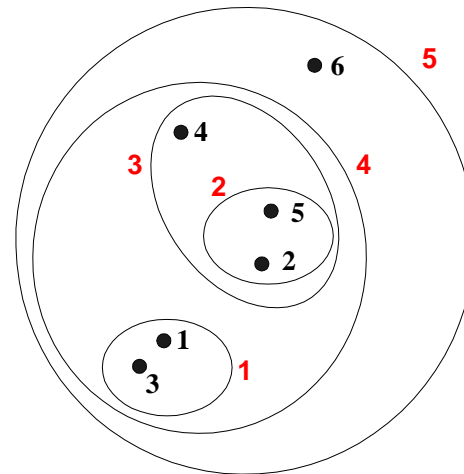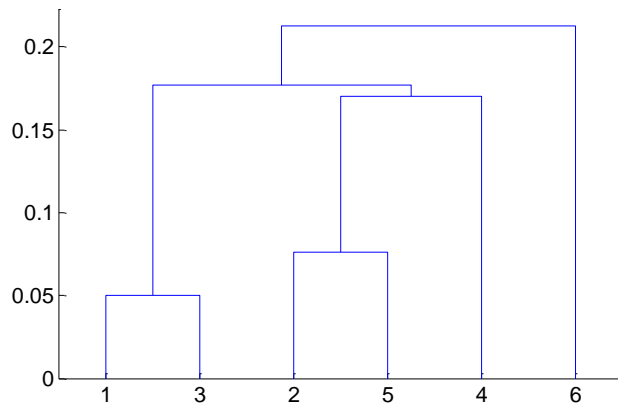# Data Mining
# Cluster Analysis: Basic Concepts and Algorithms

Introduction to Data Mining, 2$^{nd}$ Edition

by

Tan, Steinbach, Karpatne, Kumar

# Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree

- Can be visualized as a dendrogram
  - A tree like diagram that records the sequences of merges or splits

# Strengths of Hierarchical Clustering

- Do not have to assume any particular number of clusters
  - Any desired number of clusters can be obtained by 'cutting' the dendrogram at the proper level

- They may correspond to meaningful taxonomies
  - Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, …)

# Hierarchical Clustering

- Two main types of hierarchical clustering
  - Agglomerative:
    - Start with the points as individual clusters
    - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left

  - Divisive:
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains an individual point (or there are k clusters)

- Traditional hierarchical algorithms use a similarity or distance matrix
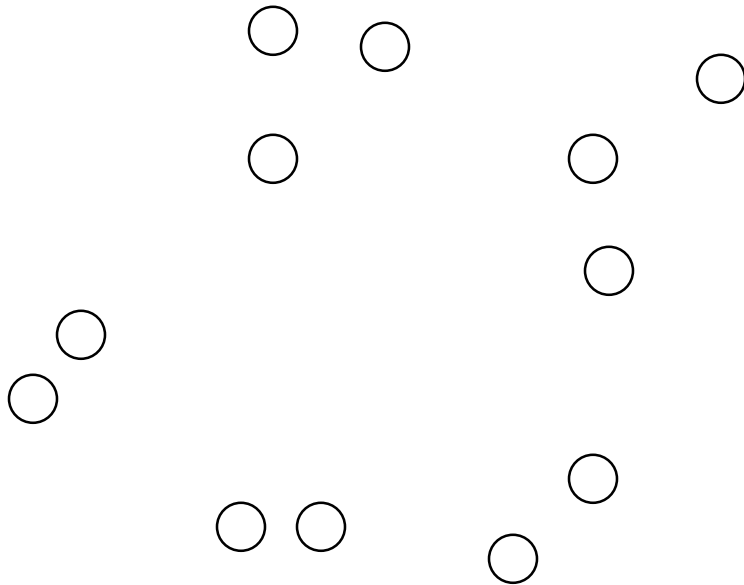  - Merge or split one cluster at a time

# Agglomerative Clustering Algorithm

- Most popular hierarchical clustering technique

- Basic algorithm is straightforward
  1. Compute the proximity matrix
  2. Let each data point be a cluster
  3. **Repeat**
  4.       Merge the two closest clusters
  5.       Update the proximity matrix
  6. **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 distinguish the different algorithms

# Starting Situation

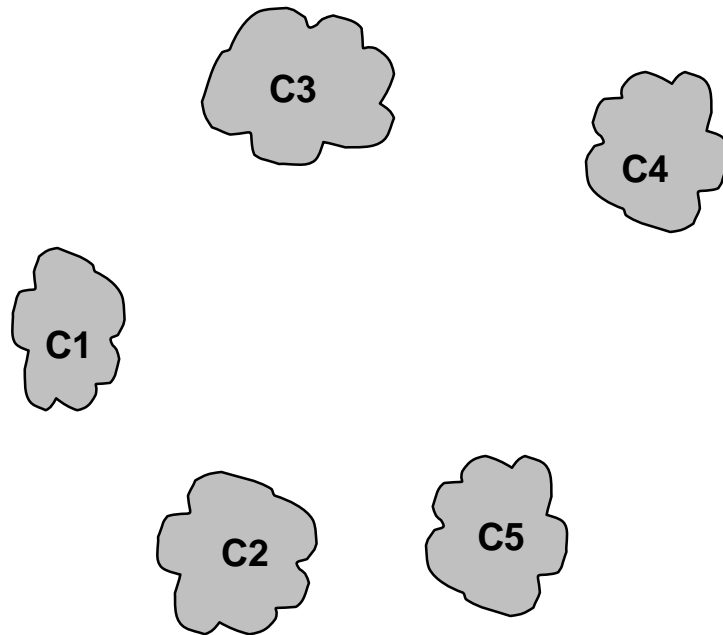- Start with clusters of individual points and a proximity matrix



