


Unit IV : Clustering

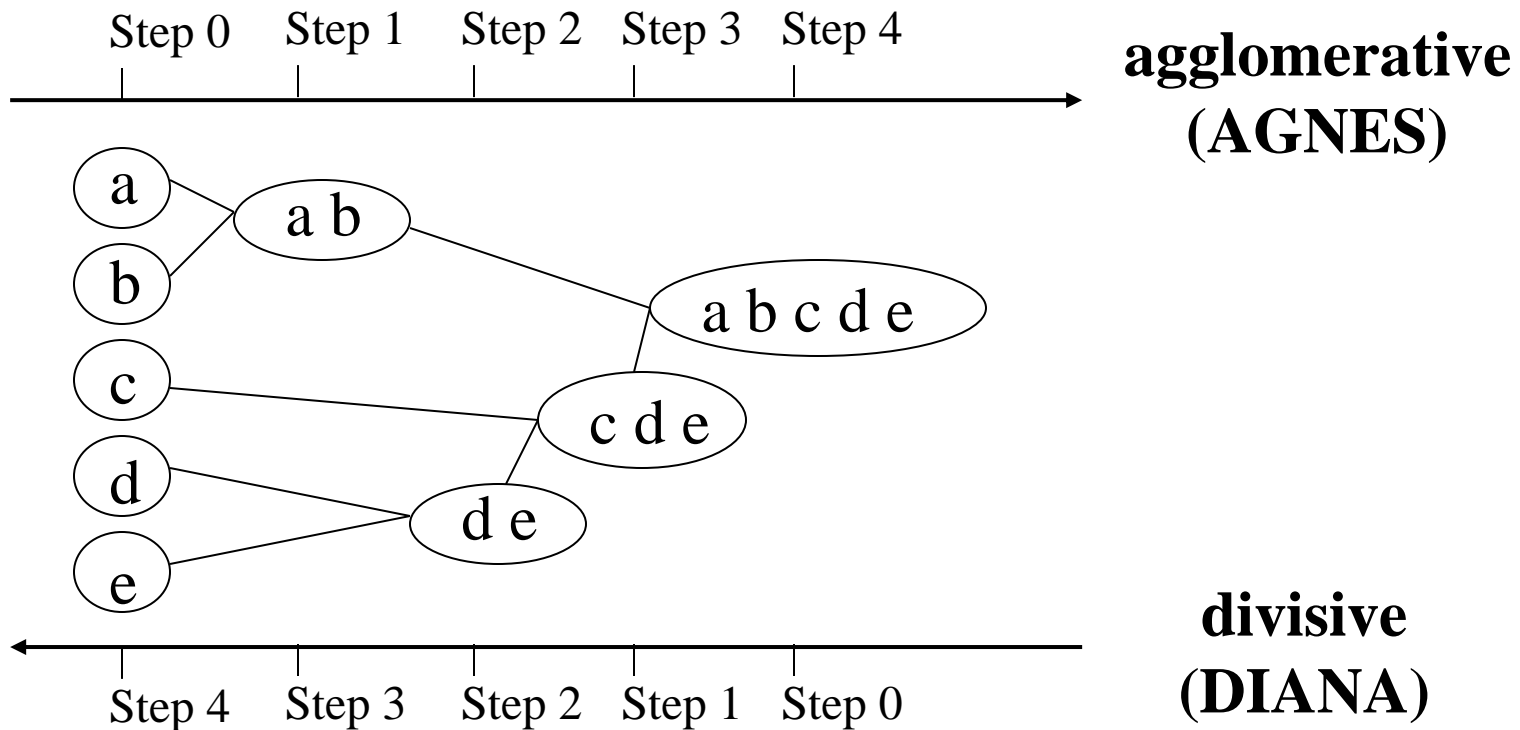
Cluster Analysis – Partitioning Methods – Hierarchical
Methods – Density Based Methods – Grid Based
Methods – Outlier Analysis

