increase in capacity from k_1 to k_2, increases equilibrium density on road A from d_1 to d_2 without changing the common travel time $\hat{t} = \bar{t}$ on both routes. The latent demand for the better road instantaneously fills any capacity melioration. If the capacity is increased enough to k_3 then nobody uses the slower road anymore ($d = 1$) and travel time is reduced.

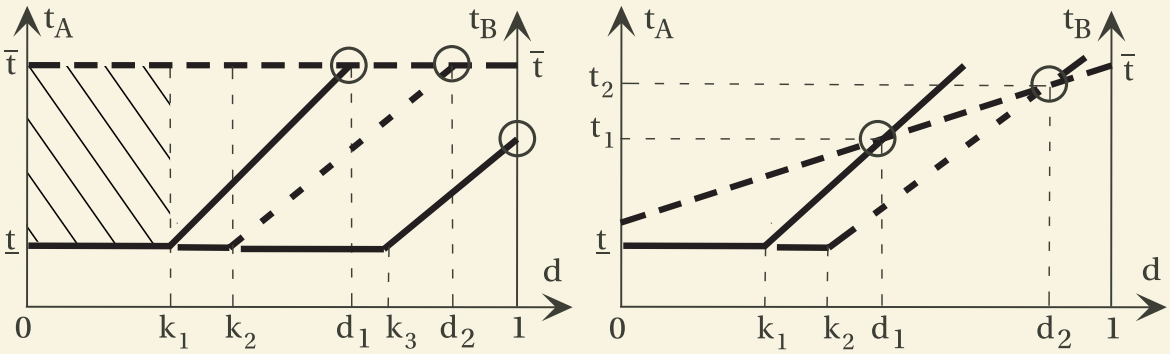


Figure 25.5: Road Paradoxes

The paradoxical result that a capacity increase might be completely ineffective is a consequence of the private behavior of commuters who ignore the cost they impose to others when switching to the fast road. The comparative statics between k_1 and k_2 tells us the following:

> A route can be summarized by two characteristics, capacity (ampleness) and quality (speed or travel time). When the alternative to a narrow and fast route is a wide and slow one, it is advisable to improve quality before capacity i.e., make the slow route faster rather than widening the narrow one.

The efficient use of the road network calls for minimizing total travel time $T = d\,t_A + (1-d)\,t_B$ or average travel time.[25@] In the open access equilibrium where both routes are used, $T = \bar{t}$, which graphically is the area of the rectangle delimited by the axes and $\bar{t}$. To internalize the negative traffic externality, one should first limit transit on the fast road A to its capacity ($d_A = k$) so that a share k of demand enjoys a shorter travel time. Graphically, the savings are the striped area on the left panel of Figure 25.5. Obviously,

additional capacity increases this potential time saving. Implementation of this scheme can bring more or less savings. If traffic lights at the entrance are used, potential users will be served on a first come basis which does imperfectly reflect their WTP for the fast road. It is thus preferable to charge for road access, the solution adopted for modern time-saving highways in many cities, at a level that will on average equate usage to capacity. If the metering technology is not too costly, it is even better to auction permits and have permit holders install a wireless transmitter to access the road faster.

Mohring (1972) makes the previous paradox even more striking by considering mass transit as the alternative to the congestion prone route. Since the train or bus operator waits for a train to be full to let it circulate (in order to maximize revenue at given cost), an increase in demand is met by increasing the frequency of trains or buses. There is thus a positive externality in consumption since the waiting time of users will be reduced. If we let the maximum travel time on route B be $\bar{t}$ (minimum train frequency), then the general formula is $t_B = \bar{t} - \beta d_B$ which is displayed on the right panel of Figure 25.5 (it is increasing because the horizontal axis shows $d = 1 - d_B$).

The equilibrium distribution between the two transportation modes continues to be driven by the no-arbitrage condition. When road capacity is increased from k_1 to k_2, car travel time drops, thus some people arbitrate and leave mass transit so that train frequency worsens and even more people switch to use their car. In equilibrium of this adjustment process, road patronage increases, as before and, but travel time $\hat{t}$ also increases from t_1 to t_2.

The paradoxes arise because there exist alternatives to the congestion prone service; the distribution of users between the available options then obey to no-arbitrage conditions which can trigger perverse effects.

25.4.4 Cost Benefit Analysis †

Data

Prud'homme and Sun (2000) assess the cost of congestion on the most used road in France, the Paris freeway ring (*boulevard périphérique*) using detailed data from 1996. The bottom of Figure 25.6, with inverted axis, displays the frequency distribution of speed across the day. We observe that free flowing traffic is the most frequent situation (55% of the minutes). At that maximum speed, slightly above the 80km/h legal limit, there is one car every 100m on each lane or one car passing trhough every 4 seconds. Another quarter of the day is congestion free although travel speeds are reduced, finally one fifth of the time is problematic (speed lesser then 50km/h).

The output of the transit system is measured by the flow (right axis, heavy line), the

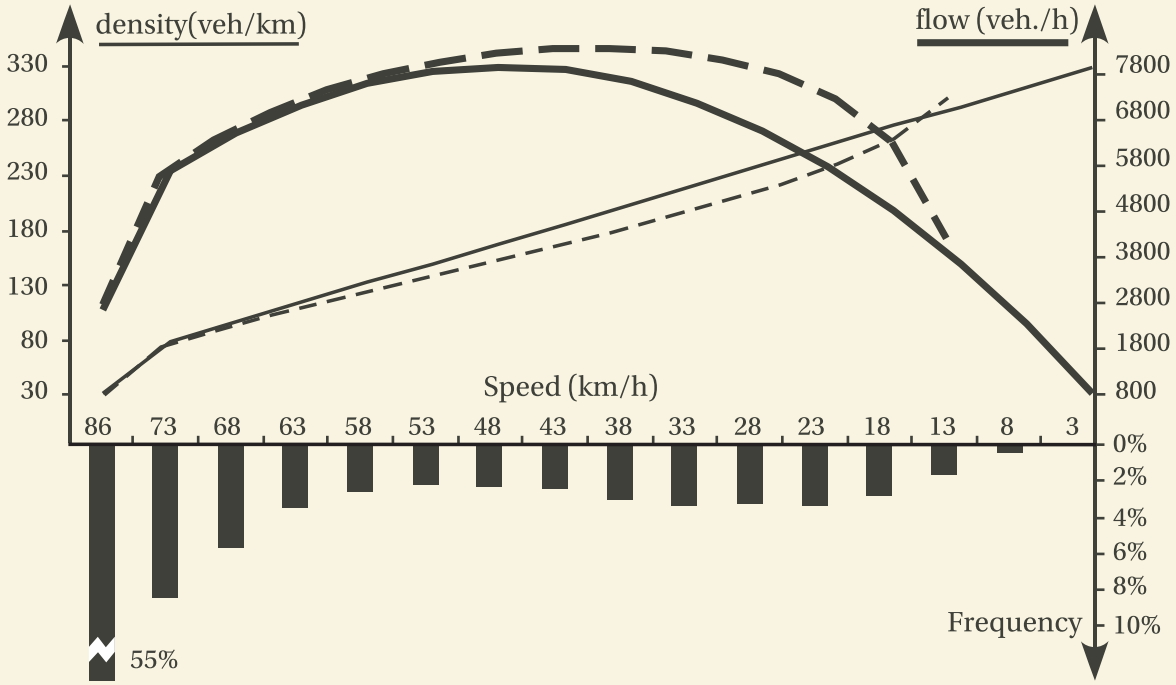


Figure 25.6: Distribution of speed on the Paris freeway ring

number of vehicles able to pass between two exits every hour; it is, at every moment, the product of speed (km/h) by density (vehicles/km) respectively displayed on the horizontal and left axis. There is congestion when the flow is forced below its maximum because the capacity of the system has been reached. Let us explain how this happens.

When more vehicles enter the freeway, starting around 6h in the morning, density increases in the sense that every vehicle comes closer to its predecessor; average speed thus starts to decrease because drivers adopt a cautious attitude. The relationship between the two magnitudes is displayed as the light curve and is almost linear i.e., putting 1 more car per lane on a one kilometer section reduces speed by roughly 1km/h. This phenomenon is probably best explained by the desire of every driver to maintain approximatively a 20m distance with the preceding vehicle. Yet the influx of new cars on the ring increases the flow steadily because traffic is nevertheless quite fluid. The maximum flow is reached for an approximate speed of 50km/h when the time interval between two cars falls down to 1.3 second which is the minimum acceptable by an average driver for his/her security.[26@] From that moment on, a 1% increase of density due to the entry of new cars, imposes a strong negative speed externality on all current users i.e., a speed reduction greater than 1%. Since flow is the product of speed by density, it must decrease which means that congestion is now present on the system.

Theory

The relation between speed s and density q is estimated by the linear regression $s = \bar{s} - \alpha q$ where $\bar{s}$ is the maximum speed on the road (it ought to be the legal limit) and α is the rate of speed decrease with density increase. Since the flow is $f \equiv s \times q$, its maximum is reached for $q = \frac{\bar{s}}{2\alpha}$, leading to a maximum flow of $\bar{f} = \frac{\bar{s}^2}{4\alpha}$ per lane. For the Paris experiment, $\bar{s} = 94$km/h and $\alpha = 0.84$ leading to $\bar{f} = 2650$ vehicles per hour per lane (the figure on the graph is trice because the Paris ring has tree lanes).

For a given hour of the day, the average user has a demand for the transit service increasing with speed or decreasing with density (since the two variables are inversely related); in other words his WTP[27@] for using the road at that moment is a decreasing function P of density q. The per km cost of using the road is the capital cost of his car r (including wear and tear) plus the fuel and time cost which are inversely proportional to speed, hence the (individual) per km cost is $c_i(q) = r + \frac{\beta}{s} = r + \frac{\beta}{\bar{s} - \alpha q}$. The social cost per km sums the cost of all users present at the moment, it is thus $C_s(q) = q c_i(q)$. The social value of using one km of road sums the WTP of all the current users, i.e., $V(q) = \int_0^q P(x)\,\mathrm{d}x$. Welfare being $W(q) = V(q) - C_s(q)$, it is maximum at q^* when $V' = C_s' \Leftrightarrow P(q) = c_i(q) + q c_i'(q) \equiv c_s(q)$ which is the social cost per km of adding one more user. Since the individual cost c_i is increasing with density, the social unit cost is greater than the individual one i.e., $c_s > c_i$.

It is now clear that the road is subject to the tragedy of the commons; in a free access equilibrium, each potential user compares his individual cost at the expected level of traffic to his own WTP ignoring that his decision to join will worsen traffic conditions for all. A commuter uses his car if $P \geq c_i$; such a behavior leads to an open access equilibrium q^o characterized by the equality $P(.) = c_i(.)$; we thus observe on Figure 25.7 an excessive density. The cost of congestion is then $W(q^*) - W(q^o)$ and since welfare is the area between the demand curve and the social cost curve, the loss due to the change from q^* to q^o is the gray area on Figure 25.7.

Notice that the social cost computed at the inefficient density $\bar{p} = c_s(q^o)$ exaggerates the efficient price p^* which is used to determine the efficient density. This observation can be refined as follows. By writing $W(q) = V(q) - C_s(q) = \int_0^q P(x)\,\mathrm{d}x - q c_i(q)$, we see that welfare is the area between the demand curve and the price, hence $W(q^*) = V(q^*) - q^* \hat{p}$ and $W(q^o) = V(q^o) - q^o p^o$. The welfare loss in the open access equilibrium has thus a second geometrical representation; it is the striped rectangle minus the striped triangle. An upper bound is then $q^*(p^o - \hat{p})$ which is considerably smaller than the very simplistic estimate $q^o(\bar{p} - p^o)$ computed using the observed data; the latter would be correct only if the demand for road use was extremely inelastic (the demand curve becomes vertical) which is not the case since empirical studies conclude to a value between -1 and -0.6.

Lastly, Figure 25.7 displays as well as an off-peak demand D_l to recall us that the issue of congestion cost makes sense at peak time only.

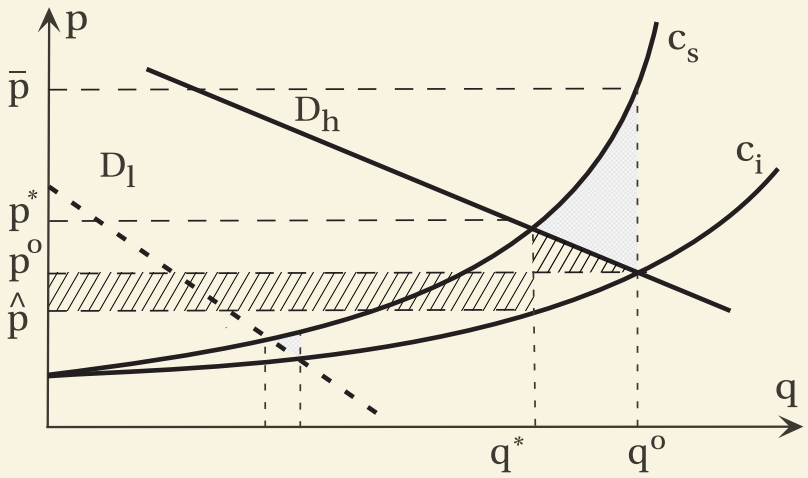


Figure 25.7: Congestion Cost Estimation

Using the above experimental data, the efficient density and flow can be computed for all classes of demand. The resulting curves are shown in dashed on Figure 25.6. In the open access equilibrium, the cost of congestion (welfare loss) represents 9.2% of the total cost of using the freeway but 17.2% of the current welfare. The aforementioned estimation error would lead to a figure three times larger than the true congestion cost. The efficient situation is simply characterized by having around 20% less cars using the freeway as soon as speed is less than 60km/h (on average density is reduced by 14%). This would guarantee that speed never falls below 14 km/h and one would achieve a 3% traffic increase, a 6% speed increase, a 16% time saving and a 9% cost reduction. If we now add the external cost of pollution at 36€ per thousand vehicle km, then efficient levels of road use are reduced by a much larger 57% with a heavy contribution (i.e., reduction) at high speeds (and low demand). This might seem incoherent but one must remember that society values very much using the ring at moments of peak demand, thus density is optimally reduced by a factor one third only. On average, accounting of pollution leads to reduce traffic by 35% together with a 48% reduction of the time spend on the system and a 44%reduction of expenditure.

Appendix

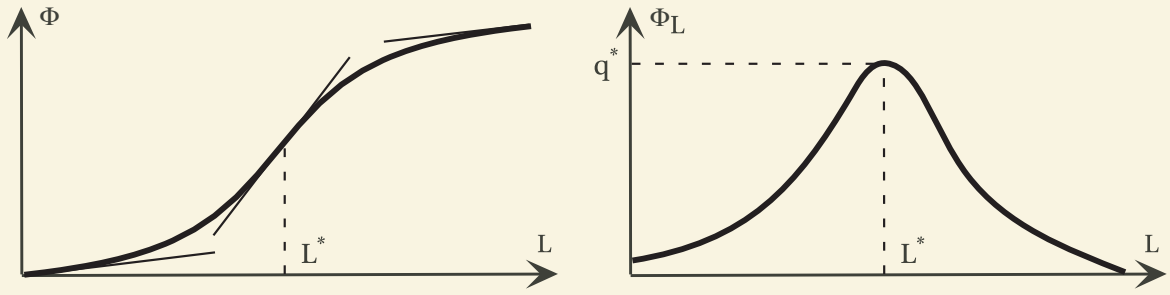
Chapter 26

Appendix

26.1 Miscellanies

26.1.1 Returns to Scale

Turgot (1844) explains how labour L (variable input) used to till a plot of land (fixed input) would initially be more and more productive; graphically, the slope of Φ which is labour productivity Φ_L increases. This process does not continue ad-infinitum because the land becomes exhausted; labour productivity then starts to decrease which constitutes the "law of decreasing productivity". Notice the difference between this concept involving partial differentiation and returns to scale which involve total differentiation.



In fact, we never observe the increasing productivity part of Φ_L because the previous curve is constructed for a small plot of land e.g., an are (100 square meters). The farmer then optimizes his total effort as $L = nL^*$ where n is the number of ares he decides to put for cultivation. If he owns a small number of ares $\underline{n}$ such that his total time is $\bar{L} > \underline{n}L^*$ then his apparent marginal productivity of labour will be constant up to $\underline{n}L^*$ and then decreasing. Obviously, plots are never of homogeneous qualities and the best ones (those with highest q^*) are used first so that visible productivity (across a region) is always decreasing.

26.1.2 Constant Elasticity of Substitution

Under CRS, if we denote $x \equiv \frac{K}{L}$ then $Q = \Phi(K,L) = L\Phi(x,1)$ so that defining $f(x) \equiv \Phi(x,1)$, the production equation reads $y = f(x)$. The marginal productivities of capital and labour are $\Phi_K = Lf'\frac{\partial x}{\partial K} = f'$ and $\Phi_L = f + Lf'\frac{\partial x}{\partial L} = f - Lf'\frac{K}{L^2} = f - xf'$. The elasticity of substitution σ of K into L, is the elasticity of $\frac{K}{L}$ with respect to the $MRS_{K/L} = \frac{\Phi_L}{\Phi_K}$ along an isoquant i.e.,

$$\begin{aligned}\frac{1}{\sigma} &= \frac{\partial MRS_{K/L}}{\partial K/L}\frac{K/L}{MRS_{K/L}} = \frac{xf'}{f-xf'}\frac{\partial}{\partial x}\left(\frac{f-xf'}{f'}\right)\\ &= \frac{xf'}{f-xf'}\frac{-xf'f'' - f''(f-xf')}{f'^2} = \frac{-xff''}{f'(f-xf')}\end{aligned}$$

Since in equilibrium the relative wage w equates the labour productivity we have $w = \Phi_L = f - xf'$, thus $\frac{dw}{dx} = -xf''$ so that using $\frac{dy}{dx} = f'$ we discover that

$$b = \frac{dy}{dw}\frac{w}{y} = \frac{dy}{dx}\frac{dx}{dw}\frac{w}{y} = \frac{f'(f-xf')}{-xff''} = \sigma$$

Let us now characterise the production functions satisfying exactly (2.10). Using $w = f - xf'$, we can write

$$\begin{aligned}&\ln y = \ln a + b\ln\left(y - xy'\right) \Leftrightarrow y^{1/b} = a^{1/b}\left(y - xy'\right)\\ \Leftrightarrow\quad & y' = \frac{a^{1/b}y - y^{1/b}}{a^{1/b}x} = \frac{y/\alpha - y^{1/b}}{x/\alpha} = \frac{y(1-\alpha y^\rho)}{x} \text{ with } \alpha \equiv a^{-1/b}, \rho = 1/b - 1\end{aligned}$$

Now, using the fact that $y' = \frac{dy}{dx}$, we have

$$\frac{dx}{x} = \frac{dy}{y\left(1-\alpha y^\rho\right)} = \frac{dy}{y} + \frac{\alpha y^{\rho-1}dy}{1-\alpha y^\rho} \Leftrightarrow \ln x = \ln y - \frac{1}{\rho}\ln\left(1-\alpha y^\rho\right) + \frac{1}{\rho}\ln\beta$$

for some integration constant β. Taking exponential, we obtain

$$\begin{aligned}x &= \beta^{\frac{1}{\rho}} y\left(1-\alpha y^\rho\right)^{\frac{-1}{\rho}} \Leftrightarrow x^\rho = \beta\frac{y^\rho}{1-\alpha y^\rho} \Leftrightarrow y^\rho = \frac{x^\rho}{\beta + \alpha x^\rho}\\ \Leftrightarrow\quad & y = x\left(\beta + \alpha x^\rho\right)^{\frac{-1}{\rho}} = \left(\beta x^{-\rho} + \alpha\right)^{\frac{-1}{\rho}}\\ \Rightarrow\quad & Q = L\left(\beta K^{-\rho}L^\rho + \alpha\right)^{\frac{-1}{\rho}} = \left(\beta K^{-\rho} + \alpha L^{-\rho}\right)^{\frac{-1}{\rho}} = \gamma\left(\delta K^{-\rho} + (1-\delta)L^{-\rho}\right)^{\frac{-1}{\rho}}\end{aligned}$$

with $\gamma \equiv (\alpha+\beta)^{\frac{1}{-\rho}} \Leftrightarrow \alpha + \beta = \gamma^{-\rho}$ and $\delta \equiv \beta\gamma^\rho$. Since the elasticity of substitution is $b = \frac{1}{1+\rho}$, ρ is called the substitution parameter, γ is the efficiency parameter (that can be normalized to unity for a single firm) and δ is the distribution parameter.

26.1.3 Rubinstein's bargaining model

There are two extreme asymmetric Nash equilibria whereby one player, either S or N stubbornly asks the whole prize and refuses anything different ad infinitum; in equilibrium, he obtains his demand. This is an unrealistic outcome because part of the prize is lost once the first offer has been refused. We thus look for a subgame perfect equilibrium (SPE).

Consider a SPE where players keep repeating the same offers (stationarity) and accept the equilibrium offer immediately (no delay). We do not know yet if it exists but we shall pinpoint it. For $i = S, N$, let δ_i and q_i be the discount factor and equilibrium offer of player i. The present value for player i of rejecting q_j is $\delta_i q_i$ since he will be able to make his equilibrium offer at the next round and it will be accepted. As we are assuming that rejection is dominated, it must be the case that $1 - q_j \geq \delta_i q_i$. However, if the inequality was strict, then player j could ask more for himself and still see his offer accepted. From the system of equation so obtained we deduce $q_i^* = \frac{1-\delta_j}{1-\delta_i\delta_j}$.

We now exhibit a strategy profile that form a no delay stationary SPE: player $\#i$ offers systematically q_i^* and accept anything lesser than $1 - \delta_i q_i^*$. When player j follows this strategy, player i gains q_i^* by offering q_i^* (which is accepted) whereas he earns less if he deviates. If player i is at the point of receiving the offer q_j^*, it is trivial to see that he better accepts it.

Lastly, it remains to be shown that there are no other SPEs. To avoid clutter we consider a common discount factor. Let p be the greatest payoff the offerer, say S, gets in any SPE starting at $t = 2$. Then at $t = 1$, in any SPE, S must accept the offer δp by N. Hence, at $t = 1$, $u_N \geq 1 - \delta p$ in any SPE. But then at $t = 0$, $u_N \geq \delta(1 - \delta p)$ so that $u_S \leq 1 - \delta(1 - \delta p)$. Since the subgames at $t = 0$ and $t = 2$ are the same, $p = \delta(1 - \delta p) \Rightarrow p \leq \frac{1}{1+\delta}$. The same reasoning with the lowest payoff q yields $q \geq \frac{1}{1+\delta}$. Since $q \leq p$ by construction, there is a unique SPE payoff $q = p = \frac{1}{1+\delta}$.

26.1.4 Coase Conjecture

Result 1 *When limited by a capacity constraint, an inter-temporal price discriminating monopolist cannot earn less than* 30% *of the rental monopoly profit.*

A durable-good monopolist sells to a continuum of consumers, each demanding a single unit. The WTP of individual with index $q \in [0, 1]$ is $v(q) \in [\bar{v}; \underline{v}]$ with $\underline{v} > 0$ and $v' < 0$.[1@] The monopolist's revenue function, $R(q) \equiv q\,v(q)$ is assumed to reach its maximum R^* at some $q^* > 0$. Both consumers and the monopolist live forever and discount the future at the rate r. Sales can occur every τ minutes. The discount rate per period is thus $\delta \equiv e^{-r\tau}$.

Let T be the last period of sales. Observe first that the price listing $(p_t)_{t \leq T}$ must be

decreasing because no one would wait to buy dearer. Next, it must be the case that $p_T = \underline{v}$ i.e., the last consumer buys the good. Indeed, if we had $p_T > \underline{v}$, then some people would be left unserved which is irrational for the firm as she can extend sales for one additional period and make an additional profit. That $p_T < \underline{v}$ cannot be true is also obvious because it would amount to make undue rebates to the last buyers. Our third observation is that the marginal buyer at time t must be indifferent between buying now and later if he derives the same net present utility. Denoting v_t the WTP of such a person, it must satisfy $v_t - p_t = \delta\left(v_t - p_{t+1}\right) \Rightarrow p_t = (1-\delta)\, v_t + \delta p_{t+1}$, hence

$$p_t = \delta^{T-t}\underline{v} + (1-\delta)\sum_{j=0}^{T-1-t}\delta^j v_{t+j} \tag{26.1}$$

From our three observations, we deduce that the only strategic choice for the monopolist is to choose the duration of sales T. A crucial feature of the Coase conjecture is the ability to sell any amount at any time. Such an untenable assumption is debunked by assuming that the firm must choose (at no cost) an instantaneous service capacity K before the start of sales; it enables to sell up to $K\tau$ units in every period. Since sales take place during T periods and the entire market of unit size is served, we have $TK\tau = 1$ so that there is only one variable to choose, either T or K.

The monopolist equilibrium behavior (capacity choice and price listing) is hard to characterize, thus we study her profit along the sub-optimal behavior consisting in selling exactly at capacity in each period i.e., $q_t = K\tau = \frac{1}{T}$. From this simple accounting statement, we deduce the WTP of the marginal buyer in each period: $v_t = v(t/T)$. The profit along the sup-optimal path (security option) is

$$\bar{\pi} \equiv \sum_{t=1}^{T}\delta^{t-1}p_t q_t \geq \frac{1-\delta}{T}\sum_{t=1}^{T}\delta^{t-1}\sum_{j=0}^{T-1-t}\delta^j v_{t+j} = \frac{1-\delta}{T}\sum_{t=1}^{T-1}\delta^{t-1}t v_t$$

as the double summation simplifies. Let $s \equiv q^* T$ be the time necessary to sell the monopoly quantity at which point $v_s = v(q^*)$. Since $T \geq s$ and v_t is decreasing, we can limit the summation index to s instead of T and substitute v_s for v_t. We thus obtain

$$\bar{\pi} \geq \frac{(1-\delta)v_s}{T}\sum_{t=1}^{s-1}t\delta^{t-1} = R^*\frac{1-\delta^s(1+s(1-\delta)/\delta)}{s(1-\delta)} = R^*\frac{1-\delta^s(1+z/\delta)}{z}$$

using $R^* = q^* v_s = \frac{s v_s}{T}$ and introducing $z \equiv s(1-\delta)$. By l'Hospitals rule, $\beta \equiv \frac{r\tau}{1-\delta} = \frac{r\tau}{1-e^{-r\tau}} \underset{\tau\to 0}{\longrightarrow} 1$, hence $\delta^s = e^{-r\tau s} = e^{\frac{-r\tau z}{1-\delta}} = e^{-z\beta} \underset{\tau\to 0}{\longrightarrow} e^{-z}$. Thus, at the limit where the time interval τ vanishes (and $\delta \to 1$),

$$\frac{\bar{\pi}}{R^*} \geq \gamma_z = \frac{1-e^{-z}(1+z)}{z}$$

whose maximum is $\gamma^* \simeq 0.3$ for $z^* \simeq 1.8$.[2@] Lastly, we can compare the capacity corresponding to this minimum profit with the monopoly sales.[3@] As $q^* T^* = s^* = \frac{z^*}{(1-\delta)}$, we have $\frac{K^*}{q^*} = \frac{1}{q^* T^* \tau} = \frac{1-\delta}{\tau z^*} = \frac{r}{\beta z^*} \underset{\tau \to 0}{\longrightarrow} \frac{r}{z^*} \simeq .56r$. Given that the subjective interest rate r is typically less than 10%, the service capacity is less than 5% of the rental monopoly sales.

Continuous time One can check that up to $1-\delta$ which vanishes to zero with τ, $\bar{\pi} = \sum_{t=1}^{T-2} \left(v_{t+1} \frac{t+1}{T} - v_t \frac{t}{T}\right) \delta^t$ i.e., the discounted value of marginal revenue. This formula exemplifies the sources of profit loss for the monopolist. Firstly, because she can't commit to stop selling at any point in time, she ends up selling too much i.e., beyond s to people with $R_m < 0$. Secondly, profitable clients (those with WTP greater than v_s) only buy gradually since K^*, the per period sales, is a tiny fraction of the straight monopoly sales q^*.

In equilibrium, buyers are ordered thus buyer Kt buys at time t paying price p_t. Since this is an optimal behavior, his utility $u(Kt) = e^{-rt}(p_t - v(Kt))$ maximizes the utility from buying at an alternative date s which is $e^{-rs}(p_t - v(Ks))$, hence, taking derivative, $u'(Kt) = e^{-rt} v'(Kt)$ must hold. Then using $p_T = \underline{v} \Rightarrow u(KT) = 0$, we integrate to get $u(Kt) = -\int_t^T e^{-rs} v'(Ks) K \,\mathrm{d}s = \int_{Kt}^{\bar{q}} e^{-rx/K} v'(x) \,\mathrm{d}x = e^{-rt} p_t - \int_{Kt}^{\bar{q}} \frac{r}{K} e^{-rx/K} v(x) \,\mathrm{d}x$ (integrating by parts). Given the definition of utility, we derive $p_t = e^{rt} \int_{Kt}^{\bar{q}} a e^{-ax} v(x) \,\mathrm{d}x$ with $a \equiv \frac{r}{K}$.

Now, firm profit is

$$\pi = \int_0^T e^{-rt} p_t K \,\mathrm{d}t = \int_0^{\bar{q}} a x e^{-ax} v(x) \,\mathrm{d}x = \int_0^{\bar{q}} R_m(x) e^{-ax} \,\mathrm{d}x = \int_0^{\bar{q}} a R(x) e^{-ax} \,\mathrm{d}x$$

using integration by parts in the second and last equalities (with $R(\bar{q}) = R(0) = 0$). Now, $\pi \geq \int_0^{q^*} \frac{ax}{q^*} R(q^*) e^{-ax} \,\mathrm{d}x$ since $v(x) = R(x)/x$ is decreasing. Thus $\frac{\pi}{R(q^*)} \geq \int_0^{q^*} \frac{ax}{q^*} e^{-ax} \,\mathrm{d}x = \frac{1 - e^{-aq^*}(1+aq^*)}{aq^*} \equiv \gamma$ which can be maximized in $z \equiv aq^*$ at $z \simeq 1.79$ yielding $\pi \geq 0.3 R(q^*)$.

26.1.5 Oligopoly

Innovation in the Cournot Oligopoly

Extending (12.5) to oligopoly using the Cournot oligopoly equilibrium quantity (5.14), we obtain

$$q_i = \frac{a - b(nc_i - (n-1)c_{-i})}{n+1} = \frac{a - bc + b(nx_i - (n-1)x_{-i})}{n+1}$$

where c_{-i} (resp. x_{-i}) is the average marginal cost (resp. R&D level) of firm i's opponents. The FOC of optimal R&D is thus $\frac{2n}{n+1} q_i = \frac{\partial \pi_i}{\partial x_i} = \frac{\partial \psi}{\partial x_i} = 2\lambda x_i$. In a symmetric equilibrium, $x_i = x_{-i} = x$, so that the FOC becomes $n(a - bc + bx) = (n+1)^2 \lambda x$ leading to the symmetric equilibrium level $\hat{x} = \frac{n(a-bc)}{\lambda(n+1)^2 - bn}$; the ensuing equilibrium quantity is then $\hat{q} = \frac{(a-bc)(n+1)\lambda}{\lambda(n+1)^2 - bn}$

with profit $\hat{\pi} = \frac{1}{b}\hat{q}^2 - \lambda\hat{x}^2 = \lambda\hat{x}^2\left(\frac{\lambda(n+1)^2-2bn^2}{2bn^2}\right)$. We obviously need to assume $\lambda > \frac{bn}{(n+1)^2}$ to avoid the trivial zero R&D corner equilibrium.

Regarding comparative statics, we note that investment, quantity and profit decrease with the quality of R&D (λ) and the number of firms (n). Indeed, we obviously have $\frac{\partial\hat{x}}{\partial\lambda} < 0$ while algebra reveals that $\frac{\partial\hat{x}}{\partial n} \propto \lambda\left(1-n^2\right) < 0$, $\frac{\partial\hat{q}}{\partial n} \propto b - \lambda(n+1)^2 < 0$ and $\frac{\partial\hat{q}}{\partial\lambda} \propto -nb(n+1) < 0$ (profit derivatives follow likewise). The effect of changes in a and b are as in the duopoly case.

Innovation in a Price Oligopoly

The original duopoly of Hotelling (1929) is extended by Raith (2003) to oligopoly using the circular city model seen in §11.1.3. His conclusions are achieved using management incentives of the form given in §20.2 and more specifically equation (20.5) (his equation (A3)). Our identification of R&D investment with many types of innovative activities thus readily generalize his findings.

When the n active equidistant firms have potentially different cost c_i, equation (11.10) characterizing the best reply of firm i is $\frac{1}{2}\left(-p_{i-1}+4p_i-p_{i+1}\right) = z_i$ where $z_i \equiv \frac{t}{n} + c_i$. In matrix form, it reads $A \times p = z$ where the entries of line i of A are $a_{ii} = 2$, $a_{i,i-1} = a_{i,i+1} = -\frac{1}{2}$ and $a_{ij} = 0$ for $j \neq i-1, i-i+1$. Matrix A is circulant i.e., each row takes the previous one and shifts it one place to the right. The inverse exists and is also circulant; its first row has coefficients $\frac{1}{\sqrt{3}}\frac{(2+\sqrt{3})^k+(2+\sqrt{3})^{n+1-k}}{(2+\sqrt{3})^{n+1}-1}$ for $k = 1$ to n.

Since A is invertible, the price equilibrium is $p = A^{-1}z$ and we can write

$$p_i = \frac{t}{n} + \gamma c_i + \delta c_{-i} \tag{26.2}$$

where c_{-i} is a weighted average of the cost of i's competitors. For instance, if $n = 3$, $\gamma = \frac{3}{5}$ and $\delta = \frac{2}{5}$ while for $n \geq 5$, $\gamma \to \frac{1}{\sqrt{3}} \simeq 0.58$ and $\delta \simeq 0.31$.[4@] What is fundamental in this equation is the presence of δ showing that the equilibrium price is sensitive to the industry cost structure, not just the own technology.

Let us plug this result into profit to investigate the optimal cost reduction. Profit is $\pi_i = (p_i - c_i)D_i$, thus $\frac{\partial\pi_i}{\partial c_i} = (\gamma-1)D_i + (p_i-c_i)\frac{\partial D_i}{\partial c_i}$ by (26.2). To compute $\frac{\partial D_i}{\partial c_i}$, observe that

$$D_i = \frac{m}{2t}\left(\frac{2t}{n} + p_{i+1} + p_{i-1} - 2p_i\right) = \frac{m}{2t}\left(\frac{2t}{n} + (\delta - 2\gamma)c_i + ...\right) \tag{26.3}$$

as the coefficient of c_i in either p_{i+1} or p_{i-1} is $\delta/2$ (cf. footnote 26.3). We thus get $\frac{\partial D_i}{\partial c_i} = (\delta/2-\gamma)\frac{m}{t}$ and $\frac{\partial\pi_i}{\partial c_i} = (\gamma-1)D_i + (\delta/2-\gamma)\frac{m}{t}(p_i-c_i)$. In a symmetric investment equilibrium, firms have identical costs thus the price equilibrium is also symmetrical so that $D_i = \frac{m}{n}$

and $p_i = \frac{t}{n} + c - x$, hence

$$\frac{\partial \pi_i}{\partial c_i} = (\gamma - 1)\frac{m}{n} + (\delta/2 - \gamma)\frac{m}{t}\frac{t}{n} = (\delta/2 - 1)\frac{m}{n} \simeq -0.85\frac{m}{n} \tag{26.4}$$

i.e., the strategic effect of R&D investment over price shrinks the direct effect $\frac{m}{n}$ by a minimum of 15% thus proving the claim made in the text. Notice though that Raith (2003), in his equation #2 uses the FOC instead of the equilibrium formula (26.2) i.e., sets $\gamma = \frac{1}{2}$ and $\delta = 0$ and thus fails to identify the strategic effect of competition upon investment incentives.

The FOC for R&D investment at a symmetric equilibrium is thus $0.85\frac{m}{n} = \lambda x \Rightarrow \hat{x} = 0.85\frac{m}{\lambda n}$. Profits are then $\hat{\pi} = \frac{m}{n^2}\left(t - \frac{0.85^2 m}{2\lambda}\right) - F$ where F is the fixed cost of operation. If the transportation cost falls ($t \searrow$) or products become less differentiated or entry cost rise ($F \nearrow$), profit falls thus generates exit ($n \searrow$) which in turn boosts the sales and innovation of the remaining active firms. A market size expansion ($m \nearrow$) boosts profits but at a less than proportional rate, thus entry occurs with an elasticity lesser than unity so that innovation, being driven by the ratio of m over n, increases (although there are new firms, old ones still manage to produce more which spurs innovation).

State of Nature

Bush and Mayer (1974) consider the original state of nature, anarchy, where private property does not exist and agents can invest some of their scarce resources to rob others. Since initial income is exogenously given, this is not a fully fledged general equilibrium model. Once the Nash equilibrium is characterized and shown to involve wasteful offensive and defensive investment, the authors look for the conditions enabling a jump out of anarchy towards a peaceful and orderly society without theft.

Each person has a differential endowment $\bar{k}_i$ of (capital) income and consumes $k_i = \bar{k}_i + q_i - q_{-i}$ where q_i is preying upon others while q_{-i} is the average preying from others. Differentiated theft ability is measured by the maximum preying resource $\bar{l}_i$. Letting $l_i = \bar{l}_i - q_i$ denote leisure and $w_i = \bar{k}_i + \bar{l}_i$ full income, the budget constraint reads $w_i = k_i + l_i + q_{-i}$ i.e., income is divided between production, leisure and loss from theft. Assuming that people only care for consumption and leisure, utility maximization leads to demands that are increasing in the net income $w_i - q_{-i}$. From $l_i = \phi_i(w_i - q_{-i})$, we deduce the appropriation best reply $q_i = \bar{l}_i - \phi_i(w_i - q_{-i})$ to the effort q_{-i} of others. A Nash equilibrium in this context is called a natural equilibrium. If $\phi_i(w_i) < \bar{l}_i$ holds true then the equilibrium features wasteful appropriative activities. This condition states that in a world of non wasteful people, one would want to spend part of his/her time to appropriate others' wealth.