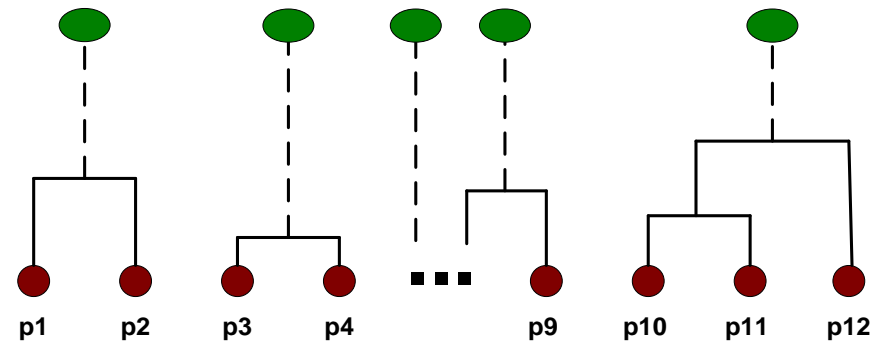| | p1 | p2 | p3 | p4 | p5 | . . . |
|------|----|----|----|----|----|-------|
| p1 | | | | | | |
| p2 | | | | | | |
| p3 | | | | | | |
| p4 | | | | | | |
| p5 | | | | | | |
| . | | | | | | |
| . | | | | | | |
| . | | | | | | |

**Proximity Matrix**

p1   p2   p3   p4   **. . .**   p9   p10   p11   p12

# Intermediate Situation

- After some merging steps, we have some clusters



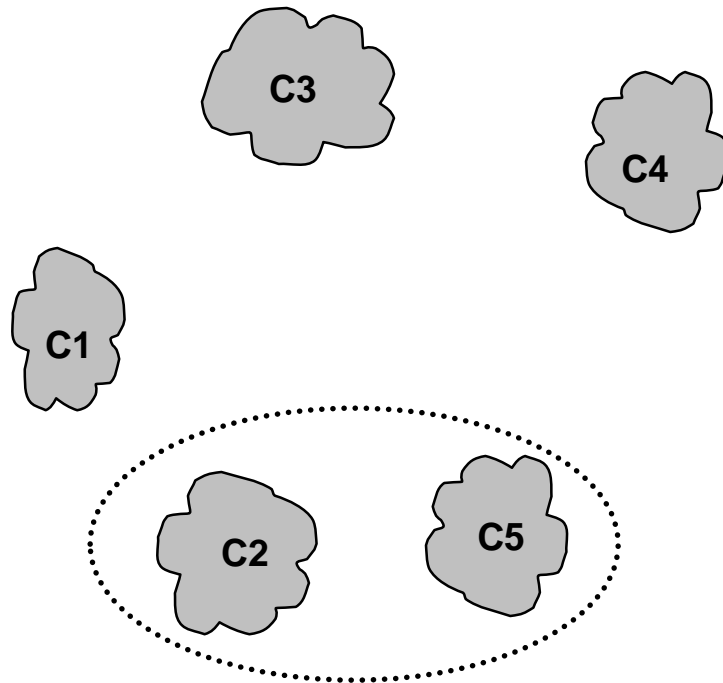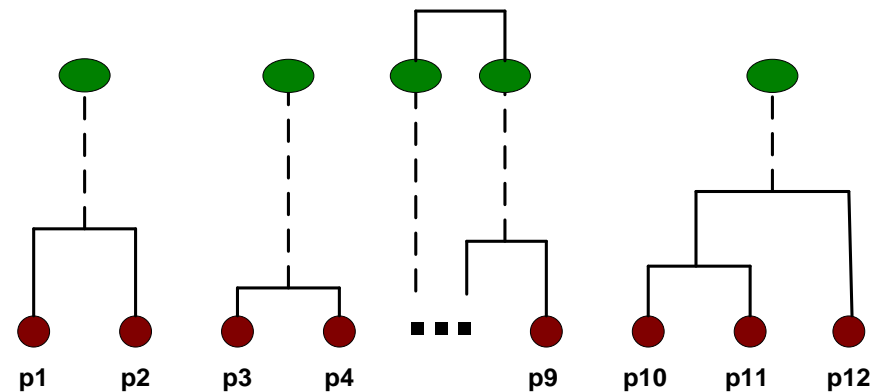**Proximity Matrix**

# Intermediate Situation

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

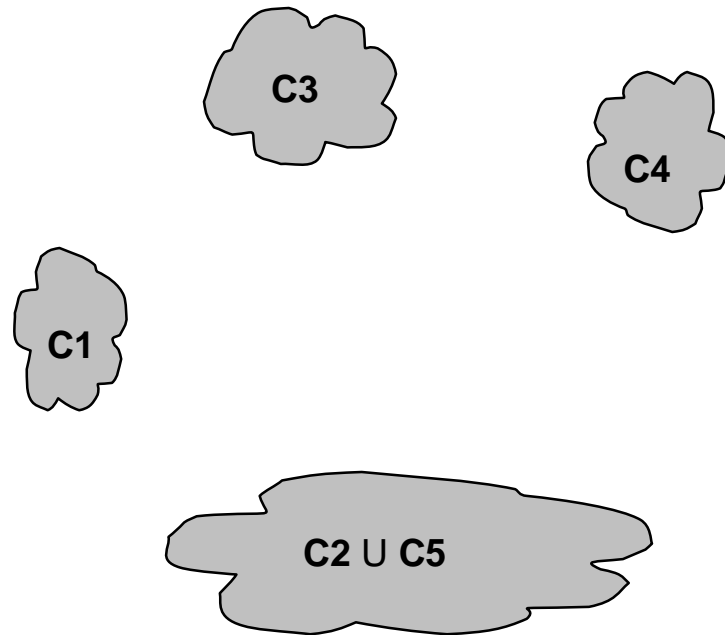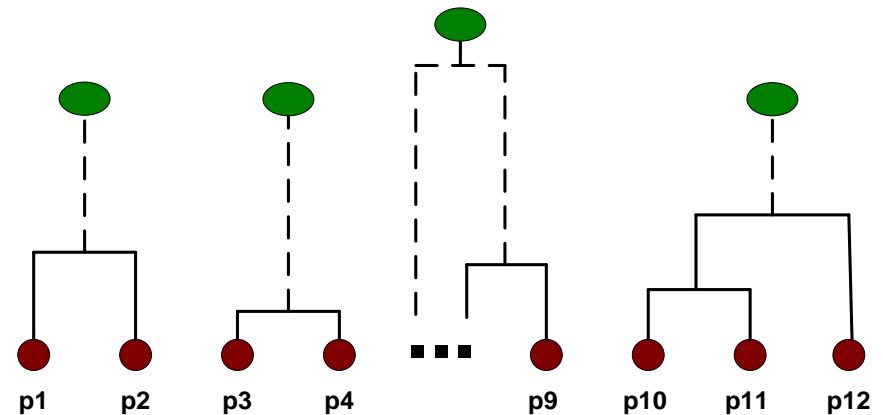|     | C1 | C2 | C3 | C4 | C5 |
|-----|----|----|----|----|----|
| C1  |    |    |    |    |    |
| C2  |    |    |    |    |    |
| C3  |    |    |    |    |    |
| C4  |    |    |    |    |    |
| C5  |    |    |    |    |    |