Cluster Analysis

1. What is Cluster Analysis?
2. Types of Data in Cluster Analysis
3. A Categorization of Major Clustering Methods
4. Partitioning Methods
5. Hierarchical Methods 
6. Density-Based Methods
7. Grid-Based Methods
8. Outlier Analysis
9. Summary

Hierarchical Clustering

- Use distance matrix as clustering criteria. This method does not require the number of clusters k as an input, but needs a termination condition

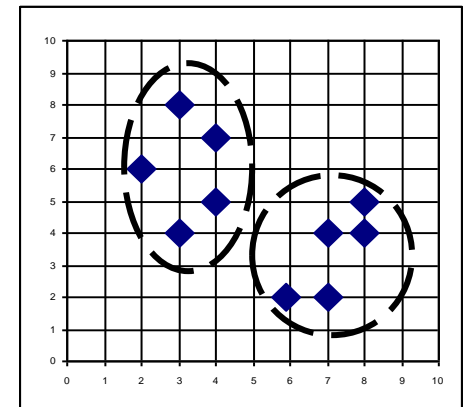
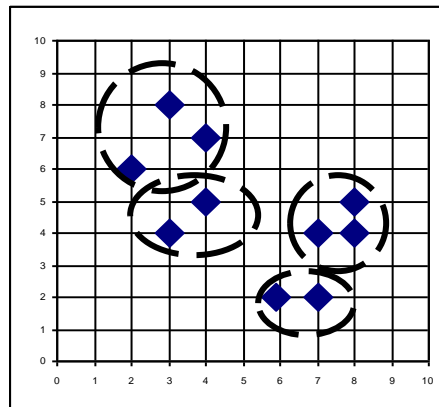
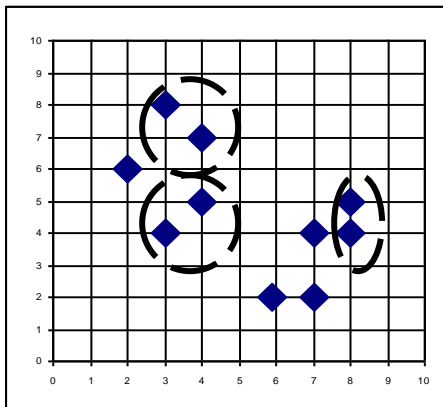


**agglomerative
(AGNES)**

**divisive
(DIANA)**

AGNES (Agglomerative Nesting)

- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical analysis packages, e.g., Splus
- Use the Single-Link method and the dissimilarity matrix.
- Merge nodes that have the least dissimilarity
- Go on in a non-descending fashion
- Eventually all nodes belong to the same cluster

