Homework 4 - ID2222 Data Mining

Group 71 - Nour Alga & Karim Haque

1st December 2025

1 Introduction

The goal of this homework was to implement and analyse spectral clustering algorithms for graph partitioning, as described in the paper "On Spectral Clustering: Analysis and an Algorithm". This algorithm was applied to two different sample graphs. For this homework, we used Matlab. An important observation is how our sample graph datasets only contain edges, contrary to the paper's "set of points". Hence, the first step in the algorithm becomes different as we can have a simpler adjacency matrix instead of the affinity matrix.

2 Implementation

It was quite straight forward to implement the algorithm in the paper. First, the dataset containing the edges of the graph was read, and an adjacency matrix A was constructed using the sparse and graph functions. The adjacency matrix was visualized using spy. In the paper the first step is to create a full affinity matrix, but this was not needed in this situation because the dataset provides the graph's edges, so each connection is treated as a binary weight.

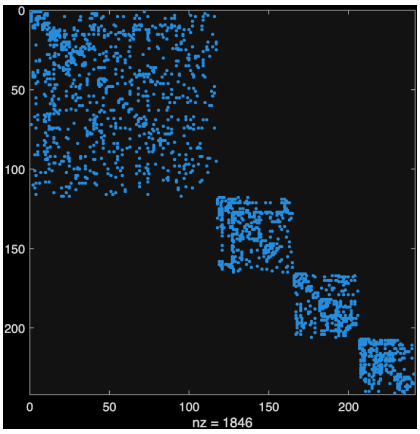
Next, the degree matrix D was formed as a diagonal matrix with entries equal to the sum of each row of A, and the normalized Laplacian $L = D^{-1/2}AD^{-1/2}$ was computed. The K largest eigenvectors of L were extracted using eigs and stacked into the matrix X.

Clustering was performed on the rows of Y using k-means, assigning each original node to a cluster corresponding to its row in Y. Additionally, the Fiedler vector (second smallest eigenvector of L) was plotted as well.

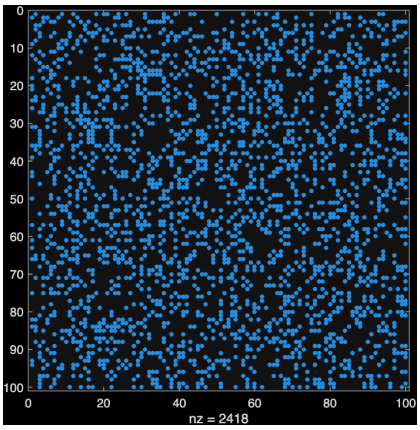
3 Results

The first dataset is a real dataset by Rob Burt while the second dataset is a synthetic one.

- Sparse Adjacency: This revealed the communities found in the graph. In the first dataset, it is clear that there are 4 separate communities, with one being larger. The second dataset is much more dense, there is no clear community structure.

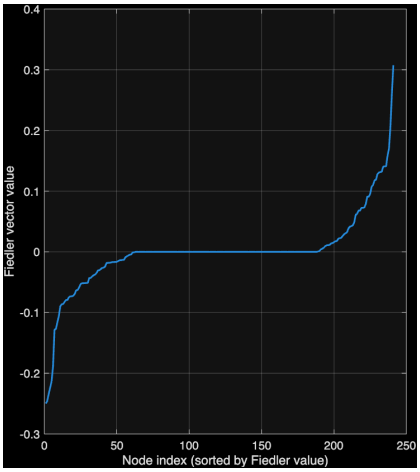


Sparse Dataset 1

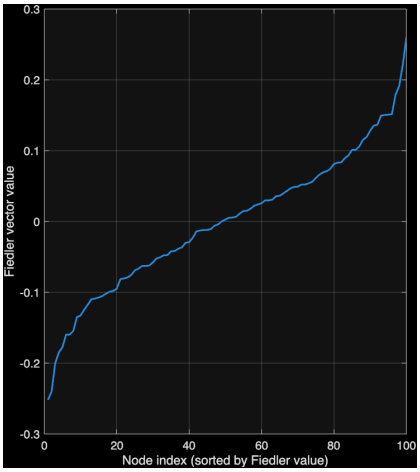


Sparse Dataset 2

- Fiedler Vectors: Dataset 1 shows distinct plateaus with sharp transitions, indicating well-separated clusters. Dataset 2 displays a smoother gradient, reflecting its uniform connectivity and less defined cluster boundaries.

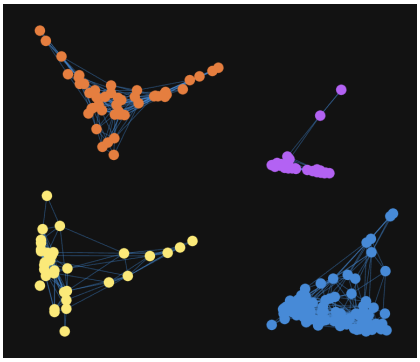


Fiedler Vector Dataset 1

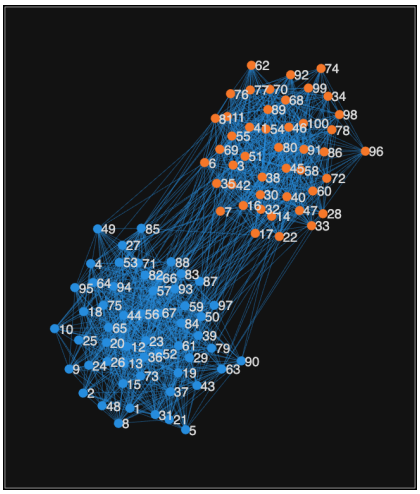


Fiedler Vector Dataset 2

- Spectral Clustering: For dataset we used cluster size = 4 given the spares pattern while we used cluster size = 2 for the second dataset. The initial approach, using the Normalized Laplacian*($L_{\text{norm}} = D^{-1/2}AD^{-1/2}$) and its K largest eigenvectors, resulted in a poor clustering due to numerical instability. It incorrectly split one visual community into two different colors. To correct this, we switched to the unnormalized Laplacian ($L = D - A$). This approach is theoretically and numerically superior for graphs with clear, well-separated components.



Spectral Clustering (Unnormalized) Dataset 1



Spectral Clustering Dataset 2

4 How to Run

To execute the program, run the hw4.m program.