**Proximity Matrix**

# After Merging

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

|         | C1 | C2 ∪ C5 | C3 | C4 |
|---------|----|---------|----|----|
| **C1**  |    | ?       |    |    |
| **C2 ∪ C5** | ? | ?    | ?  | ?  |
| **C3**  |    | ?       |    |    |
| **C4**  |    | ?       |    |    |

**Proximity Matrix**

**C3**

**C4**

**C1**

**C2 ∪ C5**

p1  p2    p3  p4          p9    p10  p11  p12

# How to Define Inter-Cluster Distance



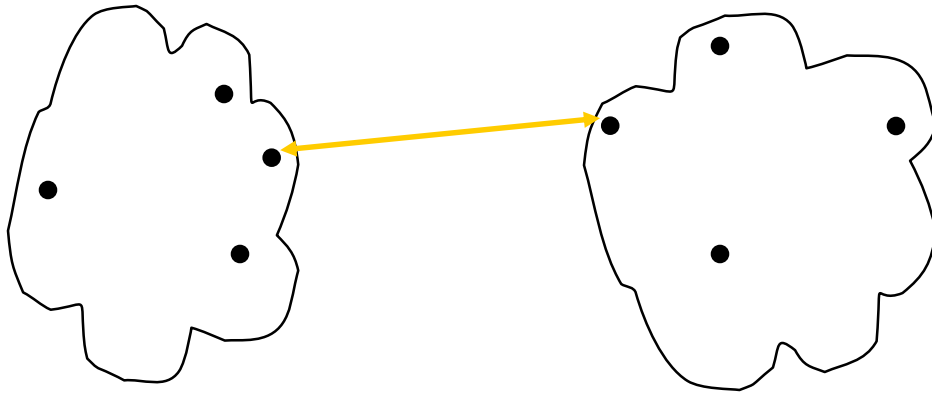|    | p1 | p2 | p3 | p4 | p5 | . . . |
|----|----|----|----|----|----|-------|
| p1 |    |    |    |    |    |       |
| p2 |    |    |    |    |    |       |
| p3 |    |    |    |    |    |       |
| p4 |    |    |    |    |    |       |
| p5 |    |    |    |    |    |       |
| .  |    |    |    |    |    |       |
| .  |    |    |    |    |    |       |
| .  |    |    |    |    |    |       |

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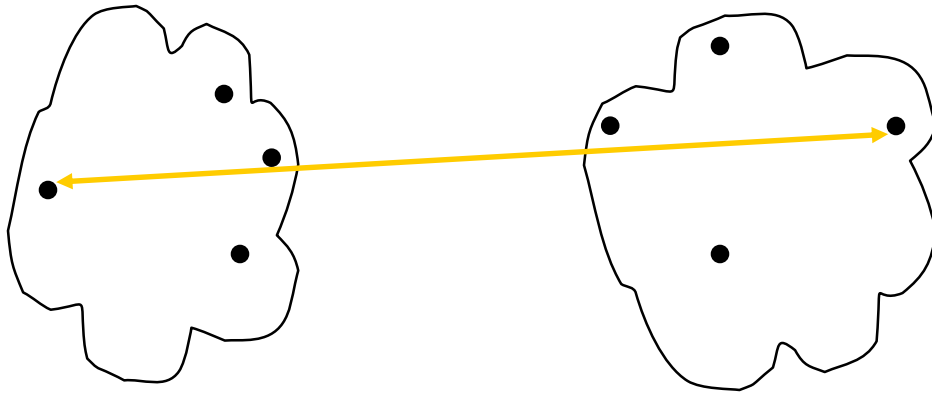
- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

# How to Define Inter-Cluster Similarity



|     | p1 | p2 | p3 | p4 | p5 | . . . |
|-----|----|----|----|----|----|-------|
| **p1** |    |    |    |    |    |       |
| **p2** |    |    |    |    |    |       |
| **p3** |    |    |    |    |    |       |
| **p4** |    |    |    |    |    |       |
| **p5** |    |    |    |    |    |       |
| **.**  |    |    |    |    |    |       |
| **.**  |    |    |    |    |    |       |
| **.**  |    |    |    |    |    |       |