In the Cobb-Douglas case, $u(k,l) = k^\alpha l^{1-\alpha}$, thus $\phi(w) = (1-\alpha)w$ so that the previous condition reduces to $(1-\alpha)/\bar{l}_i < \alpha/\bar{k}_i$ i.e., the marginal utility of leisure is lesser than that of investment (at the respective endowments). This is likely to hold true for those with little capital or very productive at appropriation or favoring consumer goods over leisure. The derivation of the equilibrium follows the Cournot method of §5.1.3. The best reply equation simplifies into $q_i = \alpha\bar{l}_i - (1-\alpha)\left(\bar{k}_i - q_{-i}\right)$. Summing over all participants yield $\hat{q} = \alpha\hat{l} - (1-\alpha)\left(\hat{k} - \hat{q}\right) \Rightarrow \hat{q} = \hat{l} - \frac{1-\alpha}{\alpha}\hat{k}$ where $\hat{\cdot}$ denotes average. Plugging $q_{-i} = \frac{nq^* - q_i}{n-1}$ into the best reply equation enables to solve for individual levels.

26.2 Risk and Uncertainty

26.2.1 Time Discounting

Result 2 *The present value of* $1€$ *to be paid at time* t *with interest rate* r *is* e^{-rt}.

Consider an asset whose value at time t is $v(t)$. The percentage increase of value over a period of length τ is $\frac{v(t+\tau)-v(t)}{v(t)}$, thus the interest rate over this period is $\frac{v(t+\tau)-v(t)}{\tau v(t)}$. Letting τ tend to zero, we obtain the instantaneous interest rate at time t as $r(t) \equiv \frac{dv/dt}{v}$. Integrating this equation yields $v(t) = v(0)e^{\int_0^t r(\tau)\,\mathrm{d}\tau}$. If we consider now a bond paying a fixed amount per period of time, then $r(t)$ is a constant r so that the return of $1€$ deposited for a length of time t is exactly $v(t) = e^{rt}$. The present value of $1€$ to be paid within time t is thus e^{-rt}.

26.2.2 Euler Equation

Result 3 *Solution of* $\max_x \int_0^\infty v(K_t, I_t)e^{-rt}\,\mathrm{d}t$ *under the restriction* $K_0 = \underline{K}$ *and* $K_T = \bar{K}$.

We use Lagrange's calculus of variations to solve $\max_x \int_0^T f(t,x,\dot{x})\,\mathrm{d}t$ s.t. $x_0 = \underline{x}, x_T = \bar{x}$ where x is a function of time, $\dot{x}$ its time derivative, $T \leq +\infty$ the horizon and $f(t,x_t,\dot{x}_t)$ the differentiable objective. Assume an optimal path x^* exists and consider y s.t. $y_0 = 0$ and $y_T = 0$ so that the distorted path $x^* + \epsilon y$ is still admissible for any ϵ. The optimality of x^* means that $J(\epsilon) \equiv \max_x \int_0^T f(t, x^* + \epsilon y, \dot{x}^* + \epsilon\dot{y})\,\mathrm{d}t$ reaches a maximum at 0 thus[5@]

$$\begin{aligned} 0 = J'(0) &= \int_0^T y f_x(t,x^*,\dot{x}^*)\,\mathrm{d}t &+ \int_0^T \dot{y} f_{\dot{x}}(t,x^*,\dot{x}^*)\,\mathrm{d}t \\ &= \int_0^T \left(f_x - \dot{f}_{\dot{x}}\right) y\,\mathrm{d}t &+ f_{\dot{x}_T} y_T - f_{\dot{x}_0} y_0 \end{aligned}$$

using integration by parts. By construction of y, the last two terms are nil, thus the integral is nil. As this is true whatever the shape of y, the parenthesis must also be nil and this constitutes the Euler equation[6@] $\frac{df}{\mathrm{d}x} = \frac{d}{dt}\left(\frac{df}{d\dot{x}}\right)$. The transversality condition

$x_T f_{\dot{x}_T} = 0$ evaluated at an optimal solution is necessary, hence when there is no reason to impose the terminal condition $x_T = \bar{x}$ (aka free boundary), we still need $f_{\dot{x}_T} = 0$. For economic applications, we have $x = K$, $\dot{x} = I$ and $f(t, x, \dot{x}) = e^{-rt} v(K, I)$, hence (18.3).

26.2.3 Risk Aversion

Result 4 *Constant absolute risk aversion (CARA) utility function:* $u(m) = -e^{-\rho m}$

If ρ is a positive constant, the differential equation $\frac{u''}{u'} = -\rho$ has solution $\ln u'(m) = a - \rho m$ where a is a constant. Next, check that $u(m) = c - be^{-\rho m}$ for constants $b > 0$ and c. Since the expected utility function is unique up to an affine transformation, the result follows.

Result 5 *Pratt (1964)'s Theorem that the following are equivalent:*

① *Mister v is more risk averse than Mister u.*

② *v is a concave transformation of u.*

③ *Mister v's ARA index is larger than Mister u's.*

④ *Mister v's risk premium is larger than Mister u's.*

• ①→② By contradiction: if $\phi(z) \equiv v(u^{-1}(z))$ is not everywhere concave, it is locally convex at some w. Let us choose gamble $\tilde{x}$ indifferent to u i.e., $\mathbb{E}[u(w + \tilde{x})] = u(w)$, then $\mathbb{E}[v(w + \tilde{x})] = \mathbb{E}\left[\phi(u)(w + \tilde{x})\right] > \phi(\mathbb{E}[u(w + \tilde{x})]) = \phi(u(w)) = v(w)$, by Jensen's inequality, thereby contradicting ① since v ought to refuse this gamble.
• ②→③: As $\phi' = \frac{v'(u^{-1})}{u'(u^{-1})}$, we derive $\phi'' \propto v''u' - v'u'' < 0 \Leftrightarrow \rho_v > \rho_u$.
• ③→④: For every fair gamble $\tilde{x}$, letting $\tilde{z} \equiv u(w + \tilde{x})$, there exits a risk premium μ_u such that $u(w - \mu_u) = \mathbb{E}[\tilde{z}]$. Likewise $\exists \mu_v$ such that $v(w - \mu_v) = \mathbb{E}[v(w + \tilde{x})] = \mathbb{E}\left[v(u^{-1}(\tilde{z}))\right] = \mathbb{E}\left[\phi(\tilde{z}))\right] \leq \phi(\mathbb{E}[\tilde{z}])$ by Jensen's inequality, hence $w - \mu_v \leq v^{-1}(\phi)(\mathbb{E}[\tilde{z}]) = u^{-1}(\mathbb{E}[\tilde{z}]) = w - \mu_u$.
• ④→①: consider the lottery $\tilde{z} \equiv \mu_v + \tilde{x}$ and initial wealth $w_0 \equiv w - \mu_u$. We know by definition of μ_u that agent u does not reject this lottery, yet $\mathbb{E}[v(w_0 + \tilde{z})] = \mathbb{E}[v(w + \tilde{x})] = v(w - \mu_v) \leq v(w - \mu_u) = v(w_0)$. If the lottery were slightly modified so that u rejects it, then v would also refuse it.

Result 6 *Proof that* $\rho'_u \leq 0 \Leftrightarrow \mu'_u \leq 0$.

For a fair gamble $\tilde{x}$, the risk premium μ solves $\mathbb{E}[u(w + \tilde{x})] = u(w - \mu)$. Differentiating yields $\mathbb{E}\left[u'(w + \tilde{x})\right] - u'(w - \mu) = -\frac{\partial \mu}{\partial w} u'(w - \mu)$. Thus $\frac{\partial \mu}{\partial w} < 0$ iff a consumer with utility $v \equiv u'$ accepts gamble $\tilde{x}$ indifferent to u i.e., he is less risk averse which implies $\rho_{u'} = \frac{-u'''}{u''} < \rho_u = \frac{-u''}{u'} \Leftrightarrow 0 > \frac{\partial \rho_u}{\partial w} = \rho_u(\rho_{u'} - \rho_u)$ after re-ordering.

Result 7 *Arrow's claims for sensible risk preferences.*

Axiom 1: given a gamble $\tilde{y}$ and wealth m, $\mathbb{E}\left[u(\tilde{y}+m)\right] \geq u(m) \Rightarrow \mathbb{E}\left[u(\tilde{y}+m+\epsilon)\right] \geq u(m+\epsilon)$ for $\epsilon > 0$ i.e., there is continued acceptance upon becoming richer.

We must prove that this property is equivalent to DARA. Assume A1 holds. Pick a fair gamble $\tilde{x}$, some wealth w and the associated risk premium μ. We have $\mathbb{E}[u(\tilde{x}+w)] = u(w-\mu) \Leftrightarrow \mathbb{E}\left[u(\tilde{y}+m)\right] = u(m)$ for $\tilde{y} = \tilde{x}+\mu$ and $m = w-\mu$. By A1, $\mathbb{E}[u(\tilde{x}+w+\epsilon)] \geq u(m+\epsilon)$ for $\epsilon > 0$. Taking the Taylor expansion leads to $\mathbb{E}\left[u'(\tilde{x}+w)\right] \geq u'(m)$ i.e., Mister u' takes the gamble, thus is less averse than Mister u. As seen above in Result 6, this means that the utility function is DARA.

Conversely, let DARA holds and consider an accepted gamble $\tilde{y}$ at wealth m. Let $\tilde{x} = \tilde{y} - E[\tilde{y}]$ and $w = m + E[\tilde{y}]$, then seek the risk premium $\mu \leq E[\tilde{y}]$ such that $\mathbb{E}\left[u(m+\tilde{y})\right] = \mathbb{E}[u(\tilde{x}+w)] = u(w-\mu) \geq u(w-E[\tilde{y}])$. Apply Result 6 to obtain $\mathbb{E}\left[u(\tilde{y}+m+\epsilon)\right] = \mathbb{E}[u(\tilde{x}+w+\epsilon)] = u(w+\epsilon-\hat{\mu}) > u(w+\epsilon-\mu) \geq u(w-E[\tilde{y}]+\epsilon) = u(m+\epsilon)$ since $\hat{\mu} < \mu$. That is to say, the richer person keeps accepting the gamble.

Axiom 2: given a gamble $\tilde{y}$ and wealth m, $\mathbb{E}\left[u(\tilde{y}+m)\right] \geq u(m) \Rightarrow \mathbb{E}\left[u(\lambda\tilde{y}+\lambda m)\right] \geq u(\lambda m)$ for $\lambda < 1$ i.e., there is continued acceptance upon scaling down gamble and wealth.

To prove that this property is equivalent to IRRA, one uses the same differentiation method.

Result 8 *The objective riskiness measures are homogeneous, sub-additive, convex and are FOC and SOD monotonic.*

Homogeneity: obvious, for the solution of $\mathbb{E}[e^{-\lambda\tilde{g}/z}] = 1$ is $z = \lambda\varphi_g$ and likewise the solution of $\mathbb{E}\left[\log[1+\lambda\tilde{g}/z]\right] = 0$ is $z = \lambda\phi_g$.

Sub-additivity: For the *de Finetti* measure φ_g, write $r = \varphi_g$, $r' = \varphi_{g'}$ and observe the decomposition $\frac{g+g'}{r+r'} = \frac{r}{r+r'}\frac{g}{r} + \frac{r'}{r+r'}\frac{g'}{r'}$. Since the exponential is convex, we have $\mathbb{E}\left[e^{-\frac{g+g'}{r+r'}}\right] \leq \frac{r}{r+r'}\mathbb{E}\left[e^{-\frac{g}{r}}\right] + \frac{r'}{r+r'}\mathbb{E}\left[e^{-\frac{g'}{r'}}\right] = 1$, hence $r + r' \geq R(g+g')$ (cf. footnote 26.3). For the *Bernoulli* measure ϕ_g, write $r = \phi_g$, $r' = \phi_{g'}$. The concavity of log yields

$$\mathbb{E}\left[\log\left(1+\frac{g+g'}{r+r'}\right)\right] \geq \frac{r}{r+r'}\mathbb{E}\left[\log\left(1+\frac{g}{r}\right)\right] + \frac{r'}{r+r'}\mathbb{E}\left[\log\left(1+\frac{g'}{r'}\right)\right] = 0$$

hence $r + r' \geq R(g+g')$ (cf. footnote 26.3).

Convexity: a corollary of Homogeneity and Sub-additivity. Indeed, if $f(\lambda x) = \lambda f(x)$ and $f(x+y) \leq f(x) + f(y)$, then $f(\lambda x + (1-\lambda)y) \leq f(\lambda x) + f((1-\lambda)y) = \lambda f(x) + (1-\lambda)f(y)$.

FOD: We say that $\tilde{g} \underset{FOD}{\succ} \tilde{g}'$ when $H_g(.) \leq H_{g'}(.)$ with at least one strict inequality i.e., the former gamble puts more weight on large prizes. Result 10 proves that $\tilde{g} \underset{FOD}{\succ} \tilde{g}' \Leftrightarrow \mathbb{E}[u(\tilde{g})] > \mathbb{E}[u(\tilde{g}')]$ for all strictly increasing u. For the *de Finetti* measure, we apply this

to the decreasing $u(x) \equiv e^{-\alpha x}$. We obtain $\mathbb{E}[e^{-\alpha\tilde{g}}] - 1 < \mathbb{E}[e^{-\alpha\tilde{g}}] - 1$ for all $\alpha > 0$. Since both are negative (cf. footnote 26.3), their zeroes are $\rho_{g'} > \rho_g \Rightarrow \varphi_g = 1/\rho_g < \varphi_{g'} = 1/\rho_{g'}$. For the *Bernoulli* measure, we use the increasing $u(x) \equiv \log(1 + x/\alpha)$ and proceed likewise to obtain $\phi_g < \phi_h$ (cf. footnote 26.3).

SOD: For two gambles with the same mean, we say that $\tilde{g} \underset{SOD}{\succ} \tilde{g}'$ when $\mathbb{E}[u(\tilde{g})] > \mathbb{E}[u(\tilde{g}')]$ for every increasing strictly concave u. For the *de Finetti* measure, we use the fact that $u(x) = e^{-\alpha x}$ is convex to obtain the same ordering as above and thus the same conclusion. For the *Bernoulli* measure, we use the fact that $u(x) \equiv \log(1 + x/\alpha)$ is concave to obtain the same ordering as above and thus the same conclusion.

26.2.4 Choice under Uncertainty

Result 9 *A less risk averse person invests more into the risky asset.*

By result 5, if ϕ is a convex function, then individual 2 with utility $\phi(u)$ is less risk averse than individual 1 with utility u. The expected utility for individuals 1 and 2 of investing a share λ into the risky asset are

$$U_1(\lambda) = \int u\left(1 + r_0 + \lambda\left(r - r_0\right)\right) h(r)\,\mathrm{d}r$$

and

$$U_2(\lambda) = \int \phi(u)\left(1 + r_0 + \lambda\left(r - r_0\right)\right) h(r)\,\mathrm{d}r$$

The FOC of maximization are

$$0 = U_1'(\lambda) = \int \psi(r) h(r)\,\mathrm{d}r \text{ where } \psi(r) \equiv (r - r_0)\,u'\left(1 + r_0 + \lambda\left(r - r_0\right)\right)$$

and

$$0 = U_2'(\lambda) = \int \psi(r)\hat{h}(r)\,\mathrm{d}r \text{ where } \hat{h}(r) \equiv \phi'\left(u\left(1 + r_0 + \lambda\left(r - r_0\right)\right) h(r)\right.$$

The change from H to $\hat{H}$ satisfies the MLRP because $\phi' = \frac{\hat{h}}{h}$ is an increasing function of r. Furthermore, $(r - r_0)\psi(r) \geq 0$ is true for all r, thus ψ satisfies the SCP. Hence applying result 12, $U_2'(\lambda) > U_1'(\lambda)$ which means that if the RHS is zero for some optimal λ_1, the LHS will be zero for some larger λ_2. ■

Result 10 *Stochastic dominance*
$\hat{\theta} \underset{FOD}{\succ} \theta \Leftrightarrow \mathbb{E}\left[u(\tilde{x}) \mid \hat{\theta}\right] \geq \mathbb{E}\left[u(\tilde{x}) \mid \theta\right]$ *for any increasing utility function* u.
$\hat{\theta} \underset{SOD}{\succ} \theta \Leftrightarrow \mathbb{E}\left[u(\tilde{x}) \mid \hat{\theta}\right] \geq \mathbb{E}\left[u(\tilde{x}) \mid \theta\right]$ *for any increasing concave utility function* u.

FOD: If u is a simple staircase with $u(.)=0$ on $]-\infty;z[$ and $u(.)=m$ on $[z;+\infty]$ then $\mathbb{E}\left[u(\tilde{x})\,|\,\hat{\theta}\right]=\int u(x)h(x|\hat{\theta})\,\mathrm{d}x=m\left(1-H(z|\hat{\theta})\right)\geq m\left(1-H(z|\theta)\right)=\int u(x)h(x|\theta)\,\mathrm{d}x=\mathbb{E}\left[u(\tilde{x})\,|\,\theta\right]$. It then remains to observe that every increasing staircase function is a sum of simple staircases and finally that every increasing function is the limit of a staircase function with many small steps.

Conversely, consider the step function $u(.)=0$ over $]-\infty,x]$ and $u(.)=1$ over $]x;+\infty]$ to get $\mathbb{E}[u(\tilde{x})\,|\,\theta]=1-H(x|\theta)$.

SOD: Using integration by part, $\mathbb{E}[\tilde{x}\,|\,\theta]=\int_a^b xh(x|\theta)\mathrm{d}x=bH(b|\theta)-\int_0^b H(x|\theta)\mathrm{d}x-aH(a|\theta)-\int_a^0 H(x|\theta)\mathrm{d}x$ which tends to $\int_0^{+\infty}(1-H(x|\theta))\mathrm{d}x-\int_{-\infty}^0 H(x|\theta)\mathrm{d}x$. Consider the function $u(x)=\min\{x,z\}$ and redo the previous computation to find $\mathbb{E}[u(\tilde{x})\,|\,\theta]=\int_0^z(1-H(x|\theta))\mathrm{d}x-\int_{-\infty}^0 H(x|\theta)\mathrm{d}x$ $z-\int_{-\infty}^z H(.|\theta)$. Hence, $\mathbb{E}\left[u(\tilde{x})\,|\,\hat{\theta}\right]\geq\mathbb{E}[u(\tilde{x})\,|\,\theta]\Rightarrow\int_{-\infty}^z H(x|\hat{\theta})\mathrm{d}x\leq\int_{-\infty}^z H(x|\theta)\mathrm{d}x$ i.e., $\hat{\theta}\underset{SOD}{\succeq}\theta$.

Conversely, $A\equiv\mathbb{E}\left[u(\tilde{x})\,|\,\hat{\theta}\right]-\mathbb{E}\left[u(\tilde{x})\,|\,\theta\right]=\int u(x)dH$ with $H=H(x|\hat{\theta})-H(x|\theta))$. Integrating by part, we obtain $A=-\int u'(x)H(x)\mathrm{d}x$ since $H=0$ at both extremes. Integrating by part again, we get $A=-uG+\int u''G$ where $G(z)=\int_{-\infty}^z H(x)\mathrm{d}x\leq 0$ by definition of $\hat{\theta}\underset{SOD}{\succ}\theta$. Since u is increasing concave, $A\geq 0$. ■

Result 11 *MLRP ($\frac{h_2}{h_1}\nearrow$) implies first order stochastic dominance (FSD) i.e., $H_2\leq H_1$.*

If $\frac{H_2}{H_1}\searrow$ was always true then by taking derivative we would have $\frac{h_2}{h_1}\leq\frac{H_2}{H_1}$. By definition of a distribution, $H_2(1)=H_1(1)=1$, thus $1\geq\frac{h_2}{h_1}(1)\geq\frac{h_2}{h_1}(x)$ by MLRP. We deduce $h_2\leq h_1$. Since $h_1\neq h_2$, there is a strict inequality for some values so that by integration $H_2(1)<H_1(1)$, a contradiction. Our initial assumption is not always satisfied, thus there is some y such that $\frac{h_2}{h_1}(y)\geq\frac{H_2}{H_1}(y)$. For any $z\geq y$, if $\frac{H_2}{H_1}(z)=\frac{h_2}{h_1}(z)$ then z is a local maximum of $\frac{H_2}{H_1}$, thus this curve will not cross the $\frac{h_2}{h_1}$ curve since the latter is always locally increasing. L'hospital rule states that $\lim_{x\to 0}\frac{H_2}{H_1}(x)=\frac{h_2}{h_1}(0)$, thus we can conclude that $y=0$ i.e., $\frac{H_2}{H_1}\leq\frac{h_2}{h_1}$ everywhere. As $\frac{H_2}{H_1}$ is increasing, the definition of a distribution implies that $\frac{H_2}{H_1}(1)=1\geq\frac{H_2}{H_1}$. ■

Result 12 *If the family of distributions $h(.|\theta)$ satisfies the MLRP and f satisfies SCP, then $\hat{\theta}>\theta\Rightarrow\mathbb{E}\left[f(\tilde{x})\,|\,\hat{\theta}\right]>\mathbb{E}\left[f(\tilde{x})\,|\,\theta\right]$.*

By MLRP, we see, checking separately for $x\gtrless y$, that $(x-y)\frac{h(x|\theta)}{h(y|\theta)}\leq(x-y)\frac{h(x|\hat{\theta})}{h(y|\hat{\theta})}$ is true; further this is strict for some values for otherwise $h(.|\hat{\theta})$ would not be different from $h(.|\theta)$. For any function g satisfying SCP, we can replace $(x-y)$ by $g(x)$ in the inequality above to obtain

$$g(x)\frac{h(x|\theta)}{h(y|\theta)}\leq g(x)\frac{h(x|\hat{\theta})}{h(y|\hat{\theta})}\Leftrightarrow g(x)h(x|\theta)\leq\frac{h(y|\theta)}{h(y|\hat{\theta})}g(x)h(x|\hat{\theta})$$

with strict inequality for some values of x. Hence, $\mathbb{E}\left[g(\tilde{x})\,|\,\theta\right]=0$ implies $\mathbb{E}\left[g(\tilde{x})\,|\,\hat{\theta}\right]>0$. The claim is proved by applying this result to the function g equal to f minus the constant number $\mathbb{E}\left[f(\tilde{x})\,|\,\theta\right]$ (if f passes SCP, g also). ■

Result 13 *A change in distribution of the risky asset increases its demand whatever the risk aversion (u) and whatever the risk-free rate (r_0) if and only if the change satisfies the MLRP.*

Consider a dominant change from H_1 to H_2. By definition of $\psi_{r_0}(r)$ defined in result 10, $(r-r_0)\psi_{r_0}(r)\geq 0$ is true, thus $\int_0^1 \psi_{r_0}(r)\,\mathrm{d}H_1(r)=0\Rightarrow\int_0^1 \psi_{r_0}(r)\,\mathrm{d}H_2(r)\geq 0$ by virtue of the previous result but this means that the optimal share λ_2 under H_2 is larger than the optimal share λ_1 under H_1.
Conversely, for any risk-free rate r_0, we can apply the lemma below to $f(r)=\psi_{r_0}(r)$ and $g(r)\equiv\frac{dH_2}{dH_1}(r)\psi_{r_0}(r)$ and claim that $\exists\alpha_{r_0},\forall r\in[0;1], g(r)\geq\alpha_{r_0}\phi_{r_0}(r)$. Developing ψ_{r_0} and simplifying by u', we obtain $\left(\frac{dH_2}{dH_1}(r)-\alpha_{r_0}\right)(r-r_0)\geq 0$ i.e., $\frac{dH_2}{dH_1}$ is equal to α_{r_0} at $r=r_0$ and is locally increasing to respect the inequality. Since this is true for any r_0, the ratio is everywhere increasing thus $\frac{dH_1}{dH_2}(x)$ is decreasing i.e., the MLRP holds.

Lemma: $\{\forall\ H,\ \int f\,\mathrm{d}H=0\Rightarrow\int g\,\mathrm{d}H\geq 0\}\Rightarrow\{\exists\alpha, g\geq\alpha f \text{ over } [0;1]\}$
Under the premises, the program $\min_{dH\geq 0}\int g\,\mathrm{d}H$ s.t. $\int f\,\mathrm{d}H=0$ and $\int \mathrm{d}H=1$ must have a non negative value. Since the constraints and the objective are linear in $\mathrm{d}H$, the Kuhn-Tucker conditions are both necessary and sufficient to characterize the optimum H^*: $\forall x\in[0;1],\ g(x)-\alpha f(x)-\beta\underset{\geq}{=}0$ if $dH^*(x)\underset{=}{>}0$ where α and β are the Lagrange multipliers (cf. §2.2.1). Integrating and using the constraints, we obtain $0=\int\left(g(x)-\alpha f(x)-\beta\right)\mathrm{d}H^*(x)=\int g\,\mathrm{d}H^*-\beta$. Lastly, the Kuhn-Tucker conditions imply that $\forall x\in[0;1], g(x)-\alpha f(x)$ is bounded below by $\beta=\int g\,\mathrm{d}H^*\geq 0$. ■

26.3 Auctions and Finance

Result 14 *The optimal selling mechanism when marginal revenues are not always increasing.*

Consider first one potential buyer. Let $f(q)\equiv\tilde{R}_m\left(H^{-1}(q)\right)\Leftrightarrow\tilde{R}_m(v)=f(H(v))$, let $F(q)\equiv\int_0^q f(x)\,\mathrm{d}x$, then let $\hat{F}\equiv\max\{G, G\leq F, G \text{ convex}\}$ be the convex envelope of F. By construction, $\hat{f}\equiv\hat{F}'\geq 0$ so that $\hat{R}_m(v)\equiv\hat{f}(H(v))$ is increasing. Observe that

$$\Delta\equiv\int_0^{\bar{v}}\varphi(v)(f-\hat{f})(H(v))\,\mathrm{d}H(v)=\left[\varphi(v)(F-\hat{F})(H(v))\right]_0^{\bar{v}}-\int_0^{\bar{v}}(F-\hat{F})(H(v))\,\mathrm{d}\varphi(v)$$

when integrating by parts. Since F and $\hat{F}$ agree at the extremes (by construction of $\hat{F}$), the first term is zero. The objective of the seller is

$$\begin{aligned}\mathbb{E}\left[\varphi\tilde{R}_m\right] &= \int_0^{\bar{v}} \varphi(v) f(H(v))\,\mathrm{d}H(v) = \int_0^{\bar{v}} \varphi(v)\hat{f}(H(v))\,\mathrm{d}H(v) + \Delta \\ &= \mathbb{E}\left[\varphi\hat{R}_m\right] - \int_0^{\bar{v}} (F-\hat{F})(H(v))\,\mathrm{d}\varphi(v)\end{aligned}$$

Returning to the case of several potential buyers, the seller's revenue (cf. eq. 22.10), up to a constant, is

$$W_S = \mathbb{E}\left[\sum_{i\geq 0}\varphi_i(\mathbf{v})\tilde{R}_{m,i}(v_i)\right] = \mathbb{E}\left[\sum_{i\geq 0}\varphi_i(\mathbf{v})\hat{R}_{m,i}(v_i)\right] - \sum_{i\geq 0}\Delta_i \tag{26.5}$$

where Δ_i is the generalization of Δ to bidder i. Since each $\Delta_i \geq 0$ by construction of $\hat{F}_i$, the maximum of W_S is reached by maximizing the first term if at this maximum $\sum_{i\geq 0}\Delta_i$ is nil. Allocating, as in the regular case, the item to the bidder with the highest adjusted marginal revenue $\hat{R}_{m,i}$ maximizes this first term but also makes the second nil. That each $\Delta_i = 0$ is true is because at every value v_i, either $H_i(v_i)$ is such that $F_i = \hat{F}_i$ or otherwise $\hat{F}_i$ is locally linear, meaning that $\hat{f}_i$ is locally constant, meaning that $\hat{R}_{m,i}$ is locally constant, implying finally, that under the proposed allocation rule, the probability of winning the item is also locally constant i.e., $d\varphi_i(v) = 0$ which makes Δ_i nil.

Result 15 *The optimal strategies in Chatterjee and Samuelson (1982) trade game are to announce* $p_b = \frac{2}{3}b + \frac{\bar{s}+3b}{12}$ *and* $p_s = \frac{2}{3}s + \frac{3\bar{s}+b}{12}$.

The player's payoffs are

$$\pi_s = \begin{cases} \frac{p_s+p_b}{2} & \text{if} \quad p_s \leq p_b \\ s & \text{if} \quad p_s > p_b \end{cases} \quad \text{and} \quad \pi_b = \begin{cases} b - \frac{p_s+p_b}{2} & \text{if} \quad p_s \leq p_b \\ 0 & \text{if} \quad p_s > p_b \end{cases}$$

For $i = b, s$ let $p_i(v_i)$ denote i's betting strategy and q_i its inverse (assuming it is increasing). The seller chooses p_s and there is trade if the random b is such that $p_s \leq p_b(b) \Leftrightarrow q_b(p_s) \leq b$; likewise the buyer chooses p_b and there is trade if the random s is such that $p_s(s) \leq p_b \Leftrightarrow s \leq q_s(p_b)$, hence each profit is a function of the unknown valuation:

$$\pi_s(p_s, b) = \begin{cases} \frac{p_s+p_b(b)}{2} & \text{if} \quad q_b(p_s) \leq b \\ s & \text{if} \quad q_b(p_s) > b \end{cases}$$

and

$$\pi_b(p_b, s) = \begin{cases} b - \frac{p_s(s)+p_b}{2} & \text{if} \quad s \leq q_s(p_b) \\ 0 & \text{if} \quad s > q_s(p_b) \end{cases}$$

The expected payoffs are:

$$\begin{aligned}(\bar{s}-\underline{s})u_s(p_s) &= \int_{\underline{s}}^{\bar{s}} \pi_s(p_s,b)\,\mathrm{d}b = \int_{\underline{s}}^{q_b(p_s)} s\,\mathrm{d}b + \int_{q_b(p_s)}^{\bar{s}} \tfrac{p_s+p_b(b)}{2}\,\mathrm{d}b \\ (\bar{b}-\underline{b})u_b(p_b) &= \int_{\underline{b}}^{\bar{b}} \pi_b(p_b,s)\,\mathrm{d}s = \int_{\underline{b}}^{q_s(p_b)} \left(b - \tfrac{p_s(s)+p_b}{2}\right)\mathrm{d}b\end{aligned}$$

The FOCs are

$$\begin{aligned}0 &= \frac{\partial u_s}{\partial p_s} = s\frac{\partial q_b}{\partial p_s} - \tfrac{p_s+p_b(q_b(p_s))}{2}\frac{\partial q_b}{\partial p_s} + \int_{q_b(p_s)}^{\bar{s}} \frac{\mathrm{d}b}{2} = \left(s-p_s\right)\frac{\partial q_b}{\partial p_s} + \tfrac{1}{2}\left(\bar{s}-q_b(p_s)\right) \\ 0 &= \frac{\partial u_b}{\partial p_b} = \left(b - \tfrac{p_s(q_s(p_b))+p_b}{2}\right)\frac{\partial q_s}{\partial p_b} - \int_{\underline{b}}^{q_s(p_b)} \frac{\mathrm{d}b}{2} = \left(b-p_b\right)\frac{\partial q_s}{\partial p_b} + \tfrac{1}{2}\left(\underline{b}-q_s(p_b)\right)\end{aligned}$$

Let us look for linear rules $p_i(v_i) = \alpha_i + \beta_i v_i \Leftrightarrow q_i(p_i) = \frac{p_i-\alpha_i}{\beta_i}$ for $i = b,s$. The FOCs now read

$$\begin{aligned}\left(s-p_s\right)\frac{1}{\beta_b} = \frac{1}{2}\left(\frac{p_s-\alpha_b}{\beta_b} - \bar{s}\right) &\Leftrightarrow p_s = \frac{\alpha_b+\beta_b\bar{s}}{3} + \frac{2}{3}s \Rightarrow \beta_s = \frac{2}{3} \\ \left(b-p_b\right)\frac{1}{\beta_s} = \frac{1}{2}\left(\frac{p_b-\alpha_s}{\beta_s} - \underline{b}\right) &\Leftrightarrow p_b = \frac{\alpha_s+\underline{b}\beta_s}{3} + \frac{2}{3}b \Rightarrow \beta_b = \frac{2}{3}\end{aligned}$$

We are left with the system $3\alpha_s = \alpha_b + \beta_b\bar{s} = \alpha_b + \frac{2}{3}\bar{s}$ and $3\alpha_b = \alpha_s + \underline{b}\beta_s = \alpha_s + \underline{b}\frac{2}{3}$ whose solution is $\alpha_b = \frac{\bar{s}+3\underline{b}}{12}, \alpha_s = \frac{3\bar{s}+\underline{b}}{12}$. ■

Result 16 *Proof that a bilateral trading mechanism cannot be at the same time efficient in trade, balanced in budget and guarantee the participation of both agents.*

The revenue equivalence formula (22.3) remains true i.e., for $i = b,s$ $u_i(i) = u_i(\underline{i}) + \int_{\underline{i}}^{i} \varphi_i(x)\,\mathrm{d}x$. As $\bar{t}_b(b) = b\varphi_b(b) - u_b(b)$, the overall expected payment is

$$\begin{aligned}&\mathbb{E}\left[t_b(b,s)\right] = \int_{\underline{b}}^{\bar{b}} \left(b\varphi_b(b) - \int_{\underline{b}}^{b} \varphi_b(x)\,\mathrm{d}x\right)\mathrm{d}H_b - u_b(\underline{b}) \\ =\ &\int_{\underline{b}}^{\bar{b}} \varphi_b(b)\left(b - \tfrac{1-H_b(b)}{f_b(b)}\right)dH_b - u_b(\underline{b}) \text{ (integrate by parts)} \\ =\ &\iint z_b(b,s)\left(b - \tfrac{1-H_b(b)}{f_b(b)}\right)\mathrm{d}H_b\,\mathrm{d}H_s - u_b(\underline{b}) \qquad (26.6)\end{aligned}$$

where $z_b \in \{0;1\}$ is the buyer's allocation rule in equilibrium of the auction. Likewise,

using $u_s(s) = u_s(\bar{s}) - \int_s^{\bar{s}} \varphi_s(x)\,\mathrm{d}x$, we get

$$\begin{aligned}
\mathbb{E}[t_s(b,s)] &= \iint z_s(b,s)\left(s + \tfrac{H_s(s)}{h_s(s)}\right) dH_s dH_b - u_s(\bar{s}) \\
&= \iint (1 - z_b(b,s))\left(s + \tfrac{H_s(s)}{h_s(s)}\right) dH_s dH_b - u_s(\bar{s}) \\
&= \bar{s} - \iint z_b(b,s)\left(s + \tfrac{H_s(s)}{h_s(s)}\right) dH_s dH_b - u_s(\bar{s}) \qquad (26.7)
\end{aligned}$$

since $\iint \left(s + \frac{H_s(s)}{h_s(s)}\right) \mathrm{d}H_s\,\mathrm{d}H_b = \int \left(s + \frac{H_s(s)}{h_s(s)}\right) h_s(s)\,\mathrm{d}s = [sH_s(s)]_{\underline{s}}^{\bar{s}} = \bar{s}$.

Summing (26.6) and (26.7), the expected total payment to the broker is

$$\bar{A} \equiv \bar{s} - u_s(\bar{s}) - u_b(\underline{b}) + \iint z_b(b,s)\left(b - \tfrac{1-H_b(b)}{h_b(b)} - s - \tfrac{H_s(s)}{h_s(s)}\right) dH_s dH_b$$

If the buyer with the lowest value $\underline{b}$ derives no extra surplus from participating then $u_b(\underline{b}) = 0$; likewise, if the seller with the highest value $\bar{s}$ derives no extra surplus from participating then $u_s(\bar{s}) = \bar{s}$. The maximal value for $\bar{A}$ is thus the integral term. We now use the efficient rule $z_b^* = 1 \Leftrightarrow b > s$ to show that $\bar{A} < 0$ i.e., the broker must bring funds.

We denote $b \wedge \bar{s} \equiv \min\{b, \bar{s}\}$, we can write $\bar{A} = \int_{\underline{b}}^{\bar{b}} A(b)\,\mathrm{d}H_b$ where

$$\begin{aligned}
A(b) &= \int_{\underline{s}}^{b\wedge\bar{s}} \left(b - \tfrac{1-H_b(b)}{h_b(b)} - s - \tfrac{H_s(s)}{h_s(s)}\right) h_s(s)\,\mathrm{d}s \\
&= \left[\left(b - \tfrac{1-H_b(b)}{h_b(b)} - s\right) H_s(s)\right]_{\underline{s}}^{b\wedge\bar{s}} = \left(b - \tfrac{1-H_b(b)}{h_b(b)} - b \wedge \bar{s}\right) H_s(b \wedge \bar{s}) \\
&= \begin{cases} -\frac{1-H_b(b)}{h_b(b)} H_s(b) < 0 & \text{if} \quad b < \bar{s} \\ b - \bar{s} - \frac{1-H_b(b)}{h_b(b)} = \tilde{R}_m(b) - \bar{s} & \text{if} \quad b \geq \bar{s} \end{cases}
\end{aligned}$$

Over $[\bar{s}; \bar{b}]$, the average of A is nil because (22.6) applies telling us that $\bar{s} = \mathbb{E}\left[\tilde{R}_m(b) | b \geq \bar{s}\right]$. Given that A is negative over $[\underline{b}; \bar{s}]$, $\bar{A}$ is negative. ■

Result 17 *Proof that the average price mechanism maximizes the gains of trade among budget-balanced mechanism guaranteeing participation of both agents.*

To avoid running a deficit, the trading mechanism must use a rule z_b such that $\bar{A} \geq 0$. Let $f_\epsilon(s,b) \equiv b - s - \epsilon\left(\frac{1-H_b(b)}{h_b(b)} + \frac{H_s(s)}{h_s(s)}\right)$ and define the transfer rule $z_\epsilon \equiv \mathbf{1}_{f_\epsilon \geq 0}$ which maximizes $\iint z f_\epsilon$ unconditionally. The optimal rule z^* solves

$$\begin{cases} \max_z \iint z f_0 \\ \text{s.t.} \iint z f_1 = 0 \end{cases} \Leftrightarrow \begin{cases} \max_z \iint z\left(f_0 + \lambda f_1\right) \\ \text{s.t.} \iint z f_1 = 0 \end{cases}$$

for some $\lambda > 0$. Observe then that $f_0 + \lambda f_1 = (1+\lambda) f_\epsilon$ for $\epsilon \equiv \frac{\lambda}{1+\lambda}$. It thus remains to exhibit one ϵ such that $g(\epsilon) \equiv \iint z_\epsilon f_1 = 0$ to have our solution. This is done by showing that $g(1) > 0$ (obvious by construction of z_1), $g(0) < 0$ (this because $\bar{A} < 0$ for the efficient rule z_0) and that g is continuous which comes from the fact that a change in ϵ, changes the bound in the integral through a change in the solution of $f_\epsilon = 0$ which is continuous in ϵ.

