CLUSTERING

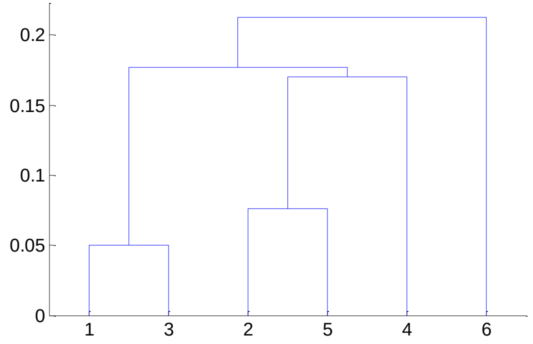
Sanjay Ranka
Distinguished Professor
Department of Computer and Information Science and Engineering
www.sanjayranka.com
sanjayranka@gmail.com
352 514 4213

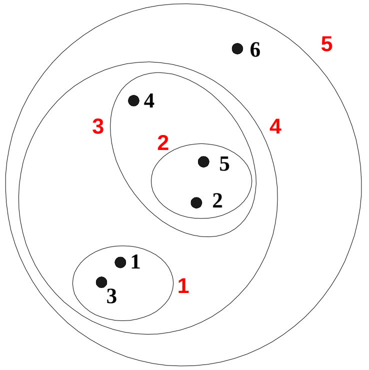
Hierarchical Clustering

- Two main types:
  - Agglomerative
    - Start with the points as individual clusters
    - Merge clusters until only one is left
  - Divisive
    - Start with all the points as one cluster
    - Split clusters until only singleton clusters remain
  - Agglomerative is more popular
- Traditional hierarchical algorithms use a similarity or distance matrix.
  - Merge or split one cluster at a time

Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree.
- Can be visualized as a dendrogram
  - Tree like diagram
  - Records the sequences of merges or splits





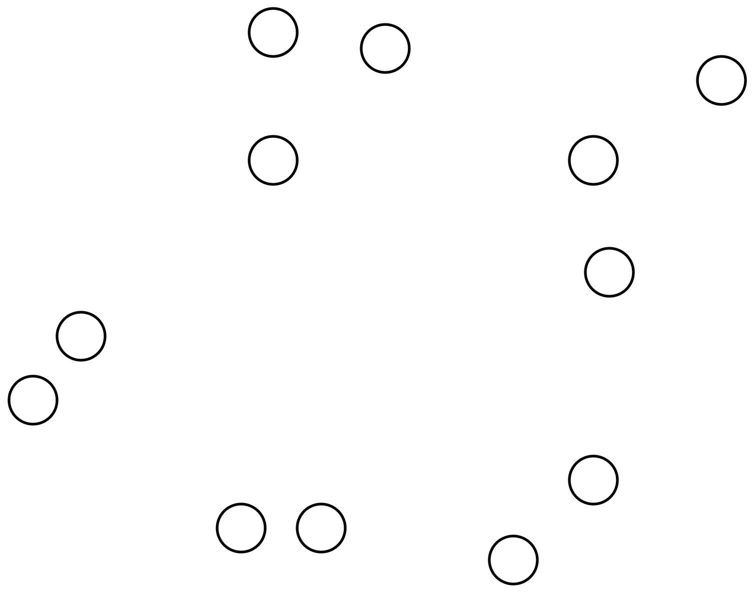
- Can 'cut' the dendrogram to get a partitional clustering

Basic Agglomerative Clustering Algorithm

- Algorithm is straightforward
  - Compute the proximity matrix, if necessary
  - Let each data point be a cluster
  - Repeat
    - Merge the two closest clusters
    - Update the proximity matrix
  - Until only a single cluster remains
- Key operation is the computation of the proximity of two clusters.
- Different approaches to defining the distance between clusters distinguishes the different algorithms.

Agglomerative Hierarchical Clustering:

- For agglomerative hierarchical clustering we start with clusters of individual points and a proximity matrix.

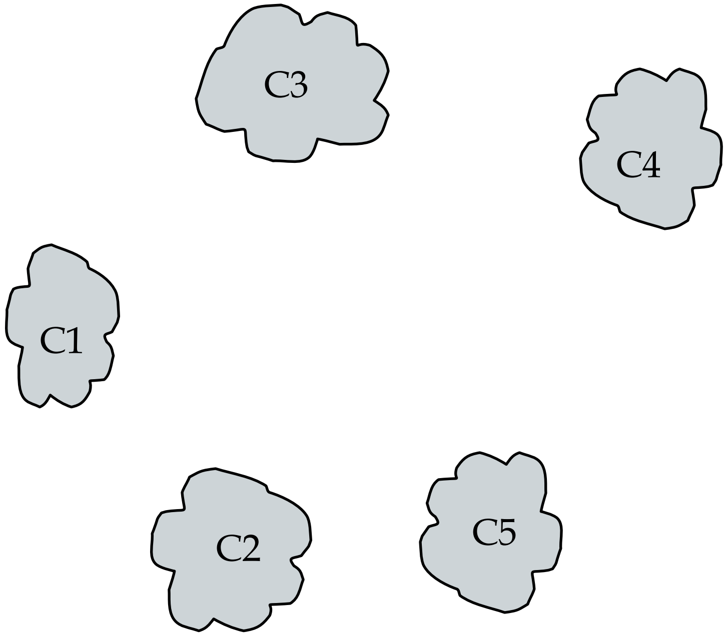


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

Agglomerative Hierarchical Clustering: Intermediate Situation

- After some merging steps, we have some clusters.

