Project on Data Analysis and Mining - part II
Analysis and Optimization of Spectral Clustering for Community Detection

April 8, 2025

1 Introduction

Community detection plays a fundamental role in network science, with applications in social networks, bioinformatics, political science, and recommendation systems. Spectral clustering, based on graph Laplacians and eigenvector transformations, has gained prominence due to its ability to detect non-convex structures that traditional clustering methods struggle with.

The goal of this project is to experimentally explore and evaluate the Spectral Clustering algorithm by Ng, Jordan, and Weiss [1] on Community Detection. The main objective is to analyze how different hyper-parameters affect clustering performance.

The following bechmark datasets for community detection will be used:

- ***Pol.Books*** [2] A network of books about US politics published around the time of the 2004 presidential election and sold by the online bookseller Amazon.com. Edges between books represent frequent co-purchasing of books by the same buyers. The network consists of $n = 105$ nodes and $m = 441$ edges (i.e. co-purchase links). Originally, books are labeled as liberal, conservative, or neutral.

- ***Football*** [3] Network of American football games between Division IA colleges during regular season Fall 2000. The network consists of $n = 115$ nodes and $m = 613$ edges. Ground-truth communities correspond to 12 conferences.

2 Research Questions

This study aims to address the performance of Spectral Clustering algorithm by Ng. et al. to detect communities in networks considering the impact of hyper-parameters:

- How does the choice of the number of clusters (k) affect the clustering results?
- How does the Gaussian kernel parameter (σ) influence community detection quality?
- What is the effect of normalized *vs* unnormalized Laplacian matrices on clustering performance?

To access the quality of found communities with different parameter settings the following evaluation metrics will used: (a) Normalized Mutual Information (NMI) [4]; and (b) Modularity Score [5, 6].

3 Methodology

3.1 Spectral Clustering Algorithm (Ng, Jordan, Weiss 2002)

The Spectral Clustering Algorithm can be summarized as follows:

Given a dataset of points:

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subseteq \mathbb{R}^d, \tag{1}$$

the normalized spectral clustering algorithm aims to partition these points into K distinct clusters by leveraging the eigenstructure of a similarity-based representation of the data. This method involves the following detailed steps:

1. **Affinity Matrix Construction:** Form the affinity matrix $W \in \mathbb{R}^{n\times n}$, where:

$$W_{ij} = \begin{cases} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right), & i \neq j \\ 0, & i = j \end{cases}, \tag{2}$$

   where σ is a scaling parameter selected by the user.

2. **Laplacian Matrix Formation:** Compute the degree matrix D, a diagonal matrix whose (i,i)-th element is the sum of the i-th row of W. Then form the normalized Laplacian matrix:

$$L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}. \tag{3}$$

3. **Eigenvector Extraction:** Determine the K largest eigenvectors of L. Construct the matrix $V \in \mathbb{R}^{n\times K}$ by stacking these eigenvectors as columns, thereby reducing dimensionality from $n \times n$ to $n \times K$.

4. **Normalization:** Construct the matrix U by normalizing each row $\mathbf{v}_i$ of V to have unit length, using:

$$u_{ij} = \frac{v_{ij}}{\left(\sum_{l=1}^{K} v_{il}^2\right)^{\frac{1}{2}}}. \tag{4}$$

5. **Clustering:** Treat each row $\mathbf{u}_i$ of matrix U as a point in $\mathbb{R}^K$. Apply a clustering algorithm (typically k-means) to cluster the points $\mathbf{u}_i$, $i = 1, 2, \ldots, n$, into clusters $C_1, C_2, \ldots, C_K$.

6. **Assignment:** Assign each original data point $\mathbf{x}_i$ to cluster A_j if and only if the corresponding row $\mathbf{u}_i$ of the matrix U was assigned to cluster C_j, yielding clusters:

$$A_j = \{\mathbf{x}_i : \mathbf{u}_i \in C_j\}, \quad j = 1, 2, \ldots, K. \tag{5}$$

This procedure effectively captures intrinsic global structures within the dataset, yielding robust clustering outcomes even when facing complex data geometries.

