
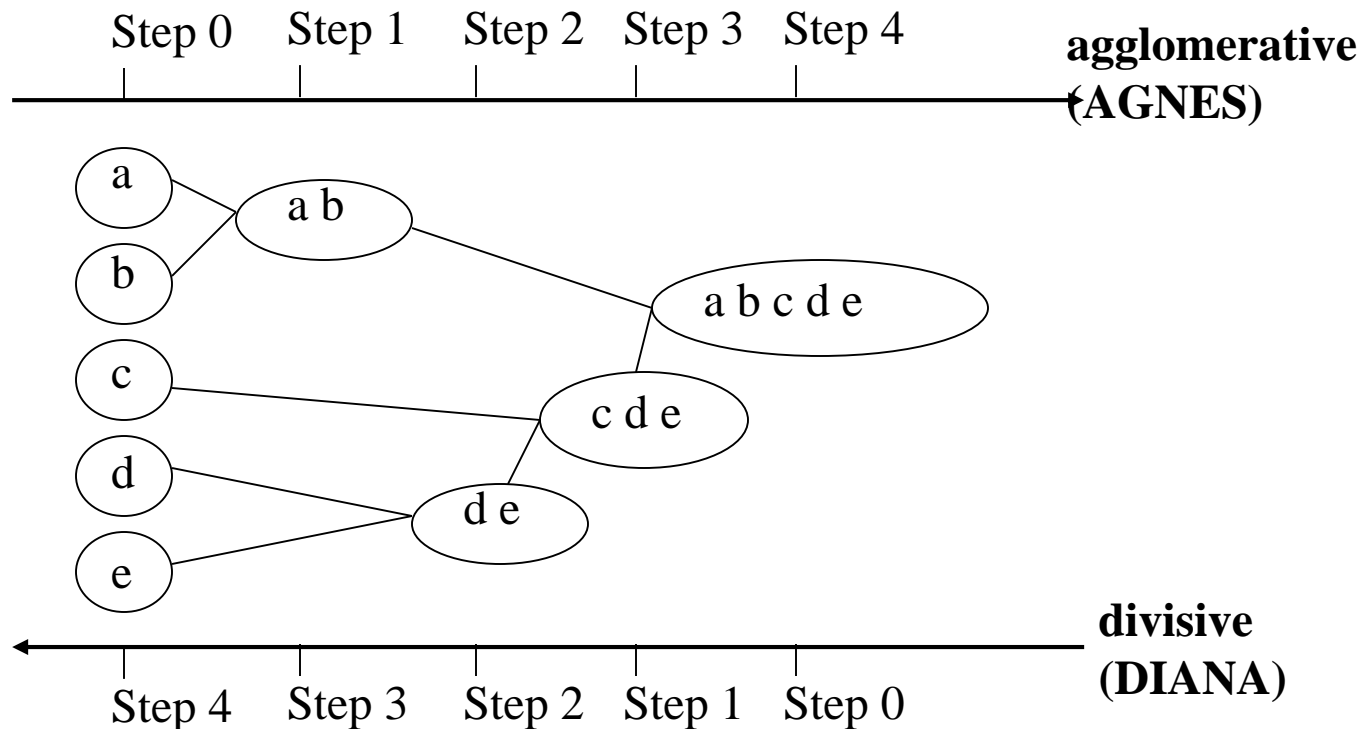


Chapter 10. Cluster Analysis: Basic Concepts and Methods

- Cluster Analysis: Basic Concepts
- Partitioning Methods
- Hierarchical Methods 
- Density-Based Methods
- Grid-Based Methods
- Evaluation of Clustering
- 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

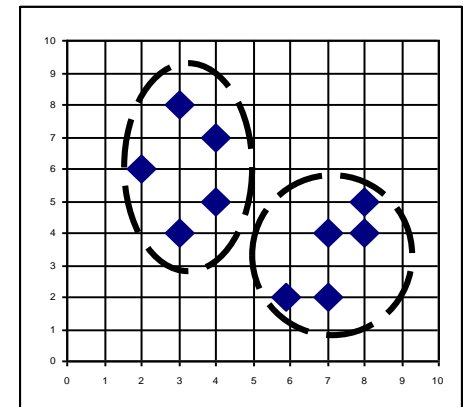
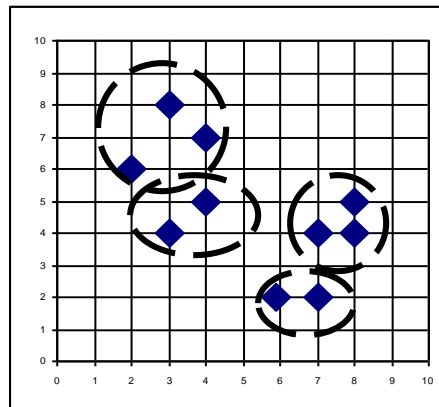
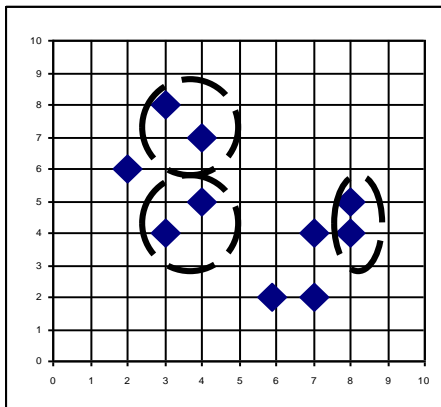