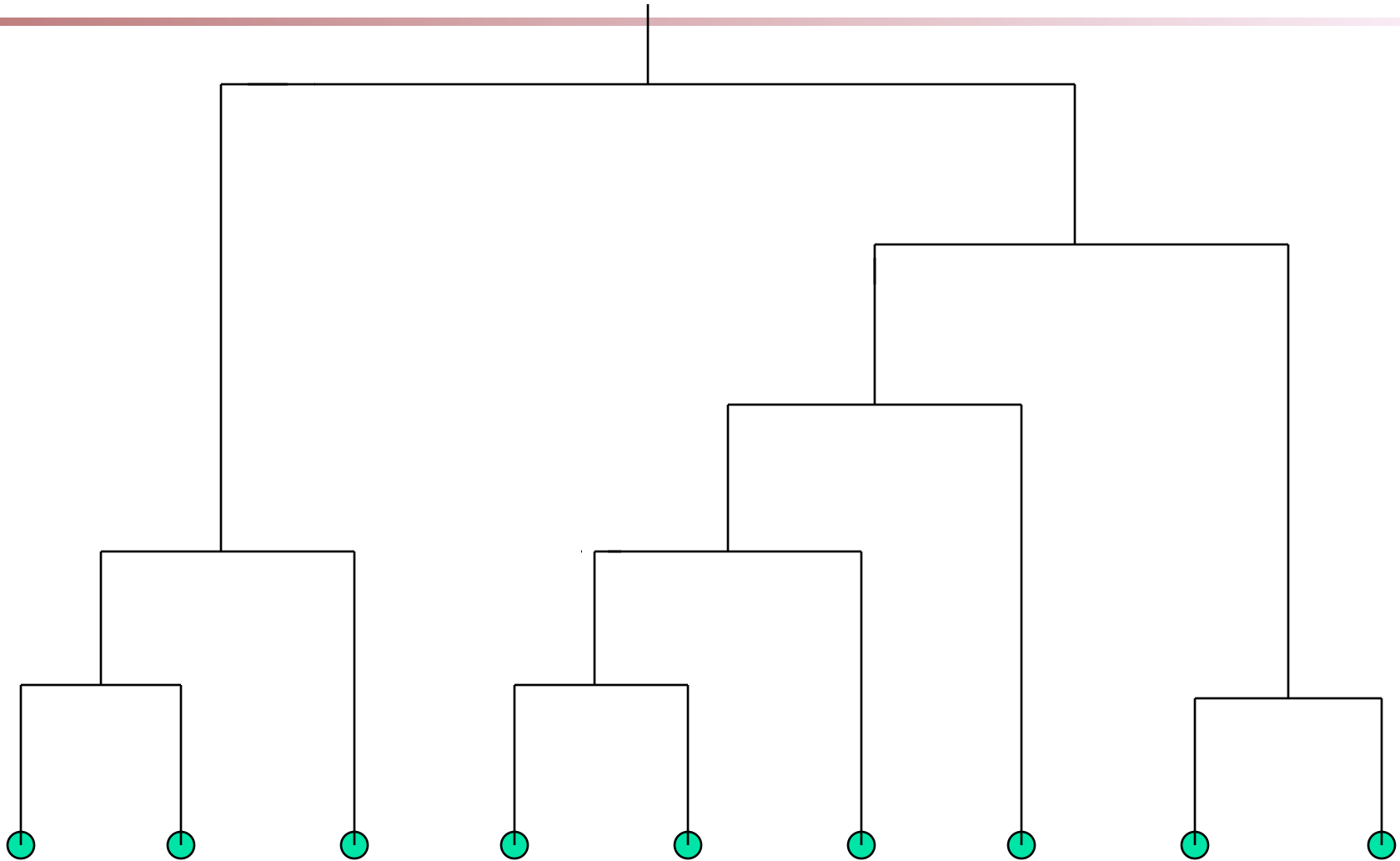


Dendrogram: Shows How the Clusters are Merged

Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram.

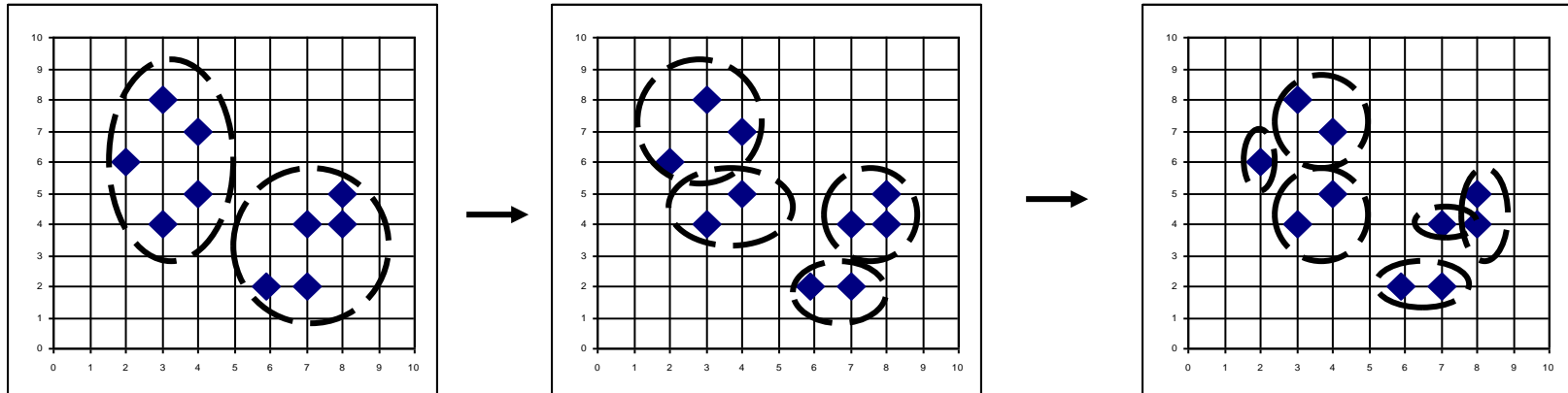
A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.