**Proximity Matrix**

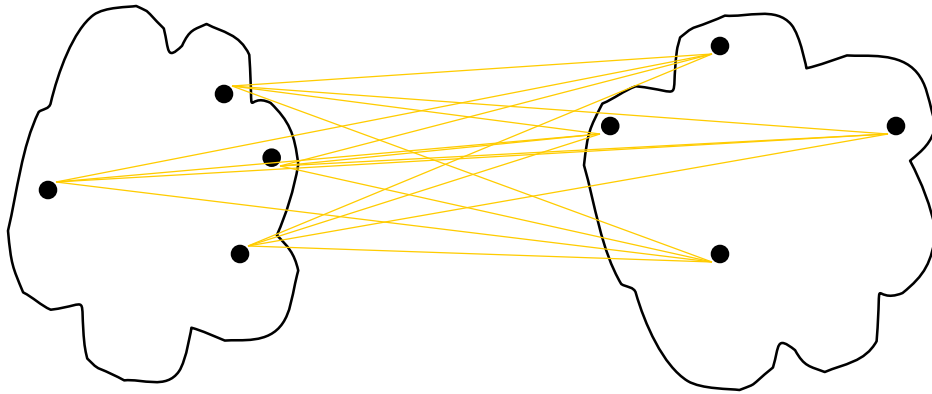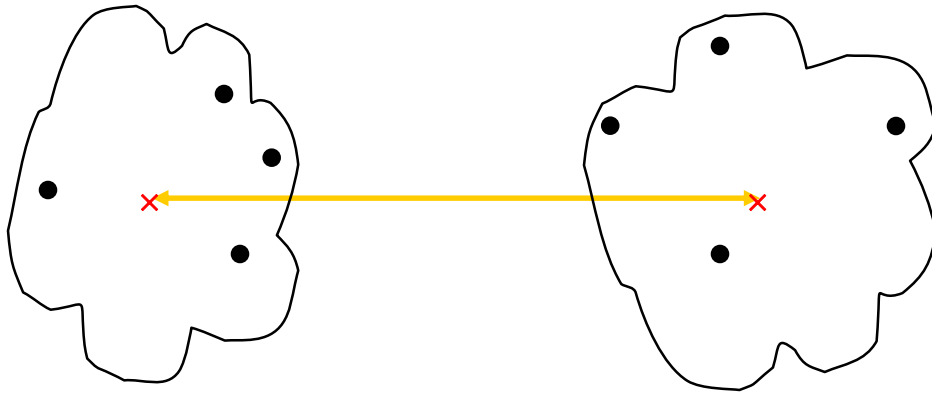- <span style="color:red">MIN</span>
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

# 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

# MIN or Single Link

- Proximity 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
- Example:
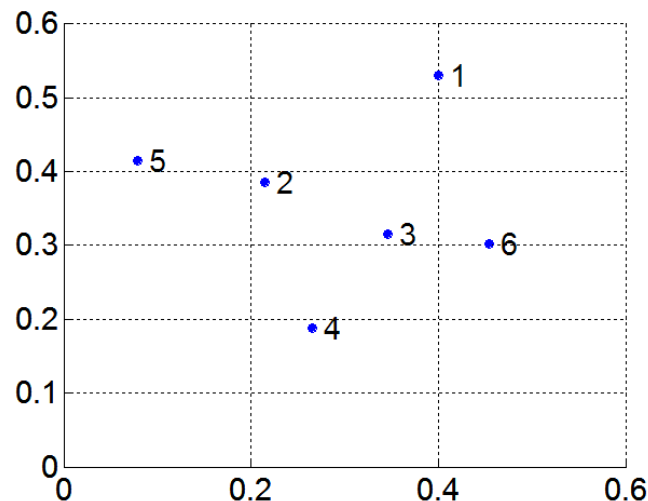


**Distance Matrix:**

|    | p1   | p2   | p3   | p4   | p5   | p6   |
|----|------|------|------|------|------|------|
| p1 | 0.00 | 0.24 | 0.22 | 0.37 | 0.34 | 0.23 |
| p2 | 0.24 | 0.00 | 0.15 | 0.20 | 0.14 | 0.25 |
| p3 | 0.22 | 0.15 | 0.00 | 0.15 | 0.28 | 0.11 |
| p4 | 0.37 | 0.20 | 0.15 | 0.00 | 0.29 | 0.22 |
| p5 | 0.34 | 0.14 | 0.28 | 0.29 | 0.00 | 0.39 |
| p6 | 0.23 | 0.25 | 0.11 | 0.22 | 0.39 | 0.00 |