Two Types of Hierarchical Clustering Alg.s

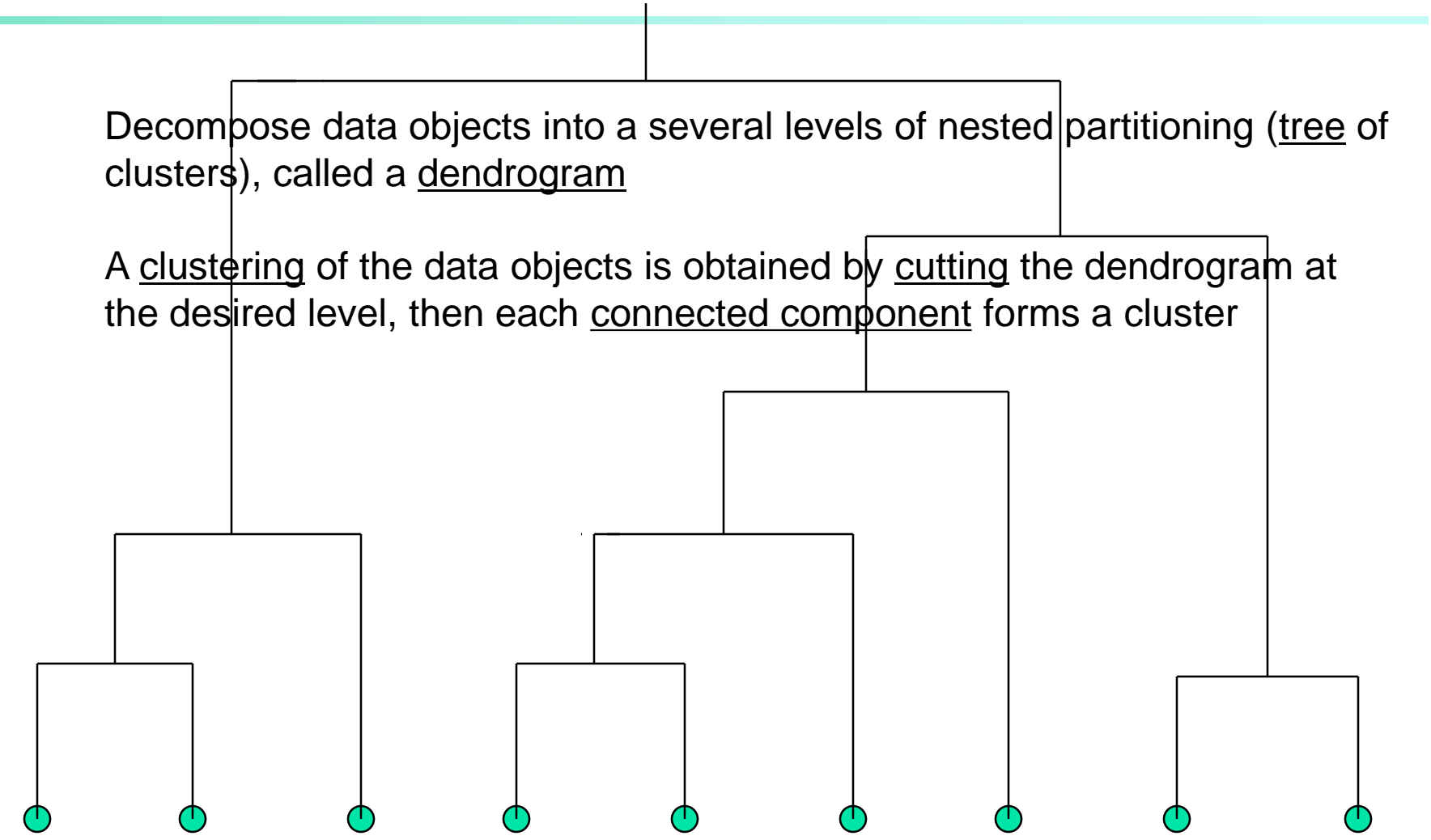
- **Agglomerative** (bottom-up): merge clusters iteratively.
 - start by placing each object in its own cluster.
 - merge these atomic clusters into larger and larger clusters.
 - until all objects are in a single cluster.
 - Most hierarchical methods belong to this category. They differ only in their definition of *between-cluster similarity*.
- **Divisive** (top-down): split a cluster iteratively.
 - It does the reverse by starting with all objects in one cluster and subdividing them into smaller pieces.
 - Divisive methods are not generally available, and rarely have been applied.