With uniform distributions over $[0;1]$, $f_\epsilon(s,b) = b - s - \epsilon(1-b-s) = (1+\epsilon)(b-s) - \epsilon$, thus $z_\epsilon = 1$ iff $s \leq b - \frac{\epsilon}{1+\epsilon}$ and $g(\epsilon) = \int_{\frac{\epsilon}{1+\epsilon}}^{1} \int_0^{b-\frac{\epsilon}{1+\epsilon}} (2(b-s)-1)\,\mathrm{d}s\,\mathrm{d}b = \frac{3\epsilon-1}{6(1+\epsilon)^3}$, thus $\epsilon = \frac{1}{3}$ leading to $b - s \geq \frac{1}{4}$ as claimed. ■

Result 18 *There is over-investment in the free-cash flow model.*

To prove that claim, we use the fact that $R_m(k^*) = 1 = \mathbb{E}[\tilde{p}]$ to write

$$1 = R_m(k^*)\mathbb{E}[\tilde{p}] = R_m(k^*) \int_0^{\hat{p}} p\,\mathrm{d}H(p) + R_m(k^*) \int_{\hat{p}}^{+\infty} p\,\mathrm{d}H(p) \tag{26.8}$$

The repayment to the lender being $\tilde{b} = \min\{\tilde{p}R(k), (1+r)k\}$, we can write

$$\frac{\mathbb{E}[b]}{k} = \frac{R(k)}{k} \int_0^{\hat{p}} p\,\mathrm{d}H(p) + (1+r)\left(1 - H(\hat{p})\right) \tag{26.9}$$

If the participation constraint $\mathbb{E}[b] \geq k$ is binding, then observe from (26.9) that

$$1 < \frac{R(k)}{k} \hat{p} \int_0^{\hat{p}} \mathrm{d}H(p) + (1+r)\left(1 - H(\hat{p})\right) = H(\hat{p}) + (1+r)\left(1 - H(\hat{p})\right)$$

hence $r > 0$.

Next, we use the fact that $\frac{R(k)}{k} > R_m(k)$ to deduce that the first term of (26.9) is greater than the first term (26.8). Given that (26.8) and (26.9) are equal, it must be the case that seconds terms are inversely ordered i.e.,

$$(1+r)\left(1 - H(\hat{p})\right) < R_m(k^*) \int_{\hat{p}}^{+\infty} p\,\mathrm{d}H(p) \Leftrightarrow (1+r) < R_m(k^*)\mathbb{E}[p|p \geq \hat{p}] \tag{26.10}$$

The risk premium of investors being $\hat{\pi} = \int_{\hat{p}}^{+\infty} \left(pR(k) - (1+r)k\right) \mathrm{d}H(p)$, we have

$$\left.\frac{\partial \hat{\pi}}{\partial k}\right|_{k^*} \propto R_m(k^*)\mathbb{E}[p|p \geq \hat{p}] - (1+r) > 0 \text{ by (26.10)}$$

meaning that the optimal investment $\hat{k}$ is greater than the efficient one k^*. ■

Notes

1.1 Since each agency uses its own typology, some figures appear widely apart when in fact they reflect slightly different aggregates.

1.2 cf. annual exchange rates.

1.3 Early versions of this work borrowed from the books by Gabszewicz (1994), Rasmusen (2006), Shy (1996) and Tirole (1988) as well as teaching notes posted on the internet.

1.4 This, as far as we know, is a novelty at the textbook level in micro-economics.

1.5 Taking an example from the realm of physics, Newton's theory of gravity can be seen as a model of the trajectory of the apple from the branch of the tree to Newton's head. It can be applied not only to apples but also to other fruits, parachutists, rockets or balls.

1.6 This means for example that bits from quantum physics are irrelevant to build Newton's theory of gravity.

1.7 Formally, game theory studies competitive activities in which participants strife against each other according to a set of rules.

1.8 For instance, behavioral economics, and in particular prospect theory, has shown the limit to the expected utility theory (cf. §19.1.3); its finding regarding non fully rational behavior are being integrated into the mainstream framework.

1.9 Introductory online courses are available at economicsnetwork; for more a more advanced level check with introecon or econphd.

1.10 Business oriented readers may wish to consult Dixit and Nalebuff (1991), Brandenburger and Nalebuff (1996) or Dixit and Skeath (1996) while more traditional introductions to game theory with economic applications are Gibbons (1992), Binmore (1992) or Osborne and Rubinstein (1994). Advanced treatments, at the graduate and post graduate levels include Fudenberg and Tirole (1991) and Myerson (1991). The website www.gametheory.net offers online courses as well as review of textbooks.

1.11 An excellent online tutorial is available at Martin Osborne's website to refresh your mind.

2.1 The limitation to only 2 factors is for exposition purposes and involves no loss of generality.

2.2 If the variations dK and dL are such that the pair $(K + dK, L + dL)$ remains in the isoquant, then the total variation of production $d\Phi = \Phi_L dL + \Phi_K dK$ is nil. A simple algebraic manipulation then yields equation (2.1).

2.3 This term is used in opposition to the market price that, as we shall see, can be deemed objective because it depends on all the technologies while this one depends on this particular technology Φ.

2.4 Like any attempt to judge and compare, this theory builds on welfare economics (cf. §2.3.3) and more specifically on Debreu (1951)'s coefficient of resource utilization.

2.5 The intermediary who provides the subcontracted activity incurs a fixed cost of maintaining a large number of employees available to meet any of his clients requests; since he bears a risk, he must charge his clients more than his own marginal cost.

2.6 The short-term cost is always greater but equal to the long-run cost at one point. As can be seen on Figure 2.2, the curves touch but do not cross, so that their tangents are equal at q_o, that is to say the marginal cost are equals.

More formally, this result is a consequence of the envelope theorem.

2.7 Letting $\lambda = 1+\epsilon$ for $\epsilon > 0$, the condition $\lambda C(q) > C(\lambda q)$ reads $\epsilon C(q) > C(q+\epsilon q) - C(q) \Leftrightarrow AC(q) = \frac{C(q)}{q} > \frac{C(q+\epsilon q)-C(q)}{\epsilon q} \underset{\epsilon\to 0}{\longrightarrow} C_m(q)$. As for the second claim, footnote 26.3 shows that the sign of $\frac{\partial AC}{\partial q}$ is that of $C_m - AC$.

2.8 $H1$–$H3$ are necessary in the short run while $H4$ has to be added only if one takes a long run perspective.

2.9 In greater generality, the factor prices w and r also hinge on the marginal cost, thus profit is a function of the prices of all outputs and inputs.

2.10 Since $AC(q) = \frac{C(q)}{q}$, the FOC defining $\bar{q}$ reads $0 = \frac{\partial AC}{\partial q} = \frac{qC_m - C(q)}{q^2} \Leftrightarrow C_m(q) = AC(q)$.

2.11 One could argue that if the marginal cost is always constant or decreasing, the minimum efficient scale is infinite but since there does not exists a technology that enables infinite production, the marginal cost must start to increase at some point, hence a finite value of $\bar{q}$ exists, although it may be very large, for instance larger than the world demand for a free product.

2.12 Since total cost can be divided into fixed and variable parts $C(q) = F + CV(q)$, the average cost has also two parts; $\frac{CV(q)}{q}$ has limit zero when production tends to zero while $\frac{F}{q}$ diverges as soon as $F > 0$ which is the case illustrated on Figure 2.3.

2.13 Utility is the economic terminology for the satisfaction or pleasure we derive from consuming goods (commodities or services).

2.14 We make the necessary assumption for u to be differentiable.

2.15 This is so because either there exists an increasing function f with $u = f(v)$ or an increasing function g with $v = g(u)$. Computing the MRS in each case reveals the equality thanks to the laws of derivation.

2.16 Under the necessary assumption that make the problem well behaved i.e., having a unique solution.

2.17 Mathematically, this is the Envelope Theorem or Maximum Theorem.

2.18 Beware if that $\mathbf{x}^*$ does maximize $\mathscr{L}(\mathbf{x}, \lambda^*)$, λ^* does not maximizes $\mathscr{L}(\mathbf{x}^*, \lambda)$, it is a saddle point. Hence Kalman (2009) argues it is a fallacy to believe that the Lagrange method reduces constrained optimization into an unconstrained one.

2.19 This statement might be false for basic food if the consumer is quite poor, a combination that we shall never consider in this book.

2.20 The exact definition are related to the income elasticity of demand: inferior if $\epsilon_m < 0$, luxury if $\epsilon_m > 1$, normal otherwise.

2.21 This *revealed preferences* reasoning is due to Samuelson (1947).

2.22 Asking Julie's neighbors how many euros they would pay for each additional video rental, we obtain a long list of integers between 0 and some maximum, say 100. Let n_i be the number of times the figure i appears in the list and $n \equiv \sum_{i\leq 100} n_i$, the maximum number of rental in this neighborhood aka. the market size $D(0)$. We can now compute the frequencies $h_i \equiv n_i/n$ and cumulative frequencies $H_i \equiv \sum_{j\leq i} h_j$ in order to express the demand, for an integer price p, as $D(p) = D(0) \times (1 - H_p)$. Passing to the continuum in prices maintains this dual view of demand and distribution of WTP which will be used in §22 on auctions.

2.23 It is possible to normalize b to unity; this is achieved by changing the physical unit from 1 to b (if $b = 4$, it amounts to sell the good in packs of 4 units) so that the price of one new unit is

bp. This trick would however prohibit a comparative statics analysis of equilibrium quantities with respect to price sensitivity of demand.

2.24 Obviously, if there is an alternative purchase that generates a higher surplus then we ought to buy this alternative first.

2.25 His idea arose from an attempt to perform a cost-benefit analysis of public works (cf. §4.2.2).

2.26 When utility is separable between money m (an index of other goods and services) and the good q i.e., $U_j(m,q) = m + u_j(q)$, the FOC of demand is $u_j'(q) = p$ which defines the WTP function as the marginal utility; (2.17) is a restatement of this FOC in integral form.

2.27 We use the letter W as in welfare because consumer surplus will be part of the market welfare.

2.28 More precisely, Willig (1976) proves that consumer surplus is the right tool to evaluate the effect of a policy upon a single market. The percentage error made by using consumer surplus instead of Hicks (1942)'s more correct but hard to estimate *equivalent variation* or *compensating variation* is bounded by $\frac{\eta}{2m}\Delta W_D$ where ΔW_D is the variation of consumer surplus brought about by the change of policy, m is the income of the individual and η his income elasticity of demand. This bound is likely to be small since either the income is much larger than the variation of consumer surplus or the income elasticity is small. On top of this theoretical finding, we can note with Hicks that measurement errors will be far greater than the conceptual ones because the quality of available statistics is often very poor.

2.29 If we write consumer surplus as a finite sum $W_D(q) = \sum_{x=1}^{q}(P(x) - P(q)) = -qP(q) + \sum_{x=1}^{q} P(x)$, then the marginal variation is $W_D(q+1) - W_D(q) = -(q+1)P(q+1) + qP(q) + P(q+1) = -q(P(q+1) - P(q))$ which is the discrete version of the derivative in the text.

2.30 He coined the term "consumer surplus" and introduced the remaing concept of this section: the supply curve, the producer surplus, the partial equilibrium and the notion of welfare.

2.31 If the technology exhibits increasing returns to scale then any active firm will try to produce an infinite quantity because the average cost tends to zero (hence goes below any positive price). This would make aggregate supply greater than demand and permanent disequilibrium would result. This particular case is studied in §17 that deals with regulation.

2.32 Consider producing a total amount q for the monopolist. We know from the multiplant problem (2.8) that the optimal distribution $(q_i)_{i\leq n}$ among the n plants satisfies $C_m(q) = C_m^1(q_1) = ... = C_m^n(q_n)$. If we denote p the cost so achieved, then the production of each plant is her competitive supply i.e., $q_i = s_i(p)$, hence the aggregate production $q = \sum_{i\leq n} q_i$ is exactly $S(p)$. We thus proven that $C_m(q) = p \Leftrightarrow q = S(p)$ which is to say that C_m is the inverse of S.

2.33 This was the original observation made by Dupuit (1844) to argue that tolls for bridges should be eliminated since their marginal cost is almost nil. This author also observed that if the current price is already inefficiently large like p_2 on the right panel of Figure 2.5, then a further increase Δp generates a much greater loss of $p_2\Delta q + \frac{1}{2}\Delta p\Delta q$ (or a large gain if we implement a reduction).

2.34 Debreu (1951) using the duality of utility and cost shows that summing consumer surpluses over markets (the utilitarian approach)

is a good proxy of an objective theoretical welfare measure that avoids making interpersonal utility comparisons. However, the existence of lump-sum transfers must be assumed in order to be able to compensate agents who would suffer from policy changes (cf. Boccard (2010a)).

2.35 There is tacit agreement in the profession since all manuals we have been able to consult gloss over the issue.

2.36 For a unique player, one speaks of decision theory.

2.37 With more strategies, we use the "domination" relation to eliminate some strategies. If one one remains, it is dominant strategy.

2.38 For ①, there are 4 cases i.e., one per matrix entry. For ②, there are 8 cases = 2 players × 2 possible dominant strategies × 2 possible optimal choice for the other. There are two cases for each of the remaining classes, making 16 in total.

2.39 There are two pure strategies equilibria and one mixed strategy equilibrium which we do not develop.

2.40 A single valued relationship is called a function while a potentially multi-valued relationship like BR_i is called a correspondence.

2.41 We are assuming the existence of money as a perfect mean to transfer utility (or satisfaction) i.e., the swapping of 1€ between farmers S and N decreases the utility of farmer S as much as it increases the utility of farmer N.

2.42 Notice that a simple translation of the axes puts the origin at the threat point as if the opportunity cost of both players was nil. In many applications of bargaining, this is done without loss of generality.

2.43 The traditional economic definition of a "quasi-rent" speaks of a payment that is received by a resource of production over the opportunity cost in the short run.

2.44 When $\delta_S = \delta_N$, we may interpret $1-\delta$ as the share of the pie wasted in having the negotiation lasting one more period.

2.45 We use the language of agency theory as this model is used in Part H.

2.46 The formal model sets $\pi_N = q_N P_N(q_N) - C_N(q_N)$ and $\pi_S = q_S P_S(q_S) - C_S(q_S, z)$ where the emission level z is a deterministic function $f(q_N)$ of N's level of activity. If N is not liable for z, he chooses $\bar{q}_N$ solving $R_m^N(q_N) = C_m^N(q_N)$, thus acting as a Stackelberg leader. In this setting, S is a follower and chooses $\bar{q}_S$ solving $R_m^S(q_S) = C_m^S(q_S, \bar{q}_N)$. Efficiency commands to maximize $\pi_N + \pi_S$ in which case the first FOC is modified into $R_m^N(q_N) = C_m^N(q_N) + \frac{\partial C^S}{\partial q_N}$. The solution involves $q_N^* < \bar{q}_N$ and $q_S^* > \bar{q}_S$; emissions are reduced by $f(\bar{q}_N) - f(q_N^*)$ but never fully eliminated.

2.47 Although it looks unequal, an efficient agreement sees S aks N to limit emissions to $z^* = f(q_N^*)$ in exchange of a payment $F = \bar{\pi}_N - \pi_N^*$. Upon accepting the offer, A would produce the maximum compatible with the agreement i.e., q_N^* and earn $\pi_N^* + F = \bar{\pi}_N$ which is enough for N to accept the initial offer. In that case, B earns $\pi_S^* - F = \bar{\pi}_S + \delta$ i.e., S grabs all the benefits from efficiency even though he still has to pay N.

2.48 We may also cite social cohesion, reputation and repeated interaction as mechanisms used to achieve cooperation.

2.49 In the light of the modern theories of commitment and reputation, we can make sense of this last argument: the individual will bond with his peer group when he recognizes that short-term IR behavior will be met with a large sanction such as lasting ostracism. Hence, the social dilemma is solved if the group has the

means to avoid free riding.

2.50 From the standpoint of modern game theory, this position is shaky for if incumbents find it beneficial to integrate the cartel, couldn't it be best for an entrant to apply also ?

2.51 Tuck (2008) ascribes this modeling choice to the 1791 revolutionary Le Chapelier law that outlawed combinations, associations and guilds.

3.1 Interestingly, he also reports that city states already sell monopoly franchises to raise finance (cf. §16.4.1).

3.2 It might be the case that the challenger will displace the incumbent and become itself the new monopoly.

3.3 Microsoft's loud speaking CFO, Steve Ballmer, once said "We don't have a monopoly; we have market share, there's a difference", a thin one we shall argue.

3.4 The terminology derives from rents earned by landlords.

3.5 Mathematically speaking, the volume effect is of first-order while the price one is of second-order.

3.6 Since P is the inverse of D, their derivative are inverse one from the other.

3.7 Equation (3.4) reads $(p - C_m)\epsilon = p \Leftrightarrow (p - C_m)P(q) = -pqP'(p) \Leftrightarrow p - C_m = -qP'(p) \Leftrightarrow C_m = R_m$ because $p = P(q)$.

3.8 By the first and second theorems of welfare, a Pareto optimum is also a competitive equilibrium but since there is only one firm, the monopoly, it sounds weird to pretend it could act as a price-taker; for that reason we prefer to speak of Pareto optimum or an efficient allocation of resources.

3.9 Dupuit (1844)'s original "utilité perdue" (lost utility) is turned into "perte sèche" (net loss) by Colson (1901-1907), then translated into "dead loss" by Edgeworth (1910). This area is also known as Harberger (1954)'s triangle for the stubborn effort of this author to measure empirically this loss at an aggregate level (cf. Hines (1999)).

3.10 A still older case seem to be an Egyptian regulation prohibiting papyrus from being planted in too many places.

3.11 Reacting and adjusting to ever changing external conditions is time and effort consuming while the reward is meager because at the maximum, profits change slowly (this is due to the envelope theorem); the firm owner may thus rationally permit some slack in the search for maximum profits.

3.12 The neutral name is chosen to combine all forms of inefficiency that are not allocative i.e., not due to the pricing behavior of the firm.

3.13 The exception would be when the quality attribute is "shinny" and signals the overall high price of the item whose ownership then serves as a social signal (cf. §24.5).

3.14 This claim cannot be transposed to a symmetrical statement contingent on quantity because the consumer surplus depends on quality in two non comparable ways as $C(s,q) = \int_0^q \big(P(s,x) - P(s,q)\big)\,\mathrm{d}x$.

3.15 Starting from the quantity–quality pair which is optimal for the monopoly, the planner can firstly increase quality to the efficient level, conditional on the monopoly quantity. Then he starts to increase quantity while adjusting quality downward because the two are substitutes. If the total increase in quantity is large enough, then quality will fall so much that it will end up lower than the original monopoly level.

3.16 This is w.l.o.g. if the cost of developing

and implementing quality are convex because a rational choice, guided by profit considerations, will always be less than what the current state of technology would be able to achieve.

3.17 Notice that our example falls outside the previous categories as $\frac{\partial^2 P}{\partial q \partial s} = 0$. This separability of q and s in the WTP will however enables to work out an analytical solution.

3.18 To do this the firm must use carriers of minimum size q/s. To simplify, we assume that demand is homogeneously spread over the period so that each carrier can travel at 100% capacity.

3.19 This does not contradict the general result enunciated above because it was contingent on price while here the contingency is quantity.

3.20 There is a slight abuse of language here since we do not account for capital cost (carrier, θ) but only variable cost (people, c).

3.21 cf. One concrete example is given in Markoff (1990), p449, who concludes that at the end of ancient regime France in 1789, the crowd voted in favor of taxes in return for effective property rights and public services but strongly opposed lump sum payments to racketeers such as Farmers-General.

3.22 Both use coercion over individuals to enforce compliance but with varied levels of physical violence.

3.23 Brennan and Buchanan (1977), and the public choice literature after them, refer to the ruler as the Leviathan in homage to Hobbes (1651).

3.24 A net tax rate of 50% has two dimensions. Firstly, it may be the outcome of taxation at 80% whereby 30% of the proceeds are used to enforce taxation i.e., quash revolts and tax evasion. Secondly, anticipating such a high level of taxation, economic agents substitute part of their time towards non taxable activities which are inherently less productive.

3.25 The first best situation implicitly assumes full cooperation among consumers-producers and the ruler for decision making and assume away wealth effects. In that case, it is viable to finance public goods via a lump sum tax t so that the people's utility is $U = (1 - p)D(p,s) - t$ whereas the ruler's net wealth is $\pi = t + pD(p,s) - s$. Welfare $W = U + \pi$ is thus as presented in the text. In this ideal setting, the ruler is a civil servant collecting taxes to finance public goods.

3.26 We assume $R_0(p^0)\Phi'(0) > 1$ to avoid a corner solution.

3.27 As $R_\alpha(.) > R_0(.)$, we have $R_0(p^0) < R_\alpha(p^0) < R_\alpha(p^\alpha)$ because p^α is the maximizer. Now, since Φ' is decreasing, the result obtains.

3.28 Since $p^\alpha < p^0$, the relevant range is $[0; p^0]$ over which $pr(p)$ is increasing.

3.29 It is also the solution of $p^\alpha r(p^\alpha)\Phi(s^\alpha) = s^\alpha \Leftrightarrow \alpha = \frac{p^\alpha}{1-p^\alpha}$ after simplifications.

4.1 This is because, under uniform pricing, the relationship between quantity and total price is linear in the mathematical sense.

4.2 It shall be proved later that it is better to offer the same conditions to anyone within a segment than randomly experimenting.

4.3 They need not be so temperamental; as rational economic agents, they have an incentive to invest time and money to avoid, by whatever means, the higher price they are being asked to pay.

4.4 Under the racial discrimination of apartheid, blacks were barred from using regular public transportation and were forced to use separate carriers; the service being worse,

it was as if they had to pay more to use the regular bus. Under economic discrimination, the reverse would happen: whites would have to pay more because their skin color reveals they belong to a wealthier group whose WTP is higher.

4.5 Article 13 of the EU Treaty allows the EC to take measures to combat discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation.

4.6 There is also a cost explanation to this pricing since the frequency of accidents is higher in urban areas.

4.7 To be sure, the manager can abate all prices by one cent in order that each rental generates a surplus of 1 cent. This makes each rental strictly superior to not renting for the client.

4.8 Club goods are studied in §5.3.2.

4.9 This is mostly a theoretical view since a firm can hardly sustain the reputation of making one and only one offer and walk away from profitable trade should it be refused.

4.10 cf. also footnote 26.3 showing that bilateral relationships are efficient when enough instruments are available.

4.11 For the case where the marginal cost is increasing, we have to determine the efficient total quantity Q^* and use $c = C_m(Q^*)$ in the previous schemes.

4.12 He also proposes as an alternative, to lower the tariff between 7h to 8h and between 17h to 18h which are the peak periods where the large worker population crosses the bridge to reach factories and go back home. In §25, we deal with *seasonality* and show that it generates congestion which in turn results in an inefficient use of the economy's scare resources. We also review the solutions to this inefficiency.

4.13 We drop the "M" superscript label and use instead "d" and "u" to indicate optimal choices under, respectively, *discriminatory* and *uniform* pricing.

4.14 This is a revealed preference argument quite common in economics.

4.15 When marginal cost is not constant, we obtain $q_i = \frac{a_i - b_i C_m(q_0)}{2}$ and summing, we get $C_m(q_0) = \frac{a_0 - 2q_0}{b_0}$ whose unique solution q_0^M determines the optimal discriminatory scheme.

4.16 As is often the case, competition among retailers is more intense in the integrated US market as compared to the scattered european one.

4.17 Because a second hand market for retrade would be active proving that the original allocation was not a Pareto optimum i.e., inefficient.

4.18 If the firm can manage to sell exclusively to the strong segment, she earns $s/4$ which dominates uniform pricing if $\delta < \frac{1}{3}$.

4.19 As we have seen in other instances, $\frac{\partial \pi}{\partial T} = 0$ at the optimum T^* and $W = \pi + W_D$ thus $\frac{\partial W}{\partial T} = \frac{\partial W_D}{\partial T} = d(T, p_2) - d(T, p_1) > 0$ at T^*.

4.20 Affirmative action, though akin to positive discrimination, works through direct regulation. The government imposes a minority quota on institutions who then use positive discrimination to motivate the participation of minority members in order to meet their quota.

4.21 The cheapening of automated (computer driven) systems generated by the IT revolution has reduced the cost of billing as well as the client's inconvenience so that we see more and more firms introducing complex schemes of differential pricing.

4.22 Regarding the consumer, under BOGO he spends 18€ to get 21€ worth of pizza, whereas if he buys only one unit at price p, he gets

14€ worth of pizza plus $18-p$€ left, thus BOGO hurts him only if $p < 14+18-21 = 11$.

4.23 The exact condition is $\frac{a_1}{b_1} > p^M \Leftrightarrow D_1(p^M) > 0$ which must hold true given that we computed p^M under the implicit assumption that both groups are consuming.

4.24 The design of the optimal three-part tariff is not over yet because the monopoly must now optimize the (F, A) combination.

4.25 From $W_{D,\mu}(p) = \int_p^\infty D_\mu(x)\,\mathrm{d}x$, we get $\frac{\partial f}{\partial \mu} = \int_p^\infty \frac{\partial D_\mu(x)}{\partial \mu}\,\mathrm{d}x = \int_p^\infty \frac{\epsilon_\mu(x) D_\mu(x)}{\mu}\,\mathrm{d}x$, hence $\bar{\epsilon}_\mu \equiv \frac{\mu}{f}\frac{\partial f}{\partial \mu} = \frac{\int_p^\infty \epsilon_\mu(x) D_\mu(x)\,\mathrm{d}x}{\int_p^\infty D_\mu(x)\,\mathrm{d}x}$ is an average of income elasticities.

4.26 This observation also known as the revelation principle is studied in §21 on adverse selection.

4.27 We do not consider pricing lists since they are absolutely equivalent to two-part tariffs in this setting.

4.28 One could think of using two-part tariffs and lowering F_h, yet this will be useless unless one unit price, either p_l or p_h is changed.

4.29 As we are dealing with marginal changes, one must imagine extremely thin trapezoidal areas.

4.30 Personalized prices depend on past actions in an implicit rather than explicit way. Sophisticated and experienced consumers are able to predict the effect their actions will have on their subsequent deals, and adjust their behavior accordingly. Dimwits or first-timers fail to account for this linkage and allow firms to exploit their myopia.

4.31 Switching cost are also what keeps us with a supermarket, a gas station, a mobile phone brand, a car brand or any other dealer service we regularly use (cf. §5.1.2).

4.32 Care must also be taken to show that $p_1 \leq p^M$ was not optimal.

4.33 The firm faces twice the same demand thus repeats the optimal one shot price whatever the degree of sophistication of consumers.

4.34 Recall also that the firm does not wish to use commit power to discriminate intertemporally.

5.1 Buyers go to the cheapest price forcing the dear firm to reduce her price if she wants to sell a positive quantity.

5.2 Rename the two farmers α and β, derive the best reply of α facing β and then set alternatively $\alpha = 1, \beta = 2$ and $\alpha = 2, \beta = 1$ to obtain the two best replies.

5.3 The mathematical definition that generalizes this example is due to Nash (1950) and hold this author's name to distinguish it from other forms of equilibrium.

5.4 By the first welfare theorem, the same welfare would obtain if the two firms were behaving as price takers.

5.5 Obviously, a "first-time" user must incur a start-up cost but in all likelihood, it is larger than any future switching cost.

5.6 Even for two identical products, we tend to adapt our preferences towards what the one we know in order to reduce cognitive dissonance.

5.7 The profusion of suppliers is due to the very small cost of entry; it is enough to set up an e-commerce solution (cf. §6.1 on entry).

5.8 In the example where $D(p) = 1-p$ and marginal cost is c, we have $p^C = \frac{1+2c}{3}$ so that the limit where the price starts to decrease is $\delta = \frac{1-c}{3}$.

5.9 Although the same phenomena took place in Europe, the fragmentation of this market

makes data collection more difficult.

5.10 All the data from Table 5.1 was gathered in the news section of the CNET website.

5.11 Footnote 26.3 proves that this is the relevant case for the equilibrium analysis.

5.12 If prices are too large then the indifferent consumer $\tilde{x}$ has a negative utility from consumption i.e., the individuals living in the middle area do not consume and the market fails to be covered. Theater A's demand is then the address solving $u_A(x) = 0$ i.e., $x = \frac{\bar{p}-p_A}{t}$. Profit is then proportional to $(p_A - c)(\bar{p} - p_A)$ leading to an optimal price $\frac{\bar{p}+c}{2}$ and sales of $\frac{\bar{p}-c}{2t}$. A sufficient condition for full market coverage in equilibrium is $\bar{p} > c+t$ in order that each firm is willing to monopolistically serve more than one half of the market.

5.13 In the case where cost are asymmetric, profit is $\pi_A = \frac{1}{2t}(p_A - c_A)(t + p_B - p_A)$, thus the best replies are $2p_A = t + c_A + p_B$ and $2p_B = t + c_B + p_A$, so that the equilibrium is $p_A^* = t + \frac{2c_A+c_B}{3}$. The margin is $p_A^* - c_A = t + \frac{c_B-c_A}{3}$ and the equilibrium profit is $\pi_A^* = \frac{1}{18t}(3t - c_A + c_B)^2 = 2tq_A^2$.

5.14 This is without loss of generality. Indeed, the efficient quantity is the demand when prices are set at marginal cost c i.e., $q^* = a - bc$; thus we can interpret a as the efficient quantity.

5.15 $\frac{1}{\beta} = \frac{b(b+2d)}{b+d} < b + d \Leftrightarrow b(b+2d) - (b+d)^2 = -d^2 < 0$.

5.16 This model remains *ad hoc* since the process of writing and negotiating contracts with potential clients is not tackled.

5.17 Recall that by definition of the competitive supply which equates price to marginal cost, units beyond that point generates losses. The rational behavior, if acceptable for consumers and authorities, is to ration demand and serve at most the competitive supply.

5.18 The laws of physics force the production and transportation of electricity to match instantaneously the household demand which is highly variable due to its correlation with weather conditions.

5.19 We choose $C_m(0) = 0$ w.l.o.g. because it amounts to study profit margin $p - c$ instead of price.

5.20 If a is very small, the equation has no solution which simply means that the market remains inactive; any price is then acceptable.

5.21 The other root being negative, the resulting supply function would be decreasing which is not permitted.

5.22 Indeed, $p^C > p^{SF} \Leftrightarrow \frac{a(1+b\delta)}{b(3+b\delta)} > \frac{a}{2\beta+b} \Leftrightarrow \beta > \frac{b}{1+\delta b} \Leftrightarrow \sqrt{b^2+4b/\delta} > \frac{2b}{1+\delta b} + b \Leftrightarrow b^2 + 4\frac{b}{\delta} > \left(\frac{2b}{1+\delta b} + b\right)^2 \Leftrightarrow 0 < \left(b^2 + 4\frac{b}{\delta}\right)(1+\delta b)^2 - b^2(3+\delta b)^2 = 4\frac{b}{\delta}$ which is true.

5.23 A club shares similarities with a public good but also differences as there is voluntary membership, congestion, exclusion of nonmembers, billing of members and collective decision of the size of the club (cf. Sandler and Tschirhart (1997)).

5.24 Even on the highway, we feel more secure from having mild traffic around us.

5.25 The Samuelson condition for a public good states that the marginal cost should equate society's WTP which sums all member's WTPs. Here, we have a "local" public good since only club members enjoy the service.

5.26 There is a solution because $\frac{C(k)}{n^2}$ tends to zero as n grow large while u_n does not.

5.27 When clients switch from a low utility club A to a high utility one B, they increase congestion at B thus reduce the utility there while they reduce congestion at A which increases utility there. Free flowing among clubs thus re-

duces the utility differential (cf. manna trick §2.1.1).

5.28 With reference to the classical Edgeworth box, the firm and the client trade over good and money. Because price is not involved in such a barter, a Pareto optimum (on the contract curve) is characterized by equality of the RMSs i.e., marginal cost equal to WTP. Money is then available to share the gain from trade in any proportion (cf. also §4.1.3).

5.29 The curve $-u_n \frac{m}{m-1}$ is above $-u_n$ thus cuts $\frac{C(k)}{n^2}$ sooner. It must be pointed out that since we treat m as a continuous variable, the free entry and efficient number of clubs may coincide.

5.30 Although a regime of perfect competition is in complete contradiction with the minimum size of clubs, we may nevertheless check that it is conducive to efficiency. Indeed when clubs are utility-takers, the market clearing equation is simply $u(k,n) - f = \bar{u}$. This leads to $u_n(k,n) = \frac{\partial f}{\partial n}$ and a fee exactly equal to the congestion cost. The free-entry equilibrium then yields the Pareto condition for club size: $TWTA = AC$.

5.31 This can be time (phone, radio, TV), software, consumables such as ink, repair services, upgrades, and other products and services that may be related to the intensity of product usage.

5.32 In a fully specified model of competition, demand depends on the strategies of all active firm.

5.33 We implicitly assume that usage volume is independent of income i.e., the marginal utility of income is constant.

5.34 Because we have $D_f = V_f \times cte$.

5.35 A correct game theoretical analysis of oligopolistic interaction of price discriminating firms must take as strategies the very tariffs that are offered by firms to customers. Spulber (1979) fails that standard when setting output as the strategic variable; his proof that perfect price discrimination is efficient in oligopoly is thus incorrect.

5.36 This shall never happen in an Hotelling model since we always assume complete cover.

6.1 In practical applications, the expenses on governmental licenses ought to be left out since they merely represent a transfer among economic agents. The pointillist may count them at the marginal cost of public funds (cf. §17.1.2).

6.2 Since there is no reason why this number should be an integer, the long-term number of competitive firms is in fact the integer part of n^* while the long-term price and the long-term individual supply are slightly larger than $\bar{p}$ and $\bar{q}$.

6.3 For any other market behavior, the indirect effect is not nil.

6.4 Since the market size is finite and demand goes to zero for very large prices, market sales are limited by some maximum volume V, so that for $n > V/F$, all firms make losses which implies that the process of free entry has a finite limit.

6.5 The multiplier is exactly 2 when the technology has constant returns to scale (cf. eq. 5.12) and is slightly greater when there are decreasing returns to scale, since it can be shown that the equilibrium price obtained in (6.8) yields the profit margin $p_n^C - c = \frac{a-bc}{b} \frac{\delta b+1}{\delta b+1+n}$ which increases when a and b duplicate.

6.6 A more complex theory would be needed to assess which ones are to leave first.

6.7 The issue of quantity competition in the presence of asymmetric technologies is treated

in §5.1.3 where it is shown that firms with the better technology increase their market share and profits upon integration. The same forces are at work but without such an extreme outcome.

6.8 A lower price means greater demand thus greater market sales; even-though individual firm sales might be smaller, this effect is more than compensated by the entry of new firms.

6.9 The French licensee Bic Cristal also did a fortune with a cheap but reliable version.

6.10 Baumol (1982)'s buoyant presentation of contestability as an uprising against heavy regulation must be interpreted in light of the political momentum of the first Reagan administration and its wave of deregulation (cf. §16.4.4).

6.11 Prices depart from marginal cost to cover the fixed cost (cf. §17.2).

6.12 An alternative hypothesis yielding the same conclusions is stickiness (H2'): the incumbent cannot adjust its price as entry occurs.

6.13 The idea goes back to Bain (1956) who states that "most analyses of how business competition works and what makes it work have given little emphasis to the force of potential or threatened competition of possible new competitors."

6.14 Furthermore, as we explain in §10.1, this behavioral assumption is incoherent within a model of interaction between rational decision-makers.

6.15 PC is like Cap Horn, a site reachable from the pacific, the atlantic or the mainland, far from Mexico but still connected to it. Contestability, on the other hand, is like Easter Island, a place filled with symbols of a long gone civilization, alien to ours and useless to understand it.

6.16 We find 20 entries for the period 1980-89 and only 4 thereafter in JSTOR with "contestability" or "contestable" in the title.

6.17 Baumol et al. (1982) initially show contempt when speaking of "the reaction functions and the other paraphernalia of standard oligopoly model". Later on Baumol and Willig (1986) more humbly recognize that "contestability... accomplishes its objectives via a process of simplification, by stripping away all barriers to entry and exit, and the strategic behavior that goes along with them both in theory and in reality". Regarding policy recommendations, they backpedal when "vehemently disagreing with a view of the world associating ... contestability with an all-pervasive laisser-faire position on the role of regulation and antitrust".

6.18 Temin (1997) recalls that Baumol was the chief economist for AT&T; he gave testimonies to the FCC from 1966 until 1996. Schechter (1996) recalls how he argued for theories and cost allocation formulas (ECPR) that benefited his employer, even though he was aware it amounted to blocking entry.

6.19 Microsoft reveals for 2004 that the "Windows" and "Office" divisions ("client" and "Information Worker" in Microsoft jargon) display rates of profits of respectively 75% and 72% which is far above the one digit values observed in the vast majority of markets (cf. also CNET report).

6.20 Whether the technology fits into the general plans of Microsoft and will effectively be used is of secondary importance.

6.21 The threshold for the fine F is $\Pi_1^C\left(q_1', q_2^S\right) - \Pi_1^C\left(q_1^S, q_2^S\right) = \left(\frac{9}{64} - \frac{1}{8}\right)\frac{(a-bc)^2}{b} = \frac{(a-bc)^2}{64b} = \frac{(q_1'-q_1^S)^2}{b}$. To estimate this amount we can use ϵ the elasticity of demand, Q the actual total sales, p the ac-

tual price and Δq the number of rationed consumers to derive $F \simeq \frac{p(\Delta q)^2}{\epsilon Q}$. Notice how it depends quadratically in the rationing created by the nonfulfillment of farmer 1.

6.22 The first truly forward market was established by the Chicago Board of Trade in 1851 for corn. There are now hundreds of forward markets all over the world.

6.23 A forward contract is a derivative security defined by three items: an underlying asset such as oil or the equity of a firm, a forward price and a maturity date. The "long" side has the obligation to buy the asset at the agreed price on the agreed date; the "short" side has the reciprocal obligation of delivering the asset to receive the payment.

6.24 If there is agreement on a price of 10€ over 17 units, the seller will pay an additional 34€to the buyer if the spot price is 12€ while the buyer will pay an additional 51€ to the seller if the spot price is 7€. Each has thus insured himself the ability to transact 17 units at the price of 10€, whatever the future spot price.

6.25 A participant to the market who is neither a producer of the good nor a final consumer.

6.26 We assume that all transactions, forward and spot, are settled in the second period so that no discount factor is needed.