# Hierarchical Clustering: MIN
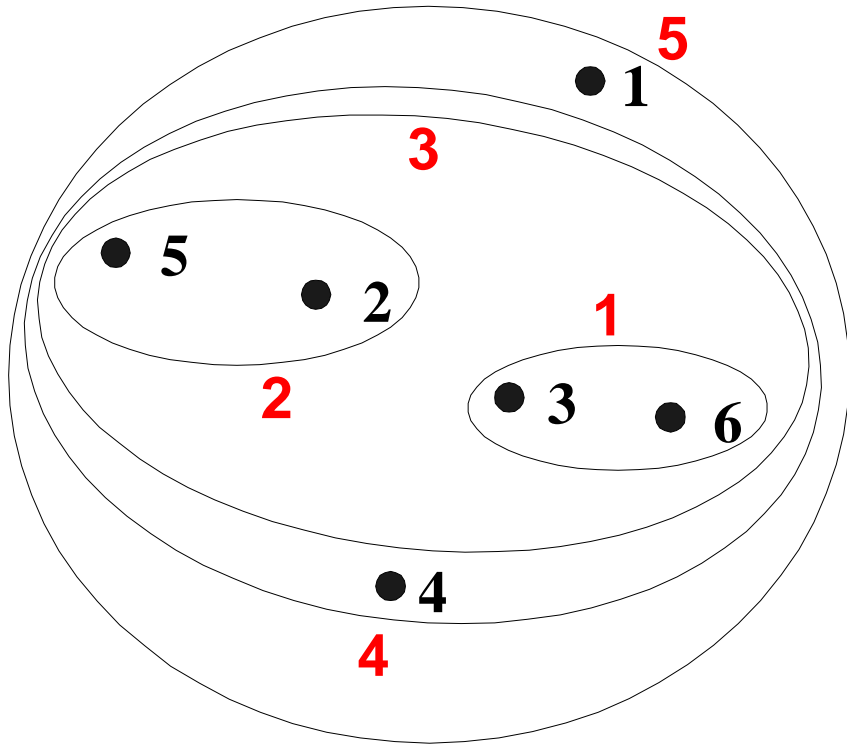


**Nested Clusters**
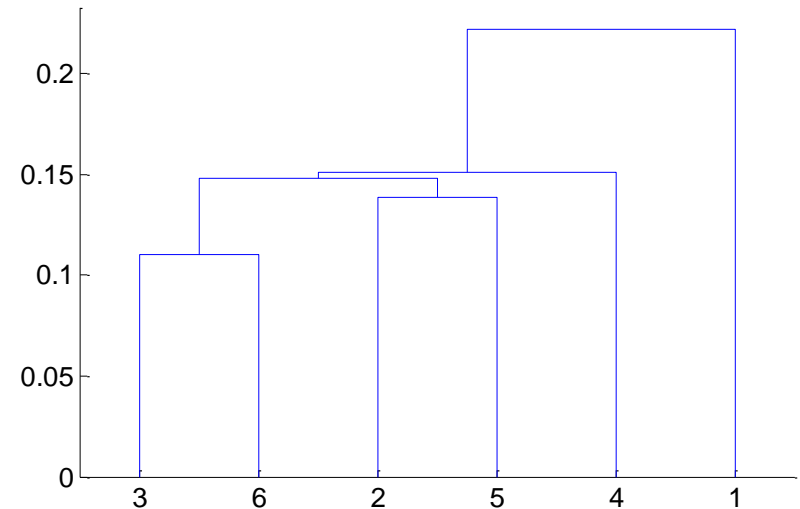
**Dendrogram**

# Strength of MIN
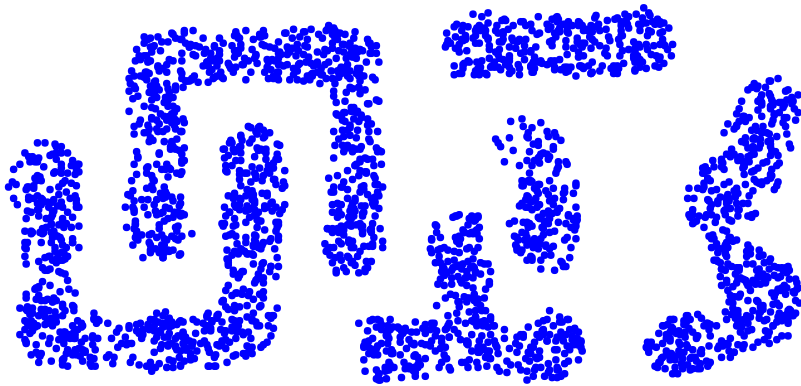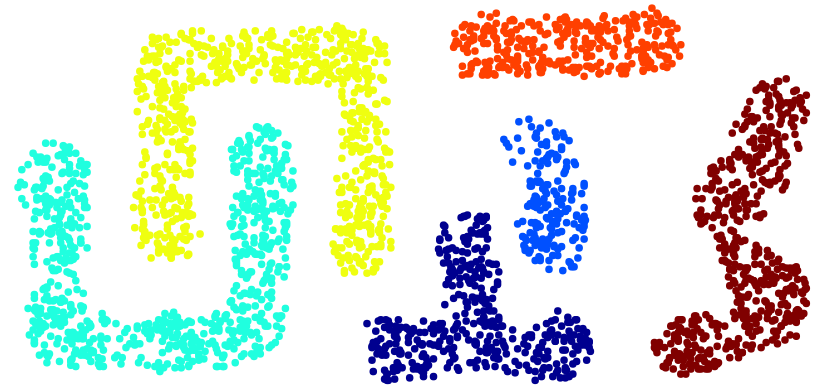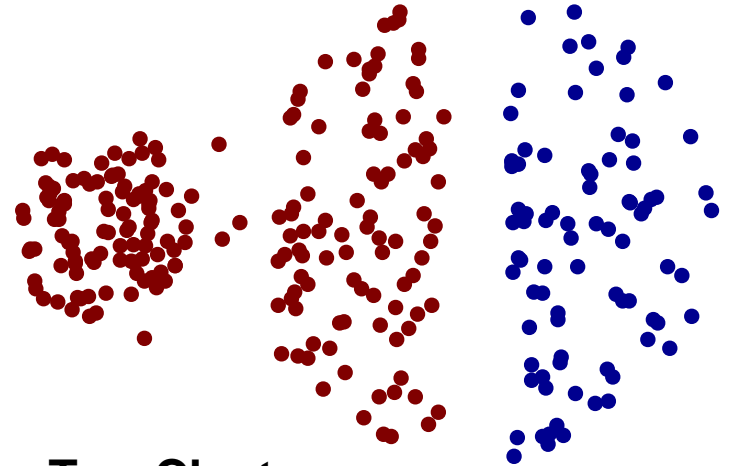


Original Points          Six Clusters

- **Can handle non-elliptical shapes**
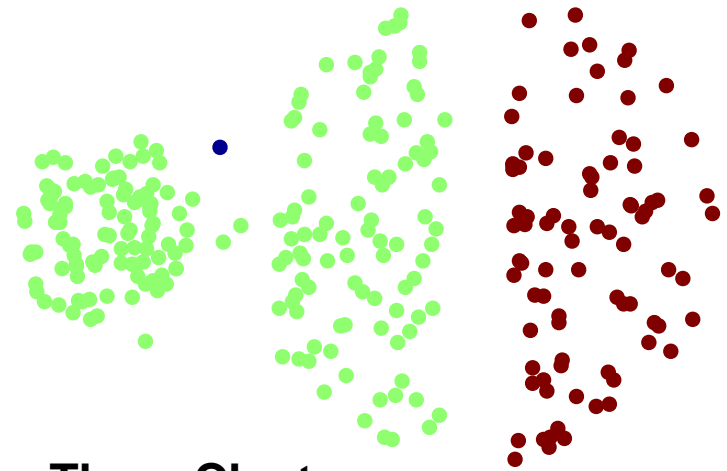
# Limitations of MIN
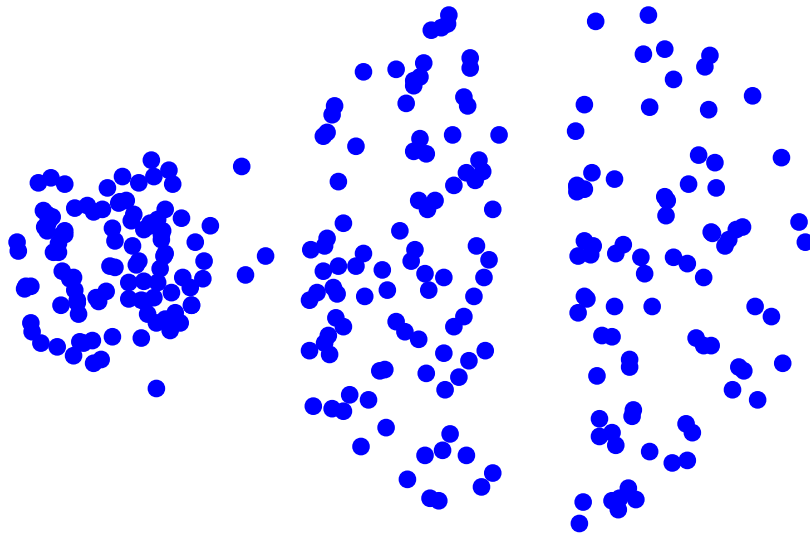