3.2 Experimental Settings

The objective is to develop an experimental pipeline to apply spectral clustering on Community Detection, to each dataset, considering the following aspects:

- **Effect of Number of Clusters (k)**: How does k impact the clustering quality?
- **Effect of Gaussian Kernel Parameter (σ)**: How does σ affect the similarity computation and clustering? Start testing $\sigma \in$ {0.1, 0.5, 1.0, 2.0, 5.0}
- **Effect of Normalized vs. Unnormalized Laplacian**: which Laplacian normalization technique yields better clustering results?

3.3 Performance Evaluation Metrics

The following evaluation measures should be applied and the results properly discussed.

3.3.1 Normalized Mutual Information (NMI)

The **Normalized Mutual Information (NMI)** [4] quantifies the similarity between two clusterings, such as a predicted clustering and a reference clustering. It ranges between 0 (no mutual information, complete independence) and 1 (perfect match).

Formally, given two partitions $U = \{U_1, U_2, \ldots, U_R\}$ and $V = \{V_1, V_2, \ldots, V_S\}$ of a set of N elements, the NMI is defined as:

$$\mathrm{NMI}(U, V) = \frac{2\, I(U; V)}{H(U) + H(V)}, \tag{6}$$

where $I(U; V)$ is the mutual information between U and V, defined as:

$$I(U; V) = \sum_{i=1}^{R} \sum_{j=1}^{S} P(i, j) \log \frac{P(i, j)}{P(i)P(j)}, \tag{7}$$

with probabilities:

- $P(i, j) = \frac{|U_i \cap V_j|}{N}$,
- $P(i) = \frac{|U_i|}{N}$,
- $P(j) = \frac{|V_j|}{N}$.

The entropies $H(U)$ and $H(V)$ are:

$$H(U) = -\sum_{i=1}^{R} P(i) \log P(i), \quad H(V) = -\sum_{j=1}^{S} P(j) \log P(j). \tag{8}$$

3.3.2 Modularity Score (Q)

The **modularity score** (Q) [5, 6] evaluates clustering quality by measuring the density of edges within clusters compared to edges between clusters. Modularity ranges between -0.5 and 1.0, with higher values indicating stronger community structure.

Given a graph with adjacency matrix A, degrees $k_i = \sum_j A_{ij}$, and total edges $m = \frac{1}{2} \sum_{ij} A_{ij}$, the modularity for a clustering $\{C_1, C_2, \ldots, C_c\}$ is:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \tag{9}$$

where:

- c_i is the community to which node i belongs,
- $\delta(c_i, c_j)$ is the Kronecker delta function:

$$\delta(c_i, c_j) = \begin{cases} 1, & \text{if } c_i = c_j \\ 0, & \text{otherwise} \end{cases}. \tag{10}$$

The term $\frac{k_i k_j}{2m}$ represents the expected number of edges between nodes i and j in a randomly configured network.

4 Report

The team work report should present experiments systematically describing any preprocessing steps, construction of the similarity matrix, and type of Laplacian used (normalized or unnormalized). Also, should explore the influence of key hyperparameters — namely, the number of clusters (k) and the Gaussian kernel width (σ) — on the clustering results.

Evaluation must be based on internal validation (Modularity) and external validation (NMI), whenever ground-truth labels are available.

For each dataset analyzed, students should report the parameter configurations tested, the best performing settings, and provide plots showing the behavior of Modularity and NMI as a function of k and σ.

It must be included an interpretation of the identified communities in relation to known metadata or real-world groupings, and conclude with a critical analysis of spectral clustering's performance and limitations observed during experimentation.

Individual REPORT (supplement to be submitted by each student)

Students must also submit a brief individual report (no more than one page), containing the following: (i) resume of the part(s) of this project you worked on; (ii) what you learned from this work.

References

[1] Ng, A.Y.; Jordan, M.I.; Weiss, Y. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, 2002, **14**, 849–856.

[2] Krebs, V. Political Books Network Visualization. Available online: `http://www.orgnet.com/`

[3] Girvan, M.; Newman, M.E.J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2002, **99**(12), 7821–7826.

[4] Lancichinetti, A.; Fortunato, S.; Kertesz, J. Detecting the overlapping and hierarchical community structure in complex networks. New J. Phys. 2009, 11, 033015.

[5] Newman, M.E.J.; Girvan, M. Finding and evaluating community structure in networks. *Physical Review E*, 2004, **69**, 026113. `https://doi.org/10.1103/PhysRevE.69.026113`

[6] Newman, M.E.J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 2006, **103**(23), 8577–8582. `https://doi.org/10.1073/pnas.0601602103`