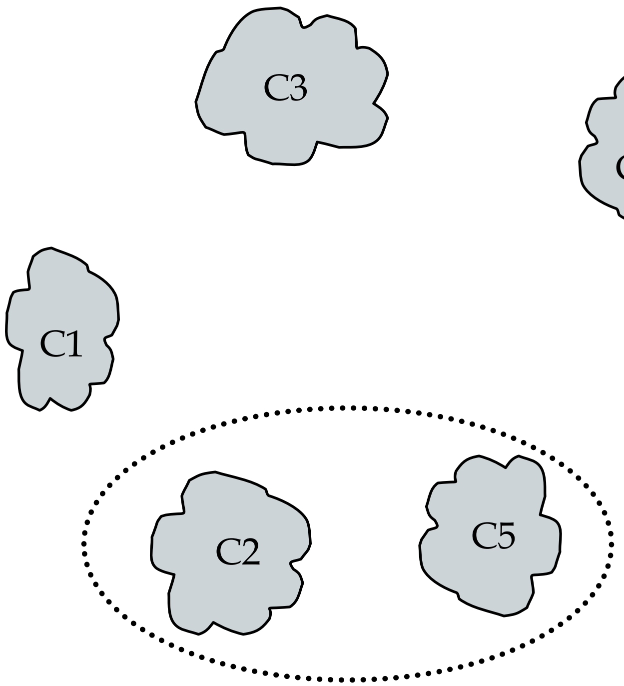


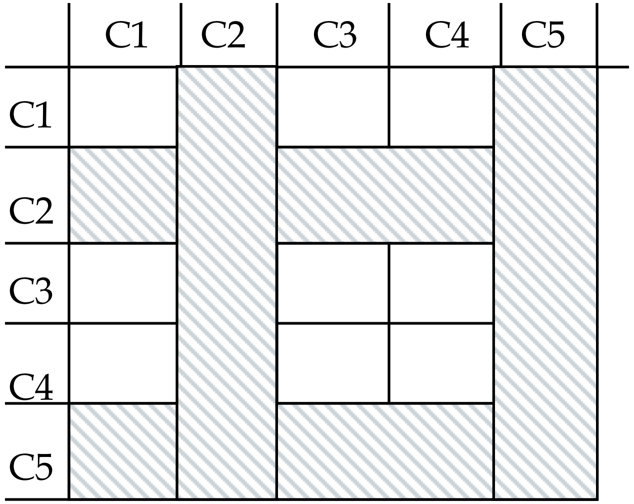
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix

Agglomerative Hierarchical Clustering: Intermediate Situation

- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.

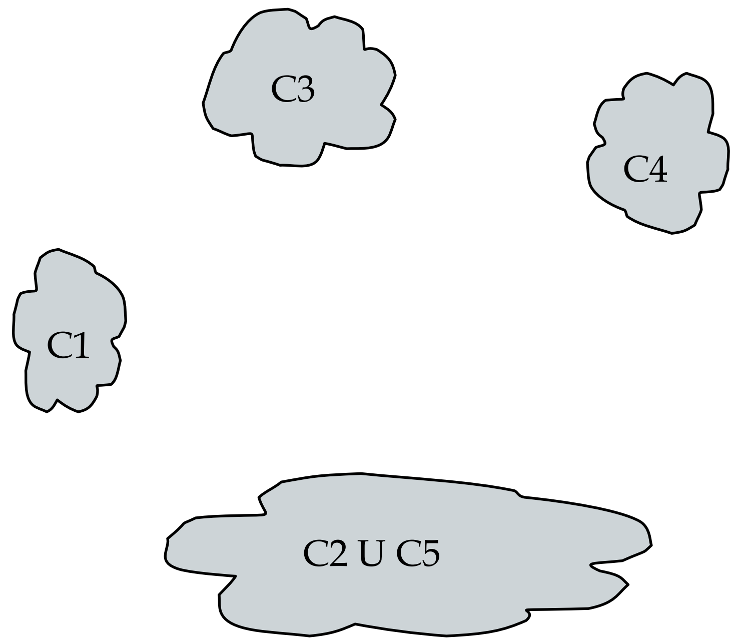




Proximity Matrix

Agglomerative Hierarchical Clustering: After Merging

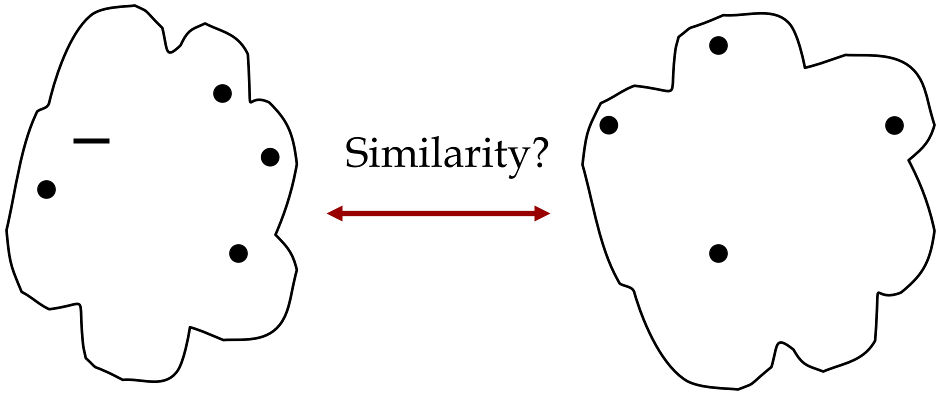
- The question is “How do we update the proximity matrix?”



