Graph Clustering

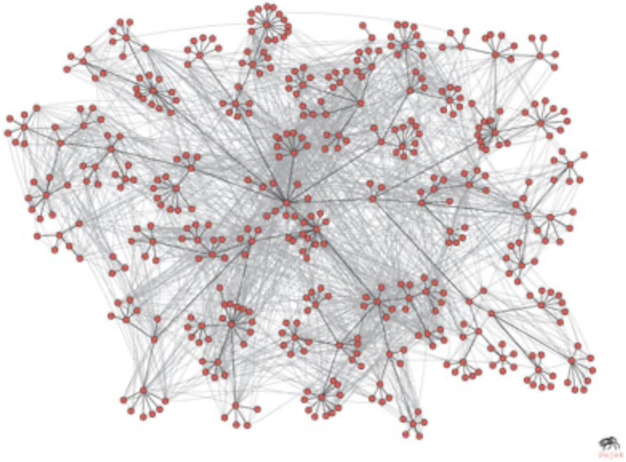
Dawei Wang

Vydhourie Thiyageswaran (Mentor)

What are networks?

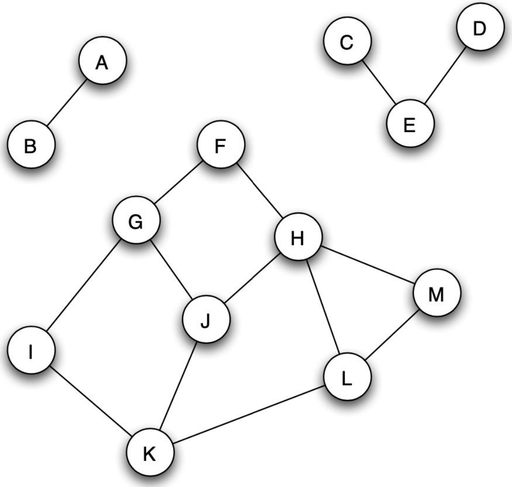
- Ideas:
  - Evaluate your actions not in isolation.
  - Cause-effect relationships can become quite subtle.
  - The dynamics of aggregate behavior.

- Related:
  - Graph theory
  - Game Theory



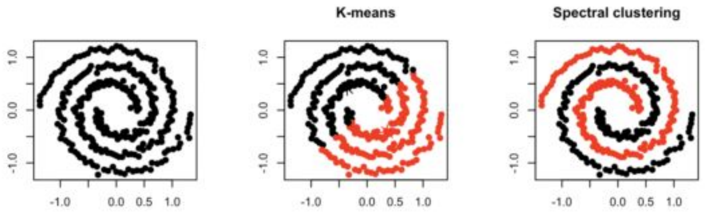
Graph

- Path and Connectivity
    - A graph is **connected** if for every pair of nodes, there is a **path** between them.
- Connected component
    - Every node in the subset has a path to every other;
    - The subset is not part of some larger set.



A graph with three connected components.

Clustering



We seek to partition observations into distinct groups so that the observations within each group are similar, while observations in different groups are different.

- **K-means clustering**:

  Partitioning a data set into K distinct, non-overlapping clusters. Each observation belongs to the cluster with the nearest mean.

- **Spectral clustering**:

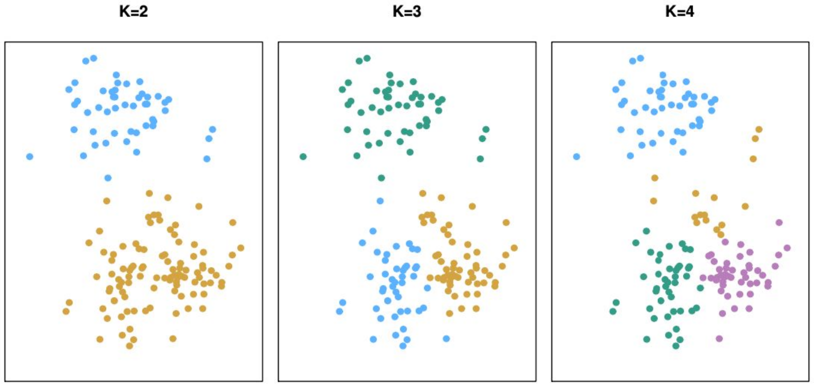
  Make use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions.

K-means clustering

1. Each observation belongs to at least one of the K clusters.
2. No observation belongs to more than one cluster.
3. **Make the within-cluster variation as small as possible.**

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}.$$

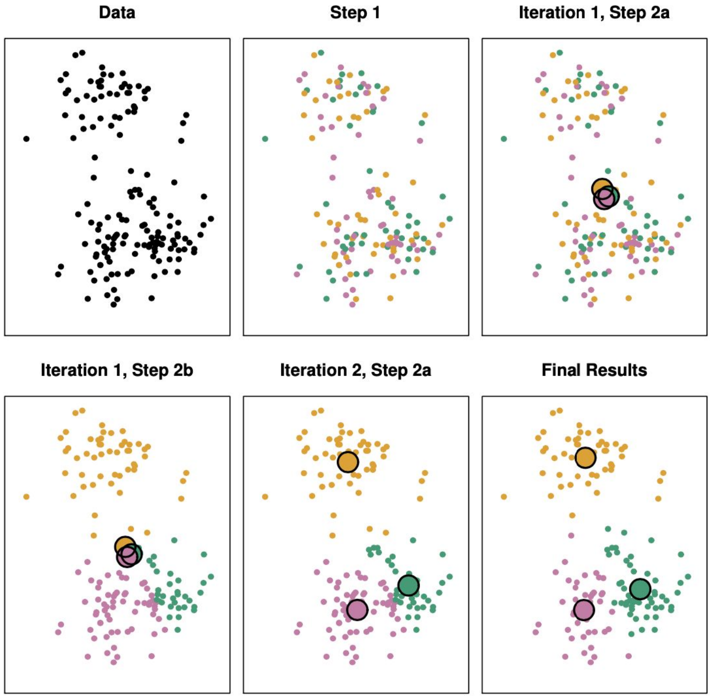
$|C_k|$ denotes the number of observations in the kth cluster.