**Original Points**
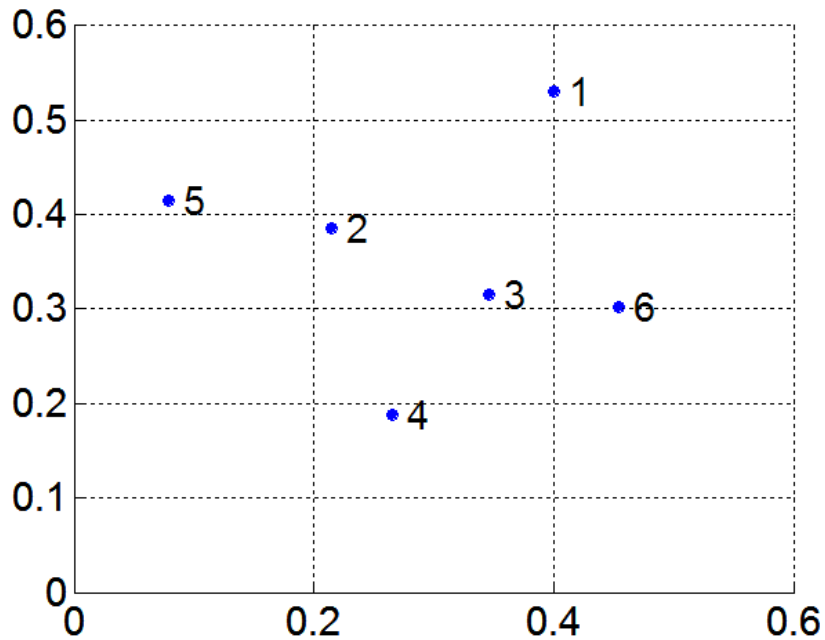
- **Sensitive to noise and outliers**

**Two Clusters**

**Three Clusters**

# MAX or Complete Linkage

- Proximity of two clusters is based on the two most distant points in the different clusters
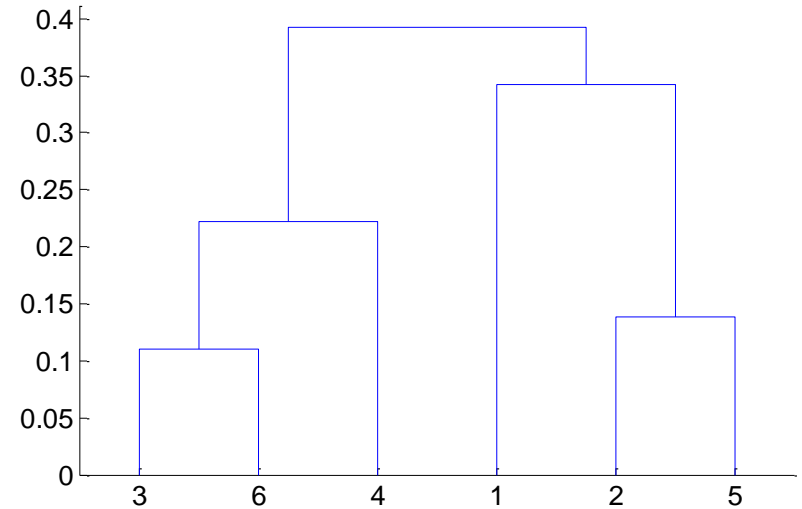  - Determined by all pairs of points in the two clusters

**Distance Matrix:**

| | p1 | p2 | p3 | p4 | p5 | p6 |
|---|---|---|---|---|---|---|
| p1 | 0.00 | 0.24 | 0.22 | 0.37 | 0.34 | 0.23 |
| p2 | 0.24 | 0.00 | 0.15 | 0.20 | 0.14 | 0.25 |
| p3 | 0.22 | 0.15 | 0.00 | 0.15 | 0.28 | 0.11 |
| p4 | 0.37 | 0.20 | 0.15 | 0.00 | 0.29 | 0.22 |
| p5 | 0.34 | 0.14 | 0.28 | 0.29 | 0.00 | 0.39 |
| p6 | 0.23 | 0.25 | 0.11 | 0.22 | 0.39 | 0.00 |