	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix

How to Define Inter-Cluster Similarity



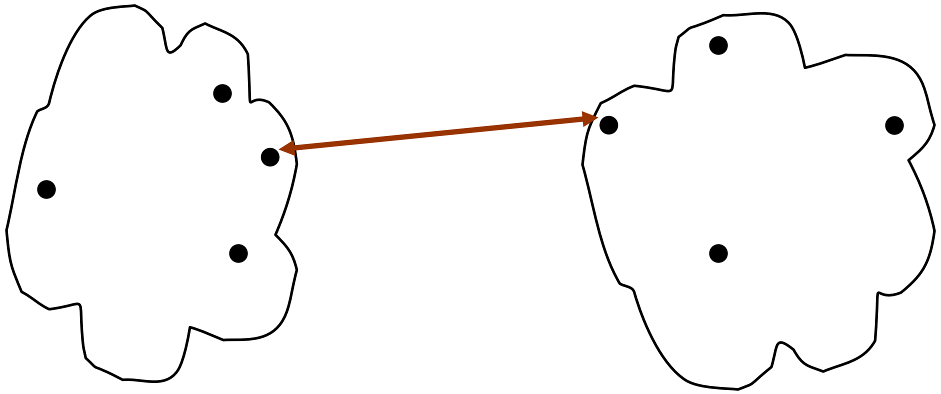
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
. . .						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

How to Define Inter-Cluster Similarity

–



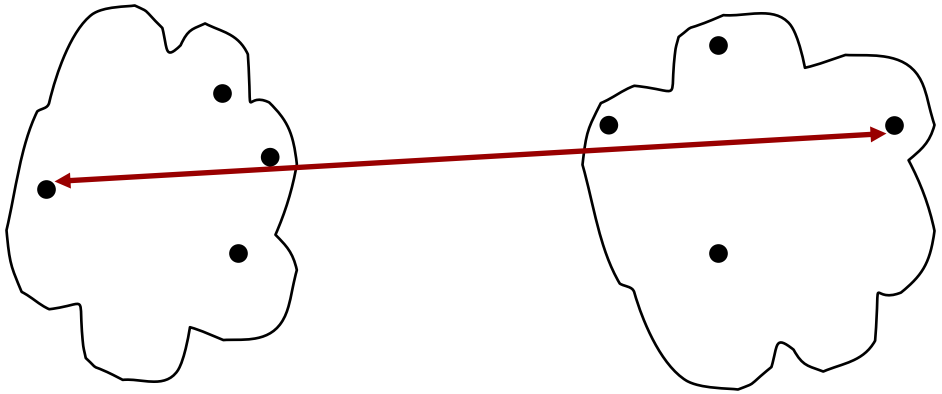
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- *MIN*
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

How to Define Inter-Cluster Similarity

–



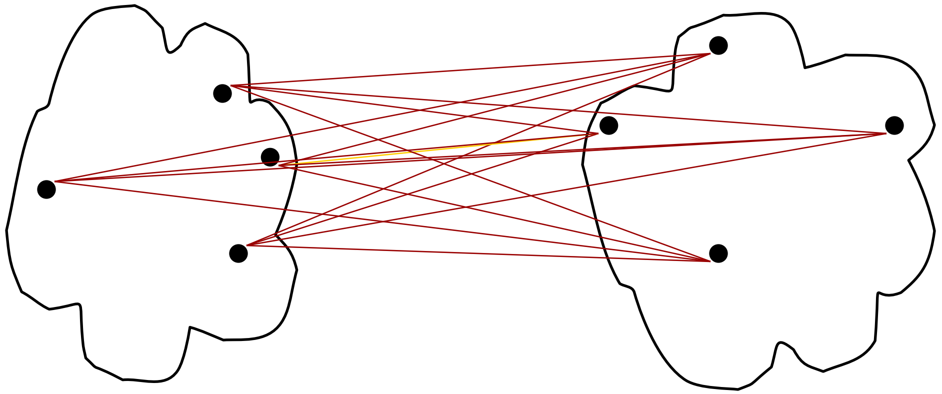
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
. . .						