Dendrogram: Shows How the Clusters are Merged



DIANA (Divisive Analysis)

- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical analysis packages, e.g., Splus
- Inverse order of AGNES
- Eventually each node forms a cluster on its own





Recent Hierarchical Clustering Methods

- Major weakness of agglomerative clustering methods
 - do not scale well: time complexity of at least $O(n^2)$, where n is the number of total objects
 - can never undo what was done previously
- Integration of hierarchical with distance-based clustering
 - BIRCH (1996): uses CF-tree and incrementally adjusts the quality of sub-clusters
 - ROCK (1999): clustering categorical data by neighbor and link analysis
 - CHAMELEON (1999): hierarchical clustering using dynamic modeling



BIRCH (1996)

- Birch: Balanced Iterative Reducing and Clustering using Hierarchies (Zhang, Ramakrishnan & Livny, SIGMOD'96)
- Incrementally construct a CF (Clustering Feature) tree, a hierarchical data structure for multiphase clustering
 - Phase 1: scan DB to build an initial in-memory CF tree (a multi-level compression of the data that tries to preserve the inherent clustering structure of the data)
 - Phase 2: use an arbitrary clustering algorithm to cluster the leaf nodes of the CF-tree
- *Scales linearly*: finds a good clustering with a single scan and improves the quality with a few additional scans
- *Weakness*: handles only numeric data, and sensitive to the order of the data record.

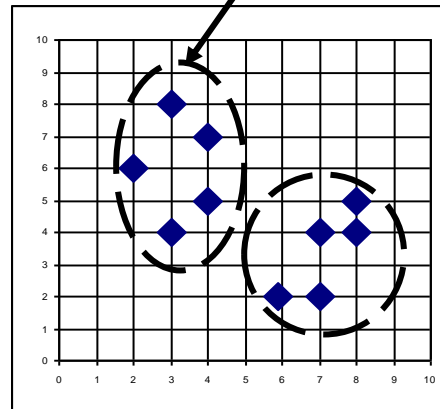
Clustering Feature Vector in BIRCH

Clustering Feature: $CF = (N, \vec{LS}, SS)$

N : Number of data points

$$LS: \sum_{i=1}^N X_i$$

$$SS: \sum_{i=1}^N X_i^2$$



$$CF = (5, (16,30), (54,190))$$

$$(3,4)$$

$$(2,6)$$

$$(4,5)$$

$$(4,7)$$

$$(3,8)$$

CF-Tree in BIRCH

- Clustering feature:
 - summary of the statistics for a given subcluster: the 0-th, 1st and 2nd moments of the subcluster from the statistical point of view.
 - registers crucial measurements for computing cluster and utilizes storage efficiently
- A CF tree is a height-balanced tree that stores the clustering features for a hierarchical clustering
 - A nonleaf node in a tree has descendants or “children”
 - The nonleaf nodes store sums of the CFs of their children
- A CF tree has two parameters
 - Branching factor: specify the maximum number of children.
 - threshold: max diameter of sub-clusters stored at the leaf nodes