K-means clustering

- Step 1: Each observation is randomly assigned to a cluster.
- Step 2(a): The cluster centroids are computed.
- Step 2(b): Each observation is assigned to the nearest centroid.
- Step 2(a) is once again performed, leading to new cluster centroids.
- Final results: the results obtained after ten iterations.

*Challenges: Specify different initial points will end with different clusters, not stable.



Spectral Clustering

L = A^T A = D - B

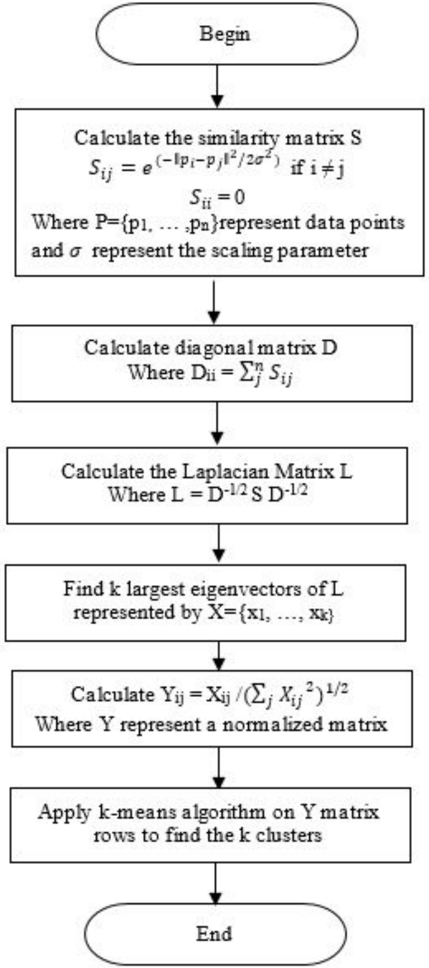
L - Symmetric positive semidefinite matrix

A - Incidence matrix

D - Diagonal matrix

B - Adjacency matrix

- Spectral clustering finds the *m* eigenvectors **Z**$_{N\times m}$ corresponding to the m smallest eigenvalues of **L**. Using a standard method (K-means), we then cluster the rows of **Z** to yield a clustering of the original data points.



Begin

Calculate the similarity matrix S
$S_{ij} = e^{(-\|p_i - p_j\|^2/2\sigma^2)}$ if i ≠ j
$S_{ii} = 0$
Where P={p_1, … ,p_n}represent data points and σ represent the scaling parameter

Calculate diagonal matrix D
Where $D_{ii} = \sum_j^n S_{ij}$

Calculate the Laplacian Matrix L
Where L = $D^{-1/2}$ S $D^{-1/2}$

Find k largest eigenvectors of L
represented by X={x_1, …, x_k}

Calculate $Y_{ij} = X_{ij} / (\sum_j X_{ij}{}^2)^{1/2}$
Where Y represent a normalized matrix

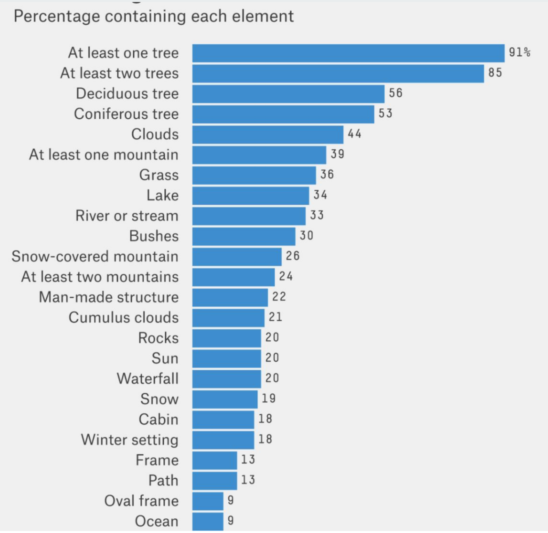
Apply k-means algorithm on Y matrix rows to find the k clusters

End

Case Study

The work of Bob Ross.

381 paintings.



Use R to run k-means clustering analysis to cluster similar paintings based on the contained elements.

Examples:

- A cluster of 50 paintings tagged “snow” and “winter”.
- A cluster of 28 paintings each with an oval white-space frame.
- A cluster of 35 paintings of ocean scenes.

Credits

Special thanks to Vydhourie for mentoring me this quarter!

https://www.cs.cornell.edu/home/kleinber/networks-book/

https://www.statlearning.com/

https://web.stanford.edu/~hastie/Papers/ESLII.pdf

https://fivethirtyeight.com/features/a-statistical-analysis-of-the-work-of-bob-ross/