Proximity Matrix

- MIN
- *MAX*
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

How to Define Inter-Cluster Similarity

–



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

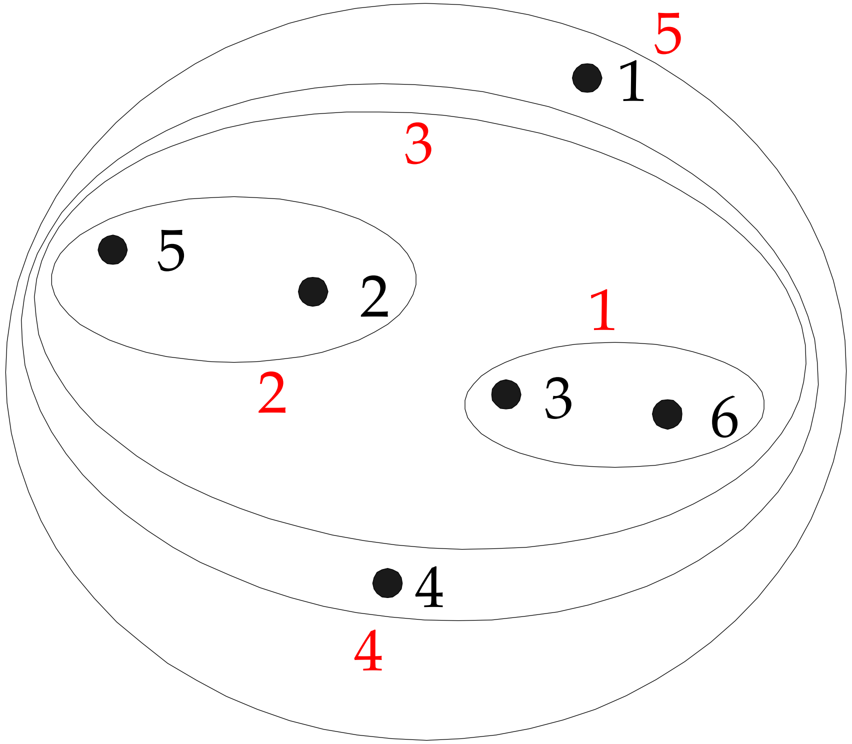
Proximity Matrix

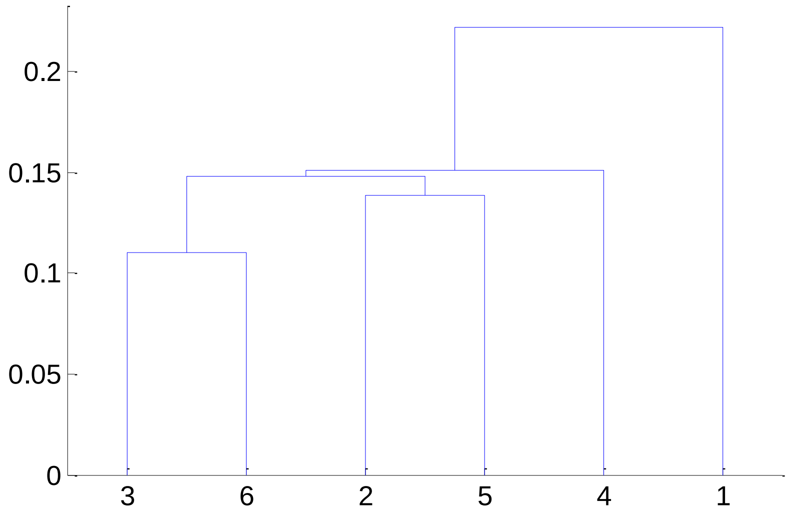
- MIN
- MAX
- *Group Average*
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

Cluster Similarity: MIN or Single Link

- Similarity of two clusters is based on the two closest points in the different clusters.
  - Determined by one pair of points, i.e., by one link in the proximity graph.
- Can handle non-elliptical shapes.
- Sensitive to noise and outliers.