AGNES (Agglomerative Nesting)

- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical 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

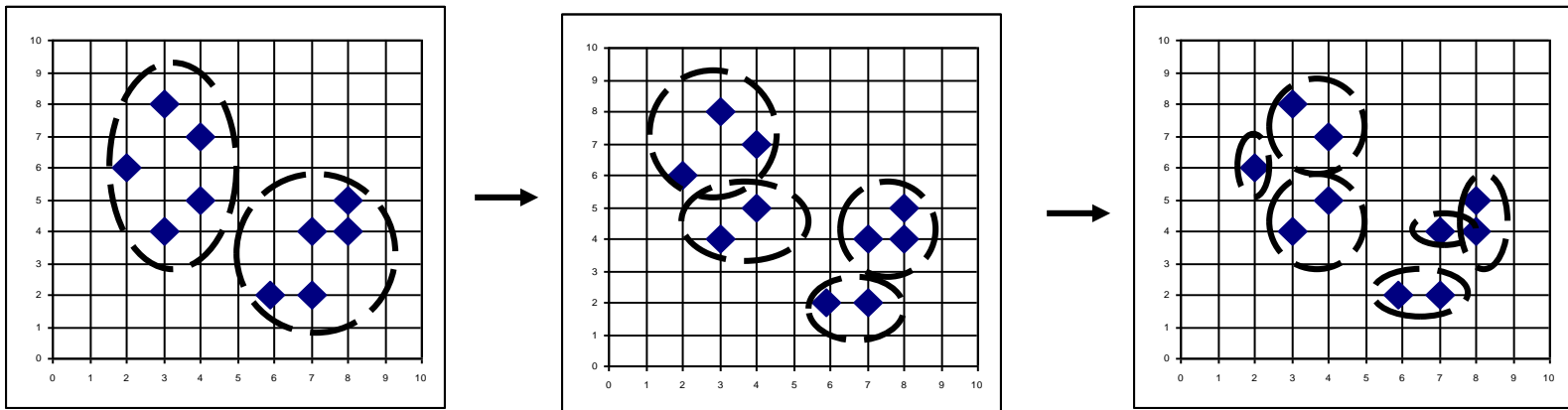


Dendrogram: Shows How Clusters are Merged

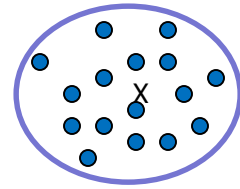
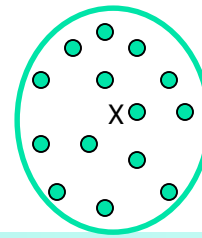


DIANA (Divisive Analysis)

- 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



Distance between Clusters

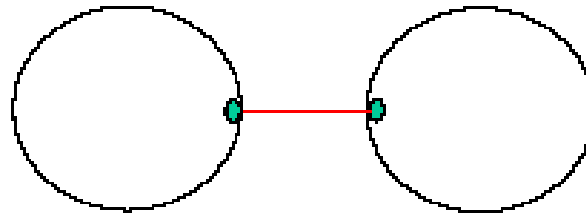


- **Single link:** smallest distance between an element in one cluster and an element in the other, i.e., $\text{dist}(K_i, K_j) = \min(t_{ip}, t_{jq})$
- **Complete link:** largest distance between an element in one cluster and an element in the other, i.e., $\text{dist}(K_i, K_j) = \max(t_{ip}, t_{jq})$
- **Average:** avg distance between an element in one cluster and an element in the other, i.e., $\text{dist}(K_i, K_j) = \text{avg}(t_{ip}, t_{jq})$
- **Centroid:** distance between the centroids of two clusters, i.e., $\text{dist}(K_i, K_j) = \text{dist}(C_i, C_j)$
- **Medoid:** distance between the medoids of two clusters, i.e., $\text{dist}(K_i, K_j) = \text{dist}(M_i, M_j)$
 - **Medoid:** a chosen, centrally located object in the cluster

