Spectral Clustering: Practical Considerations in Implementation

Ira A. Fulton Schools of Engineering
Arizona State University

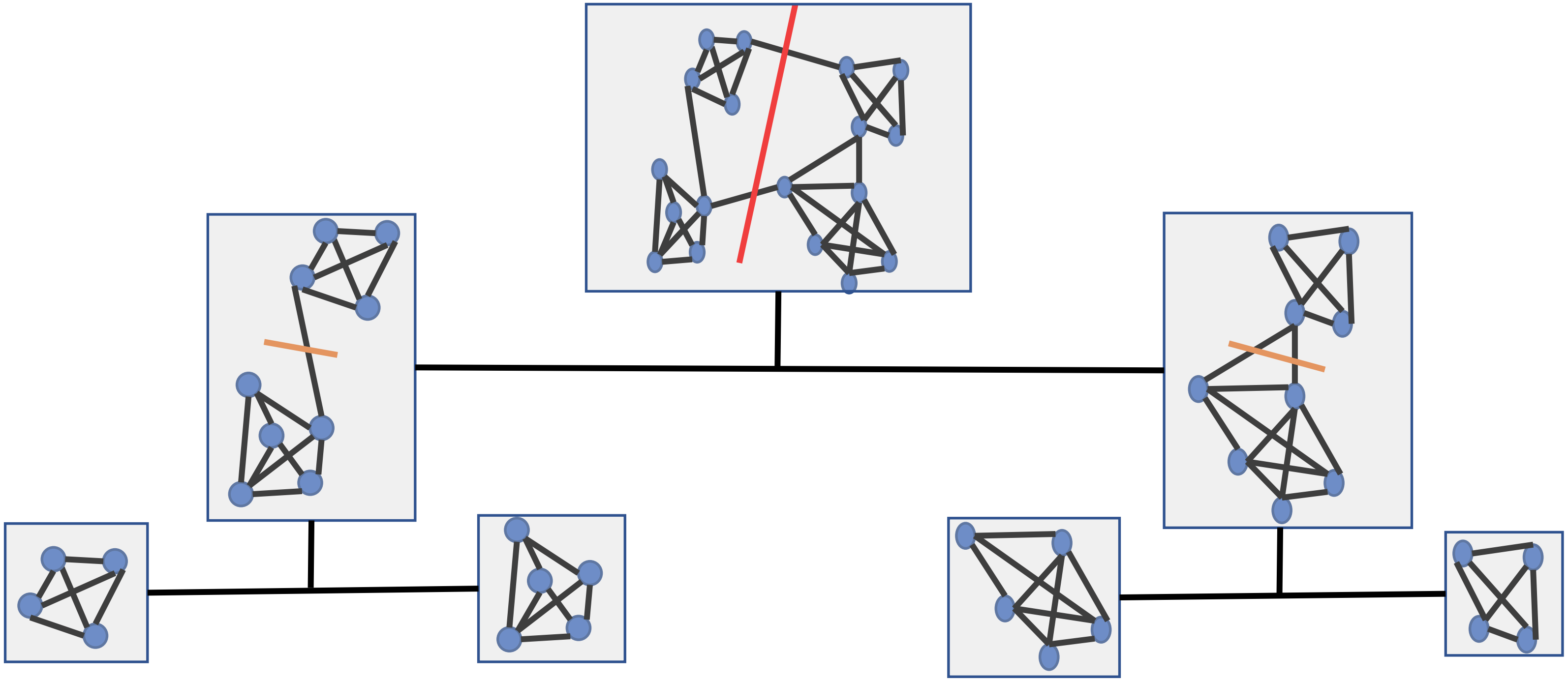
Objective

Objective

Discuss several practical implementation issues

Recursive bi-partitioning

- Recursively apply bi-partitioning algorithm in a hierarchical divisive manner.
- Disadvantages: inefficient, stability issues.



K-way graph cuts

Generalizing the 2-way objective functions :

$$J_{RatioCut}(A_1, \dots, A_k) = \sum_{i=1}^{k} \frac{Cut(A_i, \overline{A_i})}{|A_i|}$$

$$J_{NCut}(A_1, \dots, A_k) = \sum_{i=1}^{k} \frac{Cut(A_i, \overline{A_i})}{Vol(A_i)}$$

$$J_{MinMaxCut}(A_1, \dots, A_k) = \sum_{i=1}^{k} \frac{Cut(A_i, \overline{A_i})}{Cut(A_i, A_i)}$$

Implementation Considerations (1/4)

- Preprocessing: spectral clustering methods can be interpreted as tools for analysis of the block structure of the similarity matrix.
  - ➔ Building such matrices may certainly ameliorate the results.

➔ When building graphs from real data

- Calculation of the similarity matrix is not evident.
- Choosing the similarity function can highly affect the results of the following steps.
- A Gaussian kernel is often chosen, but other similarities like cosine similarity might be proper for specific applications.

Implementation Considerations (2/4)

- Graph and similarity matrix construction: Laplacian matrices are generally chosen to be positive and semi-definite thus their eigenvalues will be non-negatives.
  - A few variants

Unnormalized	$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$
symmetric	$\boldsymbol{L}_{Sy} = \boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$
Asymmetric	$\boldsymbol{L}_{As} = \boldsymbol{D}^{-1}\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{W}$

Implementation Considerations (3/4)

Computing the eigenvectors.

- Efficient methods exist for sparse matrices.

Different ways of building the similarity graphs

- ε-neighborhood graph.
- k-nearest neighbor graph.
- fully connected graph.

Implementation Considerations (4/4)

- Choosing *k*:
  - Similar to k-means, there are many heuristics to use.
  - The eigengap heuristic: to choose a k such that first k eigenvalues are very small but the (k+1)th one is relatively large.

- Clustering: simple algorithms other than k-means can be used in the last stage, such as simple linkage, k-lines, elongated k-means, mixture model, etc.

Recap: Pros and Cons of Spectral Clustering

Advantages:

- Does not make strong assumptions on the forms of the clusters.
- Easy to implement, and can be implemented efficiently even for large data sets as long as the similarity graph is sparse.
- Good clustering results.
- Reasonably fast for sparse data sets of several thousand elements.

Disadvantages:

- May be sensitive to choice of parameters for neighborhood graph.
- Computationally expensive for large datasets.