Hierarchical Clustering: MIN

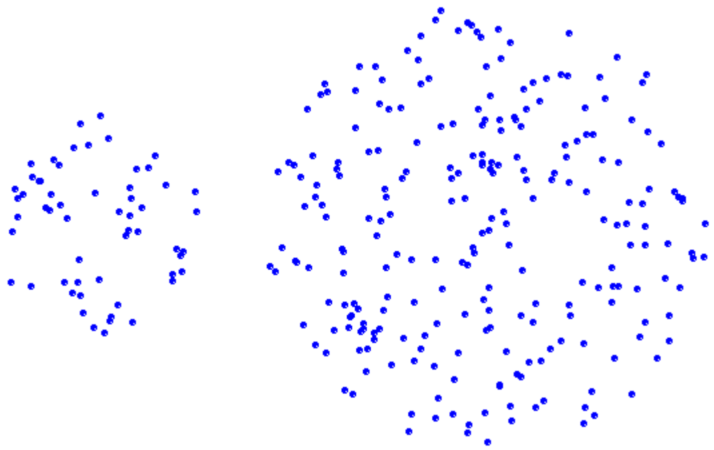




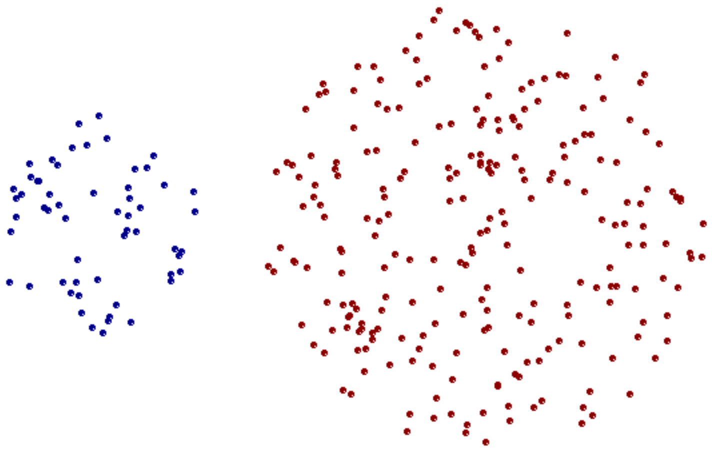
Nested Clusters

Dendrogram

Strength of MIN

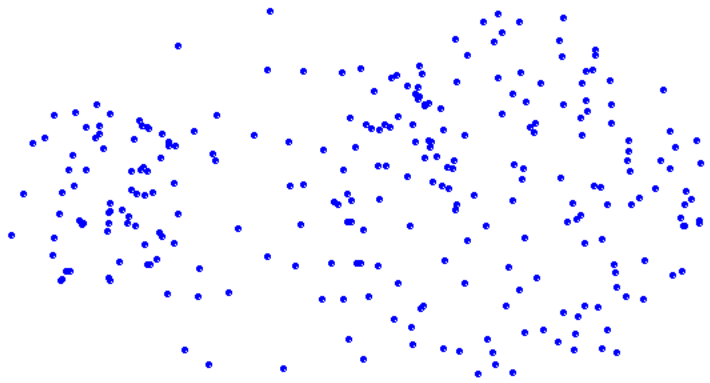


Original Points

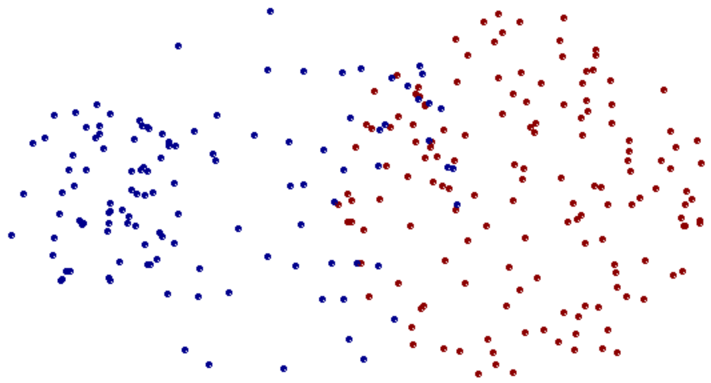


Two Clusters

Limitations of MIN



Original Points

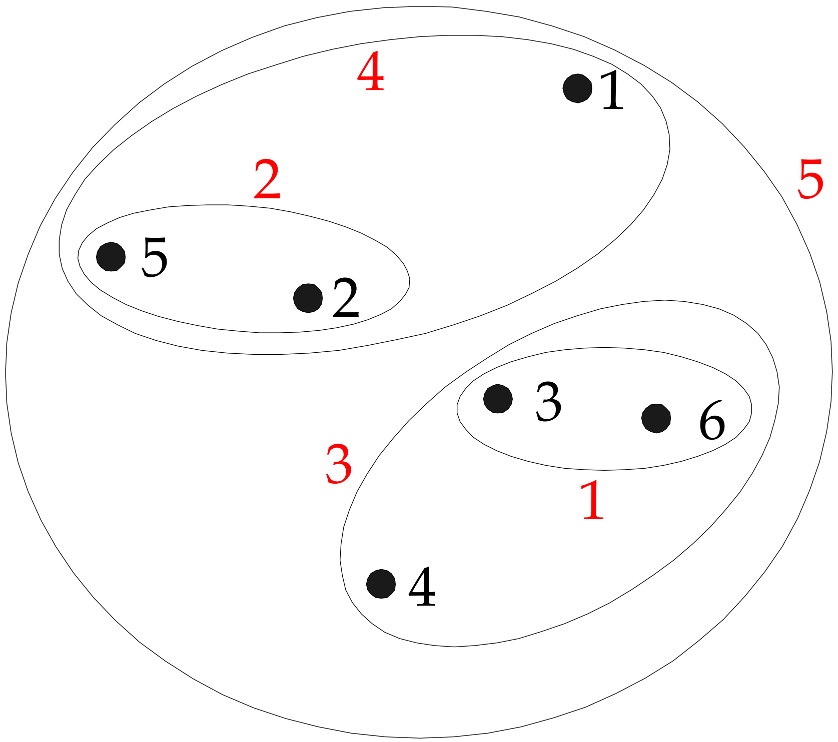


Two Clusters

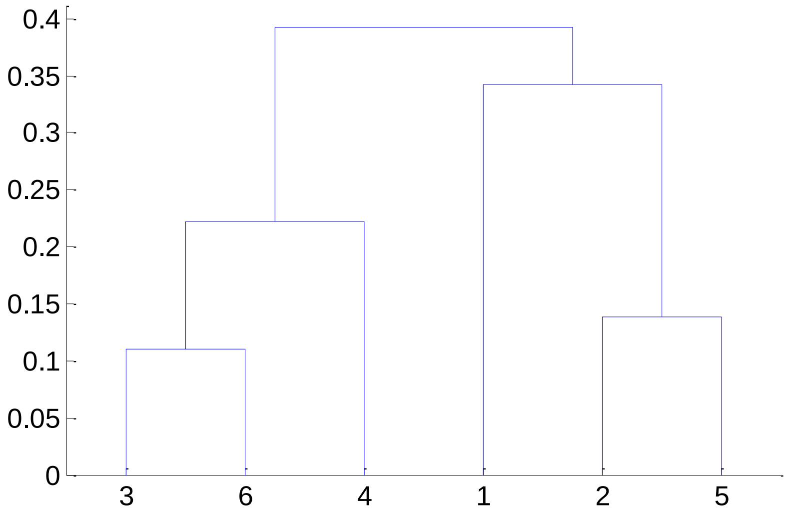
Cluster Similarity: MAX or Complete Linkage

- Similarity of two clusters is based on the two most distant points in the different clusters.
  - Determined by all pairs of points in the two clusters.
- Tends to break large clusters.
- Less susceptible to noise and outliers.
- Biased towards globular clusters.