# Hierarchical Clustering: MAX
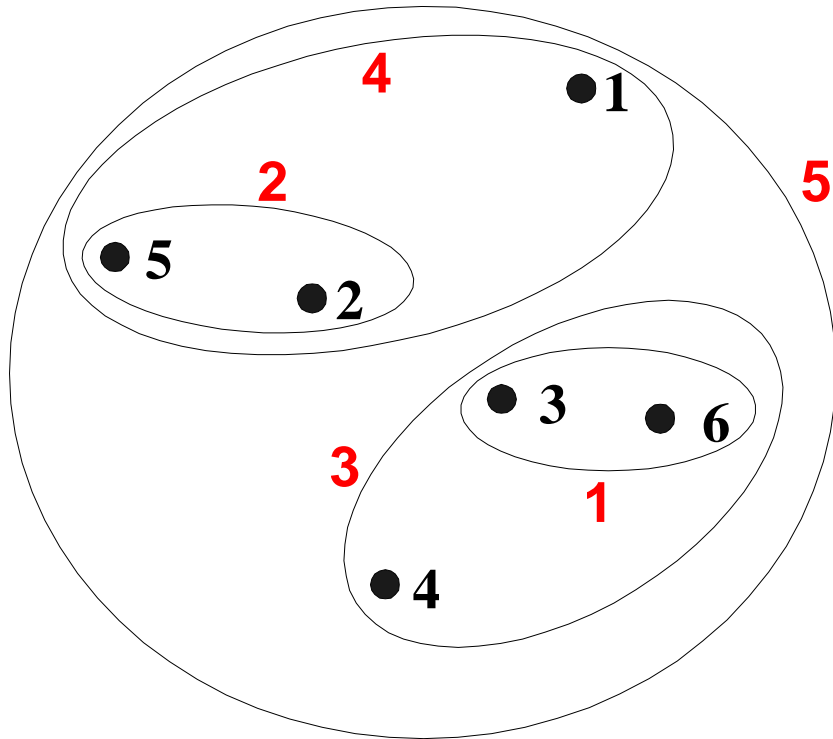


**Nested Clusters**

**Dendrogram**

# Strength of MAX



**Original Points**               **Two Clusters**
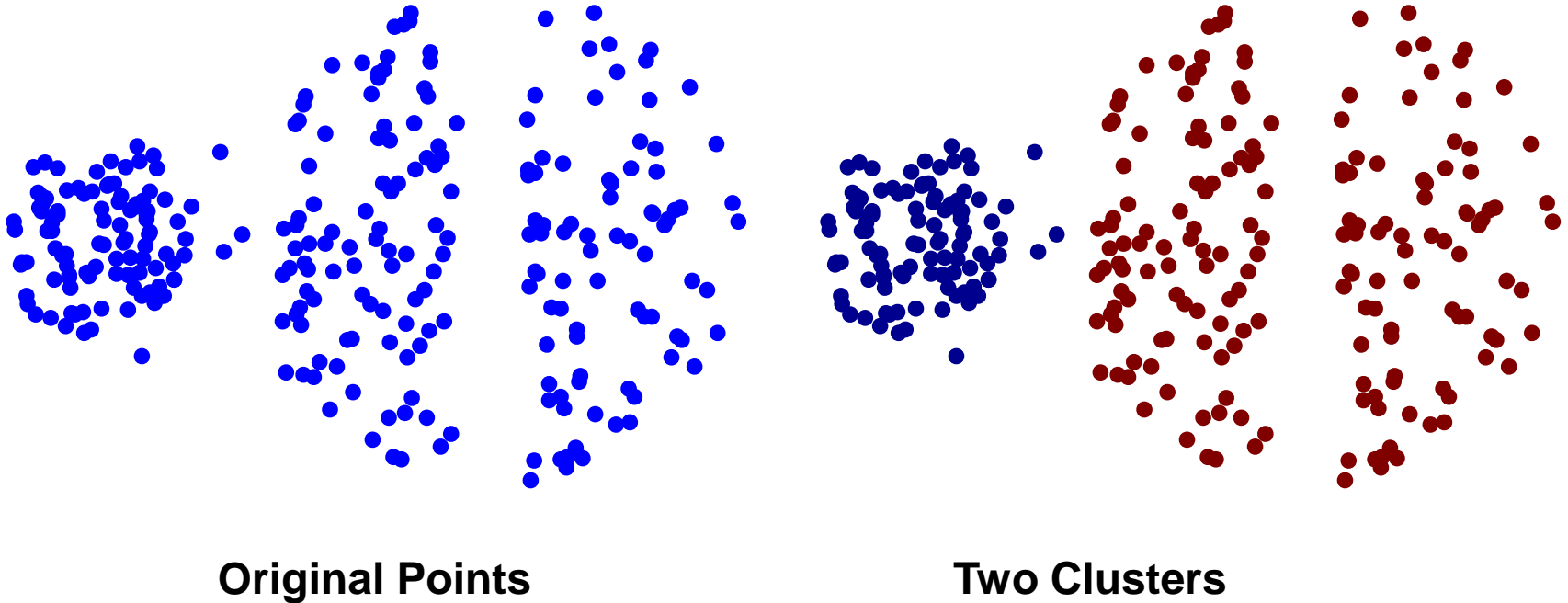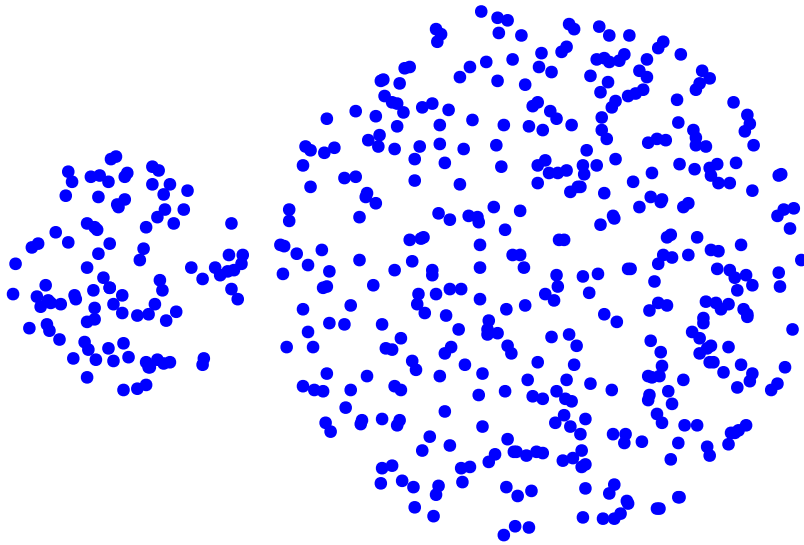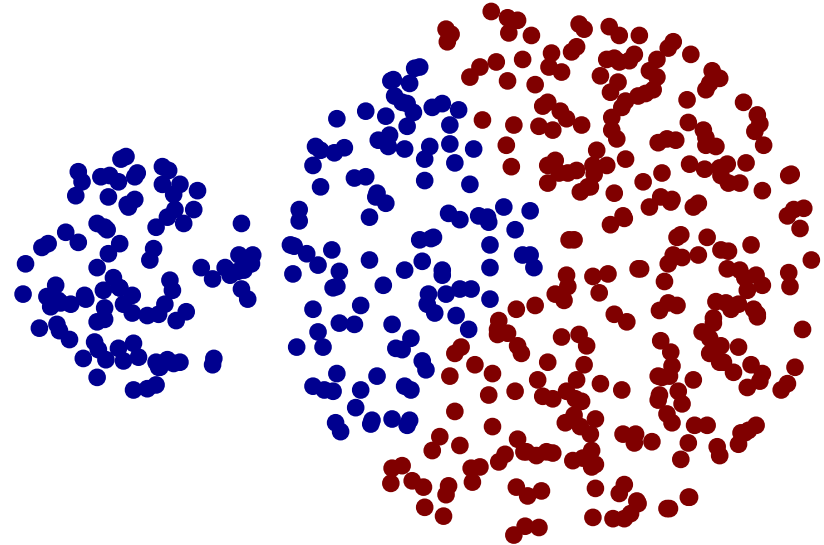
- **Less susceptible to noise and outliers**