Clustering Categorical Data: The ROCK Algorithm

- ROCK: RObust Clustering using linkS
 - S. Guha, R. Rastogi & K. Shim, ICDE'99
- Major ideas
 - Use links to measure similarity/proximity
 - Not distance-based
 - Computational complexity: $O(n^2 + nm_m m_a + n^2 \log n)$
- Algorithm: sampling-based clustering
 - Draw random sample
 - Cluster with links
 - Label data in disk
- Experiments
 - Congressional voting, mushroom data

Similarity Measure in ROCK

- Traditional measures for categorical data may not work well, e.g., Jaccard coefficient
- Example: Two groups (clusters) of transactions
 - C_1 . $\langle a, b, c, d, e \rangle$: $\{a, b, c\}$, $\{a, b, d\}$, $\{a, b, e\}$, $\{a, c, d\}$, $\{a, c, e\}$, $\{a, d, e\}$, $\{b, c, d\}$, $\{b, c, e\}$, $\{b, d, e\}$, $\{c, d, e\}$
 - C_2 . $\langle a, b, f, g \rangle$: $\{a, b, f\}$, $\{a, b, g\}$, $\{a, f, g\}$, $\{b, f, g\}$
- Jaccard co-efficient may lead to wrong clustering result
 - C_1 : 0.2 ($\{a, b, c\}$, $\{b, d, e\}$) to 0.5 ($\{a, b, c\}$, $\{a, b, d\}$)
 - C_1 & C_2 : could be as high as 0.5 ($\{a, b, c\}$, $\{a, b, f\}$)
- Jaccard co-efficient-based similarity function:

$$Sim(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$

- Ex. Let $T_1 = \{a, b, c\}$, $T_2 = \{c, d, e\}$

$$Sim(T_1, T_2) = \frac{|\{c\}|}{|\{a, b, c, d, e\}|} = \frac{1}{5} = 0.2$$

Link Measure in ROCK

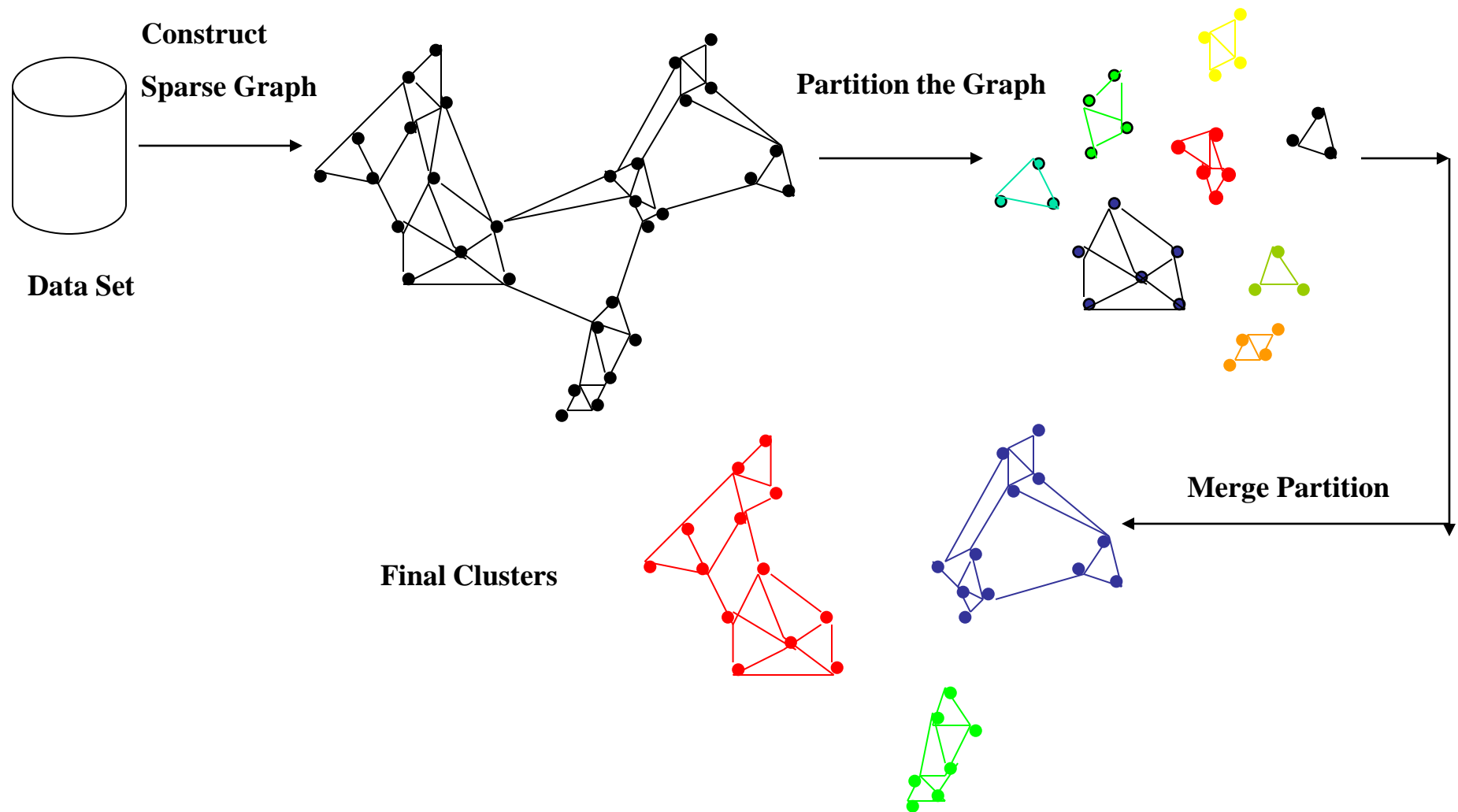
- Links: # of common neighbors
 - $C_1 \langle a, b, c, d, e \rangle$: {a, b, c}, {a, b, d}, {a, b, e}, {a, c, d}, {a, c, e}, {a, d, e}, {b, c, d}, {b, c, e}, {b, d, e}, {c, d, e}
 - $C_2 \langle a, b, f, g \rangle$: {a, b, f}, {a, b, g}, {a, f, g}, {b, f, g}
- Let $T_1 = \{a, b, c\}$, $T_2 = \{c, d, e\}$, $T_3 = \{a, b, f\}$
 - $\text{link}(T_1, T_2) = 4$, *since they have 4 common neighbors*
 - {a, c, d}, {a, c, e}, {b, c, d}, {b, c, e}
 - $\text{link}(T_1, T_3) = 3$, *since they have 3 common neighbors*
 - {a, b, d}, {a, b, e}, {a, b, g}
- Thus link is a better measure than Jaccard coefficient



CHAMELEON: Hierarchical Clustering Using Dynamic Modeling

- CHAMELEON: by G. Karypis, E.H. Han, and V. Kumar'99
- Measures the similarity based on a dynamic model
 - Two clusters are merged only if the *interconnectivity* and *closeness (proximity)* between two clusters are high *relative to* the internal interconnectivity of the clusters and closeness of items within the clusters
 - **Cure** ignores information about **interconnectivity** of the objects, **Rock** ignores information about the **closeness** of two clusters
- A two-phase algorithm
 1. Use a graph partitioning algorithm: cluster objects into a large number of relatively small sub-clusters
 2. Use an agglomerative hierarchical clustering algorithm: find the genuine clusters by repeatedly combining these sub-clusters

Overall Framework of CHAMELEON





CHAMELEON (Clustering Complex Objects)