Hierarchical Clustering: MAX

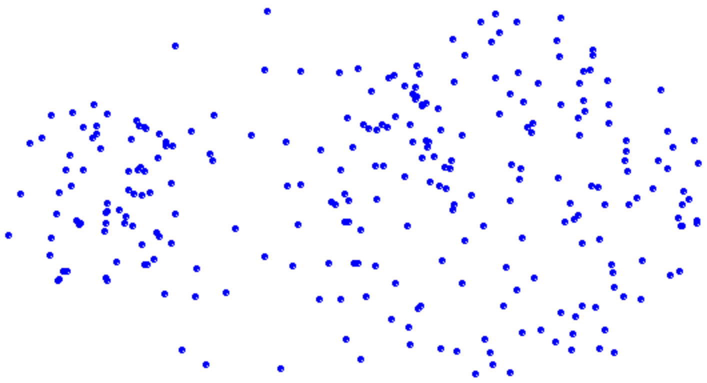


Nested Clusters

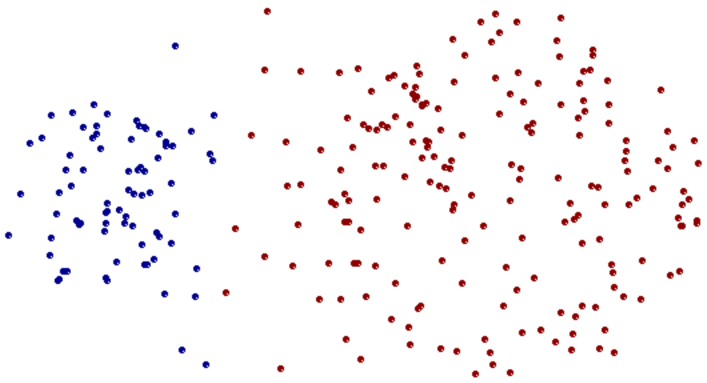


Dendrogram

Strength of MAX

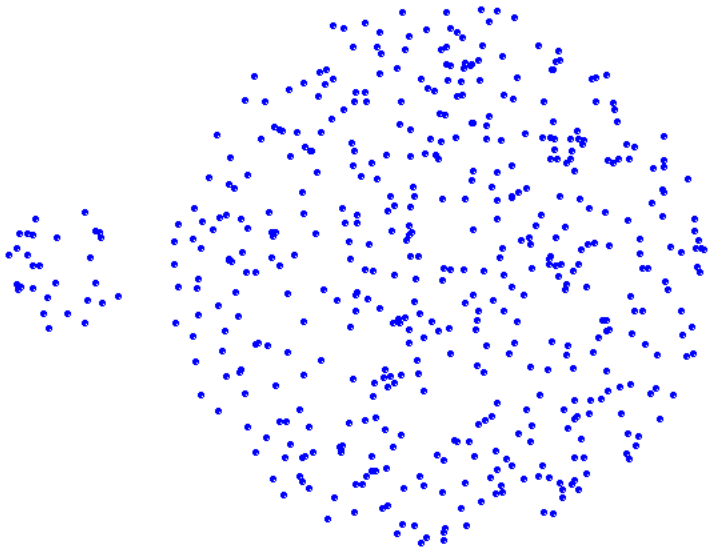


Original Points

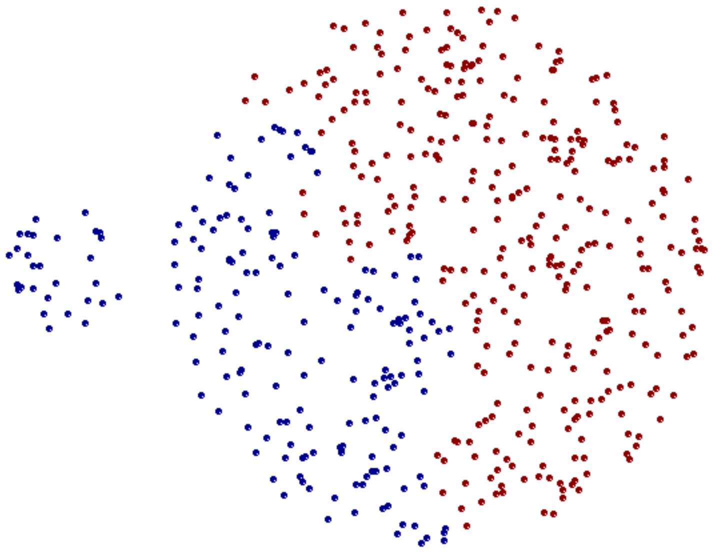


Two Clusters

Limitations of MAX



Original Points



Two Clusters

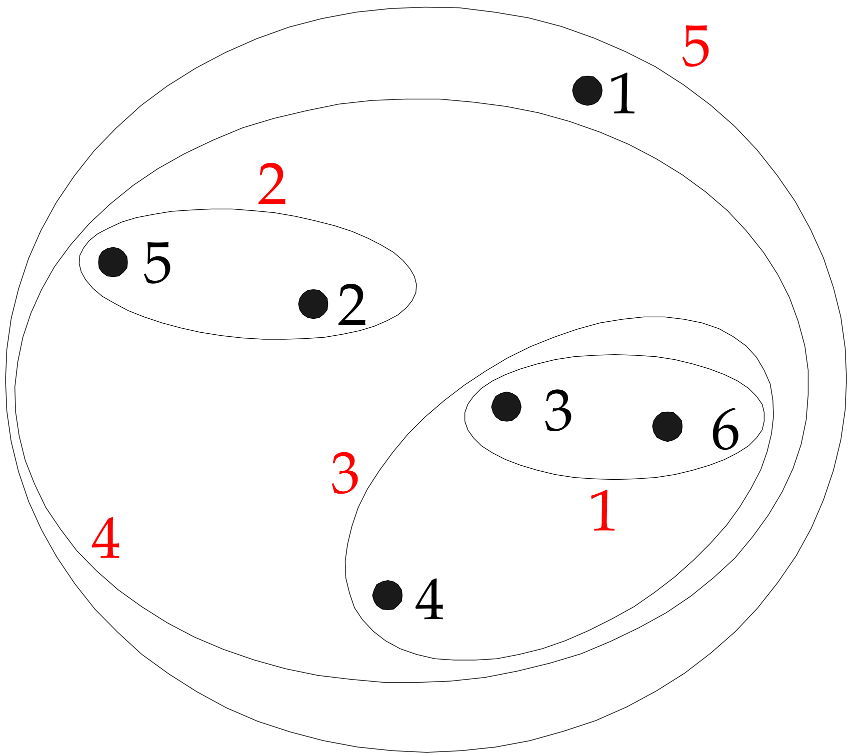
Cluster Similarity: Group Average

- Distance of two clusters is the average of pairwise distance between points in the two clusters.

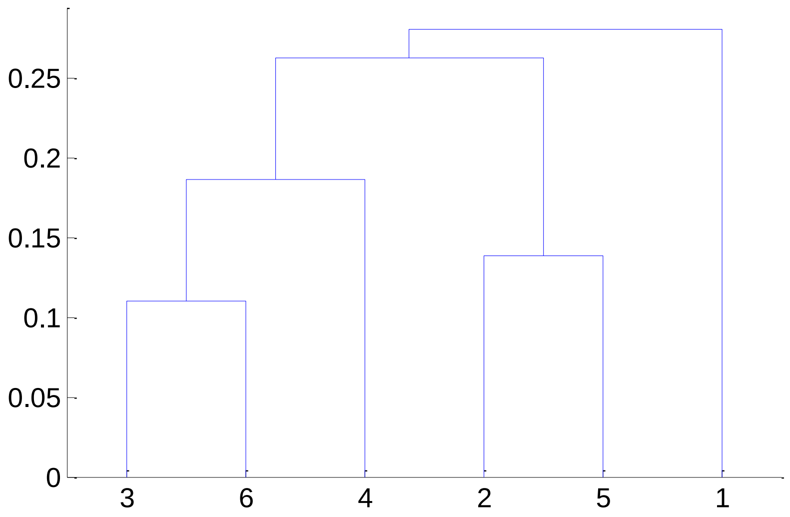
$$\text{distance}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{distance}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

- Compromise between Single and Complete Link.