6.27 We use the model of Bertrand competition of §5.2.3 where the demand addressed to firm i is $q_i = a - bp_i + dp_j$; we denote $\pi_i = p_i q_i$. With a forward position f_i, the profit is $\pi_i + (\tau_i - p_i) f_i$ and its maximization leads to $p_i = BR_i(p_j) \equiv \frac{a - f_i + dp_j}{2b}$; note that $\frac{\partial BR_j}{\partial p_i} = \frac{d}{2b} > 0$. As $f_i = \frac{\partial \pi_i}{\partial p_i} = q_i - pb$ and $\frac{\partial \pi_i}{\partial p_j} = pd > 0$, we deduce $\frac{\partial \pi_i}{\partial p_j} = \frac{d}{b}(q_i - f_i)$. Since there is no arbitrage in equilibrium, $\tau_i = p_i$ holds true so that total profit is π_i. The FOC for eliciting forward sales is $0 = \frac{d\pi_i}{df_i} = \frac{\partial p_i}{\partial f_i}\left(\frac{\partial \pi_i}{\partial p_i} + \frac{\partial \pi_i}{\partial p_j}\frac{\partial BR_j}{\partial p_i}\right) \Rightarrow 0 = f_i + \frac{d}{2b}\frac{d}{b}(q_i - f_i) \Rightarrow 0 > f_i = \frac{-d^2 q_i}{2b^2 - d^2} > -q_i$.

6.28 Recall that owners empower managers to take market related decisions because they lack information regarding demand and costs. Owners would like to make sure that managers will act so as to maximize profits. Chapter 13 and Part H analyze this difficult topic.

6.29 The absence of variables relating to the competitor can be explained by the difficulty with which this information could be obtained and above all by a legal requirement from antitrust authorities to avoid collusion (cf. §9).

6.30 If c is very small then the optimal choice is simply $\alpha_i = 0$ i.e., the manager maximizes sales without taking care of cost.

6.31 This is not a result but a coincidence due to our linear modeling; recall that forward sales distort market size while managerial compensation distorts cost.

6.32 Under integration $w_i = 0$ so that raising it a little bit is profitable since $\frac{dU_i}{dw_i} = \frac{\partial U_i}{\partial p_j}\frac{\partial p_j}{\partial w_i} = (p_i - c_i)\frac{\partial D_i}{\partial p_j}\frac{\partial p_j}{\partial w_i} > 0$.

6.33 The manufacturer might even try to use negative prices by giving for free other services to the retailer (but on a quantity basis to maintain the sales incentive.

6.34 This might be incorrect if bankruptcy generates a deadweight loss.

6.35 cf. §23 where our financial terminology is introduced.

6.36 Formally, using compact notations for partial derivatives, equation $\pi(x, y, \theta) = D$ implies $\theta_x = -\frac{\pi_x}{\pi_\sigma}$, thus $S_{xx} = \int_\theta^\infty \pi_{xx}\,\mathrm{d}H(\sigma) + \theta_x \pi_x = \int_\theta^\infty \pi_{xx}\,\mathrm{d}H(\sigma) - \frac{\pi_x^2}{\pi_\sigma} < 0$ assuming as usual that $\pi_{xx} < 0$. Now, $x_D = -\frac{S_{xD}}{S_{xx}} > 0$.

6.37 This is the famous irrelevance theorem of Modigliani and Miller (1958).

7.1 It remains puzzling that standard textbooks still offer no unified treatment of rivalry when the research literature is burgeoning with such models applied to all areas of microeconomics.

7.2 Konrad (2007) and Garfinkel and Skaperdas (2007) offer surveys of contests and conflicts respectively. The first known mathematical analysis of conflict is Borel's Colonel Blotto game regarding the optimal allocation of forces among war theaters.

7.3 Recall that in a conflict, the rules of engagement are determined by the existing environment and the winner's bounty is less than the loser's prejudice i.e., wealth is destroyed by conflict.

7.4 Thief have an opportunity cost of time, thus socially waste their effort at becoming better thieves. Likewise, victims, and the State defending them, invest into private and public protection.

7.5 Tullock (1975) originally introduces the probabilistic lottery for litigation between two agents.

7.6 As shown by Boccard (2010b), a CSF satisfying independence wrt. irrelevant alternatives and consistency with the removal of a dummy player must be separable. If we further require currency independence then the influence technology must be a power function (as used below).

7.7 When h is convex, the party having invested slightly more has a disproportionate advantage to win the prize. This is a realistic assumption for violent conflicts but not so much for economic ones.

7.8 At this point, we may reinterpret the pure influence lottery of (7.2) as the unique equiproportionate success function i.e., such that $\frac{p_i}{p_j} = \frac{e_i}{e_j}$ is true for any pair of participants.

7.9 For the record, if $\alpha = 2$, (7.2) reads $e_0^2 e_i = (e_0 - e_i)\upsilon_i$ thus $e_i = \frac{e_0 \upsilon_i}{e_0^2 + \upsilon_i}$. Summing over all participants, we derive $\hat{e}_0$ as the unique solution of the equation $1 = \sum_j \frac{\upsilon_j}{e_0^2 + \upsilon_j}$.

7.10 Recall that total effect is $\frac{d\pi_1}{dk_1} = \frac{\partial \pi_1}{\partial k_1} + \frac{\partial \pi_1}{\partial k_2}\frac{\partial k_2}{\partial k_1}$.

7.11 cf. Cooter et al. (1982) for an early exposition of this issue in the presence of asymmetric information.

7.12 Stoft (2008), in the context of carbon taxation, speaks of an untax to emphasize the disconnection to usual taxes. More on taxes as incentive instruments in §18.1.4.

7.13 Oddly enough, Samuelson (1947)'s textbook uses this colorful reference to illustrate the first and foremost principle of neoclassical economics, namely that people face tradeoff.

7.14 cf. Vahabi (2005) for a recollection of the theory developed in this hard-to-find book. Another forerunner is Bush and Mayer (1974) state of nature model where private property does not exist (cf. §26.1.5).

7.15 We use the vocable " creation" instead of " production" in order to be able to speak of " productivity" for both creation and capture .

7.16 In Hirshleifer (1988) and the literature after him, λ_i is an index of creative productivity instead of being a negative index of influence.

7.17 The rent-seeking FOC is $\frac{\partial p_i}{\partial e_i} = \lambda_i \Leftrightarrow \frac{1}{e_0} = \frac{\lambda_i}{1-p_i}$.

7.18 Letting μ_i and ϕ_i respectively denote the capture and creative productivities, the equations remain the same with $\lambda_i = \frac{\phi_i}{\mu_i}$.

7.19 In the case of violent conflict, it corroborates the observation that warlords have always enjoyed more wealth than laborers and also that war technology has always been at

the vanguard of a society's advancements. The latter stems from the fact that a marginal improvement can yield disproportionate returns when conflict technology is highly non linear.

7.20 Uniqueness arise from $\frac{1}{1+x}$ being decreasing whereas the fact that it is convex implies $\frac{1}{n}\Sigma_j\left(1+\frac{e_0}{q_0^*}\lambda_i\right)^{-1} = a > b = \left(1+\frac{e_0}{q_0^*}\frac{\Sigma_j\lambda_i}{n}\right)^{-1}$, thus we have $e_0^* < \frac{(n-1)nq_0^*}{\Sigma_j\lambda_j} = (n-1)\frac{\Sigma_j\lambda_j\bar{q}_j}{\Sigma_j\lambda_j}$ (the solution to $a=\frac{1}{n}$ is lesser than that of $b=\frac{1}{n}$).

7.21 If the appropriation process was insensitive to influence then parties would not try to influence the result and full efficiency would be achieved. To see this in the duopoly case, it is enough to change the lottery into $p_i = 1-\phi+\phi\frac{k_i}{k_i+k_j}$. We have totally insensitivity to influence for $\phi=0$ and the pure lottery for $\phi=1$. Performing the calculations, we obtain $q_0 = \frac{2}{2+\phi}$.

7.22 In violent conflicts, parties often settle to avoid destructive war but it is clear that this is an "armed peace" or a "cold war" equilibrium.

7.23 This widely cited paper uses an analytical framework quite different from the canonical one and also less straightforward.

7.24 There is a clear similarity with the "business stealing" effect in the Hotelling model where gaining one client is necessarily the loss of one for the competitor (cf. §5.2.2).

7.25 This phenomenon is quite similar to Hotelling competition. As shown in footnote 26.3, when firm A improves its technology (lowers marginal cost c_A), it lowers its price and in response so does firm B. In the new equilibrium, both firms are more aggressive, but the innovative one enjoys a greater profit at the expense of the other one. However, the change in market shares is only one sixth of the cost improvement since $\tilde{x} = \frac{1}{2}+\frac{c_B-c_A}{6t}$.

7.26 Muck like in the Tragedy of the Commons where inefficient over-exploitation ceases as soon as property rights are established and allocated to a monopolist.

7.27 Alternatively A includes the government and civil servants who are faithful to the politician since they owe their job to his safekeeping power.

7.28 If, as a matter of example, we let $G_B(\theta)-G_A(\theta) = \theta/\mu$, then $\hat{\theta}(p) = \mu p$. Since $\pi_B = v_B H\big(\mu(e_B-e_A)\big) - \frac{1}{2}\lambda_B e_B^2$, the FOC is $h\big(\mu(e_B-e_A)\big) = \frac{\lambda_B e_B}{\mu v_B}$. As $\pi_A = v_A\big(1-H(\mu p)\big) - \frac{1}{2}\lambda_A e_A^2$, the FOC is $h\big(\mu(e_B-e_A)\big) = \frac{\lambda_A e_A}{\mu v_A}$. Combining, we get $\lambda_B v_A e_B = \lambda_A v_B e_A$, very much like the cost minimizing capital labor ratio for a Cobb-Douglas technology. Substituting in either FOC, we obtain the equilibrium levels of pressure as the solution to $h(\alpha e_A) = \frac{\lambda_A e_A}{\mu v_A}$ with $\alpha \equiv \mu\left(\frac{\lambda_A v_B}{\lambda_B v_A}-1\right)$ (assumed positive to make sense). A worsening of lobby i's technology ($\lambda_i \nearrow$) leads to $e_i^* \searrow$, $e_j^* \nearrow$, $\pi_i^* \searrow$ and $\pi_j^* \nearrow$.

7.29 The best reply $\phi_B(e_A)$ satisfies $\phi_B(0) > 0$ and has slope lesser than unity because a one-to-one response only maintains net pressure while being costly. The best reply $\phi_A(e_B)$ satisfies $\phi_A(0) = 0$ and has slope lesser than unity as before. As a simple geometric representation shows, there is an equilibrium with $e_B^* > e_A^* > 0 \Rightarrow p_A > 0 \Rightarrow \theta^* \equiv \hat{\theta}(p^A) > 0$.

7.30 Formally, this lottery is identical to the differential productivities form alluded to in (7.1).

7.31 Consider an externality where parties haggle over the level q emitted by B (value function U) and weighting on A (cost function C). Whereas B would like to pick a large $\bar{q}$, A would like to pick a low level $\underline{q}$. We thus have $V_B = U(\bar{q}) - U(\underline{q})$ and $V_A = C(\bar{q}) - C(\underline{q})$.

7.32 Since $\lambda_j = 1-\lambda_i$, the numerator of k_i^* decreases while the denominator increases.

7.33 Recall that in a world of zero transaction cost, initial ownership does not matter.

7.34 For the externality, $W^* = max_q U(q) - C(q)$.

7.35 An integration by parts shows that the expected discovery time is the inverse of the hazard rate i.e., $\mathbb{E}[\tau] = \frac{1}{h(k)}$.

7.36 The density $\Pr(\tau = t) = h(x)e^{-h(x)t}$ satisfies $\int_0^t ze^{-zs}\,\mathrm{d}s = 1 - e^{-zt}$.

7.37 This is a consequence of the mixed strategy equilibrium followed by firms.

7.38 Observe now that $\pi_{10}(\vec{w}^S) + \pi_{01}(\vec{w}^S) - \pi_{11}(\vec{w}^S) - \pi_{00}(\vec{w}^S) = \frac{c(q_1-q_0)}{1-q_1} > 0$ thus $\pi_{10}(\vec{w}^S) - c > \pi_{00}(\vec{w}^S) + \pi_{11}(\vec{w}^S) - c - \pi_{01}(\vec{w}^S) \geq \pi_{00}(\vec{w}^S)$ by (7.14), hence we have shown that working dominates shirking when the other shirks. It is therefore impossible that both agents shirk and if one works the other prefers to work which proves that $\vec{w}^S$ implements mutual effort at minimum cost.

8.1 Absence of rights is often characterized by an open access regime.

8.2 For natural resources and environmental assets, one may hold a right to use, a stronger right to harvest or even the right to manage the asset.

8.3 The driver seem to be reciprocity, a universal, spontaneous, and evolutionary fit behavior. Experimental economics has shown that subjects tend to cooperate in the prisoner dilemma (as well as other social dilemmas) and that voluntary contribution for public goods are much larger than static models of game theory would predict (cf. Camerer and Loewenstein (2002) or McKenzie's contrarian point of view).

8.4 One could argue that the incumbent values the asset highly because he invested much in it but this is a sunk cost fallacy. Indeed, to maximize future returns, one ought consider only future payoffs of the asset irrespective of how much was expended on it in the past.

8.5 In the language of game theory, first possession is a correlation device that bring parties towards a correlated equilibrium without conflict (cf. bourgeois strategy in the hawk dove game or "tit-for-tat").

8.6 Shleifer (2010) is an excellent read, all the more given the "laissez-faire" stance of this author in his early days.

8.7 At a different level, the world trade organization (WTO) is an arbitration court with no formal authority upon countries whose decisions are however upheld by all. The reason here is the powerful deterrence effect of reputation; if a country refuses to abide, others will retaliate either directly with import taxes or implicitly by losing confidence.

8.8 In the ideal world of complete contracting, there is just one auditor and one judge for an entire country who sit idly since no one ever infringes the law. Their presence is nevertheless necessary to guarantee that any infraction will be detected, processed and fined.

8.9 The same applies for a faulty device that includes many components (e.g., BP Deepwater Horizon oil spill in Louisiana).

8.10 As we explain in §21 on adverse selection, market failures reduce the overall efficiency of the economy which is why the government seeks to reduce their impact.

8.11 cf. Chronology of treaties and original texts.

8.12 The former is a potential importer of agricultural products and is thus reluctant to institute the CAP. France and its ally Netherlands, are on the other hand potential exporters willing to match the opening of industrial sectors that favors Germany with the agricultural

ones. This tense negotiation also cuts back the commission's power leaving the council free to decide on all support mechanisms.

8.13 The votes of the member states are weighted in principle according to the population size of the individual EU countries but the small and medium-sized countries have a larger than proportional number of votes to safeguard their own interests.

8.14 The "de minimis" notice on agreements of minor importance which do not appreciably restrict competition under Article 81(1) of the Treaty establishing the European Community, OJ C 371, 9.12.1997.

8.15 We use the 2002 consolidated version of the 1957 treaty, now known as the *EC Treaty*, where article numbers are changed with respect to the original.

8.16 The Sherman Act was also used to prosecute labor unions who monopolized industries.

8.17 It is frequently applied by the lower federal courts to prosecute labor unions who monopolized employment in whole industries.

8.18 Some agreements or behavior are so likely to harm competition and so unlikely to produce pro-competitive benefits that they do not warrant the time and expense required for an inquiry into their effects.

8.19 For instance, setting a price range, a common list of prices for discounts, agreeing on discounts or on terms and conditions of sale, limiting supply and even agreeing on a common standard are unlawful.

9.1 As we explained in §8.3, cartels were referred to as *trust* in the US which in turn explains why the struggle against trusts by public authorities lead to the now accepted terminology of *anti-trust*.

9.2 To simplify, we disregard the issue of fixed cost.

9.3 The ceiling allocations that cover short periods are averaged to match the data on yearly production.

9.4 Iraq is not included in the data due to the eventful destiny of this country; notice also that Indonesia is now a chronic importer of oil.

9.5 There are three basic levels of licensing controlled by the government. Under *registration*, one must solely register identity and some personal information (possibly for some fee), *certification* requires an examination for minimum skills and knowledge and finally *licensure* is a harder and better protected exam for the right to practice.

9.6 The EC nevertheless calls a cartel a group of firms participating in a collusive agreement.

9.7 A related clause is the *most-favored-nation* provision whereby an (international) trading partner is promised that his import tariffs will be no higher than those offered to other trading partners (in the past or future).

9.8 This can be explained by quality uncertainty (cf. the lemons problem in §21.1.1), transaction costs to set up the resale or low residual value if already consumed.

9.9 The interaction is finite because durable goods have a finite live span and MFC or MOR clauses do not apply to next generation products.

9.10 The last period being used to retaliate, it necessarily brings a competitive outcome; this is why collusion cannot last until the end.

9.11 Indirect ways include fixing the distribution margin, conditioning promotional cost reimbursement on price fixing observance or threatening to delay or suspend deliveries.

10.1 Introduction of many variants of a basic product like automobile or breakfast cereals. Still if variants can be withdrawn cheaply there will be entry because nobody believes that the incumbent will keep so many substitutes to its own base line (cf. §11.2.1).

10.2 This legal term means to hinder, deter, impede or handicap. It originally meant the legal proceeding initiated by a creditor to repossess the collateral of a defaulting borrower.

10.3 In some formulations the postulate says that "the entrant believes the incumbent won't react". Such a belief is coherent only if experience has shown that incumbents did not react to entry (otherwise the entrant would be irrational). The postulate is thus about the incumbent's behavior and not about the perception of others.

10.4 There might be some technical impediments to immediate adaptation but if the incumbent wants to change his plans, he will be able to do so in a near future and since the relevant profit is the discounted sum of future profits, the change will be reflected. Thus, we might as well assume that marketing changes are instantaneously effective.

10.5 Their result is been extended Segal and Whinston (2000) by for price discrimination and sequential bargaining.

10.6 Since damages for breach of contract tend to be per-se illegal (cf. Posner (2001) p.232), the reader must keep in mind that we are dealing with take-or-pay contracts.

10.7 We have $p_f = v$ at $c = c_2$ and c_1 increasing with x.

10.8 Formally, there is indeterminacy for the captive and free prices upon entry since we have only one zero-profit condition. Yet, if we adopt a minimum price discrimination rule, then the captive price remains at x the value it must achieve when $c = c_1$.

10.9 The fact that there are at least two such firm competing to get the exclusivity guarantees that whoever wins, he gets no rents; hence the upstream firm reaps the whole monopoly profit.

10.10 The symmetric equilibrium solves $q^* = R((n-1)q^*)$.

10.11 In EU law it is called Open Network Provision.

10.12 Dupuit (1852), citing older French canal legislation passed in a time of "laissez-faire", is a forerunner of this position except for railways out of security concerns. Oddly enough, France now stands in complete opposition to this liberalizing practice.

10.13 Where prices are set below average total (but above variable) costs, some additional elements proving the predator's intention need to be established in order to qualify them as predatory, given that other commercial considerations, like a need to clear stocks, may lie at the heart of the pricing policy.

10.14 Some countries have non-economical motives for industrial policy such as national independence, technological autonomy, support for declining traditional activities or sustaining geographical and political cohesion.

10.15 Imitative entry occurs through managerial and labor turnover (former employee starting their own business). Knowledge is non-rival and often non-excludable which leads to positive technical externalities and increasing returns. Furthermore, new knowledge tends in the aggregate to complement existing knowledge.

11.1 The only exception to this rule has to do with quality differentiation; one can indeed find

door to door, a discounter, a standard supermarket and a delicatessen shop.

11.2 Decades after Hotelling's pioneering work, d' Aspremont et al. (1979) exhibit a flaw within his original linear city model. Fortunately, Economides (1986) shows that the flaw disappears as soon as they are several geographical dimensions which in our basic story amounts to add North and South as possible locations for firms.

11.3 If consumers are not uniformly distributed over the street but according to the cumulative distribution function $F(x)$ then profit is $\pi_A = p_A F(\tilde{x})$ so that the FOC of best reply is $F(\tilde{x}) = \frac{p_A f(\tilde{x})}{2(b-a)}$ since $\tilde{x}$ solves $p_B - p_A = (a+b-2x)(b-a)$. Using the symmetrical formula $1 - F(\tilde{x}) = \frac{p_B f(\tilde{x})}{2(b-a)}$, we derive the indifferent consumer as an implicit function of the location midpoint $\frac{a+b}{2} = \tilde{x} - \frac{1-2F(\tilde{x})}{f(\tilde{x})}$.

11.4 Imagine, after Jean François Mertens, taking the $[0;1]$ segment and bending it so as to join the extreme 1 and 0 values to form a ring.

11.5 The proper analysis requires to consider asymmetric locations, work out demands, price best responses, equilibrium and compute profits as a function of neighboors locations. One can then proceed to show that a firm fine tunes her location to be just in between her two neighbors. In equilibrium of this process symmetric locations obtain.

11.6 cf. regional economics course

11.7 It can be assumed competitive because there are many individual buyers in the city and many supplying farmers around.

11.8 Although land cultivation has decreasing returns to scale, we assume constant returns to scale for family production.

11.9 All the economic surplus of cultivation goes to landlords which is why it has been called a rent by economists.

11.10 If demand for good i increases, so does its price; the bidding curve then rises and the share of land dedicated to this crop increases to satisfy demand.

11.11 If it was not so, unhappy places would be vacated, forcing down the local housing price.

11.12 Formally, $\frac{\partial p_x}{\partial x} = \frac{-t}{q_x} < 0$ by the envelope theorem.

11.13 Suppose the price falls so much that she may maintain her initial bundle of consumption and therefore her initial utility. She would then consume more housing as it is now cheaper (cf. equality of MRS to relative price in §2.2.1) and be happier, but this cannot happen, so that the price must increase. Meanwhile this process takes place, she reduces her housing but never back to what she could afford initially so that in the end, her dwelling is larger (use a graphic to check this).

11.14 As w increases, q_x rises, thus the slope of the bidding curve $\frac{t}{q_x}$ also rises.

11.15 Executives refer to it as cannibalization; it typically occurs when a new model is introduced while the old one is still for sale although they are priced differently.

11.16 With m attributes normalized on a $[0;1]$ scale, the opportunity cost is $\sum_{i \leq m} t_i (x_i - z_i)^2$. Yet, our ability to aggregate complex information is limited by our cognitive power to the point that we value each lower rank attribute twice less than the previous one and so on; this implies that the most important attribute weights as much as all the rest and we may safely continue with two dimensions, a strong and a weak one.

11.17 Either the representative agent values diversity through the composite good or y is the

aggregate utility for different people consuming different brands. Authors also assume unit income elasticity to avoid income distribution issues in their general equilibrium model.

11.18 A lawyer who can type very fast will nevertheless let his secretary do that job to concentrate on legal studies where his skills are best rewarded.

11.19 On the one hand, he downsizes this model by performing a partial equilibrium analysis with $\alpha = 1$ but on the other hand, he allows for decreasing demand elasticity, thereby making the optimal price increasing with output (like the competitive supply function of a DRS technology). Notice that this obscuring feature fails to add anything to his findings.

11.20 In Krugman (1979)'s model with decreasing demand elasticity, increasing labour endowment leads to more variety, higher welfare, smaller per-firm output, thus a lower price or equivalently a higher wage. If trade is prohibited but workers are mobile, they'll all go to the large country that already enjoys more variety and higher wages. Thus industrial concentration might occur.

11.21 The maximal quality achievable is bounded by the technological state of the art.

11.22 For $i = A, B$, j denotes the other firm.

11.23 Danone, an important food producer was in the 1980s just one yogurt brand of a multi-activities conglomerate without much public exposure. In 1994 it decided to change its name to Danone and concentrate on quality food. Sales growth have been accompanied by much advertising.

11.24 Nestlé which was already the world's largest and most diversified food company in 2000 is pursuing in 2003 a clear strategy of concentrating on high-end food production by divesting non-core business and buying famous brands in the US where its market share is still much lower than in Europe.

11.25 cf. Brandz report on firms with highest levels of goodwill.

11.26 Since profit is concave in price, the sign of $\frac{\partial p_a^M}{\partial a}$ is that of $\frac{\partial^2 \Pi}{\partial p \partial a}$.

12.1 Among other things, it also contributed to stop whaling that was formerly a prime source of oil.

12.2 The figure expressed in year 2000 dollars uses a capitalization rate of 11% representative of the industry's profitability. Such a discounting is made necessary by the fact that the testing phases generally last more than a decade.

12.3 A reinforced cooperation is called a *Consortium*; it is made by a group of independent companies working together for the fulfillment of a specific project like a building or a motorway which requires close cooperation between engineering, planning and construction. A still greater form of cooperation is the *joint venture*.

12.4 Reverse engineering is the process of extracting knowledge from a human-made artifact i.e., starting with the known product and working backwards to divine the process which aided in its development or manufacture. International treaties on the protection of trade secrets are silent with respect to it but the software industry has recently succeeded to obtain protection in the US and Europe to fight piracy.

12.5 The iPod-iTunes bundle also paved the way to the legal selling of digital music over the internet.

12.6 Ideally, the patent office should compute the duration that enables the developer to recoup exactly his R&D costs but this would put the office under the pressure of lobbies (industry, consumers and government) with the

likely outcome that incumbent wealthy industries would get lengthier patents than the innovative and fragile industries which most need it (from a social point of view); hence it is more efficient in the long run to have the same fixed duration for all inventions whatever the sector they come from, even-though this may appear sub-optimal at first sight.

12.7 Aspirin is one of the few exceptions; it was invented more than a century ago and is still marketed as an original product by its patent holder Bayer. Notice that the name remains a trademark of Bayer.

12.8 Recall from §6.1 that there is excessive entry in most oligopoly models.

12.9 Technically, the profit at the optimum is Π_c^M satisfying $\frac{\partial \Pi_c^M}{\partial c} = -D(p_c^M)$ by the envelope theorem. Integrating, we get $V^M = \int_{\underline{c}}^{\bar{c}} D\left(p_x^M\right) \mathrm{d}x < V^*$ since $p_c^M > c$.

12.10 Technically, we have $D(\bar{c}) < D(c)$ over $[\underline{c};\bar{c}]$ hence $V^B = \int_{\underline{c}}^{\bar{c}} D(\bar{c})\,\mathrm{d}p < \int_{\underline{c}}^{\bar{c}} D(p)\,\mathrm{d}p = V^*$.

12.11 The incumbent in an oligopoly with $n > 2$ prefers ad-valorem royalties only if she can negotiate with the entire industry. If she has to negotiate separately with each firm, she prefers per-unit royalties.

12.12 Notice that $\delta \geq \bar{\delta} \Rightarrow \bar{m} \leq 1$ which is why we refer to such a situation as a drastic innovation since it allows a sole firm enjoying the innovation to exclude all obsolete ones.

12.13 More specifically, $m^F \leq n \Leftrightarrow n \geq \frac{2(\bar{m}+1)}{3}$ and $m^F \leq \bar{m} \Leftrightarrow n \leq 2\bar{m} - 2$. Recall also that we are ignoring the integer problem i.e., the fact that m is an integer.

12.14 This is only correct for $m < n$ since otherwise auction and fee are equivalent.

12.15 Unless they are subsidized but this will not happen.

12.16 Although the exact number of licenses is indeterminate, it is never n because doing so reduces the auction to a fee setting which we know to be dominated.

12.17 The Fronde revolt against the French king in the XVIIth century was in part triggered by his selling additional offices and licenses which enraged the current holders of these prerogatives.

12.18 The same problem occurs for art printing. The printer of a lithography is supposed to break the copper mold of the original once the scheduled number of copies have been printed.

12.19 In settings where the upstream firm sells a conventional input, she can build commitment by foreclosure (cf. §10.3) or vertical restraints (cf. §9.2.2).

12.20 If after entry, firms compete a la Bertrand instead of Cournot, the condition reads $\Pi_e^B + \Pi_i^B \leq \Pi_{\underline{c}}^M$ where $\Pi_i^B = 0$ and $\Pi_e^B = V^B$. It is satisfied because V^B is the monopolist's profit when pricing at $\bar{c}$ which is not the monopoly price.

12.21 If there is rivalry among at two challengers, a model of contest or conflict is needed to assess how much they would spend (cf. §7).

12.22 Gilbert and Newbery (1982) consider a time-continuous framework with discount factor r starting today (date 0). If she innovates first at date t, the challenger earns, in present value, $V^e(t) \equiv \int_t^\infty \Pi_e^C e^{-rt}\,\mathrm{d}t$ while the incumbent earns $\underline{V}^i(t) \equiv \int_0^t \Pi_{\bar{c}}^M e^{-rt}\,\mathrm{d}t + \int_t^\infty \Pi_i^C e^{-rt}\,\mathrm{d}t$. If he innovates first at date t, the incumbent earns, in present value, $\overline{V}^i(t) \equiv \int_0^t \Pi_{\bar{c}}^M e^{-rt}\,\mathrm{d}t + \int_t^\infty \Pi_{\underline{c}}^M e^{-rt}\,\mathrm{d}t$. Hence the values for innovation are $V^i = \overline{V}^i(t) - \underline{V}^i(t) = \int_t^\infty \left(\Pi_{\underline{c}}^M - \Pi_i^C\right) e^{-rt}\,\mathrm{d}t$ and $V^e = V^e(t)$ for the equilibrium date t. The difference $V^i - V^e = \int_t^\infty \left(\Pi_{\underline{c}}^M - \Pi_i^C - \Pi_e^C\right) e^{-rt}\,\mathrm{d}t$ is thus the integral from t to infinity of the instanta-

neous difference in producer surplus between monopoly and duopoly.

12.23 Management methods such as profit sharing and bonuses linked to the attainment of concrete objectives also display DRS since they force employees to bear risk; they thus become increasingly more expensive as their scale and breath increase inside the firm (cf. ch.20 on Moral Hazard).

12.24 As innovation makes firm A stronger, it forces competitors to retreat i.e., $\frac{\partial z_B^*}{\partial x_A} < 0$ whether z is price or quantity. Now recall that $\frac{\partial \pi_A}{\partial p_B} > 0$ (cf. eq. 5.16) while $\frac{\partial \pi_A}{\partial q_B} < 0$ (cf. eq. 5.2), thus the combined sign of the strategic effect.

12.25 This is done to simplify the exposition with the purpose of presenting the case opposite to Bertrand where there cannot be any market sharing.

12.26 The specific knowledge on a product or production process, often obtained through extensive and costly R&D. It is deemed to be a body of technical information that is secret, substantial and well identified (EC definition).

12.27 The term originally denoted royal decrees granting exclusive rights to certain individuals or businesses, much as a charter, the ancestor of the corporation, bestowed rights to some institutions.

12.28 There are older cases but they are more akin to privileges. In the Greek city of Sybaris in Sicily, 500BC, the king would grant a one year exclusivity to chefs who had invented a new recipe. Concessions for underground exploration and mining were granted for a limited period of time without ownership on the land itself. Privileges consisting in 5 to 20 years of exclusivity were awarded by cities to foreign craftsmen accepting to settle there to perform their art and to train local workers (e.g., Brunelleschi patent for marble transport on the Arno river, granted by Florence in 1426.)

12.29 The Uruguay Round Agreements on Multilateral Trade Negotiations of 1994 has set the duration to 20 years (it was previously 17 years in the US). All WTO member countries apply this rule.

12.30 US patent #4,873,662 for “an information handling system and terminal apparatus therefor" was filed in 1980 and awarded in 1989. In 2000, the owner British Telecom, sued some large US internet access providers claiming that its patent covered web browsing and hyperlinking. The claim was demised by courts in 2002 given the evidence of older hypertext projects.

12.31 US patent #4,698,672 “for reducing the quantity of data in image coding" was awarded in 1989. The current owner Forgent Networks sued large computer makers in 2004 for allegdly using the patent in their JPEG graphical format. Similarly, in 1996, Unisys and Compuserve began seeking royalties on their patented LZW compression method, a key element in the GIF graphic file format; it was quickly replaced by the free PNG.

12.32 A similar lawsuit against Daimler-Chrysler was dismissed late 2004.

12.33 US patent #6,460,020 for an international transaction system over the internet covering catalogue, currencies, taxes, shipping cost and addresses was filed in 1997 and awarded in 2002 to a start-up DE Technologies who sued the computer manufacturer DELL in late 2004.

12.34 US patent #5,960,411 was filed in 1997 and awarded in 1999 to Amazon for 1-click purchase described as “a method and system for placing an order to purchase an item via the Internet". Once the buyer has registered him-

self with the seller, the later sends a “cookie" to the buyer’s computer to ease future transactions. The idea is identical to opening a bank account, obtaining a credit card and using the later in a shop afterwards. Amazon sued immediately Barnes & Nobles for devising a computer code implementing the very same idea; the case was settled privately.

12.35 US patent #6,782,370 covers a “System and Method for Providing Recommendation of Goods or Services Based on Recorded Purchasing History” was filed in 1997 and awarded in 2004 to Cendant Corporation, a hotel and car rental company, that immediately sued Amazon for its practice of recommending books to viewers.

12.36 The original convention of 1886 was changed into the Universal Copyright Convention in 1952 at Geneva and revised in 1971 at Paris.

12.37 The author agrees with US federal court judge Posner that the sheer majority of works which have a very low value should be made freely available to constitute the base of new creations. The very fact that a formal (even free) agreement must be obtained from the copyright holder to use his work creates transaction costs (cf. §13.3.3) thwarting the creation process. He also remarks that “adding 20 years to the already long 50 years copyright term has only negligible effects on the incentive to create a copyrighted work in the first place. Who would refuse to create a work unless he could count on his heirs’ receiving income from it more than 50 years after his death?”

12.38 Note: a “brat” is spoiled child. As explained by Duncum (2007), the reason why Bratz captured a large market share within a few years of its introduction (unlike previous attempts over the 50 years reign of barbie) is “age compression”, the fact that modern girls become preoccupied with adult life sooner and thus leave traditional dolls sooner (aka Kids Getting Older Younger). The more trendy Bratz, then, offers a bridge between childhood and adolescence (cf. also the staggering success of the Titeuf comic (1M books/year) geared at the same age category).

12.39 The jury did not award any punitive damages and found that MGA acted willfully when employing the designer.

12.40 The argument of lower quality sometimes used to bar import from poor countries to richer ones does not apply here since a US-made pair of Levis is THE genuine product.

12.41 Region codes correspond to the following areas of the globe: 1 for Canada and USA, 2 for Japan, Middle East, South Africa, Western Europe, 3 for East Asia, Southeast Asia, 4 for Pacific and other americans countries, 5 for Africa, Eastern Europe, India, Mongolia, North Korea, 6 for China and 8 for airplanes, and cruise ships.

12.42 The difficult coding of the singer’s voice was first tested on Suzanne Vega’s song tom’s diner, encoded here with Advanced Audio Coding at 20 kbits/sec i.e., $1/72^{th}$ of the space needed on a CD.

12.43 In a memorable quote in front of the US congress, the head of the MPAA said, “the VCR is to the American film producer and the American public as the Boston Strangler is to the woman home alone.”

13.1 Our data is limited to the US and France as Eurostat does not yet compile exhaustively this information. The VA percentages are obtained as the ratio of value added (inside firms) to gross output (market sales of final products).

The sectors differ in the two tables for statistical institutes have a differing focus of analysis. The sectors weights, as % of GDP, are reported in table 1.1.

13.2 Incorporation is the legal process through which a company receives a charter from the State that allows it to operate as a corporation. The articles of incorporation describe the regulations that govern the running of the company, they include the main business and purposes of the company, the shareholder's voting rights, the directors duties, the general working and management practices.

13.3 This is called a single tier system; there also exists a two-tier system where stockholders elect a supervisory board beside the administrative one.

13.4 Basing compensation on performance is one frequent form of contest (cf. §7).

13.5 Notice that once docility is present, society may exploit it by teaching values that are truly altruistic; that is, which contribute to the society's fitness, but not to the individual's (like going to war singing).

13.6 In not-for-profit organizations like public administrations, the authority structure and the prestige associated to high levels make the influence cost inefficiency particularly binding. The rational response is thus bureaucracy, a set of rigid rules that sometimes do not make sense at the individual level but overall protect the institution from abuses.

13.7 This meritocratic conception is directly inspired by the chinese mandarin system, dating from the 7^{th} century whereby public service positions were filled through written examination open to all males. This scheme, as well as knowledge of the chinese political and economical system was brought to Europe by jesuits during the 17^{th} century and strongly influenced the enlightenment thinkers. Adoption occurred during the 18^{th} in France (early through the revolution), England and the US.

13.8 It must be noted that during the middle ages, official positions were few compared with the aristocratic population (itself very small). To fill in the positions awarded by rulers, families somehow competed with their most able siblings thereby guaranteeing a minimal meritocracy. As the State grew in size and became involved with technical activities (war machinery, boats, roads, canals), adequate education became indispensable; it lead to the creation of specialized graduate schools.

13.9 It is customary to refer to functions as units and thus to the organization as the "U-form". The organization around products is said to be multidivisional, so that one speaks of the "M-form".

13.10 A likewise dichotomy seem to apply to China vs. USSR when seen as large centralized economic producers. Whereas the defunct Soviet economy consisted of dozens of specialized ministries, China comprises a large number of reasonably self-sufficient regions.

13.11 For instance, Chakravarty and MacLeod (2009) analyze the case of building contracts in the US and show they are based on a standard template with only a few terms modified to deal with special circumstances.

13.12 There exists a theory of complete contracts stemming from the even more abstract implementation theory which proves that complex contracts can achieve almost anything in ideal environments (including the first-best). As soon as one of the previous issues start to bite, complete contracts becomes less useful and may be dominated by simpler ones.

13.13 Many specific bargains are cheap to strike, but since most relationships involve a large number of details, the small costs can add up to a substantial level if we are to negotiate on everything.

13.14 Like all in behavioral works, the theory is fairly simple since there is no strategic interaction, only external social forces that push individual in a certain direction; the force of this approach lies in its empirical validation that purely theoretical works lack.

13.15 Interestingly, the norm is not imposed by society upon the parties but endogenously designed so as to fit their objective.

13.16 Notice that such a scheme requires a common cultural understanding in order for the chosen norm to have the same meaning to each party.

13.17 Gneezy and Rustichini (2000) explain how parents are usually ashamed when arriving late to pick-up their children at school because the master has a duty to stay (without extra pay). Once fines for latecomers were introduced, many parents started arriving extremely late, happily paying the price; they had turned the school into a rather inexpensive day-care center.

13.18 Mathematically, the joint payoff is a convex function of the degree of complexity so that the maximum is achieved at a corner.

13.19 Most people take pride in doing the right thing and do not suffer from the extra effort it may require wrt. the legal minimum.

13.20 It may even be an informal contract enforced by social pressure.

13.21 The tendency to believe that one brings special skills to the relationship, or has invested into human capital specific to it and deserves to be rewarded for these things.

13.22 The empirical analysis also uncovers a new behavioral force: ex-ante competition legitimizes the terms of the contract, and aggrievement occurs mainly about outcomes within the contract and not about the contract itself.

13.23 Of course, if the vehicle hovered over Africa, rural China or India, the green areas would be much smaller, and there would be large spaces inhabited by the little black dots we know as families and villages. But the red lines would be fainter and sparser in this case, too, because the black dots would be close to self-sufficiency, and only partially immersed in markets. This implies that the study of rural economies cannot be performed with the standard framework that was developed by economists of the industrialized countries to respond to the preoccupations of these specific societies.