Generic Methods for Computing Distance Between Clusters

■ Single link

- The distance between one cluster and another cluster is equal to the shortest distance from any member of one cluster to any member of the other cluster. For categorical data, use the greatest similarity from any member of one cluster to any member of the other cluster.
- i.e., $\text{dist}(K_i, K_j) = \min(t_{i0}, t_{i1})$

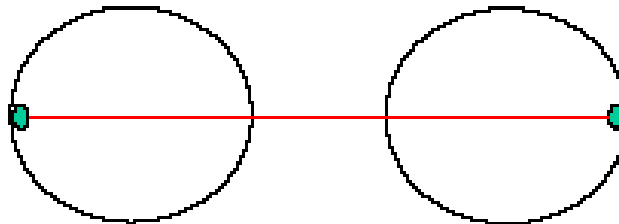


$$\text{dist}(K_t, (K_i \cup K_j)) = \min \{ \text{dist}(K_t, K_i), \text{dist}(K_t, K_j) \}$$

Generic Methods for Computing Distance Between Clusters

■ Complete link

- In *complete-link* clustering (also called the *diameter* or *maximum* method), the distance between one cluster and another cluster is equal to the greatest distance from any member of one cluster to any member of the other cluster.
- $\text{dist}(K_i, K_j) = \max(t_{ip}, t_{jq})$

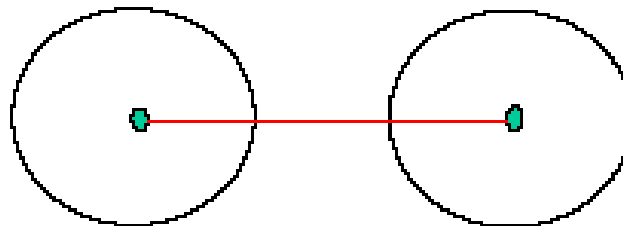


$$\text{dist}(K_t, (K_i \cup K_j)) = \min \{ \text{dist}(K_t, K_i), \text{dist}(K_t, K_j) \}$$

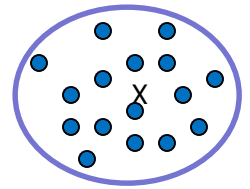
Generic Methods for Computing Distance Between Clusters

- Average link

- In *average-link* clustering, the distance between one cluster and another cluster is equal to the average distance from any member of one cluster to any member of the other cluster.
- $\text{dist}(K_i, K_j) = \text{avg}(t_{ip}, t_{jq})$



Centroid, Radius and Diameter of a Cluster (for numerical data sets)



- Centroid: the “middle” of a cluster

$$C_m = \frac{\sum_{i=1}^N (t_{ip})}{N}$$

- Radius: square root of average distance from any point of the cluster to its centroid

$$R_m = \sqrt{\frac{\sum_{i=1}^N (t_{ip} - c_m)^2}{N}}$$

- Diameter: square root of average mean squared distance between all pairs of points in the cluster

$$D_m = \sqrt{\frac{\sum_{i=1}^N \sum_{i=1}^N (t_{ip} - t_{iq})^2}{N(N-1)}}$$

Example: Agglomerative clustering with single-link

■ Distance matrix

1.

| | a | b | c | d | e | f |
|---|---|---|----|----|----|-----|
| a | 0 | 4 | 13 | 24 | 12 | 8 |
| b | | 0 | 10 | 22 | 11 | 10 |
| c | | | 0 | 7 | 3 | 9 |
| d | | | | 0 | 6 | 18 |
| e | | | | | 0 | 8.5 |
| f | | | | | | 0 |

2.

| | a | b | {c,e} | d | f |
|-------|---|---|-------|----|-----|
| a | 0 | 4 | 12 | 24 | 8 |
| b | | 0 | 10 | 22 | 10 |
| {c,e} | | | 0 | 6 | 8.5 |
| d | | | | 0 | 18 |
| f | | | | | 0 |

Example: Agglomerative clustering with single-link

3.

| | {a,b} | {c,e} | d | f |
|-------|-------|-------|----|-----|
| {a,b} | 0 | 10 | 22 | 8 |
| {c,e} | | 0 | 6 | 8.5 |
| d | | | 0 | 18 |
| f | | | | 0 |

4.