- Need to use average connectivity for scalability since total connectivity favors large clusters.
- Less susceptible to noise and outliers.
- Biased towards globular clusters.

Hierarchical Clustering: Group Average



Nested Clusters

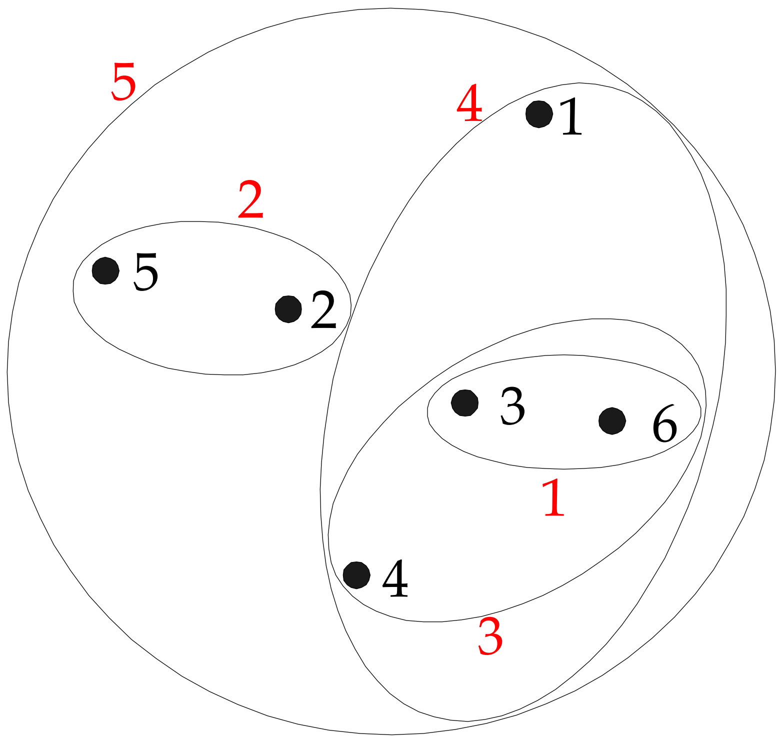


Dendrogram

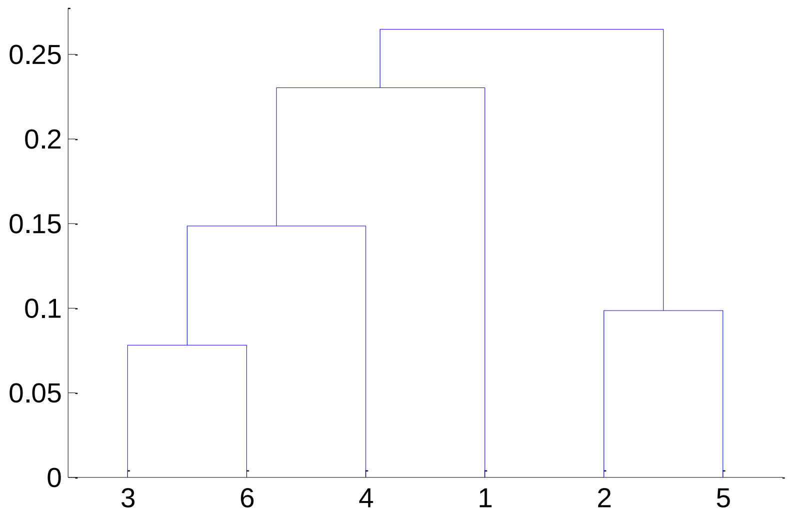
Cluster Similarity: Ward's Method

- Similarity of two clusters is based on the increase in squared error when two clusters are merged.
  - Similar to group average if distance between points is distance squared.
- Less susceptible to noise and outliers.
- Biased towards globular clusters.
- Hierarchical analogue of K-means
  - But Ward's method does not correspond to a local minimum
  - Can be used to initialize K-means