13.24 The changes at this level are either peaceful and slow like the EU construction or occur suddenly through revolutions (e.g., French Republic or USSR) or in the aftermath of wars (e.g., US independence or post 1945 Japan).

13.25 The primary purpose of institutions is to reduce uncertainty by establishing a stable framework (but not necessarily efficient) for exchange. The ability of a society to maintain social peace and foster economic performance is now widely linked to the choice of its institutions such as political regime, customs, laws and economic regulations (cf. North et al. (2007, 2006) for a synthetic introduction).

13.26 If it was possible to write contracts covering every meaningful future event then no litigation would ever occur. Indeed, if one party tries to act opportunistically, the other party calls on to the judge; the latter is able to ver-

ify what is happening, look at what the contract says and forces the wrongful party to perform. Anticipating, this deterministic course of event, each party is compelled to respect every single aspect of the contract. As a result, courts are inactive.

13.27 If the managerial effort e (counted in €'s of opportunity cost) participates to decrease the cost $C(q,e)$ then the manager-owner will expand the effort e^* solving $\frac{\partial C}{\partial e} = -1$ which is the efficient effort at the society level.

13.28 For instance, money (coins, bill, bank check, credit card) eliminates the transaction cost of barter which requires a double coincidence of wants.

13.29 We deliberately use the vocabulary of incentives and agency as it is the theory that will be used to shed light on this issue in the next chapter.

13.30 Nickname of perquisite, an extra allowance or privilege enjoyed as a result of one's position.

13.31 This would be the case if the parties can use a rigid mechanism to share any benefit since they will be unable to influence the division.

13.32 This author thought that negotiation would resolve efficiently this issue but failed to understand the importance of sequentiality and commitment.

13.33 The human capital of employees is not transferable as slavery is not an option. However employment contracts customarily include a clause specifying that the output of intellectual activity like software code or industrial designs remains the exclusive property of the firm.

13.34 Leontief (1946) compares these two schemes and elegantly demonstrates that the price-only negotiation is akin to standard monopoly pricing, thus inefficient, while the price-quantity negotiation is akin to a perfect discrimination, thus efficient (cf. §4.1.3).

14.1 For instance, most companies contract office cleaning to specialized firms.

14.2 Marshall (1890) agreed with the collective rationality argument but claimed that the price mechanism would avoid indeterminacy even though he failed to explain clearly how and why.

14.3 The error is probably due to the obsession with the market price mechanism that is still widespread in the economics profession.

14.4 Reversing the roles amounts to let the buyer set the price and then allow the seller to choose the quantity. The seller thus maximizes $pq - C(q)$ and demands the competitive quantity $s(p)$ solving $p = C_m$. The buyer then maximizes $R(s(p)) - ps(p) = R(q) - qC_m(q)$ using the change of variable $q = s(p) \Leftrightarrow p = C_m(q)$. The FOC is thus $\frac{\partial R}{\partial q} = \frac{\partial C}{\partial q} + q\frac{\partial^2 C}{\partial q^2} > \frac{\partial C}{\partial q}$ since cost is convex; the buyer then selects a quantity lesser than the optimal one $\bar{q}$.

14.5 When $\pi_S = F - C(\bar{q})$, $\frac{\partial \pi_S}{\partial k_S} = -C_k = \frac{\partial \pi}{\partial k_S}$ and likewise $\pi_B = R(\bar{q}) - F$ implies that $\frac{\partial \pi_B}{\partial k_B} = R_k = \frac{\partial \pi}{\partial k_B}$.

14.6 The original reason why S and B engage in a bilateral relationship is because they expect the complex design of the item to be worthwhile i.e., the surplus δ is greater than the sum of default profits (cf. footnote 26.3 for a graphical illustration).

14.7 Notice that any proportion $\lambda \in]0;1[$ would support the results to come.

14.8 Alternatively, one party deposit a financial hostage that does not produce interest; the release occurs when the trade is agreed so that the party becomes impatient and loses most bargaining power.

14.9 Thanks to their close relation each can observe the action chosen by its partner but not a court as would be required to enforce a contract stipulating a specific level. More on this in §20.1.

14.10 The proper theory of incentives is the object of ch. 20.

14.11 A unit boss may be biased toward its workforce because sustained contact with workers fosters friendship and empathy. Hence she cares about the satisfaction of her subordinates.

14.12 The CEO of a large corporation has a wage linked to performance but he hardly owns more than 1% of its firm's stock.

14.13 Some targeted investments can be written into a contract, but human resources dedicated at learning one's partner needs or getting more flexible, will always escape control or audit at a reasonable cost.

14.14 We only consider this case which is the most frequent to avoid adding another design whose symmetry would not yield any significative additional insight.

14.15 We use the principal-agent terminology of part H.

14.16 The architect of the Sydney Opera House did not foresee the need for air conditioning. Since this omission was discovered lately and standard noisy units were out of the question, an expensive ice hockey floor was installed as a means of cooling the building. (borrowed from Preston McAfee).

14.17 We mean that he will rationally take the opportunity to advance his objective, profit. No moral connotations is involved.

14.18 We assume $F_e < V$ in order that continuation of the project be always the efficient action.

14.19 For a convex design cost d with $\frac{\partial d'}{\partial \alpha} > 0$, we still get $\underline{\tau}_\alpha > \bar{\tau}_\alpha$ while the comparison $\pi_{FP} > \pi_{C+}$ involves $d(\underline{\tau}_\alpha) - d(\bar{\tau}_\alpha) - \underline{F}(\underline{\tau}_\alpha - \bar{\tau}_\alpha) + (1-\underline{\tau}_\alpha)\sigma(V-\underline{F})$ which, using super modularity, is shown to increase with α.

14.20 This feature of modern corporations is at the root of some large bankruptcies (or public rescues when "too big to fall").

14.21 Regulation is optimal provided than the social damage is large enough so that the savings on damages overcome the loss on safe firms forced to adopt the standard. Alternatively, it is enough to assume that risky firms are numerous enough.

14.22 Contracting upon the input K is obviously even more difficult (i.e., costly) than using output. Since q and K are in a one-to-one relationship, it would be a dominated strategy. Introducing uncertainty as we do in §20.2 could make it a worthwhile consideration.

14.23 The model makes sense only if the cost function $c(.)$ is convex, includes a fixed cost and $K_0 < K^*$.

14.24 Our setting is formally equivalent to short term vs. long term cost when considering capital as a flexible or inflexible factor.

15.1 The EU Treaty has no provision for merger so that the Court of Justice judged in 1973 that there is abuse of a dominant position if an undertaking already holding such a position strengthens it by acquiring a competitor.

15.2 Note though that some very large firms remain in private hands i.e., away from regulators and stock markets.

15.3 We count the percentage change between the first two years and the last two years.

15.4 Forbes' ranking gives equal weight to rankings on sales, profits, assets and market

capitalization. We use the same method as in the above footnote and compute % up and down before taking the top tiers for each.

15.5 Their ranking is solely based on revenue.

15.6 Excluding firms with less than three years in the list; small firms have less than 20 thousand employees.

15.7 Contrary to our standard notation, subscripts refer here to the number of active firms not the label of a particular one.

15.8 This can be achieved by linking their wage bonus to their brand's profit and not to the parent company overall profit.

15.9 They even compete in the world rally championship (WRC) for image building.

15.10 A conglomerate holding k brands out of a total $n-1$ finds it profitable to introduce a new brand if $k\pi_{n-1} < (k+1)\pi_n \Leftrightarrow k < \frac{n}{2+1/n}$, roughly meaning that it should not control more than half of the brands, which is true in the European car markets as can be seen in Table 15.18 which contains 18 firms to which we can add 3 brands from VW, 2 from Fiat, 2 from Daimler and 1 from PSA.

15.11 The intuition lies in the fact that individual profit is convex increasing in output, thus convex decreasing in marginal cost. This means that every firm is a risk-lover when it comes to draw a technology from a distribution with fixed mean (cf. Février and Linnemer (2004) for a generalization).

15.12 It turns out that in our simple model, production reshuffling and knowledge diffusion are identical. While the former is a productive efficiency, the latter points at dynamic efficiency.

15.13 The first step to show that the combined post merger output $\bar{q}_1+\bar{q}_2$ is lesser than the sum $\hat{q}_1+\hat{q}_2$ of pre-merger outputs. We use two facts, firstly the best reply function $b(.)$ in the Cournot model is decreasing at less than a one-to-one rate and secondly the merged firm reduces both of its plants' outputs below their unmerged best responses since it internalizes the externality that each plant's output has on its other plant: $\bar{q}_i \leq b(\bar{q}_j + \bar{q}_{-12})$ for $i = 1,2$ where $\bar{q}_{-12}$ is the pre-merger aggregate output of the outsiders to the merger. Three cases appear: If $\bar{q}_i < \hat{q}_i$ for $i = 1,2$ then the result is trivial. If $\bar{q}_i > \hat{q}_i$ for $i = 1,2$ then the second observation implies $\bar{q}_i \leq b(\bar{q}_j + \bar{q}_{-12}) < b(\hat{q}_j + \bar{q}_{-12}) = \hat{q}_i$ for $i = 1,2$, a contradiction. Lastly, if $\bar{q}_1 \leq \hat{q}_1$ but $\bar{q}_2 > \hat{q}_2$, then by the first observation, $\bar{q}_1 + \bar{q}_2 \leq \bar{q}_1 + b(\bar{q}_1 + \bar{q}_{-12}) < \hat{q}_1 + b(\hat{q}_1 + \bar{q}_{-12}) = \hat{q}_1 + \hat{q}_2$, the desired conclusion. The second step uses again the first observation but applied this time to the rest of the industry to conclude that overall output has shrunk: $\bar{Q} = \bar{q}_1 + \bar{q}_2 + b(\bar{q}_1 + \bar{q}_2) < \hat{q}_1 + \hat{q}_2 + b(\hat{q}_1 + \hat{q}_2) = \hat{Q}$.

15.14 To ensure that all firms participate, we need only assume that the worst does.

15.15 The "real" entropy is $H \equiv e^{\Upsilon}$; in physics its measures the amount of "disorder" of a system.

15.16 Under the approximation of §15.3.4, we have $s_i = \frac{\gamma_n}{i}$ for all $i \leq n$ with $\gamma \simeq \frac{1}{0.577+\ln(n)}$, thus $H = \gamma_n^2 \sum_{i \leq n} i^{-2} \simeq \gamma_n^2 \times \frac{\pi^2}{6}$ which computed in Table 15.13 for a sample of market sizes.

15.17 Sales data, in thousands, come from the European Automobile Manufacturers Association. Notice that the name correspond to the industrial group; it may contain several brands. For instance, Rover and Land-Rover belonged to BMW in 1992; the former went independent again while the latter is now a Ford brand. Volvo is now also a Ford brand.

15.18 It is called survival function when X designates a time.

15.19 A random variable with zero mean and unit variance. In the simplest case of the random $\epsilon_t = \pm 1$ with equal probability while in the continuous time notation, the temporal evolution is modeled by geometric Brownian motion: $\frac{dX}{X} = \mu dt + \sigma dw$ where μ is the systematic component (trend) and dw the increment of a Wiener process.

16.1 We adapt Boadway (1997)'s panorama as well as the terminology of *public economics* to our industrial organization perspective.

16.2 Smith (1776) emphasizes the role of the State for some public works and institutions, without substantiating much the eligibility criteria.

16.3 Musgrave (1957) originally refers to "merit goods" but the literature also uses labels such "mixed goods", "quasi-private goods" or "specific egalitarianism". Consumptions deemed immoral (e.g., gambling, prostitution) or harmful (e.g., drugs, tobacco, junk food) are demerit goods whose treatment is opposite.

16.4 Again, the criteria for defining these express value judgments at society's level.

16.5 More precisely, Samuelson (1954) use two characteristics: absence of rivalry in consumption and the (practical) impossibility of exclusion.

16.6 Conversely, it would appear that for developing countries, every dollar spend on the rule of law is worth the while; the problem is whether the political elite think likewise.

16.7 Interestingly, the century long experience of the welfare state in the OECD shows that the key factors for its emergence are democratization, aging population, prosperity, openness to trade, and an homogenous culture.

16.8 Our emphasis on the mostly developed member countries of the OECD is driven by the availability of historical statistics (cf. also Schuknecht and Tanzi (2005)).

16.9 Social security funds is included on top of central, state and local governments revenue (cf. Government Finance Statistics). Since expenditure rise to 47.4% (resp. 46.6%) of GDP, deficit is around 2% of GDP.

16.10 Core mission includes general services (legislative organs, central ministries), public order, defense and environment but without debt interest. Although economic affairs include the regulation of industries, most of the amounts relate to support programmes, subsidies and public infrastructure spending.

16.11 Defense, in advanced economies, serves geo-strategic interests rather then protection against invasion. Debt whose service is between 2 and 4% of GDP represents a burden shifted onto future generations by current leaders to enable larger expenses on redistribution.

16.12 If my neighbor patrols the neighborhood, he protects his house and mine at the same time. I have no incentive to reward him for the lost time and he cannot threaten not to protect my house since that would leave his own unprotected; thus, unless I am compelled by a sense of duty, I shall free ride on his effort.

16.13 It is our opinion, from reading the literature, that proponents of this view seek every possible model (even feeble ones) to justify the current role of the State in a manner that avoids value judgment so as not to reveal their left-wing agenda.

16.14 As noted by Shleifer and Vishny (1994), the market socialism controversy, quite active during the first half of the XXth century, fell into the same trap. Discussing the pros and cons of an economy based on prices (market) or quantities (socialist command-and-control)

does not make a lot of sense if one must assume that both a socialist or a capitalist government would want to advance the welfare of its citizens given the overwhelming evidence to the contrary (cf. §16.2.4).

16.15 Their interactions is however assumed to take place in a free country i.e., one where violence is not an option.

16.16 This school of thought does not hide its right-wing political agenda towards shrinking the State. The sincerity is appreciated although it makes it hard to disentangle objective findings from mantras.

16.17 Their analysis is for the US but seem to apply fairly well to other western participative democracies.

16.18 This is a libertarian view of the "good" society, inspired by the history of local communities in the early days of the US.

16.19 The creation of a common agricultural market in Europe responded to the need to establish compensatory measures to smooth out the initial large productivity differences between member states agricultural sectors. Since then, the program has remained, generated perverse effects and appreciable welfare losses. However, little of that 50bn€ rent is dissipated in rent-seeking activities, thus this program is mostly redistributive although its fairness is more and more questioned.

16.20 In the same fashion, Glaeser and Shleifer (2003) recall that Enron's collapse provoked a public uproar and a call for new accounting rules. The new law was later influenced by the lobbying of industries but it seems clear that no one provoked the Enron scandal as a mean to get a new favorable regulation.

16.21 Since half of the global war effort is defensive, a small percentage of conquerors is enough to generate frequent conflicts.

16.22 This phenomenon is more acute in Europe because the demise of the Roman Empire left more political fragmentation than elsewhere. Indeed, when every territory is surrounded by enemies, rivalry is bound to be fiercer so that military edge acquires a greater value which in turn propels military innovation (cf. similarity with R&D patent races in §12). This also explains why Europe was able to colonize the rest of the world once it .

16.23 cf. Rabelais (1534)'s Coin is the sinews of war and Baumol (2004).

16.24 Epithets often adjunct to the word State are *Leviathan, rentier, autocratic, predatory, proprietary, grabbing or exploitative*.

16.25 As recalled by Glaeser and Shleifer (2003), the US at the times of the Gilded age came to be dominated by a small economic and political elite, operating the country in a manner akin to rent-planning. The progressive era that ensued with the empowering of the federal (central) government can be understood as a restoration of public order and a leveling of the competitive playing field.

16.26 For instance, Besley and Persson (2009) recall that income taxation was created in countries on the verge of war.

16.27 Notice that the revolutions in England, France and USA predating the advent of democracy were all triggered by conflict between ruler and council w.r.t. a tax increase (as well as demands for political rights).

16.28 Quite often, the ruler starts by prohibiting an activity (on moral grounds) in order to later sell selective exemptions.

16.29 As recalled by Acemoglu and Robinson (2000), this goal was achieved by forcing the

traditional ruling elite to accept more and more voting members from the bourgeoisie.

16.30 Policies favoring exports and limiting imports practiced by all European countries. It lasted longer in France or Spain as compared to the Netherlands or England because parliament in the latter countries stripped the ruler of his/her granting privileges sooner.

16.31 According to Irwin (2010), this US doctrine retaliated UK discriminatory trade practices. Bairoch (1972) shows it has been practiced by all advanced economies well into last century.

16.32 Colonialism, which involves military occupation or its threat (cf. gunboat diplomacy), is not motivated by rent-planning alone but has undoubtedly the effect of forcing foreign occupied markets to open to national industries; it basically negates the colonized the right to exercise protectionism themselves.

16.33 The UK, paragon of free trade, had many binding such restrictions.

16.34 As we explain formally in §3.3.3, the progressive inclusion of the bourgeoisie into the elite (taking place sooner in the Netherlands and England) changes the State's attitude towards the economy from pro-monopoly to pro-competitive.

16.35 Rent-planning is only sympathetic to drastic innovations that expand activity into new directions and require the establishment of new cliques. The improvement of an existing process or the novel ability to substitute a good or service will be opposed by the clique currently in charge unless it can appropriates it. Thus, depending on the bargaining power of this particular clique within the entire elite, the innovation will be more or less retarded.

16.36 As noted by Vargas Llosa in de Soto et al. (1986), "When legality is a privilege available only to those with political and economic power, those excluded—the poor—have no alternative but illegality".

16.37 Pirenne (1914) recalls how guilds or workers opposed technological innovations with destruction of machines. One can likewise interpret the slowdown of the US and UK economies in the 70s as the result of entrenchment of pressure groups at the head of the ruling elite.

16.38 Religious zeal (including the soviet brand of communism) motivated some advances in military technology but we dot count these as advancing human welfare. Fundamental research, on the other hand, being a public good is equally subsidized by all affluent states.

16.39 We believe rent-planning to be more accurate for the developing world since the ruling elite leads the relationship and uses coercion, if not outright violence, to achieve its objectives.

16.40 This is fully consistent with the commitment problem of the State who promised the private managers of the privatized firms that their markets would be protected forever. Once a new government steps in, it feels free to renege on its promises to cater to voters' interests.

16.41 A recent and extreme example would be Spanish air control.

16.42 Beyond psychology and moral sentiments of justice, risk aversion plays a role since the return to owned assets is known whereas that of "would be" assets is, be definition, uncertain.

16.43 The euractiv think-tank provides a comparison of the two systems. According to a recent interview from the think-tank cfoeurope with financial officers of large firms, the lobbying in Brussels is on the rise.

16.44 A bribe is not a welfare neutral transfer because a fraction is always lost in transaction costs. For instance, money laundering does not convert dirty dollars into clean ones on a one-to-one basis. There is here an exact analogy with taxes and the cost of public funds (cf. §17.1.2).

16.45 It is itself subject to collusion or bid-rigging.

16.46 The payment to the State is a transfer that bears no wealth effect if the winner is a firm, not an individual.

16.47 He rescues from oblivion a French (and to a lesser extent German) practice brought to British attention by Chadwick (1859) (cf. §16.4.1 on concessions).

16.48 In some countries, tenants are required to post a month or two of rent in a blocked bank account to guarantee that when they leave there is enough to rehabilitate the flat to its original state.

16.49 Lately, upon change of majority in the town hall, some contracts have been interrupted and, on average, public provision is now cheaper. Since the private utilities benefit from scale economies and public provision is deemed productively less efficient by the same report, the previous stylized fact indicate that private providers enjoy a rent.

16.50 In US parlance, privatization refers to the provision of a public service by a private (regulated) firm as opposed to in-house provision by a public institution while in European parlance, it refers to the sale of a State Owned Enterprise (SOE) to private investors.

16.51 Quite often, the ruler starts by prohibiting an activity (on moral grounds) in order to later sell selective exemptions.

16.52 This is called Tax Farming. The maximal incentives we refereed to above have the undesirable consequence, that the tax-farmer is lead to abuse his coercive power to extract more than the nominal tax rate or amount.

16.53 His account regards France but seem to apply equally to the rest of Europe.

16.54 This scheme is risky thus promotes productive efficiency yet, in the case of a service sold to the public it has a poor record for allocative efficiency since monopoly pricing will take place.

16.55 When the price is zero, one speaks of expropriation. Notice that the SOE is an old concept found in many civilizations such as Pharaonic Egypt or the Inca empire.

16.56 Notice that these economic transactions do not belong to the standard we use in this book since one party enters unwillingly the negotiation.

16.57 Bel (2006) recalls that the very first privatizations were undertaken by Nazi Germany during the 1930s.

16.58 Rising unemployment triggers the cost of welfare benefits, social security programs are still being extended, pensions become more costly because of progress in longevity.

16.59 This is what theory recommends but not necessarily what happens since there are cases where the regulator was quickly captured by the incumbent and did nothing to foster entry.

16.60 Typically, the liabilities related to pensions or health care are taken over by the State which means that they will be paid by future generations and are not to be discounted from the receipts of the sale.

16.61 Inspired by the private initiative of Englishman Freddie Laker in 1977 (cf. Examples in §10.4).

17.1 The validity of this analysis took long to

be recognized since Marshall (1890) and Pareto (1906), to cite a few, opposed it.

17.2 He mentions two railways companies operating between Paris and Versailles on the opposite banks of the Seine river who ended-up merging quickly.

17.3 He sees it essential to the community in the areas of national defense, geographical cohesion, political transparency and economic progress.

17.4 "Competition between a limited number of entrepreneurs is rationally nothing but a passing phase after which there is the definitive creation of a sole monopoly based on the ruin of the others, or a monopoly of all of them or of some of them in coalition."

17.5 At the same time, luck often bestows someone with a monopoly position or considerable augment the value of an asset. Profits so derived are then justly seen as windfall, but in the absence of a clearcut method to distinguish the two ways, we should permit the latter to embolden the former.

17.6 To understand this position, it is useful to ponder the opposite polar case when, for instance, Apple charges preposterous prices for its gadgets to devotees and no one complains.

17.7 An important element put forward by contestability theory are sunk cost since in §6.1.7 it is shown that with zero sunk cost, the monopoly is disciplined to efficient pricing at average cost by the threat of entry.

17.8 The task is easier in a city where the network has many rays so that failing one is not so costly; however the fact that inputs are complementary remains.

17.9 For instance, in 1822, the Parisian authority gave monopoly franchises over boroughs to city gas producers against an obligation to serve street with minimum demand.

17.10 Unbridled competition in the UK railways lasted until 1865 and, as such, represents a test of the natural monopoly. Foreman-Peck (1987) shows it had higher cost of service (when compared to other European countries) due to unnecessary duplications. In retrospect, the pro-active continental policies for development planning proved to be more economical.

17.11 Crocker and Masten (1996-01-01) explain how the relationship evolved from performance to relational contracts with a dedicated commission. The switch to State regulation was therefore only a change of scale, not of contractual nature.

17.12 As an alternative, Hotelling (1938) suggests inheritance or land taxes but they fall into the problem of putting incomparable services in the same bag.

17.13 Recall that we deal here with public services not public goods (cf. §16.1.1).

17.14 This is the case for R&D ventures (cf. §12.1.2).

17.15 Although this issue is more pragmatic and carry less ideological weight, it turns out to be more important because it is empirically more relevant.

17.16 Even in the best case where the nature of the service allows competition among producers (e.g., cultural clubs and services not based on an immobile facility), the client is only taking the membership binary decision; social objectives will be fulfilled only if the industry is regulated.

17.17 Boiteux (1956)'s equation (17.2) is identical to Ramsey (1927)'s optimal indirect taxation scheme. Baumol and Bradford (1970) re-

late both works to the general theory of welfare alluded in §2.3.3.

17.18 Letting $\mathbf{q} \equiv (q_i)_{i \leq n}$ denote a bundle of quantities, the solution of $\max_{\mathbf{q}} W(\mathbf{q})$ under the constraint $W_S(\mathbf{q}) \geq F$ ranges from the perfect competition (efficient) outcome $\mathbf{q}^*$ when $F = -\infty$ (the constraint never binds), to the monopoly outcome $\mathbf{q}^M$ when $F = W_S(\mathbf{q}^M)$ (this is seen by a simple contradiction argument). The case $F = 0$ is generally understood as "second best".

17.19 Beyond basic cases such as water, energy, transportation or telecommunications whose justification has been given before, one can add preventive health and education as contributing to homogenize opportunities within the entire population (cf. §16.1.1).

17.20 This setting covers the monopoly as well as the perfect competition paradigm.

17.21 Beware that the pure monopolist tends to restrict output; thus, in the case of a benin negative externality, the monopolist still produces less that the efficient level i.e., as Hotelling (1931) noted, he can be the environmentalist's best friend.

17.22 Laffont (1977) refines the analysis by showing that technological uncertainty (relevant information unknown to everyone) does not play any role for choosing a policy instrument as long as the expectations of actors are the same.

17.23 More venal employees will seek greater budget in order to get higher salaries, perks, esteem, economic power, patronage capacity, management leeway.

17.24 Unless TWTP is extremely large and marginal cost rises slowly, $2q^* < a$, the bureau absolute ideal.

17.25 Unless bureaucrats can rob from the budget (in which case the concentration of wealth is still undesirable), they tend to spend over items that are only imperfect substitutes to cash payment. For examples, they provide employment or award contracts to friends who are not the cheapest alternatives in the market. They thus raise costs which justify greater budgets in the future.

17.26 Against the discretionary use of the budget, we may mention the control exercised by users (seeking maximum output) and employees (seeking maximum wage bill).

17.27 Notice that with average cost pricing, each revenue is equal to the attributable cost plus a common markup since after simplification we obtain $C_i = \frac{C}{C-F_0}\left(c_i q_i + F_i\right)$.

17.28 The obtainment of extraordinary (economic) profits ($\Pi > 0$) is equivalent to beating the market in terms of rate of return on invested capital; it attracts new investors in the activity.

17.29 Since the cost of capital r is an average over markets with free entry where long-run competitive prices are observed, it is very likely that $\rho^* > r$ holds true given that the present market is monopolized.

17.30 Analytically, $\operatorname{sign}\left(\frac{\partial \rho}{\partial K}\right) = -R + \omega L - \omega K \frac{\partial L}{\partial K} = -R + \omega L + \omega K \frac{\Phi_K}{\Phi_L} < -R + \omega L + \omega \frac{r}{\omega} = -\Pi < 0$ because the MRTS $\frac{\Phi_K}{\Phi_L}$ is equal to the price ratio at (K^*, L^*) and then decreases as K increases.

17.31 By (17.5), the "eye-drop" is in fact a zero profit curve computed with $\bar{\rho}$ as the price of capital. It is convex because if we take the average of two input mixes then the cost is exactly the average due to linearity while the revenue is more than the average because the revenue function is concave in quantity. Hence the average mix yields a greater profit i.e., corresponds to a greater rate of return.

17.32 Some people go as far as arguing that during the old days of the AT&T monopoly over telephone in the US, the price cap was higher than the monopoly price.

17.33 Universal Service was a commercial strategy of extensive geographical coverage by AT&T to compete against the older telegraph service of Western Union.

17.34 State Aid is permitted for regional development, R&D, environmental protection, restructuring of firms, promoting culture, preserving heritage and to fight unemployment. There exists block exemptions for small and medium sized enterprises, training and employment. According to Eurostat, State aid has been halved between 1994 and 2004 from an initial level of around one percentage point of GDP (cf. also).

18.1 Our effort saves the reader from being buried under literally hundred of pages of Hamiltonians and useless retelling of optimal control theory by economists.

18.2 We use $\dot{}$ to designate temporal derivatives.

18.3 Depending on application, it is called the Golden Rule savings rate or the fundamental equation of renewable resources.

18.4 In some later models, population L grows at rate ξ and the objective is total utility $Lu(q/L)$ so that we need to subtract ξ from r but this threaten convergence of the intertemporal objective. Others (Cass, Koopmans) have directly considered maximizing per-capita utility in which case it is obvious that in the steady state equation α is replaced by $\alpha+\xi$ which amounts to add α to the discount rate r in the FOC!

18.5 Technically u is concave. At the individual level, this is monetary risk aversion but at the aggregate level, this is aversion to income inequality. cf. debate on climate change policies where protecting the world for future generations comes at the cost of maintaining current income inequalities.

18.6 Another intuitive derivation ask the following: given current capital k, should we move to a different stationary state? Assuming that capital can be bought at market rate r, the answer is found by maximize current value $v(k+\xi,\epsilon)$ under the budget constraint $r\xi-\epsilon=0$. The FOC of this standard optimization is that the RMS be equal to the price ratio i.e., $r=\phi(k)$; the solution is exactly $(k,\epsilon)=(0,0)$ when $k=k^*$.

18.7 More precisely, the easier extraction is, the larger the quantity extracted at every instant.

18.8 If $T=+\infty$ then $q_t=q^*$ for all t, so that total extraction diverges while if $T=0$, total extraction is nil. The equation under study has therefore a minimal positive solution for each k.

18.9 In short, he/she has no regards for his people' felicity, an eagerness for ostentatious immediate consumption and is under the constant threat of demise through a coup d'état.

18.10 cf. Long (1975) and Sinn (2008).

18.11 When r is small, $\underline{T}\simeq 2\frac{k}{q^*}$.

18.12 The fact that $T^m>T^*$ tends to increase output but it is a small effect compared to the twice factor.

18.13 We do not distinguish between exhaustible and renewable resources i.e., we pool fisheries, land (for hunting and grazing but not toiling) with the underground soil (e.g., oil, gas, minerals, water).

18.14 As shown by Nobel laureate Ostrom (2003), the management of commons is often

successfully organized by the involved community.

18.15 Recall that DRS is equivalent to decreasing average return or marginal return lesser than average return.

18.16 This does not mean that all resources are over exploited; yet it can be shown that $\tau_A > \tau_B \Rightarrow \hat{l}_A > l_A^*$ i.e., the better resource is over exploited.

18.17 Such a phenomenon is frequently observed in developing economies trying to harness efficiency gains through the privatization of commons.

18.18 Such a use of taxes is known as repricing or the double dividend in the field of public finance.

18.19 Kotchen and Salant (2009) nevertheless show that if the tragedy is severe, a mild tax can reduce exploitation and increase profits at the same time although the device will fall short of restoring full efficiency. See also §5.1.3 where we tackle the effect of a tax over the entire industry.

18.20 The fact that in the first round, the total distributed exceeded the aggregate output was a plain mistake.

18.21 Minerals are geologically speaking renewable resources but we call them *exhaustible* because their growth is unnoticeable at the scale of human activity.

18.22 The 200 miles Exclusive Economic Zone zone was created in the 1980s when satellite and aviation technology enabled governments to enforce their exclusivity over the said area.

18.23 Typically newborns minus deaths in the period.

18.24 It fits quite well the data for many wild species over a short period of time.

18.25 In ecology, one speaks of positive recruitment or more births than deaths per period of time.

18.26 Set $y = 1/k$ and use $\dot{y} = -\dot{k}/k^2$ to get $\dot{k} = -\frac{\dot{y}}{y^2}$. The equation is now $\frac{-\dot{y}}{y} = \left(1 - \frac{1}{y}\right) r \Leftrightarrow -r = \frac{\dot{y}}{y-1} = \frac{\partial \ln(y-1)}{\partial t}$ whose solution is $y - 1 \propto e^{-rt} \Leftrightarrow \frac{y_t - 1}{y_0 - 1} = e^{-rt}$.

18.27 For $q > \hat{q}$, the bell curve $\rho(k)k - q$ is entirely below the axis i.e., the rate of change is $s = \dot{k} < 0$ meanwhile $k_t > 0$, so that population is ever decreasing.

18.28 This aggregate formula is valid at the individual firm level under the quite acceptable proviso that all firms are equally efficient i.e., $q_i = kL_i$ for any firm i.

18.29 The open access equilibrium for a renewable biomass is called the bionomic equilibrium.

18.30 Analytically, $k_\eta = \frac{(1-\eta)p+1+\sqrt{8p\eta+((1-\eta)p+1)^2}}{4p} \leq 1 \Leftrightarrow 8p(1+\eta)(1-p) \leq 0 \Leftrightarrow p \geq 1$.

18.31 Note that the static oligopoly with n firms correspond to dynamic extraction by a monopoly with discount factor solving $k_\eta = k^n$.

18.32 This feature leads voters to see the tax as a disguised mean to increase general taxation and government revenue.

19.1 It is called *risk-free* because the securities emitted by governments of countries who never failed to pay their obligations are considered to be without risk of default (cf. Sovereign Ratings).

19.2 For a 1M€ loan at 5% over 20 years, the difference between the approximate and exact rates would be 61€ per month.

19.3 The support of H is $[\underline{x};\overline{x}]$ where $\underline{x}$ is the largest solution of $H(x) = 0$ and $\overline{x}$ is the smallest solution of $H(x) = 1$.

19.4 Bounds are removed when they do not convey useful information for the computation.

19.5 Let $\zeta \equiv \sum_{k\geq 1} \frac{1}{2^k}$, then $\frac{\zeta}{2} = \sum_{k\geq 2} \frac{1}{2^k} = \zeta - \frac{1}{2}$, thus $\zeta = 1$.

19.6 Let $\xi \equiv \sum_{k\geq 1} \frac{k}{2^k}$, then $\xi = \sum_{k\geq 1} \frac{1}{2^k} + \frac{1}{2}\sum_{k\geq 2} \frac{k-1}{2^{k-1}} = 1 + \frac{1}{2}\xi$ by the result of footnote 26.3, thus $\xi = 2$.

19.7 This is Jensen inequality. To prove it, observe that the concavity of u is equivalent to the property $\frac{u(x)-u(x_0)}{x-x_0} < u'(x_0)$ (the chord is flatter than the tangent at x_0); taking $x_0 = \mathbb{E}[\tilde{x}]$ and integrating we obtain $\mathbb{E}[u(\tilde{x})] < \mathbb{E}\left[u(x_0) + u'(x_0)(\tilde{x} - x_0)\right] = u(x_0) + u'(x_0)\mathbb{E}[\tilde{x} - x_0] = u(\mathbb{E}[\tilde{x}])$.

19.8 Observe that the function ρ completely characterizes the preferences since it enables to recover the utility function u up to an affine transformation (cf. proof).

19.9 The same problem plagues the concept of Pareto dominance in that both define only a partial order over the objects one would like to rank. This is why public economics rely on welfare functions to aggregate preferences and build a social preference out of individual ones.

19.10 The original authors do not put a great emphasis on naming their proposal. Hart (2010) calls them *economic* and *operational* which does not convey an intuitive meaning. Our naming choice reflects the intellectual debt we owe to these forerunners of risk studies.

19.11 An additional necessary condition is that no gamble can be always accepted by one of the above people. It is equivalent to require utility unboundedly negative when approaching bankruptcy.

19.12 This innocuous hypothesis is necessary for the technical analysis to go through.

19.13 We drop the ˜ from random variables in the footnotes for clarity. The function $f(\alpha) \equiv 1 - \mathbb{E}[e^{-\alpha g}]$ starts at 0 for $\alpha = 0$, then increases since $f'(0) = \mathbb{E}[g] > 0$ but tends to some negative value for α large because the gamble achieves a loss with positive probability. It must have another zero ρ_g. Another definition would be $\lim_{\epsilon\to 0} \sup\{\alpha s.t. |f(\alpha)| \leq \epsilon\}$. This way the riskiness of a sure win is zero while that of a sure loss is infinity.

19.14 Being ARRA, the richer they are the bolder they act as $\rho' \leq 0$; but IRRA limits the force of this effect since $-\rho/x \leq \rho'$. CRRAs are those with maximum such effect.

19.15 Let φ be the common riskiness, we have $1 = \mathbb{E}[e^{-g/\varphi}] \times \mathbb{E}[e^{-g'/\varphi}] = \mathbb{E}[e^{-g/\varphi} e^{-g'/\varphi}] = \mathbb{E}[e^{-(g+g')/\varphi}]$.

19.16 The function $\hat{f}(\alpha) \equiv \mathbb{E}\left[\log[1 + \alpha g]\right]$ which starts at 0 for $\alpha = 0$, then increases as $\hat{f}'(0) = \mathbb{E}[g] > 0$ and goes to $-\infty$ for α close to $1/L$ where L is the maximal loss of the gamble. It must have another zero and ϕ_g is its inverse.

19.17 More formally, ϕ_g is the wealth required but also sufficient to avoid bankruptcy when playing repetitively the gamble. Notice that bankruptcy must be understood as losing all the money one has allowed himself to invest into risky prospects i.e., keeping enough to attend living expenses.

19.18 The weight onto negative values being so small, the α parameter of footnote 26.3 must tend to $1/L$ in order to push the log towards $-\infty$.

19.19 This is to be expected from a measure named after an insurance specialist.

19.20 This is to be expected since Bernoulli was involved with gambling in the first place.

19.21 The previous reasoning is false for a financial firm since it always takes care to di-

versify its assets across markets and activities in order that a loss somewhere be always compensated by a gain elsewhere; also it tries to build large portfolio of similar clients in order to smooth out the effect of chance (or bad luck) down to the mean of the underlying risk.

19.22 We assume that the resulting profit is non negative.

19.23 If $Y = f(X)$ where $f' < 0$, then $0 \geq (X - X_0)(Y - Y_0)$ whether $X < (>) X_0$ because the second term is always of the opposite sign of the first. Integrating conserves the inequality.

20.1 There is a subtle difference in the literature regarding the action of the agent: it can be *observable* by the principal only or *verifiable* by the judge (on top of being observable); this distinction is sometimes referred to as soft vs. hard evidence.

20.2 What is meant here is that although it would be technically possible to perform such a monitoring, it would be so costly that it is better not undertaken.

20.3 Notice that minimal effort can take several meanings. When the job is painful it is presence at the work place that will determine this minimum level (absence is easily verified). If the agent loves his work, the minimum level is in fact the threshold above which he refuses to sacrifice family life without additional compensation.

20.4 Although the equation to solve is still $q'u' = c'$, it is evaluated at a different income level; hence wealth effects could alter the first best effort but it remains first-best conditional on $\underline{\pi}$, just like e^* was efficient conditional on $\underline{u}$.

20.5 The market price can be normalized to unity by scaling adequately the output unit.

20.6 The case of yardstick evaluation is similar. One replaces e_2 by the fixed objective $\bar{q}$ to obtain the same FOC of effort. Setting $\bar{q} = \beta h(0)$ makes $e = \bar{q}$ the solution of the FOC. The only difference to derive the cost formula is that $1/h(0)^2$ is half the previous value because H is the law of $\tilde{\epsilon}_1$ only.

20.7 Lazear and Rosen (1981) nevertheless show that if the worker displays decreasing absolute risk aversion, the tournament may be the preferred alternative.