# Limitations of MAX



**Original Points**                    **Two Clusters**

- **Tends to break large clusters**

- **Biased towards globular clusters**

# Group Average

- Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

$$proximity(Cluster_i, Cluster_j) = \frac{\sum\limits_{\substack{p_i \in Cluster_i \\ p_j \in Cluster_j}} proximity(p_i, p_j)}{|Cluster_i| \times |Cluster_j|}$$
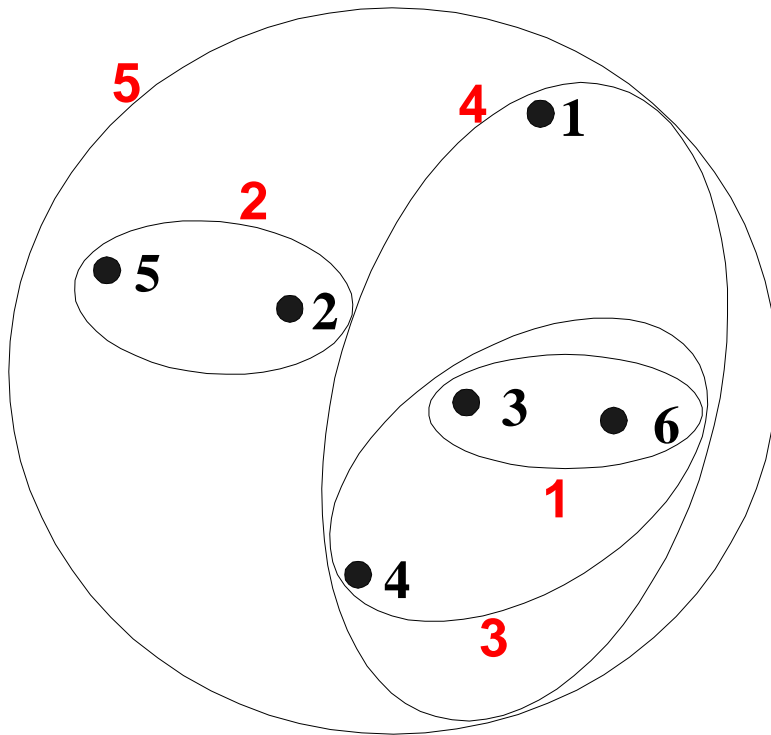
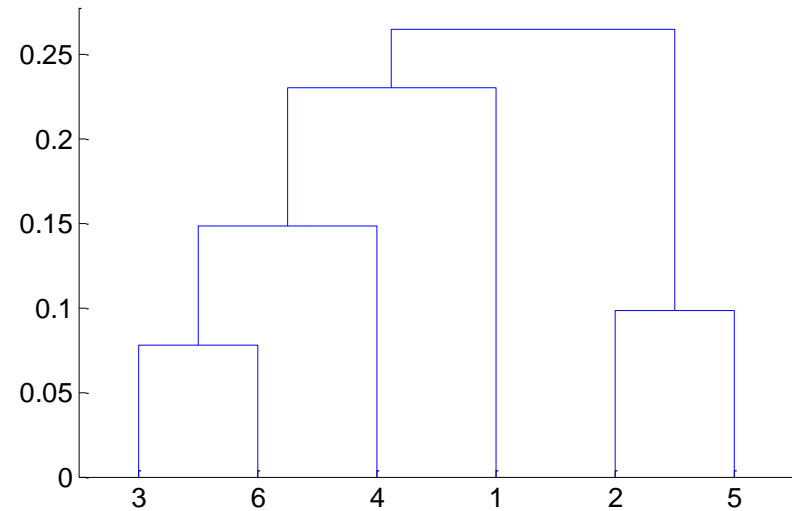- Need to use average connectivity for scalability since total proximity favors large clusters
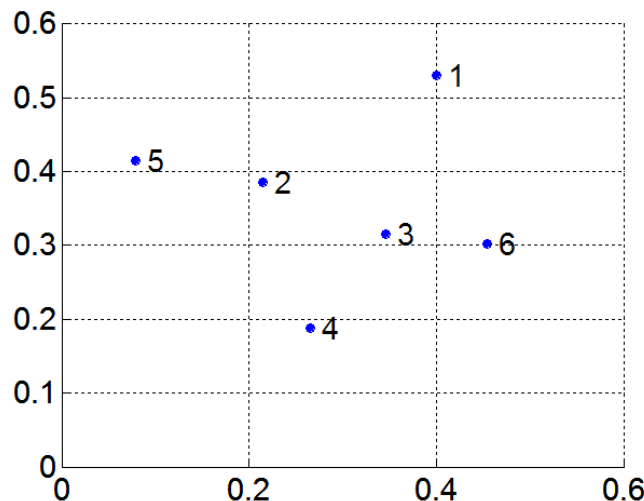
**Distance Matrix:**

|     | p1   | p2   | p3   | p4   | p5   | p6   |
|-----|------|------|------|------|------|------|
| p1  | 0.00 | 0.24 | 0.22 | 0.37 | 0.34 | 0.23 |
| p2  | 0.24 | 0.00 | 0.15 | 0.20 | 0.14 | 0.25 |
| p3  | 0.22 | 0.15 | 0.00 | 0.15 | 0.28 | 0.11 |
| p4  | 0.37 | 0.20 | 0.15 | 0.00 | 0.29 | 0.22 |
| p5  | 0.34 | 0.14 | 0.28 | 0.29 | 0.00 | 0.39 |
| p6  | 0.23 | 0.25 | 0.11 | 0.22 | 0.39 | 0.00 |

# Hierarchical Clustering: Group Average



**Nested Clusters**

**Dendrogram**

# Hierarchical Clustering: Group Average

- Compromise between Single and Complete Link

- Strengths
  - Less susceptible to noise and outliers

- Limitations
  - Biased towards globular clusters