Hierarchical Clustering: Ward's method

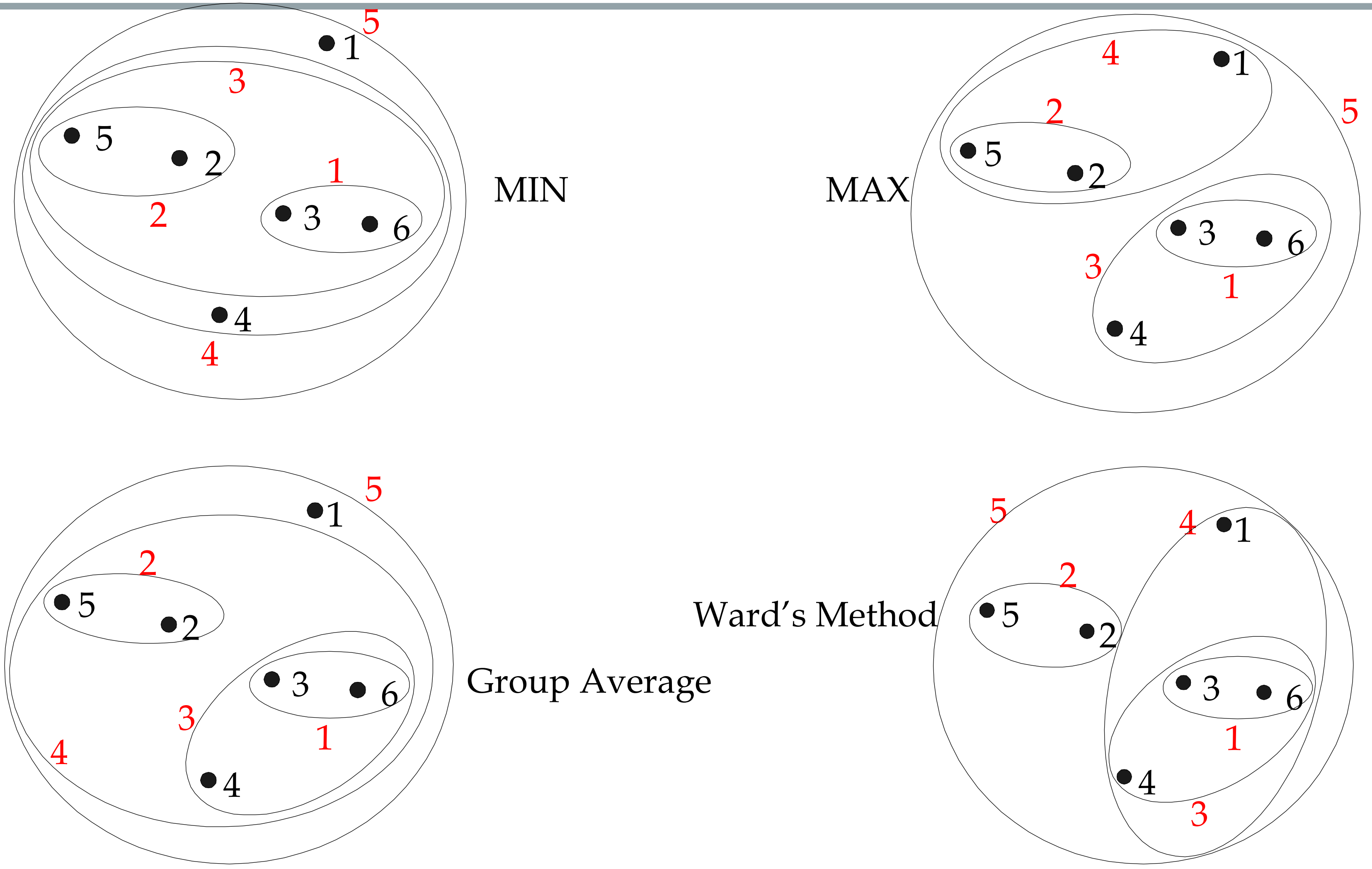


Nested Clusters



Dendrogram

Hierarchical Clustering: Comparison



Hierarchical Clustering: Time and Space requirements

- $O(N^2)$ space since it uses the proximity matrix.
  - N is the number of points.
- $O(N^3)$ time in many cases.
  - There are N steps and at each step the proximity matrix (size N^2) must be updated and searched.
  - By being careful, the complexity can be reduced to $O(N^2 \log N)$ time for some approaches.

Hierarchical Clustering: Problems and Limitations

- Once a decision is made to combine two clusters, it cannot be undone.
- No objective function is directly minimized.
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise and outliers.
  - Difficulty handling different sized clusters and convex shapes.
  - Breaking large clusters.