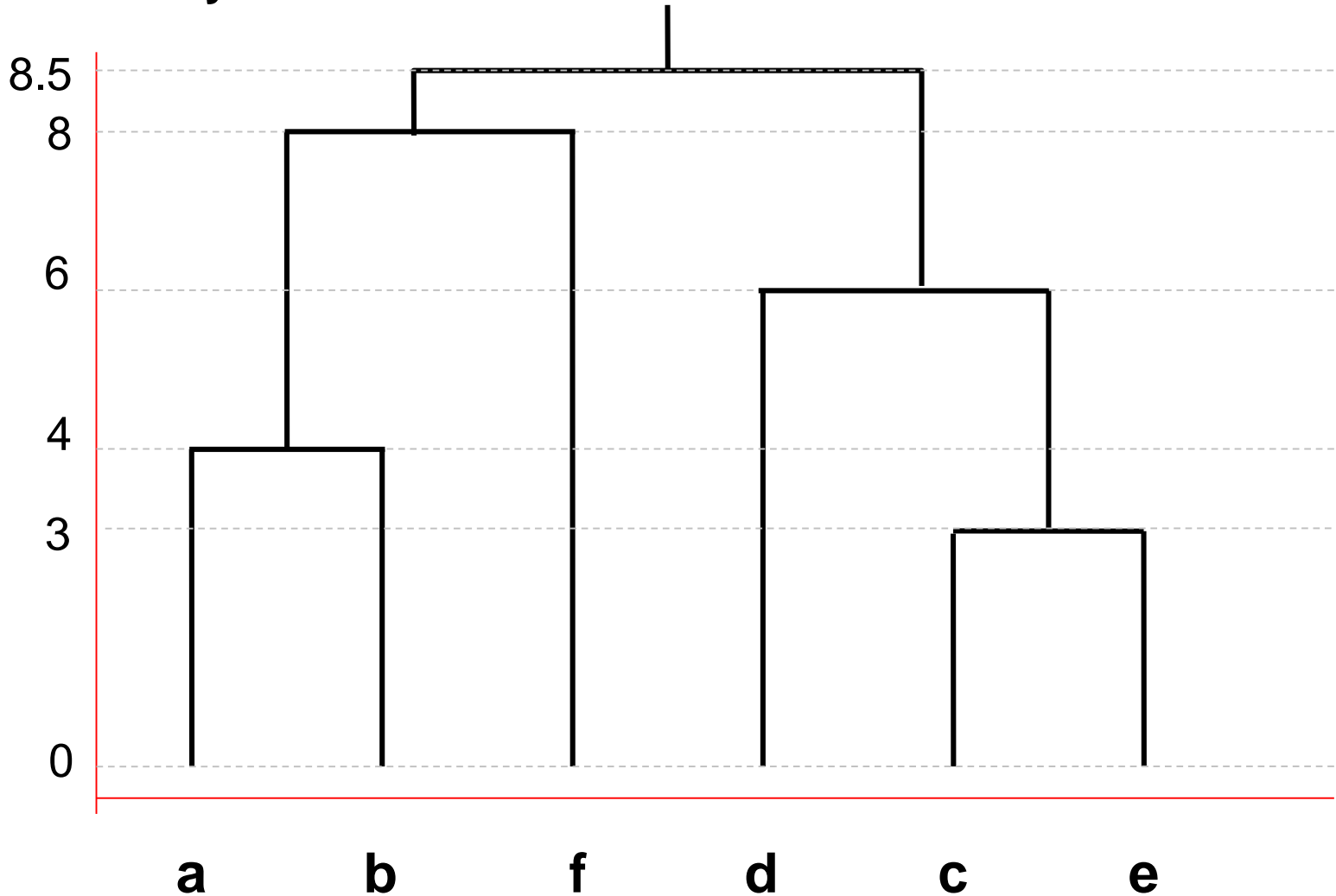
| | {a,b} | {c,e,d} | f |
|---------|-------|---------|-----|
| {a,b} | 0 | 10 | 8 |
| {c,e,d} | | 0 | 8.5 |
| f | | | 0 |

5.

| | {a,b,f} | {c,e,d} |
|---------|---------|---------|
| {a,b,f} | 0 | 8.5 |
| {c,e,d} | | 0 |

Example: Dendrogram

— Similarity between clusters



Extensions to Hierarchical Clustering

- Major weakness of agglomerative clustering methods
 - Can never undo what was done previously
 - Do not scale well: time complexity of at least $O(n^2)$, where n is the number of total objects
- Integration of hierarchical & distance-based clustering
 - BIRCH (1996): uses CF-tree and incrementally adjusts the quality of sub-clusters (**Project for Student**)
 - CHAMELEON (1999): hierarchical clustering using dynamic modeling (**Project for Student**)