20.8 W.l.o.g. that the same agent is in charge of the activity whose productivity is one while the other looks after the activity whose productivity is γ. This can be checked from the final profit Π_{team} formula.

20.9 Indeed there is no point to relate his pay to the result of line β since this would only add a risk burden given that he does not work on that line.

20.10 W.l.o.g. $\sigma = 1$, $\Pi_{\text{ind}} - \Pi_{\text{team}}$ is proportional to $2\rho^2 - (\gamma-1)^2\gamma + 2\rho(\gamma^2 + 1 - \gamma)$. The meaningful solution to the quadratic equation in ρ is is a convex increasing function $f(\gamma)$ satisfying $f(1) = 0$ and asymptotic from above to $\frac{\gamma-3}{4}$. Hence $\Pi_{\text{team}} > \Pi_{\text{ind}} \Leftrightarrow \rho < f(\gamma)$ which we approximate in the text by the asymptote.

20.11 Absent from the model is the discount factor for future earning; clearly, the more impatient (eager to live by the day) is the employee, the lesser the implicit incentives.

20.12 The rest of the time can be dedicated to shirk or pursuing a personal objective i.e., different from that of the principal.

20.13 We assume here that the principal is risk neutral i.e., cares for profit only; otherwise his objective would be the utility of ex-post profit.

20.14 Given that u and π are concave, $\frac{\pi'(q-w)}{u'(w)}$ is increasing so that equation (20.12) has at most one solution in w; as the LHS diminishes with

q this solution increases with q. We might however have $\omega^*(q) = 0$ over some interval of low q's and $\omega^*(q) = q$ over some interval of large q's.

20.15 We can safely assume that the IR constraint is binding ($\lambda > 0$) whenever c is convex, q is bounded and conceivable effort is unbounded, for under these hypothesis $U(q,e)$ diverges to $-\infty$ as effort increases.

20.16 As in footnote 26.3, a higher q diminishes the LHS of (20.11) but raises the RHS so that the solution must increase. Notice that this added hypothesis is the MLRP seen in §19.4. Indeed, $\frac{h_e}{h} = \frac{\partial \ln h}{\partial e}$ implies that for $e < e'$ we have $\ln\left(\frac{h(e,q)}{h(e',q)}\right) = \int_e^{e'} \frac{h_e}{h}(x,q)\,\mathrm{d}x$, thus the MLRP is either that the LHS is increasing in q or that $\frac{h_e}{h}$ is increasing in q.

20.17 The equation determining μ is $\int \pi(q - \omega(q)) h_e(e,q)\,\mathrm{d}q = -\mu\left\{\int u(\omega(q)) h_{ee}(e,q)\,\mathrm{d}q - c''(e)\right\}$ where the term in braces is the second order condition of utility maximization for the agent, which is necessarily non positive.

20.18 This reversal is not as extraordinary as it may seem: having provided effort the agent has become indispensable to the principal and might have acquire the bargaining power.

21.1 If two coins have the same face value but are made from metals of unequal value, the cheaper will tend to drive the other out of circulation; the more valuable coin will be hoarded or used for foreign exchange instead of for domestic transactions. Although British authors refer to this fact as Gresham Law, Nobel prize winner Robert Mundell recalls us that ancient greeks philosophers suggested the law which was clearly enunciated in the middle ages by Oresme and Copernicus.

21.2 Certification requires a State (justice and police) strong enough to deter impersonators from defrauding clients and the very certifiers from either abusing their dominant position or selling false certificates. Examples still exists with the auditors in the Enron case or the credit rating agencies in the subprime meltdown.

21.3 Letting $\bar{c} \equiv \frac{2(1-\lambda)(1-w)^2}{(2-\lambda)^2}$, it is possible to check that the solutions to equation $\pi_A = \pi_B$ are $c = \bar{c}$ where both profits are nil (which defines λ_0 implicitly) and $c = (1-\lambda)^2\bar{c}$ which defines λ_1 implicitly.

21.4 This idea is independent but reinforces the traditional motive of education which is to increase one's own ability.

21.5 The net present value of future payments of 1€ per period starting tomorrow is $1/r$.

21.6 This is obviously an extreme assumption only used for its simplicity.

21.7 An exhaustive formal reference is Laffont and Tirole (1993).

21.8 We follow the model of agency relationship seen in §20.1.

21.9 As in §4.3.2 for price discrimination, excluding inefficient agents in order to improve the performance of efficient ones is possible but if V is large this is never optimal.

21.10 This technique can be illustrated with the following simple question: who is the tallest Dutch? If we lack a Dutch database we can still search a European database and use our intuition that Dutch are the tallest people of Europe. Once we encounter the tallest European, we only need to check that he/she is a Dutch citizen.

21.11 The general model requires only that $P_h(.) - P_l(.) > 0$ is valid for the entire domain.

21.12 In the general case, the IC constraints are $\int_0^{q_h}(P_h - P_l) \geq u_h - u_l \geq \int_0^{q_l}(P_h - P_l)$.

21.13 In the general case, the equation is $(1-\alpha)(P_l - c) = \alpha(P_h - P_l) \Leftrightarrow P_l = \hat{c} \equiv (1-\alpha)c + \alpha P_h > c$.

21.14 The proper demonstration uses the fact that types vary continuously and plays with the IC condition as follows: for any ν, we must have $u\left(\theta,q(\theta)\right)\geq u\left(\theta,q(\nu)\right)\Leftrightarrow\theta U\left(q(\theta)\right)-T\left(q(\theta)\right)\geq\theta U(q(\nu))-T(q(\nu))\Rightarrow\theta U'q'-T'q'\geq 0$ by letting ν tend to θ. Inverting the roles of the two variables, we obtain $\nu U'q'-T'q'\leq 0$, hence the optimal scheme $q(\theta)$ must satisfy $\theta U'(q(\theta))=T'(q(\theta))$.

21.15 Properly speaking, differentiate (21.13) and use the SOC associated to (21.13) to prove that $q'(.)>0$.

21.16 If the solution to the FOC is not increasing, then some "ironing" is needed to force non decreasing-ness (cf. Fudenberg and Tirole (1991), ch. 7)).

21.17 If $k\geq\frac{a-bc}{2}$, the capacity is large enough to meet demand at the monopoly price $p^M(a)=\frac{a+bc}{2b}$; the announcement $\hat{a}=k+\frac{a+bc}{2}>a$ does the trick. If demand is so large that there is congestion at the monopoly price, the firm simply tells the truth and sells all her capacity at the maximum price $p_k(a)$.

21.18 Formally we also have to specify a zero subsidy if the price is set below marginal cost in order to force loss should the firm ever choose to price that way.

21.19 In the present context, λ is an ad-hoc parameter while in the general theory, its value is obtained (computed) at the equilibrium. Alternatively, one can assume that the regulator has a strict preference for consumer surplus over firm rent.

21.20 Note that (IC_h) is satisfied since it reads $0\geq\delta(\hat{q}_h-\hat{q}_l)$ and we saw that $\hat{q}_h<q_h^*<q_l^*=\hat{q}_l$.

21.21 In a continuous setting, the opportunity value to the agent of reducing the item cost by e is $d(e)$ where d is increasing convex. Let e^* be the efficient effort solving $1=\dot{d}$ and $\beta\equiv e^*-d(e^*)$ be the maximal technology saving.

21.22 He won't however do more because he supports the financial cost $d(e)$.

21.23 By giving more risk sharing to bad types an insurer improves efficiency thus he can improve its per-capita profit over bad types. Imitation by competing insurers yields the result.

22.1 The word auction derives from the latin "auctus" which means increase. British authors refer to the traditional *Roman* auction as the *English* auction; we have not been able to find an explanation for this bizarre praxis avoided by the landmark article of Vickrey (1961).

22.2 For instance, professional auctioneers who used the candle auction were granted a monopoly by Henri II of France in 1556.

22.3 Ask prices (supply) are ranked lowest to highest while bids (demand) are ranked highest to lowest. The demand and supply profiles so generated are then matched to determine an equilibrium price and an amount of trade.

22.4 As argued in §16.4.3 & §3.3.3, the financial motive is more reaslistic.

22.5 Since most participants to these lotteries have limited liability, the price asked by the government ought to be nearly zero and in any case far from the true highest value; as a consequence, the government is relinquishing part of its rent in favor of a single firm or agent and will be forced to resort to distortionary taxes to make up for the difference. This constitutes a first drawback of this method.

22.6 If the price in a Roman auction increases by very small amounts, the last bidder ends up paying the second highest bid plus a very small amount just as in the Japanese version.

22.7 Attrition is a second-price format whereas

the (standard) all-pay auction is a first-price format. Indeed, the highest bidder need only expand a little more than his last opponent. The terminology derives from warfare and was first theorized in biological competition.

22.8 We consider a Nash equilibrium of the auction game where bidders use optimal strategies.

22.9 The payment is not necessarily linked to being awarded the item; in an all-pay auction for instance, all bidders pay their bid.

22.10 Given his private value v, the bidder builds a bidding strategy σ_v that is optimal in this auction. Since he may learn many different values, he ends up with a long list of strategies. However, upon learning v, σ_v is better (at least not worse) than $\sigma_{\hat{v}}$ while upon learning $\hat{v}$, $\sigma_{\hat{v}}$ dominates σ_v. This is another instance of the revelation principle.

22.11 Check that with a starting price s, the expected revenue is $\frac{30-s}{30-10}s$ and is maximum for $s = 15$.

22.12 A monopoly sells many units while the auctioneer has typically a single item for sale.

22.13 For most statistical distributions, the hazard rate $\frac{h(.)}{1-H(.)}$ is increasing so that $\hat{R}_m(.)$ is also increasing. Furthermore, we have constructed H from a regular demand function admitting a decreasing marginal revenue.

22.14 Recall that profit is $\Pi = W_S - v_0$.

22.15 Since we assumed $\hat{R}_{m,i}$ increasing, a higher value v_i leads to a higher marginal revenue, thus a higher or equal probability of winning; this means that the necessary condition (22.2) $\varphi' \geq 0$ is satisfied. The general case is treated in appendix.

22.16 Notice that the reasoning we've used here was identical to that shown in Table 22.1.

22.17 The public policy implication is that governments should only aim at providing education to those who cannot pay for it (on top of its regalian missions). This way, the most brilliant and hard-working people will successfully bid for the control the scarce economic resources, generate the highest added value and in the mean time grab for themselves a fair reward.

23.1 Likewise, the entrepreneur "does the right thing" whereas a (good) *manager* "does things right". As wittily stated by Marshall (1907) in a related matter, "*the government could print a good edition of Shakespeare's works, but it could not get them written*".

23.2 Syverson (2010) says that "productivity is quite literally a matter of survival for businesses". Capital accumulation and the expansion of the labor force matter also but are less desicive.

23.3 Capital is of the circulating kind i.e., it completely depreciate in the process of value creation.

23.4 If we choose to express profit in terminal value then the revenue is simply R but the economic cost is then $(1+r_0)k$ to account for the opportunity cost of time.

23.5 We also assume $R_m(0) > 1$ to guarantee that k^* is positive, thereby avoiding trivialities.

23.6 It is customary in the literature to use the letter "u" as in utility to denote the objective function of the agent.

23.7 From $R_m(k_\alpha) = \frac{1}{1-\alpha}$, we deduce $\frac{\partial k_\alpha}{\partial \alpha} = \frac{1}{(1-\alpha)^2 R''} < 0$ and $\frac{\partial \pi_\alpha}{\partial \alpha} = \frac{\partial k_\alpha}{\partial \alpha}(R_m - 1) < 0$ since $R_m > 1$ over the range $k < k_0$ i.e., i.e., investment and agency value decrease with α the outsider participation.

23.8 Since we assumed $1 < R_m(0) = \lim_{\epsilon \to 0} \frac{R(\epsilon)}{\epsilon}$, it must be true that $1 > \lim_{\alpha \to 1} \frac{k_\alpha}{R(k_\alpha)}$.

23.9 Indeed, $\frac{\partial\varphi}{\partial\alpha} = \frac{\partial k_\alpha}{\partial\alpha}\frac{R-kR_m}{R^2} = \frac{\partial k_\alpha}{\partial\alpha}\frac{(1-\alpha)R-k}{(1-\alpha)R^2}$ since $R_m = \frac{1}{1-\alpha}$ at k_α; now k_α being the maximizer of $(1-\alpha)R-k$, the latter expression is positive at k_α, hence $\frac{\partial\varphi}{\partial\alpha} < 0$.

23.10 At best for her, the participation constraint of the investor is binding i.e., $F = \alpha R(k_\alpha)$ so that her final profit is at best $\pi_\alpha - F = (1-\alpha)R(k_\alpha) - k_\alpha$ whose derivative is $-R(k_\alpha) < 0$ by the envelope theorem (k_α is an optimum). Hence her payoff is maximum at $\bar{\alpha}$.

23.11 What really matters is that the firm has not enough cash or marketable securities or risk-free debt to cover the cost of the project.

23.12 Regarding existence of V_{no} and V_{go} it is enough to assume an infinite tail for the statistical distribution of x. This implies that $\mathbb{E}[\tilde{x}|\text{go}] = \mathbb{E}[\tilde{x}|\tilde{x} < (1+r)V_{\text{go}}]$ is an increasing concave function of V_{go}, thus the equation $v - rk = \mathbb{E}[\tilde{x}|\tilde{x} < (1+r)v]$ has a unique solution in v so that V_{go} is uniquely determined; it is increasing in the project size k. Now, $V_{\text{no}} = \mathbb{E}[\tilde{x}|\tilde{x} > (1+r)V_{\text{go}}]$ is also uniquely determined.

23.13 $\underline{\alpha} \le \bar{\alpha} \Leftrightarrow \sqrt{\mu(\mu+4\sigma^2)} - \mu \le 2\sigma^2\sqrt{\mu/\sigma^2} = 2\sqrt{\mu\sigma^2}$

$$\Leftrightarrow \quad 4\mu\sigma^2 \ge \left(\sqrt{\mu(\mu+4\sigma^2)} - \mu\right)^2 = \mu(\mu+4\sigma^2) + \mu^2 - 2\mu\sqrt{\mu(\mu+4\sigma^2)}$$
$$\Leftrightarrow \quad 2\mu\sqrt{\mu(\mu+4\sigma^2)} \ge \mu(\mu+4\sigma^2) + \mu^2 - 4\mu\sigma^2 = 2\mu^2$$
$$\Leftrightarrow \quad \sqrt{\mu(\mu+4\sigma^2)} \ge \mu \Leftrightarrow \mu(\mu+4\sigma^2) \ge \mu^2 \Leftrightarrow \mu + 4\sigma^2 \ge \mu \Leftrightarrow 4\sigma^2 \ge 0!$$

23.14 An example of different preferences would be Italian football teams which seem more interested in avoiding an overwhelming defeat than clinching a victory.

23.15 Football in major tournaments is rather like the war of attrition where each team waits for an error or a risk taking by the other side to counter-attack and secure an advantage (cf. §7.4.2).

23.16 The debt-financed entrepreneur earns marginally more. Indeed, the average return $\frac{R(k)}{k}$ of a DRS technology is greater than its marginal $R_m(k)$, thus $\frac{1}{\hat{p}} = \frac{R(k)}{k} > R_m$ implies that $1 > \hat{p}R_m(k)$ thus $pR_m(k) - 1 < 0$ for $p \in [0;\hat{p}[$ and therefore $\frac{\partial\hat{\pi}}{\partial k} > \int_0^{+\infty}\left(pR_m(k)-1\right)\mathrm{d}H(p) = \frac{\partial\pi^*}{\partial k}$.

23.17 Alternatively, the random component could be the cash-flow generated by current assets out of which the entrepreneur will finance part of her new investment.

23.18 This presentation is inspired by Berkovitch and Kim (1990) and Yossi Spiegel's teaching notes.

23.19 She can emit new equity paying a fixed dividend to avoid moral hazard.

23.20 Conditional on the cash flow realization, the optimal investment is a corner solution at x.

23.21 Assuming that the law $H(k,x)$ admits a density $h(k,x)$, the change from k to $\hat{k}$ satisfies the *monotone likelihood ratio property* (MLRP) if $\hat{k} > k \Rightarrow \frac{h(\hat{k},x)}{h(k,x)} \nearrow$ in x. The family of distributions $h(k,.)$ is said to satisfy MLRP if the previous property is true for all parameter values.

23.22 As shown in §19.4, the MLRP property implies FSD (first order stochastic dominance), $H_{\hat{k}} \le H_k$ and it is easy to check that the expectation of an increasing function increases with a FSD change.

23.23 Our claim is the contraposition of the following property: for any $k < k^*, \pi(\phi_\gamma, k) \ge \pi(\phi_d, k)$. To prove the property, we use MLRP: $\mathbb{E}[\gamma_d|k^*] = \mathbb{E}[\gamma|k^*] \Rightarrow \mathbb{E}[\gamma_d|k] \ge \mathbb{E}[\gamma|k]$. This says that every renegotiation offer ϕ acceptable after γ_d is also acceptable after γ (both conditional on investment k), hence the maximum of π for renegotiation following γ cannot be less than the maximum of π for renegotiation following γ_d.

23.24 The notion is second order stochastic

dominance.

23.25 Since that $R(\tilde{x})$ is concave, $-R$ is convex, it is thus enough to apply the above characterization of riskiness to see that riskier applicants repay less on expectation.

23.26 Draw $\pi(\tilde{x}) = \max\{\tilde{x} - (1+r)d, 0\}$ as a function of $\tilde{x}$ to check the effect of increasing r.

23.27 This critique is not given prominent space because the model does not generalize beyond the success-failure model used by these authors.

23.28 Other typical but also imperfect measures to align the objective of managers with that of shareholders are salary incentives (participation to profits) and stock-options.

23.29 This is the famous irrelevance theorem of Modigliani and Miller (1958).

23.30 The remnant could also be more brutally diverted into perquisites i.e., an immediate consumption of no value for the firm.

23.31 Instead of an additive uncertainty, we could equally use a multiplicative market price uncertainty like in previous models at the cost of heavier formulas.

23.32 We assume that H is one of the many distributions whose hazard rate is increasing. Since the LHS of the FOC is increasing with k while the RHS is decreasing, there is indeed a unique solution $\hat{k}$.

24.1 Notice that there shouldn't be too many of them because in a world with a million standards, there is no "real" standard!

24.2 Have you ever tried to remove the motor of your car and replace it by a different one ?

24.3 One speaks of "one-stop-shopping" when consumers enjoy a discount for purchasing multiple services from a single supplier (cf. §5.3.3 & §24.1).

24.4 Proof: Lowering p_{AB} below $p_A^* + p_B^*$ is akin to raising p_B above p_B^* but since the latter is at its optimal level, the loss in sales (of good B) is exactly compensated by the increased revenues (on good B). However, there is a spillover effect in that some people who would cease to buy good B switch to the bundle, thus raising profits.

24.5 For example, in the context of bundling computer programs it does not seem farfetched to assume that selling components separately would require substantial extra programming costs in order to guarantee compatibility of the components with older softwares, costs that could be avoided if the new programs are bundled.

24.6 If the consumers leisure budget is exogenous, the long term equilibrium determines the number of active firms (or software) as the ratio of revenues to development cost of one software.

24.7 Game theory has a branch studying repeated interaction; it has shown that cooperation can emerge in a non cooperative environment.

24.8 Historically a circus wagon carrying a musical band; today a current or fashionable trend.

24.9 Their model reframe previous works of Rohlfs (1974) and Farrell and Saloner (1985).

24.10 Solving for $\frac{dP}{dq} = 0$, $q^0 = \frac{1}{2}(1-s)$; the maximal sustainable price is $c^0 = \frac{1}{4}(s+1)^2$.

24.11 We have $W'(b) = s - c - \frac{1}{2}b(3b + 2s - 4)$ and $q^* = \frac{2-s+\sqrt{s(s+2)+4-6c}}{3}$.

24.12 $n^M(c) = \frac{1-s+\sqrt{s(s+2)+1-3c}}{3}$

24.13 An externality is pecuniary if its transmission vector is the price mechanism, otherwise it is technological.

24.14 If innovations in the modern sector can

be protected by IPRs such as patents or strategic entry barriers, then monopoly ensue (cf. §10.2). If imitation is easy then competition involves equal cost thus leads to zero producer surplus which will bar a challenger from entering given the presence of the fixed cost (cf. §6.1).

24.15 If there were n varieties, instead of a continuum, we would have to replace L by L/n from this point on.

24.16 He can only grab the surplus of the weakest sector in each sector.

24.17 In this simple setting, there is a multiplicity of equilibria and the one that actually takes place depends on considerations outside the model.

24.18 The same goes for boats since the conductor steers with his right arm (cf. venetian paddlers), the tiller is on the right side and to protect it when crossing incoming traffic one has to keep on the right side.

24.19 It is also argued, with less force, that the revolutionary Robespierre countered the papal advice on political grounds with a mandatory opposite custom.

24.20 In this market Microsoft's products achieved the best ratings and quickly reached a 80% share; somehow "Mac" users taught MS how to make friendlier products for the PC world.

24.21 CNET reported cost estimates for the Xbox from 320 to 400$, the initial public price being 400$ later slashed down to 180$; it is also noticeable that a assembly plant in Hungary was shut down to favor cheaper production from China.

24.22 Analytically, letting z denote consumption of positional goods, the net utility of a conceited person can be stated as $u(z) - \delta(\hat{z} - z)$ where u is the private utility function and $\hat{z}$ is the average spending of the consumer's friends or relatives. Then the optimal individual consumption solves $u' = -\delta$ while the socially optimal one solves $u' = 0$. If positional goods display diminishing returns (i.e., $u'' < 0$), the result obtains.

24.23 To avoid corner solutions, it is necessary to assume αn small enough (cf. Grilo et al. (2001)).

24.24 Although long time clients tend to get served without delay, new ones (at the marginal) are often being rationed.

24.25 In the past, passengers rights were inexistent and no airline had an incentive to compete with others by increasing this side of customer service; thus the equilibrium level was zero i.e., upon late arrival, you could stay on the ground without any compensation whenever overbooking generated an excess demand at boarding time. In Europe the 2005 directive forces airlines to pay damages to passengers suffering refusal to board. It is now cheaper for them to auction off the few missing seats.

24.26 Disney has maintained a reputation of releasing video or DVD of their old movies in the consumer market (not the rental one) for 6 months only with a promise to stop selling it for 5 full years, a time long enough for kids to become uninterested by the feature.

24.27 Equivalently, if some immediate second-hand market was put in place to permit exchange among all potential clients.

24.28 A more realistic model would describe the hierarchical pyramid model where the probability of escalating is π.

25.1 This is close to the *non-rivalry* property of public goods but the property of *non-excludability* is not true here since connection

to the network is controlled by the operator and can therefore be conditioned to the payment of a subscription. Exception are radio and TV if they are transmitted by air waves without encryption.The possibility to exclude entrance at school or hospital may seem outrageous to most but is feasible at low cost; hence neither of these services qualify as a public good!

25.2 An alternative is a choice between two things, so that if one is taken, the other must be left.

25.3 Some deliberately create queues to increase the word of mouth effect and attract even more buyers.

25.4 Under the now ubiquitous ADSL or cable connection, traffic is free and we only suffer the opportunity cost of time.

25.5 It would take two-part pricing to restore efficiency; by setting p^* consumers are led to demand the welfare maximizing quantity which can then be recuperated through a subscription.

25.6 The name comes from the analogy with a bicycle wheel, which consists of a number of spokes jutting outward from a central hub.

25.7 Consumption of electricity is measured in Watts (power) per hour, while the instantaneous demand is measured in Watts. A Giga Watt (GW) is a billion Watts, a Tera Watt (TW) is a thousand GW.

25.8 The effect of 1C temperature variation on French demand are nearly 2% in winter and 1% in summer. Haziness effect from clarity to obscurity can boost demand by up to 6%.

25.9 In recent years, as much as 21% of the French park was unavailable due to maintenance. Indeed, the intensive use of nuclear power plants requires tighter security measures, thus longer and more frequent periods of maintenance.

25.10 In France however, nuclear power accounts for 54% of the installed power and 77% of generation; recent studies puts the share of capital cost around two thirds of total costs for a nuclear power station.

25.11 Notice that on days of extremely high internal demand, exports are reduced by a factor 2 or 3 using market mechanisms.

25.12 It comes as no surprise that this analysis was performed by an engineer from an electric utility, the French monopoly EDF (cf. Drèze (1964) for details).

25.13 Similar ideas were considered in the UK but France had a higher share of industrial clients with whom price discrimination was negotiable; as we already commented, price discrimination with residential clients was almost unanimously opposed by political forces. This is why peak load pricing took more time to make its way in the UK.

25.14 In line with our findings on two-part tariffs, the yearly subscription is set at a higher level (190€ vs. 120€); there is also a higher installation cost for the second (night) meter.

25.15 This is so because the peak price is difficult to change on short notice.

25.16 In the last 5 years, EDF never declared more than 21 red days because it is prudent to keep one in reserve for an unexpected event.

25.17 The elasticity of surface with respect to radius is 2 since for a radius R, surface is $S = 2\pi R^2$, thus $dS = 2\pi R$ and $\epsilon = \frac{dS/S}{dR/R} = \frac{dS}{dR}\frac{R}{S} = 2$. This means that a one percent increase in city radius generates a 2% increase in city surface.

25.18 These areas cluster 182 millions inhabitants or 62% of the total US population in 2003

which was 291 millions (income and population data from the US Bureau of Economic Accounts).

25.19 The US Federal Highway Administration (FHWA) estimates the remaining causes to be accidents (25%), works (15%), bad weather (10%) and poor signaling (5%).

25.20 It is enough to narrow a 2 lanes street into one lane over 20 meters to generate a strict one lane street in terms of traffic fluidity.

25.21 In french, but the instructive table 12 comparing Paris and London in the conclusion is easy to understand.

25.22 Competitive pressure means a potentially unlimited pool of road users ready to pay a price $\bar{p}$ to use the road. This turns the WTP function flat at the level $\bar{p}$ thus equating the welfare and monopoly optimal choices.

25.23 The fact that most public infrastructure are financed from general taxation turns revenues from congestion pricing into manna that many see fit to be invested into capacity expansion. This is obviously unwarranted for these revenue should only pay for current maintenance and the interest charge on the debt contracted out at construction time.

25.24 The fact that capacity is measured by lanes which is an integer creates a tuning problem that is alleviated by pooling road geographically (cross-subsidization) or temporally (because traffic tend to grow).

25.25 In our simple setting where demand is normalized to unity, these are the same.

25.26 Notice however that in such tight conditions, a sudden use of breaks can create a chain reaction that can completely stop traffic.

25.27 This is in fact the excess over the WTP for his next best alternative. Since the latter cannot be estimated, it is normalized to zero.

26.1 The assumption $\underline{v} > 0$ known as the "gap case" is necessary to avoid running into technical difficulties.

26.2 Choosing $s \geq 1$ implies that $z > 1 - \delta$ must hold and indeed it holds at the proposed solution.

26.3 Beware that this calculation is done while neglecting the cost of installing the service capacity.

26.4 The general treatment ($n > 3$) involves separating the coefficients of c_{i-1} and c_{i+1} from the other ones. Parameter γ is the entry (i,i) of A^{-1}, while δ is twice the entry $(i,i-1)$. Studying the equation $A \times A^{-1} = Id$, one finds $2\gamma - \delta/2 = 1$.

26.5 Initially, we have $J'(0^+) \leq 0$ and $J'(0^-) \geq 0$ but since J is continuously differentiable, the two expressions converges one to the other.

26.6 Take $y = gz$ where g is smooth positive and admissible, we get $0 = \int gz^2$, thus $z = f_x - \dot{f}_{\dot{x}}$ is uniformly nil. This is similar to stating that $< y, z >= 0$ for any y i.e., z is in the orthogonal of L^2 which is the zero singleton since L^2 is its own dual.

Bibliography

Abito J. M. and *Wright J.*, Exclusive dealing with imperfect downstream competition, *International Journal of Industrial Organization*, 26(1):227–246, 2008. 261

Acemoglu D., Why not a political coase theorem social conflict commitment and politics, *Journal of Comparative Economics*, 31(4):620–652, 2003. 225

Acemoglu D., Oligarchic versus democratic societies, *Journal of European Economic Association*, 6(1):1–44, 2008. 457

Acemoglu D. Political economy lecture notes. Technical Report chap 1, MIT, 2010. 220, 454

Acemoglu D. and *Robinson J. A.*, Why did the west extend the franchise, *Quarterly Journal of Economics*, 115(4):1167–1199, 2000. 754

Acemoglu D., *Kremer M.*, and *Mian A.*, Incentives in markets firms and governments, *Journal of Economic Behavior & Organization*, 24(2):273–306, 10 2008. 359

Adams W. and *Yellen J.*, Commodity bundling and the burden of monopoly, *Quarterly Journal of Economics*, 90(3):475–498, 1976. 648

Aghion P. and *Bolton P.*, Contracts as barriers to entry, *American Economic Review*, 77(3):388–401, 1987. 262, 266

Akerlof G., The market for lemons quality uncertainty and the market mechanism, *Quarterly Journal of Economics*, LXXXIV(3):488–500, 1970. 529, 572, 573, 623, 635

Alchian A. and *Demsetz H.*, Production information costs and economic organization, *American Economic Review*, 62(5):777–795, 1972. 376

Allaz B. and *Vila J. L.*, Cournot competition forward markets and efficiency, *Journal of Economic Theory*, 59(1):1–16, 1993. 174

Armstrong M. Recent developments in the economics of price discrimination, chapter 4. Volume 2 of Blundell et al. (2007), 2007. 121, 151, 305

Armstrong M. and *Porter R.*, editors, Handbook of industrial organization, volume 3. Elsevier, 2007. 772, 780, 793, 797

Arrow K. The rate and direction of inventive activity, chapter Economic Welfare and the Allocation of Resources for Invention. Princeton University Press, 1962. 323, 324, 331

Arrow K., Aspects of the theory of risk bearing. Helsinki: Yrj Hahnsson Foundation, 1965. 540, 542, 546

Arrow K. and *Seppo H.*, editors, Frontiers of economics. Basil Blackwell, Oxford, 1985. 772

Arrow K., *Chenery H. B.*, *Minhas B. S.*, and *Solow R. M.*, Capital labor subsitution and economic efficiency, *Review of Economic Studies*, 43(3):225–251, 1961. 38

Arundel A., The relative effectiveness of patents and secrecy for appropriation, *Research Policy*, 30(4):611–624, 2001. 340

Auerbach A. J. and *Feldstein M.*, editors, Handbook of public economics, volume Volume 4. Elsevier, 2002. ISBN 1573-4420. 795

Auerbach F., Das gesetz der bevolkerungskonzentration, *Petermann's Geographische Mitteilungen*, 59:74–76, 1913. 441

Aumann R. What is game theory trying to accomplish, pages 28–76. In Arrow and Seppo (1985), 1985. 50

Aumann R. and *Serrano R.*, An economic index of riskiness, *Journal of Political Economy*, 116 (5):810–836, 10 2008. 542

Averch H. and *Johnson L.*, Behavior of the firm under regulatory constraint, *American Economic Review*, 52(5):1053–1069, 1962. 490, 491

Bagwell K. The economic analysis of advertising, chapter 28. Volume 3 of Armstrong and Porter (2007), 2007. 309

Bailey E. E. and *Malone J. C.*, Resource allocation and the regulated firm, *Bell Journal of Economics*, 1(1):129–142, 1970. 493

Bain J., Barriers to new competition. Harvard University Press, Cambridge, MA, 1956. 257, 258, 736

Bairoch P., Free trade and european economic development in the 19th century, *European Economic Review*, 3(3):211–245, 1972. 755

Bajari P. and *Tadelis S.*, Incentives versus transaction costs a theory of procurement contracts, *RAND Journal of Economics*, 32(3):387–407, 2001. 402

Baron D. and *Myerson R.*, Regulating a monopolist with unknown costs, *Econometrica*, 50(4): 911–930, 1982. 588

Baumol W., Entrepreneurship productive unproductive and destructive, *Journal of Political Economy*, 98(5):893–921, 1990. 457

Baumol W., Contestable markets an uprising in the theory of industry structure, *American Economic Review*, 72(1):1–15, 1982. 166, 736

Baumol W., Red queen games arms races rule of law and market economies, *Journal of Evolutionary Economics*, 14(2):237–247, 2004. 754

Baumol W. and *Bradford D.*, Optimal departures from marginal cost pricing, *American Economic Review*, 60(3):265–283, 1970. 757

Baumol W. and *Willig R.*, Contestability developments since the book, *Economic Journal*, 38: 9–36, 1986. 164, 736

Baumol W., *Panzar J.*, and *Willig R.*, Contestable markets and the theory of industry structure. Harcourt College, 1982. 163, 736

Bays J. and *Jansen P.* Prizes a winning strategy for innovation. Technical report, Mckinsey, 2009. 187

Becker G. S., A theory of social interactions, *Journal of Political Economy*, 82(6):1063–1093, 1974. 670

Becker G. S., A note on restaurant pricing and other examples of social influence on prices, *Journal of Political Economy*, 99(5):1109–1116, 1991. 677

Becker G. S., A theory of competition among pressure groups for political influence, *Quarterly Journal of Economics*, 98(3):371–400, 1983. 201

Bel G., The coining of privatization and germanys national socialist party, *Journal of Economic Perspectives*, 20(3):187–194, 2006. 756

Bergman M., *Coate M. B.*, *Jakobsson M.*, and *Ulrick S. W.* Atlantic divide of gulf stream convergence merger policies in the european union and the united states. Technical report, U.S. Federal Trade Commission (FTC), 2009. 419

Bergstrom T. and *Varian H.*, Two remarks on cournot equilibria, *Economics Letters*, 19(1):5–8, 1985. 131

Berkovitch E. and *Kim E. H.*, Financial contracting and leverage induced over and underinvestment incentives, *Journal of Finance*, 65(3):765–94, 1990. 766

Berliner J. S., Factory and manager in the ussr. Harvard University Press, 1957. 499

Bernoulli D., Specimen theoriae novae de mensura sortis, *Commentarii Academiae Scientiarum Imperiales Petropolitanae*, 5:175–192, 1738. 537, 538

Bertrand J., Theorie des richesses revue de theories mathematiques de la richesse sociale par leon walras et recherches sur les principes mathematiques de la theorie des richesses par augustin cournot, *Journal des Savants*, 67:499–508, 1883. 65, 124, 133

Besley T. and *Persson T.*, The origins of state capacity property rights taxation and politics, *American Economic Review*, 99(4):1218–44, 2009. 754

Bezançon X. Lhistoire des concessions. Technical report, EGF-BTP, 2005. 465

Bhaskar V. and *To T.*, Is perfect price discrimination really efficient an analysis of free entry, *RAND Journal of Economics*, 35(4):762–776, 2004. 152

Binmore K., Fun and games a text on game theory. D C Heath & Co., 1992. 726

Binmore K. Interpersonal comparison of utility. Technical report, ELSE, 2007. 46

Blackstone E. A., Restrictive practices in the marketing of electrofax copying machines and supplies, *Journal of Industrial Economics*, 23(3):189–202, 1975. 108

Blundell R., *Newey W. K.*, and *Persson T.*, editors, Advances in economics and econometrics theory and applications ninth world congress, volume 2 of *Econometric Society Monographs (No. 42)*. Cambridge University Press, 2007. 771

Boadway R., Public economics and the theory of public policy, *Canadian Journal of Economics*, 30(4):753–772, 1997. 753

Boccard N. On efficiency concentration and welfare. Technical Report 40, Core Discussion Paper, 2009. 427

Boccard N. Duality of welfare and profit maximization. Technical report, SSRN eLibrary, 2010a. 483, 729

Boccard N. Rent seeking vs production conflict axiomatization and equivalence. Technical report, SSRN eLibrary, 2010b. 738

Boccard N. On royalty licensing. Technical report, SSRN eLibrary, 2010c. 327

Boccard N., Renegotiation in agency contracts menus vs simple contracts, *Spanish Economic Review*, 4(4):261–280, 2002. 549

Boccard N. and *Wauthy X.*, Bertrand competition and cournot outcomes further results, *Economics Letters*, 68(3):279–285, 2000. 134

Boiteux M., Sur la gestion des monopoles astreints a lequilibre budgetaire, *Econometrica*, 24(1): 22–40, 1956. 482, 757

Boiteux M., La tarification des demandes en pointe application de la theorie de la vente au cout marginal, In *Revue Générale de l'électricité, trans. in* Izzard (1960), pages 321–40. 692

Boiteux M., Le tarif vert de lelectricite de france, In *Revue Française de l'énergie, cf.* Chick (2002), pages 1–16. 694

Bolton P. and *Dewatripont M.*, Contract theory. MIT Press, 2005. 362

Bonanno G. and *Vickers J.*, Vertical separation, *Journal of Industrial Economics*, 36(3):251–259, 1988. 179

Bork R. H., The goals of antitrust policy, *American Economic Review (Papers and Proceedings)*, 57(2):242–253, 1967. 235, 261

Brandenburger A. and *Nalebuff B.*, Co opetition. Doubleday, 1996. 726

Brander J. and *Lewis T.*, Oligopoly and financial structure the limited liability effect, *American Economic Review*, 76(5):956–970, 1986. 180

Brander J. and *Spencer B.*, Strategic commitment with protectrd the symmetric case, *Bell Journal of Economics*, 14(1):225–235, 1983. 334, 336

Brennan G. and *Buchanan J. M.*, Towards a tax constitution for leviathan, *Journal of Public Economics*, 8(3):255–273, 1977. 731

Broz J. L., The origins of central banking solutions to the free rider problem, *International Organization*, 52(02):231–268, 1998. 455

Brueckner J. K. The structure of urban equilibria a unified treatment of the muth mills model, chapter 20, pages 821–845. Volume 2 of Mills (1987), 1987. 290

Brueckner J. K., Network structure and airline scheduling, *Journal of Industrial Economics*, 52 (2):291–312, 2004. 688, 689

Brueckner J. K., *Thisse J. F.*, and *Zenou Y.*, Why is central paris rich and downtown detroit poor an amenity based theory, *European Economic Review*, 43(1):91–107, 1999. 292

Buchanan J. M., An economic theory of clubs, *Economica*, 32(125):1–14, 1965. 480

Buchanan J. M. and *Tullock G.*, The calculus of consent logical foundations of constitutional democracy. University of Michigan Press, 1962. 452

Bulow J. and *Roberts J.*, The simple economics of optimal auctions, *Journal of Political Economy*, 97(5):1060–1090, 1989. 608, 610, 616

Bulow J., *Geanakoplos J.*, and *Klemperer P.*, Multimarket oligopoly strategic substitutes and complements, *Journal of Political Economy*, 93(3):488–511, 1985. 171

Burstein M. L., The economics of tie in sales, *Review of Economic Studies*, 42(1):68–73, 1960. 646

Bush W. C. and *Mayer L. S.*, Some implications of anarchy for the distribution of property, *Journal of Economic Theory*, 8(4):401–412, 1974. 715, 738

Camerer C. F. and *Loewenstein G.* Behavioral economics past present future. Technical report, 2002. 740

Chadwick E., of competition for the field as compared with competition within the field, *Journal of Statistical Society of London*, 22(3):381–420, 1859. 464, 756

Chakravarty S. and *MacLeod W. B.*, Contracting in the shadow of the law, *RAND Journal of Economics*, 40(3):533–557, 2009. 748

Chamberlin E., Duopoly value where sellers are few, *Quarterly Journal of Economics*, 44(1):63–100, 1929. 66

Chamberlin E., The theory of monopolistic competition. Harvard University Press, 1933. 66, 296

Chandler A., Organizational capabilities and the economic history of the industrial enterprise, *Journal of Economic Perspectives*, 6(3):79–100, 1992. 359

Chatterjee K. and *Samuelson W.*, Bargaining under incomplete information, *Operations Research*, 31(5):835 –851, 1982. 615, 722

Chick M., Le tarif vert retrouve the marginal cost concept and the pricing of electricity in britain and france 1945 1970, *The Energy Journal*, 23(1), 2002. 774

Chung T.-Y., Incomplete contracts specific investments and risk sharing, *Review of Economic Studies*, 58(5):1031–1042, 1991. 390

Clark C. W., Profit maximization and the extinction of animal species, *Journal of Political Economy*, 81(4):950–961, 1973. 525

Clark C. W. and *Munro G. R.*, The economics of fishing and modern capital theory a simplified approach, *Journal of Environmental Economics and Management*, 2(2):92–106, 12 1975. 523

Clarke E. H., Multipart pricing of public goods, *Public Choice*, 11(1):19–33, 1971. 615

Clarke R. and *Davies S.*, Market structure and price cost margins, *Economica*, 49(195):277–287, 1982. 425, 426

Coase R., The nature of the firm, *Economica*, 4(16):386–405, 1937. 224, 370, 375, 392, 400

Coase R., Durability and monopoly, *Journal of Law and Economics*, 15(1):143–149, 1972. 120, 270

Coase R., The marginal cost controversy, *Economica*, 13(51):169–182, 1946. 94

Coase R., The problem of social cost, *Journal of Law and Economics*, 3:1–44, 1960. 63, 222, 223

Cobb C. and *Douglas P.*, A theory of production, *American Economic Review*, 18(1):138–165, 1928. 28, 31, 38

Colson C., Cours deconomie politique. P. Gauthier-Villars et Alcan, 1901-1907. 730

Congleton R. D. and *Lee S.*, Efficient mercantilism revenue maximizing monopoly policies as ramsey taxation, *European Journal of Political Economy*, 2008. 85

Conway P., *Janod V.*, and *Nicoletti G.* Product market regulation in oecd countries 1998 to 2003. Technical report, OECD Economics Department Working Paper, No 419, 2005. 280

Cooter R., *Marks S.*, and *Mnookin R.*, Bargaining in the shadow of the law a testable model of strategic behavior, *Journal of Legal Studies*, 11(2):225–251, 1982. 738

Corts K. S., Teams versus individual accountability solving multitask problems through job design, *RAND Journal of Economics*, 38(2):467–479, 2007. 353, 558

Cournot A., Recherches sur les principes mathematiques de la theorie des richesses. Paris: Ha-

chette, 1838. 14, 44, 65, 88, 124, 125, 167, 242, 251, 381, 424, 430, 474

Courty P., Some economics of ticket resale, *Journal of Economic Perspectives*, 17(2):85–97, 2003. 674

Cowling K. and *Waterson M.*, Price cost margins and market structure, *Economica*, 43(171):267–74, 1976. 425

Crandall R. W. and *Winston C.*, Does antitrust policy improve consumer welfare assessing the evidence, *Journal of Economic Perspectives*, 17(4):3–26, 2003. 255

Cremer H., *Gasmi F.*, *Grimaud A.*, and *Laffont J.-J.*, Universal service an economic perspective, *Annals of Public and Cooperative Economics*, 72(1):5–43, 2001. 501

Crew M. A. and *Kleindorfer P. R.* Pricing and regulatory innovations under increasing competition, chapter Price Caps and Revenue Caps: Incentives and Disincentives for Efficiency. Kluwer, 1996. 494

Crocker K. J. and *Masten S. E.*, Regulation and administered contracts revisited lessons from transaction cost economics for public utility regulation, *Journal of Regulatory Economics*, 9(1): 5–39, 1996-01-01. 757

Currie J. and *Gahvari F.*, Transfers in cash and in kind theory meets the data, *Journal of Economic Literature*, 46(2):333–383, 2008. 449

d' Aspremont C. and *Jacquemin A.*, Cooperative and noncooperative protectrd in duopoly with spillovers, *American Economic Review*, 78(5):1133–1137, 1988. 320

d' Aspremont C., *Gabszewicz J. J.*, and *Thisse J. F.*, On hotellings stability in competition, *Econometrica*, 47(5):1145–1150, 1979. 743

Dal Bo E. and *Di Tella R.*, Capture by threat, *Journal of Political Economy*, 111(5):1123–1152, 2003. 203

Dansby R. and *Willig R.*, Industry performance gradient indexes, *American Economic Review*, 69 (3):249–260, 1979. 434

Dasgupta P. Facts and values in modern economics. Technical report, Cambridge University, 2007. 49

Dasgupta P. Modern economics and its critics, chapter 3. In Mäki (2002), 2002. 21

Davidson C. and *Deneckere R.*, Incentives to form coalitions with bertrand competition, *RAND Journal of Economics*, 16(4):473–486, 1985. 422

de Borda J.-C. Memoire sur les elections au scrutin, 1781. 66

de Condorcet N. Essai sur la constitution et les fonctions des assemblees provinciales, 1788. 64

de Finetti B., Sulla preferibilita, *Giornale degli Economisti e Annali di Economia*, 11:685–709,

1952. 540, 542

de Meza D. and *Webb D.*, Too much investment a problem of asymmetric information, *Quarterly Journal of Economics*, 102(2):281–292, 1987. 637

de Soto H., *Ghersi E.*, and *Ghibellini M.*, El otro sendero la revolucion informal. El Barranco, Lima, 1986. 755

Debreu G., The coefficient of resource utilization, *Econometrica*, 19(3):273–292, 1951. 726, 728

Demsetz H., Toward a theory of property rights, *American Economic Review (Papers and Proceedings)*, 57(2):347–359, 1967. 218

Demsetz H., Information and efficiency another viewpoint, *Journal of Law and Economics*, 12(1): 1–21, 1969. 452

Demsetz H., Industry structure market rivalry and public policy, *Journal of Law and Economics*, 16:1–10, 1973. 424

Demsetz H., Why regulate utilities, *Journal of Law and Economics*, 11:55–65, 1968. 463

Demski J. and *Sappington D.*, Resolving double moral hazard problems with buyout agreements, *RAND Journal of Economics*, 22(2):232–240, 1991. 391, 393

DiMasi J., *Hansen R.*, and *Grabowski H.*, The price of innovation new estimates of drug development costs, *Journal of Health Economics*, 22(24):151–185, 2003. 319

Director A. and *Levi E.*, Law and the future trade regulation, *Northwestern University Law Review*, 51:281–96, 1956. 261

Dixit A., The role of investment in entry deterrence, *Economic Journal*, 90(357):95–106, 1980. 267

Dixit A., A model of duopoly suggesting a theory of entry barriers, *Bell Journal of Economics*, 10 (1):20–32, 1979. 259

Dixit A., Strategic behavior in contests, *American Economic Review*, 77(5):891–898, 1987. 192

Dixit A. and *Nalebuff B.*, Thinking strategically. W. W. Norton & Co., 1991. 726

Dixit A. and *Norman V.*, Advertising and welfare, *Bell Journal of Economics*, 9(1):1–17, 1978. 314

Dixit A. and *Olson M.*, Does voluntary participation undermine the coase theorem, *Journal of Public Economics*, 76(3):309–335, 2000. 224

Dixit A. and *Skeath S.*, Games of strategy. W. W. Norton & Company, 1996. 726

Dixit A. and *Stiglitz J.*, Monopolistic competition and optimum product diversity, *American Economic Review*, 67(3):297–308, 1977. 296

DoJ. Horizontal merger guidelines. Technical report, U.S. Department of Justice and the Federal

Trade Commission, 1997. 237

Dorfman R. and *Steiner P.*, Optimal advertising and optimal quality, *American Economic Review*, 44(5):826–36, 1954. 313

Downs A., An economic theory of democracy. Harper and Row, New York, 1957. 452

Downs A., The law of peak hour expressway congestion, *Traffic Quarterly*, 16:393–409, 1962. 702

Drèze J., Some postwar contributions of french economists to theory and public policy with special emphasis on problems of resource allocation, *American Economic Review*, 54(4):1–64, 1964. 482, 769

Duncum P., Aesthetics popular visual culture and designer capitalism, *International Journal of Art & Design Education*, 26(3):285–295, 2007. 747

Dupuit J., De la mesure de lutilite des travaux publics, *Annales des Ponts et Chaussées*, 8(2): 332–75, 1844. 42, 45, 88, 97, 116, 477, 479, 728, 730

Dupuit J., De linfluence des peages sur lutilite des voies de communication, *Annales des Ponts et Chaussées*, 11:7–31, 1849. 115

Dupuit J. Dictionnaire de leconomie politique, volume 2, chapter Péage, Voies de communication, pages 339–344, 846–54. Guillaumin, Paris, 1852. 474, 742

EC. Report on raw materials. Technical report, European Commission, 2008. 505

EC. Report on competition policy. Technical report, European Commission, 2005. 253

EC. Report on merger and acquisitions. Technical report, European Commission, 2001. 418

Economides N., Nash equilibrium in duopoly with products defined by two characteristics, *RAND Journal of Economics*, 17(3):431–439, 1986. 743

Economides N., Quality choice and vertical integration, *International Journal of Industrial Organization*, 17(6):903–914, 1999. 655

Economides N. and *Himmelberg C.* Critical mass and network size with application to the us fax market. Technical report, NYU working paper, 1995. 658

Edgeworth F., Mathematical psychics an essay on the application of mathematics to the moral sciences. C. Keegan Paul and Co., 1881. 33, 58, 65, 184, 185

Edgeworth F., Contribution to the theory of railway rates iv, *Economic Journal*, 23(90):206–26, 1913. 33, 475

Edgeworth F. Papers relating to political economy vol 1, chapter The Theory of Pure Monopoly. MacMillan, New York, 1925. 133

Edgeworth F., Applications of probabilities to economics, *Economic Journal*, 20(78,80):284–304, 441–465, 1910. 481, 730

Edlin A. and *Hermalin B.*, Contract renegotiation and options in agency problems, *Journal of Economic Behavior & Organization*, 16(2):395–423, 2000. 391, 393

Einav L. and *Levin J. D.*, Empirical industrial organization a progress report, *Journal of Economic Perspectives*, 24(2):145–62, 2010. 16

Esping-Andersen G. and *Myles J.* The welfare state and redistribution. Technical report, UPF, 2008. 448

Esteban J. and *Ray D.*, Collective action and the group size paradox, *American Political Science Review*, 2001. 205

Esty B. and *Ghemawat P.* Airbus vs boeing in superjumbos a case of failed preemption. Technical Report Working Paper No. 02-061, Harvard Business School, Strategy Unit, 2002. 294

Fang H. and *Norman P.*, To bundle or not to bundle, *RAND Journal of Economics*, 37(4):946–963, 2006. 649

Farrell J. and *Klemperer*. Coordination and lock in competition with switching costs and network effects, pages 1967–2072. Volume 3 of Armstrong and Porter (2007), 2007. 129

Farrell J. and *Saloner G.*, Standardization compatibility and innovation, *RAND Journal of Economics*, 16(1):70–83, 1985. 652, 654, 767

Farrell J. and *Shapiro C.*, Horizontal mergers an equilibrium analysis, *American Economic Review*, 80(1):107–126, 1990. 426

Farrell M., The measurement of productive efficiency, *Journal of Royal Statistical Society*, 120 (III):253–281, 1957. 29

Feldstein M. S., Equity and efficiency in public sector pricing the optimal two part tariff, *Quarterly Journal of Economics*, 86(2):176–187, 1972. 484

Fershtman C. and *Judd K.*, Equilibrium incentives in oligopoly, *American Economic Review*, 77 (5):927–941, 1987. 177

Février P. and *Linnemer L.*, Idiosyncratic shocks in an asymmetric cournot oligopoly, *International Journal of Industrial Organization*, 22(6):835–848, 2004. 752

Fisher F. M., The social costs of monopoly and regulation posner reconsidered, *Journal of Political Economy*, 93(2):410–416, 1985. 193, 462

Fisher F. M. Innovation and monopoly leveraging. Technical report, MIT, 1999. 274

Foreman-Peck J. S., Natural monopoly and railway policy in the nineteenth century, *Oxford Economic Papers*, 39(4):699–718, 1987. 757

Foster D. and *Hart S.*, An operational measure of riskiness, *Journal of Political Economy*, 117(5): 785–814, 2009. 542

Fraiman N., *Singh M.*, *Arrington L.*, and *Paris C.* Zara case study. Technical report, Columbia Business School, 2003. 293

Friedman M. and *Savage L.*, The utility analysis of choices involving risk, *Journal of Political Economy*, 56(4):279–304, 1948. 539

Frye T. and *Shleifer A.*, The invisible hand and the grabbing hand, *American Economic Review*, 87(2):354–358, 1997. 460

Fudenberg D. and *Tirole J.*, The fat cat effect the puppy dog ploy and the lean and hungry look, *American Economic Review (Papers and Proceedings)*, 74(2):361–368, 1984. 171, 172

Fudenberg D. and *Tirole J.*, Game theory. MIT Press, 1991. 726, 764

Fujita M. and *Thisse J. F.*, Economics of agglomeration. Cambridge University Press, 2002. 292

Gabor A., A note on block tariffs, *Review of Economic Studies*, 23(1):32–41, 1955. 94

Gabszewicz J. J., La concurrence imparfaite. La Découverte Paris, 1994. 726

Gaffeo E., *Gallegati M.*, and *Palestrini A.*, On the size distribution of firms additional evidence from the g7 countries, *Physica A: Statistical Mechanics and its Applications*, 324(1-2):117–123, 6 2003. 441

Garfinkel M. R. and *Skaperdas S.* Economics of conflict an overview, chapter 22. Volume 2 of Sandler and Hartley (2007), 2007. 738

Gasmi F., *Laffont J.-J.*, and *Vuong Q.*, Econometric analysis of collusive behavior in a soft drink market, *Journal of Economics and Management Strategy*, 2(1):277–311, 1992. 312

Gelman J. and *Salop S.*, Judo economics capacity limitation and coupon competition, *Bell Journal of Economics*, 14(2):315–325, 1983. 276

Ghemawat P., Capacity expansion in the titanium dioxide industry, *Journal of Industrial Economics*, 33(2):145–163, 1984. 267

Gibbons R., Game theory for applied economists. Princeton University Press, 1992. 726

Gibbons R., Four formalizable theories of the firm, *Journal of Economic Behavior & Organization*, 58(2):200–245, 2005. 367, 396, 404

Gibrat R., Les inegalites economiques. Libraire du Recueil Sirey, Paris, 1931. 434, 441, 443

Giebe T. and *Wolfstetter E.*, License auctions with royalty contracts for winners and losers, *Games and Economic Behavior*, 63(1):91–106, 5 2008. 328

Gilbert R. J., Exclusive dealing preferential dealing and dynamic efficiency, *Review of Industrial Organization*, 16(2):167–184, 2000. 260

Gilbert R. J. and *Klemperer P.*, An equilibrium theory of rationing, *RAND Journal of Economics*, 31(1):1–21, 2000. 675

Gilbert R. J. and *Newbery D.*, Preemptive patenting and the persistence of monopoly, *American Economic Review*, 72(3):514–526, 1982. 331, 745

Gintis H., The evolution of private property, *Journal of Economic Behavior & Organization*, 1 (64):1–16, 2007. 220

Glaeser E. L. and *Shleifer A.*, The rise of the regulatory state, *Journal of Economic Literature*, 41 (2):401–425, 2003. 235, 406, 407, 754

Gneezy U. and *Rustichini A.*, A fine is a price, *Journal of Legal Studies*, 29(1):1–17, 2000. 749

Goldberg V. P., Regulation and administered contracts, *Bell Journal of Economics*, 7(2):426– 448, 1976. 375, 464

Goolsbee A. and *Chevalier J. A.* Measuring prices and price competition online amazon and barnes and noble. Technical report, NBER, 2002. 135

Gordon H. S., The economic theory of a common property resource the fishery, *Journal of Political Economy*, 62(2):124–142, 1954. 517, 522

Gossen H., Entwickelung der gesetze des menschlichen verkehrs und der daraus fliessenden regeln fur menschliches handeln. Braunschweig: Vieweg, 1854. 27

Grafton R. Q., *Adamowicz W.*, *Dupont D.*, *Nelson H.*, *Hill R. J.*, and *Renzetti S.*, The economics of the environment and natural resources. Basil Blackwell, 2004. 218

Gray L. C., Rent under the assumption of exhaustibility, *Quarterly Journal of Economics*, 28(3): 466–489, 1914. 506

Grilo I., *Shy O.*, and *Thisse J. F.*, Price competition when consumer behavior is characterized by conformity or vanity, *Journal of Public Economics*, 80(3):385–408, 2001. 671, 768

Grossman G. M. and *Shapiro C.*, Informative advertising with differentiated products, *Review of Economic Studies*, 51(1):63–81, 1984. 315

Grossman S., Nash equilibrium and the industrial organization of markets with large fixed costs, *Econometrica*, 49(5):1149–1172, 1981. 141, 142, 143

Grossman S. and *Hart O.*, The costs and benefits of ownership a theory of vertical and lateral integration, *Journal of Political Economy*, 94(4):691–719, 1986. 376, 377, 396, 397

Grossman S. and *Hart O.* The economics of uncertainty, chapter Corporate Financial Structure and Managerial Incentives. University of Chicago Press, 1982. 639

Grout P., Investment and wages in the absence of binding contracts a nash bargaining approach, *Econometrica*, 52(2):449–60, 1984. 378

Groves T., Incentives in teams, *Econometrica*, 41(4):617–631, 1973. 615

Haavelmo T., A study in the theory of economic evolution. North-Holland, Amsterdam, 1954. 199

Harberger A. C., Monopoly and resource allocation, *American Economic Review (Papers and Proceedings)*, 44(2):77–87, 1954. 75, 730

Harford J., What drives merger waves, *Journal of Financial Economics*, 77(3):529–560, 2005. 416

Harsanyi J., Games with randomly disturbed payoffs a new rationale for mixed strategy equilibria, *International Journal of Game Theory*, 2(1):1–23, 1973. 56

Hart O., Firms contracts and financial structure. Oxford University Press, 1995. 370, 395

Hart O. and *Moore J.*, Property rights and the nature of the firm, *Journal of Political Economy*, 98(6):1119–1157, 1990. 376

Hart O. and *Moore J.*, Contracts as reference points, *Quarterly Journal of Economics*, 123(1): 1–48, 2008. 366

Hart S. Comparing risks by acceptance and rejection. Technical report, Hebrew University of Jerusalem, 2010. 542, 761

Hau T. D. Congestion charging mechanisms for roads. Technical Report 1071, World Bank, 1992. 701

Herfindahl O. Concentration in the steel industry. PhD thesis, Columbia University, 1950. 433

Hicks J., The theory of monopoly, *Econometrica*, 3(1):p1–20, 1935. 75

Hicks J., Consumers surplus and index numbers, *Review of Economic Studies*, 9(2):126–137, 1942. 728

Hicks J., The four consumers surpluses, *Review of Economic Studies*, 11(1):31 – 41, 1943. 93

Hillman A. and *Riley J.*, Politically contestable rents and transfers, *Economics and Politics*, 1(1): 17–39, 1989. 191

Hines J. R., Three sides of harberger triangles, *Journal of Economic Perspectives*, 13(2):167–188, 1999. 730

Hirschmann A. O., The paternity of an index, *American Economic Review*, 54(5):761, 1964. 433

Hirshleifer J., The analytics of continuing conflict, *Synthese*, (76):201–33, 1988. 199, 738

Hirshleifer J. and *Osborne E.*, Truth effort and the legal battle, *Public Choice*, 108(1):169–195, 07 2001. 193

Hobbes T., Leviathan. Andrew Crooke, London, 1651. 57, 219, 731

Holmes T. J., The effects of third degree price discrimination in oligopoly, *American Economic Review*, 79(1):244–250, 1989. 153

Holmstrom B., Moral hazard in teams, *Bell Journal of Economics*, 13:324–340, 1982a. 621

Holmstrom B., Managerial incentive problems a dynamic perspective republished in 1999, *Review of Economic Studies*, 66(1):169–182, 1982b. 560

Holmstrom B., Moral hazard and observability, *Bell Journal of Economics*, 10(1):74–91, 1979. 564, 565

Holmstrom B. and *Tirole J.* The theory of the firm, volume 1, chapter 2. Elsevier, 1989. 351, 415

Holt C. A. and *Scheffman D. T.*, Facilitating practices the effects of advance notice and best price policies, *RAND Journal of Economics*, 18(2):187–197, 1987. 251

Hotelling H., The general welfare in relation to problems of taxation and of railways and utility rates, *Econometrica*, (6):242–269, 1938. 757

Hotelling H., The economics of exhaustible resources, *Journal of Political Economy*, 39(2):137–175, 1931. 509, 758

Hotelling H., Stability in competition, *Economic Journal*, 39(153):41–57, 1929. 136, 137, 283, 284, 285, 289, 714

Hotelling H., A general mathematical theory of depreciation, *Journal of the American Statistical Association*, 20(151):340–353, 1925. 533

Hume D., A treatise of human nature. John Noon and Thomas Longman, 1740. 57, 64

Ikeda T. and *Toshimitsu T.*, Third degree price discrimination quality choice and welfare, *Economics Letters*, 106(1):54–56, 1 2010. 102

Innes R., Limited liability and incentive contracting with ex ante action choices, *Journal of Economic Theory*, 52(1):45–67, 1990. 632

Irmen A. and *Thisse J. F.*, Competition in multi characteristics spaces hotelling was almost right, *Journal of Economic Theory*, 78(1):76–102, 1998. 295

Irwin D. Revenue or reciprocity founding feuds over early us trade policy. In Irwin and Sylla (2010), 2010. 755

Irwin D. and *Sylla R.*, editors, Founding choices american economic policy in the 1790s. University of Chicago Press, 2010. 784

Izzard H., Peak load pricing, *The Journal of Business*, 33(2):157–179, 1960. 774

Jaffe A. and *Lerner J.*, Innovation and its discontents. Princeton University Press, 2004. 340

Jensen M., Agency costs of free cash flow in corporate finance and takeovers, *American Economic Review*, 76(2):323–329, 1986. 629

Jensen M. and *Meckling W.*, Theory of the firm managerial behavior agency costs and ownership structure, *Journal of Financial Economics*, 4(3):305–360, 1976. 621, 627, 631

Jerath K. and *Zhang Z. J.*, Store within a store, *Journal of Marketing Research*, 47(4):748–763,

2010. 261

JFTC. State of corporate groups in japan. Technical report, Japan Fair Trade Commission, 2001. 414

Jing R. and *Winter R.* Exclusionary contracts. Technical report, 2010. 265

Joskow P., Vertical integration and long term contracts the case of coal burning electric generating plants, *Journal of Law Economics and Organization*, 1(1):33–79, 1985. 375

Kadiyali V., Entry its deterrence and its accommodation a study of the us photographic film industry, *RAND Journal of Economics*, 27(3):452–478, 1996. 312

Kaldor N., The economic aspects of advertising, *Review of Economic Studies*, (18):44926, 1950. 311

Kalman D., Leveling with lagrange an alternate view of constrained optimization, *Mathematics Magazine*, 82(3), 2009. 727

Kambil A. and *van Heck E.* Competition in the dutch flower markets. Technical report, New York University and Tilburg University, 1996. 603

Kamien M. I. and *Tauman Y.*, Patent licensing the inside story, *Manchester School*, 70(1):7–15, 2002. 326, 329

Kerr S., On the folly of rewarding a while hoping for b, *Academy of Management Journal*, 18(4): 769–783, 1975. 356

Kessler J. and *Leider S.* Norms and contracting. Technical report, Harvard University, 2010. 365

Khaldun I., The muqaddimah. Tunis, 1377. 84

Kihlstrom R. E. and *Riordan M. H.*, Advertising as a signal, *Journal of Political Economy*, 92(3): 427–450, 1984. 316, 579

Klein B., *Crawford R.*, and *Alchian A.*, Vertical integration appropriable rents and the competitive contracting process, *Journal of Law and Economics*, 21(2):297–326, 1978. 375

Klemperer P., Auctions theory and practice. Princeton University Press, 2003. 599

Klemperer P. and *Mayer M.*, Supply function equilibria in oligopoly under uncertainty, *Econometrica*, 57(6):1243–1277, 1989. 143, 144

Kleven H. J. and *Kreiner C. T.* The marginal cost of public funds in oecd countries hours of work versus labor force participation. Technical report, CESifo Working Paper No. 935, 2003. 477

Knight F., Risk uncertainty and profit. Houghton Mifflin, Boston, 1921. 66

Knight F., Some fallacies in the interpretation of social cost, *Quarterly Journal of Economics*, 38 (4):582–606, 1924. 701, 702

Knight F. H., The economic organization. University of Chicago, Chicago, 1933. 20

Knittel C. R., The adoption of state electricity regulation the role of interest groups, *Journal of Industrial Economics*, 54(2):201–222, 2006. 476

Kocas C. Online price competition within and between heterogeneous retailer groups. Technical report, Department of Marketing and Supply Chain Management, Michigan State University, 2005. 135

Koerner J. The dark side of coffee price war in the german market for roasted coffee. Technical report, Department of Food Economics and Consumption Studies, University of Kiel, 2002. 133

Konrad K. A. Strategy in contests an introduction. Technical Report Markets and Politics Working Paper No. SP II 2007-01, WZB, 2007. 738

Kotchen M. and *Salant S.* A free lunch in the commons. Technical report, NBER, 2009. 195, 760

Kreps D. and *Scheinkman J.*, Quantity precommitment and bertrand competition yields cournot outcomes, *Bell Journal of Economics*, (14):326–337, 1983. 134

Krueger A. O., The political economy of the rent seeking society, *Journal of Political Economy*, 64 (3):291–303, 1974. 462

Krugman P., Increasing returns monopolistic competition and international trade, *Journal of International Economics*, 9(4):469–479, 1979. 298, 744

Krugman P., Intraindustry specialization and the gains from trade, *Journal of Political Economy*, 89(5):959–973, 1981. 299

Laffont J.-J., More on prices vs quantities, *Review of Economic Studies*, 44, 1977. 758

Laffont J.-J. and *Martimort D.*, The theory of incentives the principal agent model. Princeton University Press, 2002. 528

Laffont J.-J. and *Tirole J.*, A theory of incentives in procurement and regulation. MIT Press, 1993. 594, 763

Laffont J.-J. and *Tirole J.*, Using cost observation to regulate firms, *Journal of Political Economy*, 94(3):614–641, 1986. 588, 592

Lafontaine F. and *Slade M.*, Vertical integration and firm boundaries the evidence, *Journal of Economic Literature*, 45(3):629–685, 2007. 350

Lagrange J. L., Theorie des fonctions analytiques contenant les principes du calcul. Imprimerie de la Republique, Paris, 1797. 41

Lancaster K., A new approach to consumer theory, *Journal of Political Economy*, (74):132–157, 1966. 292

Law M. T. and *Kim S.*, Specialization and regulation the rise of professionals and the emergence of occupational licensing regulation, *Journal of Economic History*, 65:723–756, 2005. 246

Lazear E. P. and *Rosen S.*, Rank order tournaments as optimum labor contracts, *Journal of Political Economy*, 89(5):841–864, 1981. 557, 762

Lee G. K., Understanding the timing of fast second entry and the relevance of capabilities in invention vs commercialization, *Research Policy*, 38(1):86–95, 2 2009. 167

Leibenstein H., Allocative efficiency versus x efficiency, *American Economic Review*, (56(3):392–415, 1966. 75, 76

Leibenstein H., Bandwagon snob and veblen effects in the theory of consumers demand, *Quarterly Journal of Economics*, 69(4):619–625, 1955. 670

Leijonhufvud A., Towards a not too rational macroeconomics, *Southern Economic Journal*, 60(1): 1–13, 1993. 23

Leland H. and *Pyle D.*, Informational asymmetries financial structure and financial intermediation, *Journal of Finance*, (32):371–387, 1977. 625

Leontief W., The pure theory of the guaranteed annual wage contract, *Journal of Political Economy*, 54:76–79, 1946. 96, 750

Lerner A., The concept of monopoly and the measurement of monopoly power, *Review of Economic Studies*, (1):157–175, 1934. 73, 434

Lerner A. P., The economics and politics of consumer sovereignty, *American Economic Review*, 62 (1/2):258–266, 1972. 185

Levenstein M. C. and *Suslow V. Y.*, What determines cartel success, *Journal of Economic Literature*, 44(1):43–95, 2006. 243

Levhari D. and *Mirman L. J.*, The great fish war an example using a dynamic cournot nash solution, *Bell Journal of Economics*, 11(1):322–334, 1980. 525

Levin J. and *Tadelis S.*, Profit sharing and the role of professional partnerships, *Quarterly Journal of Economics*, 120(1):131–172, 2005. 572, 575

Levin J. and *Tadelis S.*, Contracting for government services theory and evidence from us cities, *Journal of Industrial Economics*, 58(3):507–541, 2010. 408

Liebowitz S. and *Margolis S.*, The fable of the keys, In *Journal of Law and Economics* Spulber (2002), pages 1–25. 669

Liebowitz S. and *Margolis S.*, Winners losers and microsoft competition and antitrust in high technology. The Independent Institute, Oakland California, 1999. 667, 668

Lindert P. H. Welfare states markets and efficiency. Technical report, University of California - Davis, 2007. 449

Littlechild S. C., Two part tariffs and consumption externalities, *Bell Journal of Economics*, 6(2): 661–670, 1975. 111

Littlechild S. C., Regulation of british telecommunications profitability. Secretary of State, Department of Industry, London, UK, 1983. 495

Long N. V., Resource extraction under the uncertainty about possible nationalization, *Journal of Economic Theory*, 10(1):42–53, 1975. 759

Loury G. C., Market structure and innovation, *Quarterly Journal of Economics*, 93(3):395–410, 1979. 210

Machiavelli N., Il principe. Antonio Blado d'Asola, 1532. 224

Macho-Stadler I. and *Pérez-Castrillo D.*, An introduction to the economics of information. Oxford University Press, 1996. 549

Mahenc P. and *Salanié F.*, Softening competition through forward trading, *Journal of Economic Theory*, 116:282–293, 2004. 176

Mäki U., editor, Fact and fiction in economics models realism and social construction. Cambridge University Press, 2002. 777

Malthus T. R., The nature of rent. John Murray, London, 1815. 474

Mankiw G. and *Whinston M.*, Free entry and social inefficiency, *RAND Journal of Economics*, 17 (1):48–58, 1986. 160

Markides C. and *Geroski P. A.*, Racing to be second, *Business Strategy Review*, 15(4):25–31, 2004. 53, 167

Markoff J., Peasants protest the claims of lord church and state in the cahiers de doleances of 1789, *Comparative Studies in Society and History*, 32(3):413–454, 1990. 731

Marshall A., Principles of economics. London: Macmillan and Co., 1890. 14, 46, 47, 66, 311, 375, 430, 475, 750, 757

Marshall A., The social possibilities of economic chivalry, *Economic Journal*, 17(65):7–29, 1907. 765

Martimort D. and *Stole L.*, Contractual externalities and common agency equilibria, *Advances in Theoretical Economics*, 3(1), 2003. 203

Mas-Collel A., *Winston M. D.*, and *Green J. R.*, Microeconomic theory. Oxford University Press, 1995. 451

Maskin E. and *Riley J.*, Monopoly with incomplete information, *RAND Journal of Economics*, 15: 171–196, 1984. 113

Maskin E., *Qian Y.*, and *Xu C.*, Incentives information and organizational form, *Review of Economic Studies*, 67(2):359–378, 2000. 360

Masten S. E., The organization of production evidence from the aerospace industry, *Journal of*

Law and Economics, (27):403–417, 1984. 374, 404

Matraves C., Market structure protectrd and advertising in the pharmaceutical industry, *Journal of Industrial Economics*, (2):169–194, 1999. 437

Matraves C., European integration and market structure in the soft drinks industry, *International Journal of Economics of Business*, 3(9):295–310, 2002. 437

Matthews S. A., Renegotiating moral hazard contracts under limited liability and monotonicity, *Journal of Economic Theory*, 97(1):1–29, 2001. 634

McAfee P. and *McMillan J.*, Bidding rings, *American Economic Review*, 82(3):579–599, 1992. 247, 602

Mcafee R. P. and *Wiseman T.*, Capacity choice counters the coase conjecture, *Review of Economic Studies*, 75(1):317–332, 2008. 120

McAfee R. P., *McMillan J.*, and *Whinston M. D.*, Multiproduct monopoly commodity bundling and correlation of values, *Quarterly Journal of Economics*, 104(2):371–383, 1989. 648

McGuire M. C. and *Olson, Mancur J.*, The economics of autocracy and majority rule the invisible hand and the use of force, *Journal of Economic Literature*, 34(1):72–96, 1996. 83

Mead J., Price and output policy of state enterprise, *Economic Journal*, (LIV):215–216, 321–340, 1944. 496

Megginson W., Introduction to the special issue on privatization, *International Review of Financial Analysis*, 16(4):301–303, 2007. 470

Megginson W. and *Netter J.*, From state to market a survey of empirical studies on privatization, *Journal of Economic Literature*, 3(2):321–389, 2001. 469

Milgrom P., Putting auction theory to work. Cambridge University Press, 2004. 599

Milgrom P., Good news and bad news representation theorems and applications, *RAND Journal of Economics*, 12(2):380–391, 1981. 547, 634

Milgrom P. and *Roberts J.*, Limit pricing and entry under incomplete information an equilibrium analysis, *Econometrica*, (50):443–459, 1982. 277

Milgrom P. and *Weber R. J.*, A theory of auctions and competitive bidding, *Econometrica*, (50): 1080–1122, 1982. 605

Mill J. S., The principles of political economy with some of their applications to social philosophy. Longmans, Green and Co, London, 1848. 65, 474

Mills E., editor, Handbook of regional and urban economics, volume 2. Elsevier, 1987. 775

Milne C. Universal service for users recent research results an international perspective. Technical report, Antelope consulting, 1997. 502

Mirrlees J. Essays in equilibrium behavior under uncertainty, chapter Notes on Welfare Economics, Information and Uncertainty. North Holland, 1974. 564

Modigliani F., New developments on the oligopoly front, *Journal of Political Economy*, 66(3): 215–232, 1958. 258

Modigliani F. and *Miller M.*, The cost of capital corporation finance and the theory of investment, *American Economic Review*, 48:261–297, 1958. 737, 767

Mohring H., Optimization and scale economies in urban bus transportation, *American Economic Review*, 62(4):591–604, 1972. 704

Mohring H. and *Harwitz M.*, Highway benefits an analytical framework. Northwestern University Press, Evanston, Illinois, 1962. 701

Monteverde K. and *Teece D.*, Supplier switching costs and vertical integration in the automobile industry, *Bell Journal of Economics*, (13):206–213, 1982. 374, 404

Morrison S. and *Winston C.* Deregulation of network industries whats next, chapter The Remaining Role of Government Policy in the Deregulated Airline Industry, pages 1–40. AEI-Brookings Joint Center for Regulatory Studies, 2000. 471

Mosca M., On the origins of the concept of natural monopoly economies of scale and competition, *European Journal of the History of Economic Thought*, 15(2):317–353, 2008. 474

Mueller M., Universal service competition interconnection and monopoly in the making of american telephone system. MIT Press, Cambridge, MA, 1997. 503

Müller H. and *Warneryd K.*, Inside vs outside ownership a political theory of the firm, *RAND Journal of Economics*, 32(3):527–541, 2001. 206

Murphy K. M., *Shleifer A.*, and *Vishny R. W.*, Industrialization and the big push, *Journal of Political Economy*, 97(5):1003–1026, 1989. 661

Murphy K. M., *Shleifer A.*, and *Vishny R. W.*, Why is rent seeking so costly to growth, *American Economic Review*, 83(2):409–414, 1993. 664

Musgrave R. A., The voluntary exchange theory of public economy, *Quarterly Journal of Economics*, 53(2):213–237, 1939. 66

Musgrave R. A., A multiple theory of budget determination, *Finanzarchiv*, 25(1):33–43, 1957. 753

Mussa M. and *Rosen S.*, Monopoly and product quality, *Journal of Economic Theory*, (18):301–317, 1978. 81, 301

Myers S., Determinants of corporate borrowing, *Journal of Financial Economics*, (5):147–176, 1977. 630, 631

Myers S. and *Majluf N.*, Corporate financing and investment decisions when firms have information that investors do not have, *Journal of Financial Economics*, (13):187–221, 1984. 623

Myerson R., Game theory analysis of conflict. Harvard University Press, 1991. 726

Myerson R., Incentive compatibility and the bargaining problem, *Econometrica*, 47(1):61–74, 1979. 582, 606

Myerson R., Nash equilibrium and the history of economic theory, *Journal of Economic Literature*, (37):1067–82, 1999. 24

Myerson R., Optimal auction design, *Mathematics of Operations Research*, 6:58–73, 1981. 606

Myerson R. and *Satterthwaite M.*, Efficient mechanisms for bilateral trade, *Journal of Economic Theory*, (28):265–281, 1983. 616

Nash J., Equilibrium points in n person games, *Proceedings of National Academy of Sciences*, 36: 48–49, 1950. 733

Nelson R., Increased rents from increased costs a paradox of value theory, *Journal of Political Economy*, 65(5):387–393, 1957. 131

Netessine S. and *Shumsky R.*, Introduction to the theory and practice of yield management, *INFORMS Transactions on Education*, 3(1), 2002. 96

Nickerson J. A. and *Vanden Bergh R.*, Economizing in a context of strategizing governance mode choice in cournot competition, *Journal of Economic Behavior & Organization*, 40(1):1–15, 9 1999. 373

Niskanen W. A., The peculiar economics of bureaucracy, *American Economic Review*, 58(2):293–305, 1968. 359, 487

North D. C., Institutions institutional change and economic performance. Cambridge University Press, 1990. 368

North D. C., *Weingast B. R.*, and *Wallis J. J.* A conceptual framework for interpreting recorded human history. Technical report, World Bank, 2006. 219, 531, 749

North D. C., *Wallis J. J.*, *Webb S. B.*, and *Weingast B. R.* Limited access orders in the developing world a new approach to the problems of development. Technical report, World Bank, 2007. 749

Numa G., Dupuit and walras on the natural monopoly in transport industries, *History of Political Economy*, 43(4), 2011. 474

Nussbaum H. International cartels and multinational enterprises. In Teichova et al. (1986), 1986. 243

Odlyzko A. Economics of information security, chapter Privacy, Economics, and Price Discrimination on the Internet. Kluwer, 2004a. 87, 96

Odlyzko A., The evolution of price discrimination in transportation and its implications for the internet, *Review of Network Economics*, 3(3):323–346, 2004b. 97

OECD, editor. Highlights of public sector pay and employment trends 2002 update, 2005. OECD. 450

Oi W., A disneyland dilemma two part tariffs for a mickey mouse monopoly, *Quarterly Journal of Economics*, 85(1):77–96, 1971. 108, 110

Olson M., The logic of collective action public goods and the theory of groups. Harvard University Press, 1965. 65, 205, 452, 454

OPEC. Annual statistical bulletin. Technical report, OPEC, 2004. 244

Osborne M. and *Rubinstein A.*, A course in game theory. MIT Press, 1994. 726

Ostrom E., Coping with tragedies of the commons, *Annual Review of Political Science*, 2(1):493–535, 11 2003. 759

Pareto V., La courbe de la repartition de la richesse. Viret-Genton, 1896. 441

Pareto V., Manuale di economia politica con una introduzione alla scienza sociale. A. Milani, Padova o Libraria, 1906. 185, 192, 458, 757

Pareto V., The new theories of economics, *Journal of Political Economy*, 5(4):485–502, 1897. 20

Pareto V., Cours deconomie politique. F. Rouge, Lausanne, 1896-97. 65

Pauly M., The economics of moral hazard, *American Economic Review*, 58(3), 1968. 529

Pigou A. C., The economics of welfare. Macmillan and Co., London, 1920. 88, 92, 93, 96, 451, 683, 701, 702

Pirenne H., The stages in the social history of capitalism, *American Historical Review*, 19(3): 494–515, 1914. 755

Poincaré H., Science and hypothesis. Walter Scott Publishing, London, 1905. 21

Posner R. A., Social costs of monopoly and regulation, *Journal of Political Economy*, (83):807–828, 1975. 462

Posner R. A., Natural monopoly and its regulation, *Stanford Law Review*, 21(3):548–643, 1969. 464, 470, 475

Posner R. A., Antitrust law. University of Chicago Press, 2001. 235, 742

Pratt J., Risk aversion in the small and in the large, *Econometrica*, 32(1):122–36, 1964. 540, 717

Priest G. L., The origins of utility regulation and the theories of regulation debate, *Journal of Law and Economics*, 1(36):289–323, 1993. 476

Prud'homme R. and *Bocarejo J. P.*, The london congestion charge a tentative economic appraisal, *Transport Policy*, 12:279–287, 2005. 701

Prud'homme R. and *Sun Y. M.*, Le cout economique de la congestion du peripherique parisien

une approche desagregee, *Cahiers Scientifiques du Transport*, 37:59–73, 2000. 704

Prud'homme R., *Kopp P.*, and *Bocarejo J. P.*, Evaluation economique de la politique parisienne des transports, *Transports*, 436, 2006. 700

PWC. Mining deals. Technical report, Price Waterhouse Coopers, 2008. 505

Qian Y., *Roland G.*, and *Xu C.*, Coordination and experimentation in m form and u form organizations, *Journal of Political Economy*, 114(2):366–402, 2006. 360

Rabelais F., Gargantua. 1534. 754

Raith M., Competition risk and managerial incentives, *American Economic Review*, 93(4):1425–36, 2003. 76, 714, 715

Ramsey F. P., A mathematical theory of saving, *Economic Journal*, 38(152):543–559, 1928. 506

Ramsey F. P., A contribution to the theory of taxation, *Economic Journal*, 37(145):47–61, 1927. 757

Rasmusen E., Games and information an introduction to game theory. Blackwell, fourth edition edition, 2006. 549, 565, 726

Rasmusen E., *Ramseyer M.*, and *Wiley J. S.*, Naked exclusion, *American Economic Review*, 81(5): 1137–1145, 1991. 261, 266

Reed W. J., The pareto zipf and other power laws, *Economics Letters*, 74:15–19, 2001. 443

Reiffen D. and *Kleit A. N.*, Terminal railroad revisited foreclosure of an essential facility or simple horizontal monopoly, *Journal of Law and Economics*, 33(2):419–438, 1990. 274

Rey P. and *Tirole J.* A primer on foreclosure, chapter 33. Volume 3 of Armstrong and Porter (2007), 2007. 274

Ricardo D., Principles of political economy and taxation. John Murray, London, 1817. 298

Riordan M., On delegating price authority to a regulated firm, *RAND Journal of Economics*, 15 (1):108–115, 1984. 587

Robinson J., Economics of imperfect competition. London: MacMillan., 1933. 98, 283, 311

Robinson M. S., Collusion and the choice of auction, *RAND Journal of Economics*, 16(1):141–145, 1985. 602

Robson A. and *Skaperdas S.*, Costly enforcement of property rights and the coase theorem, *Economic Theory*, 36(1):109–128, 2008. 207

Rochet J.-C. and *Tirole J.*, Platform competition in two sided markets, *Journal of European Economic Association*, 1(4):990–1029, 2003. 669

Rodrik D. Industrial policies for the twenty first century. Technical report, KSG, Harvard University, 2004. 279

Rogerson W., Simple menus of contracts in cost based procurement and regulation, *American Economic Review*, 93(3):919–926, 2003. 592

Rohlfs J., A theory of interdependent demand for a communication service, *Bell Journal of Economics*, (5):16–37, 1974. 767

Roover R. D., Monopoly theory prior to adam smith a revision, *Quarterly Journal of Economics*, 65(4):492–524, 1951. 232

Ross S., The determination of financial structure the incentive signaling approach, *Bell Journal of Economics*, (8):23–40, 1977. 638

Rothschild M. and *Stiglitz J.*, Equilibrium in competitive insurance markets, *Quarterly Journal of Economics*, (90):629–49, 1976. 594

Rothschild M. and *Stiglitz J.*, Increasing risk i a definition, *Journal of Economic Theory*, 2(3): 225–243, 1970. 541, 636

Rousseau J.-J., Discours sur lorigine et les fondements de linegalite parmi les hommes. Marc-Michel Rey, Genève, 1755. 53, 64

Rubinstein A., Perfect equilibrium in a bargaining model, *Econometrica*, 50:97–109, 1982. 61

Ruggles N., The welfare basis of the marginal cost pricing principle, *Review of Economic Studies*, 17(1):29–46, 1949. 477

Salanié B., The economics of contracts a primer. MIT Press, 2005. 362

Salant S. W. and *Shaffer G.*, Unequal treatment of identical agents in cournot equilibrium, *American Economic Review (Papers and Proceedings)*, 89(3):585–604, 1999. 425, 426

Salant S. W., *Switzer S.*, and *Reynolds R.*, Losses from horizontal merger the effects of an exogenous change in industry structure on cournot nash equilibrium, *Quarterly Journal of Economics*, 98(2):185–99, 1983. 420

Salop S. C., Monopolistic competition with outside goods, *Bell Journal of Economics*, 10(1):141–156, 1979. 152, 288

Samuelson P., The pure theory of public expenditure, *Review of Economics and Statistics*, 36(4): 387–389, 1954. 753

Samuelson P., Foundations of economic analysis. Harvard University Press, Cambridge, 1947. 727, 738

Sandler T. and *Hartley K.*, editors, Handbook of defense economics, volume 2. Elsevier, 2007. 781

Sandler T. and *Tschirhart J.*, Club theory thirty years later, *Public Choice*, 93(3):335–355, 12 1997. 734

Sandmo A., On the theory of the competitive firm under price uncertainty, *American Economic*

Review, 61:65–73, 1971. 545

Sauter W., Services of general economic interest and universal service in eu law, *European Law Review*, 2, 2008. 501

Schaefer M., A study of the dynamics of the fishery for yellowfin tuna in the eastern tropical pacific ocean, *Bulletin Inter-American Tropical Tuna Commission*, 2:247–285, 1957. 519

Schechter P., Customer ownership of the local loop a solution to the problem of interconnection, *Telecommunications Policy*, 20(8):573–584, 10 1996. 736

Schelling T., Models of segregation, *American Economic Review (Papers and Proceedings)*, 59(2): 488–493, 1969. 670, 675

Schmitz P. W. Partial privatization and incomplete contracts the proper scope of government reconsidered. Technical report, SSRN eLibrary, 2000. 399

Schnitzer M., Dynamic duopoly with best price clauses, *RAND Journal of Economics*, 25(1):186–197, 1994. 251

Schrank D. and *Lomax T.* The 2005 urban mobility report. Technical report, Texas Transportation Institute, 2005. 699

Schuknecht L. and *Tanzi V.* Reforming public expenditure in industrialised countries are there trade offs. Technical Report 435, European Central Bank, 2005. 753

Schumpeter J. A., Capitalism socialism and democracy. Harper & Row, New York, 1942. 318, 424, 457, 476, 618

Schwartz M., Third degree price discrimination and output generalizing a welfare result, *American Economic Review*, 80(5):1259–1262, 1990. 100

Scotchmer S., Profit maximizing clubs, *Journal of Public Economics*, 27(1):25–45, 1985a. 148

Scotchmer S., Two tier pricing of shared facilities in a free entry equilibrium, *RAND Journal of Economics*, 16(4):456–472, 1985b. 146

Scotchmer S. Local public goods and clubs, chapter 29, pages 1997–2042. Volume Volume 4 of Auerbach and Feldstein (2002), 2002. ISBN 1573-4420. 145

Scott A., The fishery the objectives of sole ownership, *Journal of Political Economy*, 63(2):116–124, 1955. 523

Seade J. Profitable cost increases and the shifting of taxation equilibrium response of markets in oligopoly. Technical report, University of Warwick, Department of Economics, 1985. 131, 196

Segal I. and *Whinston M.*, Naked exclusion comment, *American Economic Review*, 90(1):296—309, 2000. 742

Selten R., Reexamination of the perfectness concept for equilibrium points in extensive games,

International Journal of Game Theory, (4):25–55, 1975. 57

Shannon C., A mathematical theory of communication, *Bell System Technical Journal*, (27):379–423, 623–656, 1948. 433

Shapiro C. and *Stiglitz J.*, Equilibrium unemployment as a worker discipline device, *American Economic Review*, 74(2):433–444, 1984. 640

Shepard A., Contractual form retail price and asset characteristics in gasoline retailing, *RAND Journal of Economics*, 24(1):58–77, 1993. 387

Shleifer A., A theory of yardstick competition, *RAND Journal of Economics*, 16:319–327, 1985. 496

Shleifer A., Understanding regulation, *European Financial Management Review*, 2005. 452

Shleifer A. Efficient regulation. Technical report, Harvard University, 2010. 740

Shleifer A. and *Vishny R. W.*, The politics of market socialism, *Journal of Economic Perspectives*, 8(2):165–176, 1994. 753

Shubik M. and *Levitan R.*, Market structure and behavior. Harvard University Press, Cambridge, Massachusetts, 1980. 138

Shy O., Industrial organization theory and applications. MIT press, 1996. 726

Sidak G. and *Baumol W.*, Stranded costs, *Harvard Journal of Law & Public Policy*, 18(3):835–849, 1995. 500

Simon H., Organizations and markets, *Journal of Economic Perspectives*, 5(2):25–44, 1991. 351, 367

Simpson J. and *Wickelgren A. L.*, Naked exclusion efficient breach and downstream competition, *American Economic Review*, 97(4):1305–1320, 2007. 261

Singh N. and *Vives X.*, Price and quantity competition in a differentiated duopoly, *RAND Journal of Economics*, (Vol. 15):546–554, 1984. 41, 138, 141

Sinn H.-W., Public policies against global warming a supply side approach, *International Tax and Public Finance*, 15:360–394, 2008. 759

Slade M. E., Market power and joint dominance in uk brewing, *Journal of Industrial Economics*, 52(1):133–163, 2004. 274

Smith A., An inquiry into the nature and causes of the wealth of nations. London, 1776. 58, 65, 224, 232, 240, 246, 298, 454, 456, 474, 528, 753

Solow R. M., A contribution to the theory of economic growth, *Quarterly Journal of Economics*, 70(1):65–94, 1956. 21

Spence M., Monopoly quality and regulation, *Bell Journal of Economics*, 6(2):417–429, 1975. 80

Spence M., Competition in salaries credentials and signaling prerequisites for jobs, *Quarterly Journal of Economics*, 90(1):51–74, 1976a. 160

Spence M., Job market signaling, *Quarterly Journal of Economics*, (87):355–74, 1973. 577, 625, 638

Spence M., Product selection fixed costs and monopolistic competition, *Review of Economic Studies*, 43(2):217–235, 1976b. 151, 296

Spulber D., Famous fables of economics. Blackwell, 2002. 666, 787

Spulber D. F., Non cooperative equilibrium with price discriminating firms, *Economics Letters*, 4 (3):221–227, 1979. 735

Sraffa P., The laws of returns under competitive conditions, *Economic Journal*, 36:535–550, 1926. 283, 475

Stackelberg H., Marktform und gleichgewicht. Springer, (trans. as "The Theory of the Market Economy", Oxford University Press, 1952), 1934. 154, 167, 384

Stigler G., The xistence of x efficiency, *American Economic Review*, 66(1):213–216, 1976. 76

Stigler G., The theory of economic regulation, *Bell Journal of Economics*, 2(1):3–21, 1971. 453

Stigler G. J., The theory of price. Macmillan, New York, 1966. 224

Stiglitz J. E. and *Weiss A.*, Credit rationing in markets with imperfect information, *American Economic Review*, 71(3):393–410, 1981. 635, 637

Stinchcombe M. Notes for honors game theory. Technical report, U. Texas, 2007. 50

Stoft S. E., Carbonomics how to fix the climate and charge it to opec, *SSRN eLibrary*, 2008. 516, 738

Stole L. Price discrimination and competition, chapter 34. Volume 3 of Armstrong and Porter (2007), 2007. 151, 153

Sun G.-z. Exploiting gossens second law a simple proof of the euler equation and the maximum principle. Technical Report 14/05, Aug 2005. 508

Svensson J., Eight questions about corruption, *Journal of Economic Perspectives*, 19(3):19–42, 2005. 460

Sylos-Labini P., Oligopoli e progresso tecnico. Giuffré, 1957. 258

Syverson C. What determines productivity. Technical report, NBER, 2010. 765

Tanzi V. and *Schuknecht L.*, Public spending in the 20thcentury. Cambridge University Press, Cambridge, 2000. 449

Teichova A., *Lévy-Leboyer M.*, and *Nussbaum H.*, editors, Multinational enterprise in historical perspective. Cambridge University Press, 1986. 791

Temin P. Entry prices in telecommunications then and now. Technical report, MIT, 1997. 736

TfL. Congestion charging fourth annual monitoring report. Technical report, Transport for London, 2006. 701

TfL. Congestion charging sixth annual monitoring report. Technical report, Transport for London, 2008. 701

Theil H., Economics and information theory. North-Holland, Amsterdam, 1967. 433

Tilly C., Coercion capital and european states ad990 1990. Basil Blackwell, Cambridge, MA, 1990. 455

Tirole J., The theory of industrial organization. MIT press, 1988. 270, 301, 726

Tirole J., Incomplete contracts where do we stand, *Econometrica*, 67(4):741–781, 1999. 364

Tuck R., Free riding. Harvard University Press, Cambridge, MA, 2008. 65, 730

Tullock G., The welfare costs of tariffs monopolies and theft, *Western Economic Journal*, 5:224–232, 1967. 462

Tullock G., On the efficient organization of trials, *Kyklos*, 28(4):745–762, 1975. 738

Tullock G. Toward a theory of the rent seeking society, chapter Efficient Rent Seeking, pages 97–112. Texas A&M University Press, 1980. 189

Turgot A.-R.-J. Oeuvres de turgot, chapter Réflexions sur la formation et la distribution des richesses (1766), pages 1–66. Eugène Daire et Hippolyte Dussard, 1844. 709

Tyagi R., On firms preferences for product differentiation, *Economics Bulletin*, 2007. 308

Usher D., Theft as a paradigm for departures from efficiency, *Oxford Economic Papers*, 39(2): 235–252, 1987. 188

Usher D., The coase theorem is tautological incoherent or wrong, *Economics Letters*, 61(1):3–11, 1998. 224

Usher D., How dreadful life used to be. Wiley InterScience, New York, qed.econ.queensu.ca/pub/faculty/usher/book/Chapter1.pdf 2003. 219

Vahabi M., The value of destructive power, *Crossroads*, 2005. 738

Veblen T., The theory of the leisure class. Dover, 1899. 670

Verboven F. Quantitative study to define the relevant market in the passenger car sector. Technical report, Catholic University of Leuven, 2002. 430

Verhulst P. F., Notice sur la loi que la population poursuit dans son accroissement, *Correspondance mathématique et physique*, 10:113–121, 1838. 518

Vickers J., Delegation and the theory of the firm, *Economic Journal*, 95(Supplement: Conference

Papers):138–147, 1985. 177

Vickers J. and *Yarrow G.*, Privatization an economic analysis. MIT Press, Cambridge, Mass, 1988. 469

Vickrey W., Counterspeculation auctions and competitive sealed tenders, *Journal of Finance*, 16: 8–37, 196, 1961. 603, 606, 615, 764

Vickrey W., Some implications of marginal cost pricing for public utilities, *American Economic Review (Papers and Proceedings)*, 45(2):605–620, 1955. 477

Vickrey W., Some objections to marginal cost pricing, *Journal of Political Economy*, 56(3), 1948. 482

Vickrey W., Pricing and resource allocation in transportation and public utilities, *American Economic Review*, 53(2):452–465, 1963. 700

Vishwasrao S., Royalties vs fees how do firms pay for foreign technology, *International Journal of Industrial Organization*, 25(4):741–759, 8 2007. 327

Von Neumann J. and *Morgenstern O.*, Theory of games and economic behavior. Princeton University Press, Princeton, 1944. 538

von Pufendorf S., De jure naturae et gentium. (The Law of Nature and of Nations), Lund, 1672. 74

von Thünen J. H., Der isolierte staat in beziehung auf landwirtschaft und national okonomie. Jena, Hamburg, 1826. 289

von Weizsäcker C. C., A welfare analysis of barriers to entry, *Bell Journal of Economics*, 11(24): 399–420, 1980. 158, 187

Waldman M., Eliminating the market for secondhand goods an alternative explanation for leasing, *Journal of Law and Economics*, 40(1):61–92, 1997. 110, 118

Wallace D., Monopolistic competition and public policy, *American Economic Review (Papers and Proceedings)*, 26(1):77–87, 1936. 257

Wallis J. J., Institutions organizations interests and impersonality, *Journal of Economic Behavior & Organization*, Forthcoming, 2011. 219

Walras L., Principe dune theorie mathematique de lechange, *Journal des Economistes*, 1874. 603

Walras L. Letat et le chemin de fer. Technical report, Université de Lausanne, 1875. 475

Waterson M. Beer the ties that bind. Technical Report 930, University of Warwick, 2010. 274

Weber M. Wirtschaft und gesellschaft economy and society trans 1978, chapter chap. 6 (part III), pages 650–78. University Presses Of California, Columbia And Princeton, 1922. 358, 447

Weitzman M., Prices vs quantities, *Review of Economic Studies*, 41:477 – 491, 1974a. 486

Weitzman M., Free access vs private ownership as alternative systems for managing common property, *Journal of Economic Theory*, 8(2):225–234, 1974b. 514

Weitzman M., Income wealth and the maximum principle. Harvard University Press, 2003. 506, 508

Wellisz S. H., Regulation of natural gas pipeline companies an economic analysis, *Journal of Political Economy*, 71(1):30–43, 1963. 490

Wernerfelt B., Renegotiation facilitates contractual incompleteness, *Journal of Economics & Management Strategy*, 16(4):893–910, 2007. 365

Whinston M., On the transaction cost determinants of vertical integration, *Journal of Economic Behavior & Organization*, 9(1):1–23, 2003. 373, 397

Williamson O., Managerial discretion and business behavior, *American Economic Review*, 53(5): 1032–1057, 1963. 488

Williamson O., Economies as an antitrust defense the welfare trade offs, *American Economic Review*, 50(1):18–36, 1968. 424, 425, 426

Williamson O., The new institutional economics taking stock looking ahead, *Journal of Economic Literature*, (38):595–613, 2000. 367

Williamson O., Markets and hierarchies analysis and antitrust implications. New York Free Press, 1975. 360, 372

Williamson O., The economic institutions of capitalism. New York Free Press, 1985. 24, 185, 372

Williamson O., The vertical integration of production market failure considerations, *American Economic Review*, 61:112–23, 1971. 375, 386

Willig R., Consumers surplus without apology, *American Economic Review*, 66(4):589–597, 1976. 46, 728

Wilson C., A model of insurance markets with incomplete information, *Journal of Economic Theory*, (12):167–207, 1977. 597

Wilson R. Strategic models of entry deterrence, chapter 10. Elsevier, 2002. 257

Wilson R., Nonlinear pricing. Oxford University Press, 1993. 113

Wiseman J., The theory of public utility price an empty box, *Oxford Economic Papers*, 9(1):56–74, 1957. 479

Yotopoulos P. A., The rip tide of privatization lessons from chile, *World Development*, 17(5):683–702, 1989. 468

Zuniga M. P. and *Guellec D.* Who licenses out patents and why lessons from a business survey. Technical report, OECD, Directorate for Science, Technology and Industry, 2009. 326

Index

Abuse of Dominant Position, 233, 253
Adverse Selection, 20, 376, 499, 529, 572, 623, 636, 637
Advertising, 30, 55, 170, 181, 185, 186, 258, 276, 277, 292, 308, 341, 377, 394, 458, 577
Aftermarket monopolization, 149
Agency Relation, 212, 392, 401, 498, 528, 622
Agency relationship, 763
Agent, *see* Agency Relation
Agreement Enforcement, 196
Airlines, 82, 91, 96, 274, 276, 312, 471, 473, 688
Anti-competitive, 18, 89, 119, 240, 242, 260, 265, 320
Antitrust, 14, 18, 69, 118, 166, 217, 233, 237, 240, 272, 274, 411, 433, 503, 504, 602, 736
Apple, 260, 322, 345, 346, 667, 670
Appropriation, 185
Arbitrage, 27, 37, 78, 90, 91, 100, 104, 105, 174, 430, 482, 601, 674, 682–684
Arm's Length Relationship, 176, 402
Asset Substitution, 180, 627
Asymmetric Information, 19, 405, 499, 503, 528, 532, *see* Private Information
AT&T, 248, 503, 759
Attrition (War of), 186, 211, 275, 343, 603, 766
Auction, 20, 86, 99, 104, 132, 187, 243, 348, 455, 599, 684, 727
Authority, 388
Averch-Johnson effect, 491

Backward Induction, 57, 562
Bain-Sylos, 165, 258, 262, 277
Bandwagon, 653, 654, *see* Free Rider
Bargain-then-rip-off, 112, 149
Bargaining, 58, 378, 385, 392, 395, 551, 553, 569, 614
Barrier to Entry, 609
Barter, 59, 151, 224, 750
Beauty Contest, 187, 601
Benchmarking , *see* Yardstick
Bertrand Paradox, 54, 133, 138, 142, 157, 163, 174, 186, 212, 250, 263, 271, 285, 290, 304, 321, 324, 422, 551, 596, 667
Best Price Clause, 247, 249
Big Push, 661
Bilateral Monopoly, 59, 173, 371, 380, 397, 611, 616
Block Exemption, 234, 252, 320, 437, 759
Block Tariff, 94, *see* Two-part Tariff
Brand, 77, 251, 258, 292, 309, 377
Brand Loyalty, 129
Bribe, *see* Corruption
Bundling, 94, 96, 135, 164, 645, 647, 652
Bureaucracy, 351, 358, 373, 487, 748

Career Concerns, 356
Cartel, 18, 100, 159, 195, 196, 198, 226, 233, 235, 236, 240–244, 247, 248, 455, 456, 463, 464, 513, 527, 602
Cartelization, *see* Cartel
Chicago School, 164
Club good, 145
Coase Theorem, 63, 207, 222, 223
Coercion, 405, 449
Collusion, 14, 18, 233, 240, 247, 252, 312, 320, 471, 602, *see* Cartel, 737, 756

Commitment, 500, 577, 675
Common Agricultural Policy, 227
Common Pool Resource, 188, 194, *see* Commons, 739
Commons, 194, 196, 198, 513, 515, 517
Competition for the Market, 463
Complementarity, 19, 42, 77, 97, 100, 233, 310, 314, 322, 360, 394, 415, 637, 647, 650, 655, 656
Concentration, 154, 231, 233, 421, 433, 437
Concession, 359, 456, 465, 466, 468, 474, 476, 500, *see* License, 746, 756
Conflict, 185, *see* Contest
Congestion, 105, 341, 587, 683, 685, 688, 732
Congestion Charge, *see* Peak Load Pricing
Constant Elasticity of Substitution, 38
Constrained Pareto optimality, 147
Consumer Surplus, 45–47, 92, 93, 95, 97, 110, 127, 132, 155, 159, 160, 242, 323, 337, 481, 495, 580, 728
Contest, 185, 354
Contestability, 15, 163, 166, 262, 322, 462, 757
Contract, 77, 250, 251, 336, 352, 495, 498, 500, 550, 582, 587, 596, 615, 621
Contract Theory, 15, 142, 376
Copycat, 321, *see* Patent
Corruption, 171, 235, 406, 460, 601
Cost Benefit Analysis, 704
Cost-Plus Contract, 589
Cream Skimming, 503, 595, 597, 598, *see* Free Rider
Credibility, 58, 142, 214, 262, 267, 271, 275, 623, 626, 629, *see* Signaling
Cross-Subsidization, 78, 276, 411, 478, 489, 502

Deadweight Loss, 74, 94, 114, 325, 364, 497, 587, 737
Debt Overhang, 569, 630
Decreasing Marginal Utility, *see* Decreasing Returns to Scale
Decreasing Returns to Scale, 42, 334
Depletion, 525
Deregulation, 416, 463, 470, 655, 697
Deterrence, 221
Differential Pricing, *see* Discrimination
Differentiation, 14, 18, 28, 88, 115, 136, 251, 282, 283, 285, 287, 304, 309, 310, 315, 650
  Characteristics, 292
  Quality, 300
Discrimination, 17, 87, 122, 240, 250, 253, 256, 273, 280, 304, 344, 346, 385, 387, 483, 580, 587, 610, 684, 692, 698, 750, 769
Divestiture, 274, 367, 419, 744
Dominant Position, 15, 18, 69, 231, 233, 257, 270, 411, 646
Double Marginalization, 251, 380, 386, 387, 648, 655
Durable Goods, 94, 97, 104, 117, 118, 120, 166, 250, 251, 270, 288, 330, 647, 744, 768
DVD, 100, 345

Economic variable, 443
Economies of Scale, 28, 32, 79, 164, 257, 311, 314, 359, 415, 420, 473, 496, 528, 545, 670, 681, 728
Economies of Scope, 37, 164, 292, 415
Effect on Trade, 233, 418
Efficiency, 16, 49, 60, 127, 382, 393, 477, 568
Eminent Domain , *see* Expropriation
Entrepreneur, 20, 28, 351, 510, 618
Entry Barrier, 14, 18, 33, 155, 164, 234, 257, 280, 332, 416, 431, 458, 652, 736
Entry Deterrence, *see* Entry Barrier
Equi-marginal Principle, 27, 31, 37, 40, 508, 510
Equilibrium, 33
Essential Facility, 270, 274
Euler Equation, 506
European Commission, 15, 69, 225, 226, 233,

429–431, 433, 497
European Union, 320, 691
Exclusion, 110, 763
Exhaustion Principle, 232, 342
Exogenous Shock, 195
Exploitation Paradox, 199
Expropriation, 449, 476, 700
Externality, 186, 188, 448, 473, 485, 701
Extinction, 525

Facilitating Practices, 247, *see* Best Price Clause
First Best, 551, 555
First Welfare Theorem, 261, 451
Fixed Cost, 18, 30, 32, 35, 37, 70, 80, 89, 94, 155, 161, 165, 259, 267, 268, 278, 309, 336, 369, 403, 415, 477, 483, 495, 586, 594, 630, 655, 687
Foreclosure, 120, 257, 258, 272, 274, 458, *see* Entry Barrier, 745
Franchise, 252, 330, 385, 387, 463, 466, 553, 587, *see* License, 730, 757
Free Cash Flow, *see* Quasi-Rent
Free Rider, 54, 64, 188, 202, 205, 206, 242, 321, 351, 352, 357, 448, 451, 454, 630, 652

Game Theory, 50
Goodwill, 165, 276, 312, 341, 503
Guilds, 246

Hedging, 173
Herfindhal-Hirschmann Concentration Index, *see* Concentration
Hidden Action, 529, *see* Moral Hazard
Hidden Information, 529, *see* Adverse Selection
Hold-up, 62, 330, 352, 372, 375, 378, 380, 385, 388, 389, 456, 464, 589, 655
Homo Economicus, 22
Horizontal agreement, 320, 757
Horizontal Integration, 19, 360, 411, 473
Hotelling, 655, 671, *see* Differentiation

Incentives, 19, 57, 167, 179, 187, 242, 246, 270, 287, 323, 338, 351, 352, 354, 356, 371, 391, 395, 402, 411, 488, 493, 498, 528, 549, 573, 587, 602, 613, 621, 627, 637, 645, 731, 753
Incentive Compatibility, 562
Incorporation, 206
Increasing Returns, 475
Individual Rationality, 14, 22, 60, 125, 259, 262, 275, 368, 392, 563, 584, 671
Information Rent, 113, 114, 572, *see* Private Information
Innovation, 14, 18, 76, 115, 154, 166, 186, 200, 234, 259, 292, 318, 320, 331, 338, 340, 344, 348, 461, 521, 601, 652, *see* R&D
Insurance, 91, 104, 377, 448, 529, 532, 539, 549, 561, 565, 594, 597
Integration, *see* Horizontal Integration, *see* Market Integration, *see* Vertical Integration
Intellectual Property Rights, 18, 338, 670
Interbrand Competition, 292, 438
iPod, 322, 670, *see* Apple
IPR, 768
ISO, 574, 649

Joint Venture, 757

Know-How, 55

Laffer Curve, 207
Lagrange, 41, 721
Lagrangian Optimization, *see* Lagrange
Laissez Faire, 742
Law of one price, 89, 733
Lerner Index, 73, 74, 77, 89, 98, 111, 127, 130, 150, 313, 434, 483, 586, 594
Leverage, 190
Liberalization, 233, 464
License, 69, 155, 330, 339, 345, 600, 604, 609, *see* Concession, *see* Franchise, 735
Licensed Occupations, 246

Limited Liability, 405, 627, 630
Litigation, 186, 221, 406
Lobbying, 132, 192, 461, 498, 601
Location, 285, 292
Long Tail, 443
Loss-Leader, 150, 669
Lumpy Cost, 30, 146, 479, 623

Make or Buy, 19, 221, 224, 373, 400, 402, 404, 408, *see* Vertical Integration
Manna Trick, *see* Equi-marginal Principle
Marginal Cost of Public Funds, 202, 338, 449, 477, 483, 499, 587, 588, 735
Marginal Cost Pricing, 477
Market Failure, 280, 310, 322, 568, 575, 643, 651, 666, 667
Market Integration, 161, 691
Market Power, 14, 15, 17, 33, 54, 58, 70, 166, 234, 247, 434
Mathematical Optimization, 25
Maximum Differentiation Principle, 284, 287
Meet or Release, 250
Mercantilism, 86
Merger Paradox, 241, 411, 415, 420
Metering, 108
Microsoft, 69, 119, 135, 166, 260, 322, 340, 344, 436, 462, 647, 667, 669
Minimum Efficient Scale, 33, 35, 162, 476
Monopolization, *see* Dominant Position
monopolization, 474
Monopoly, 17, 68, 338, 471, 609
Moral Hazard, 20, 351, 352, 528, 529, 639
Moral Hazard in Team, 354
Most Favored Customer, 119, 250
Multi-plant, 36, 77, 241, 421

Nash Bargaining Solution, 61
Nash Equilibrium, 56
Natural Monopoly, 20, 164, 359, 448, 463, 464, 473, 481, 492, 496, 500, 601, 616
Natural Resources, 19, 505
Network Externality, 320, 415
Non Linear Pricing, 112, 583
Normative Analysis, 48

Oligopoly, 14, 17, 370
  Duopoly, 124, 262, 651
Open access, 740
Opportunity Cost, 30, 75, 177, 294, 551–553, *see* Sunk Cost, 750
Outsourcing, 30, 380

Paradox of power, 200
Pareto, 147, *see* Efficiency
Patent, 321, 322, 330, 334, 338, 339, 768
Peak Load Pricing, 479, 681, 683, 690, 694, 699, 701
Peak-Load Pricing, 95, 105, 107, 732
Pecuniary Externalities, 661
Peer Pressure, 23, 64, 354, 670
Perfect Competition, 17, 33
Performance based Compensation, 187, 212
Perks, *see* Perquisite
Perquisite, 177, 199, 470, 629, 750, 758
Piece Rate, 552
Positive Discrimination, 103
Potential Competition, 164, 166, 431
Poverty Trap, 663
Predation, 275
Preemption, 258, 332, 651
Price Discrimination, 231, 236, 385, 646, 647, 663, 674, 675, *see* Discrimination
Principal, 212, 392, 401, 528, 552, 561, 563, 568, 572, 582, *see* Agency Relation
Prisoner Dilemma, 52, 64, 196, 740
Prisoner's Dilemma, 122, 175, 177, 180, 306, 307, 335, 513
Private Information, 498, 534
  Perfect Information, 33
Private vs. Public, 408
Privatization, 465
Privilege, 401

Procurement, 132, 185, 187, 370, 374, 401, 466, 480, 489, 601, 603
Procurement Auction, *see* Auction, *see* Procurement
Procurement auction, 401
Producer Surplus, 34
Productivity Analysis, 28, 76, 457
Property Rights, 63, 186, 200, 207, 223, 368, 373, 388, 395
Public choice theory, 186
Public Funds, *see* Marginal Cost of Public Funds
Public Good, 83, 145, 146, 320, 448, 594, *see* Public Service, 734, 768
Public Service, 83, 233, 359, 447, 448, 450
Public Service Obligation, 497, 501
Purchase History, 121

Quality, 79, 655, 683
Quality of Service, 478
Quasi Rent, 72, 247, 629
Quasi-Rent, 61, 364, 366, 384, 388–390, 400

R&D, 30, 79, 318, 338
Ramsey-Boiteux, 616
Ratchet Effect, 497, 499, 568
Rationing, 160, 587, 635, 673, 677, 683, 737
Regalian State, 765
Regulation, 19, 29, 74, 94, 453, 473, 568
Regulatory Capture, 132, 340, 498, 503, 504, 601
Relevant Market, 49, 234, 411, 429, 430
Renegotiation, 364
Rent Maximization, 520
Rent-Seeking, 19, 63, 70, 75, 186, 372, 406, 458, 601
Replacement Effect, 324
Resale Price Maintenance, 252, 272
Residual claimant, 352, 373, 389, 390, 456, 552, 589, 590
Restraint, *see* Vertical Restraint
Returns to Scale, 28, 709, *see* Economies of Scale
Revealed Preference, 99
Revelation Principle, 582, 606, 733, 765
Revelation principle, 596
Reverse Engineering, 321, 346
Risk, 173, 377

Sample Selection Bias, 374, 405
Scale, *see* Economies of Scale
Scope, *see* Economies of Scope
Screening, 187
Search, 325
Seasonality, 37, 105, 732
Second Best, 164, 480, 482, 492, 499, 563, 583, 586, 616, 633, 758
Self-Selection, 88, 107, 582, *see* Incentives
Service of General Economic Interest, *see* Public Service
Signaling, 259, 275, 577, 638
Social interaction, 357, 670
Special Interest Group, 201, 280, 452, 453, 457
Specific Investment, 59
Stackelberg, 729
Standard, 645
  Compatible, 645
Standardization, 401
Start-up costs, 30
State, 280
State Aid, 227, 231, 233, 254, 502, 759
State Owned Enterprise, 448
State Regulation, 78, 185, 186, 190, 203, 222, 230, 246, 316, 368, 399, 406, 408, 446, 453, 459, 463, 478, 501, 575, 641, 748, 756
Stranded Cost, 500, *see* Sunk Cost
Strategic Complements, 176, 179
Strategic Substitutes, 175, 176, 180, 335
Sunk Cost, 30, 155, 156, 165, 211, 318, 322, 332, 375, 389, 500, 675, *see* Fixed Cost
Switching Cost, 129, 212, 372, 429, 651, 733

Take or pay, 92

Take-or-pay, 262
Taxation, 131, 760
Team Work, 194
Technical progress, 60
Tournament, *see* Performance based Compensation
Trade, 297
Trademark, 104, 230, 338, 341, 458
Transaction Cost, 514
Transaction cost, 105
Transaction Costs, 223, 355, 370–373, 376, 601, 747
Trust, 235, 236, 408, 423, 741
Two-part Tariff, 94, 106, 107, 111, 331, 385, 386, 588, 647, 769
Tying, 236, 274, 344, 646, 648

Uncertainty, 401, 532, 537, 545, 614, 627, 631, 635, 637, 638
Undertaking, 233, 274, 751
Universal Service Obligation, *see* Public Service Obligation
Untax, 516
Urban Economics, 292
Utility, 39, 43, 55, 137, 213, 285, 301, 339, 727, 729

Variety, 37, 88, 161, 292, 320
Versioning, 304
Vertical Integration, 176, 179, 252, 270, 320, 330, 370, 373, 380, 386, 388, 395, 399, 408
Vertical Restraint, 19, 238, 240, 251, 346
Voluntary Export Restraint, 244

Welfare, 17, 24, 48, 49, 74, 93, 95, 96, 127, 131, 140, 155, 156, 158, 162, 164, 175, 177, 182, 235, 242, 265, 272, 314, 319, 322, 325, 336, 337, 388, 398, 480, 482, 496, 497, 510, 572, 589, 600, 609, 614, 651, 660, 667, 706, 728, 730, 736
Willingness to pay, 27, 48, 581, 659, 660

X-efficiency, 75, 202, 352, 358

Yardstick, 29, 76, 187, 325, 356, 496, 557