Introduction

The code used in this lab reproduces the algorithm found in the paper "On Spectral Clustering: Analysis and an algorithm" by Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. The sample graphs used to test the implementation are "example1" and "example2".

How to run

- Download the sample graphs from the links found above
- Ensure that Python3, NumPy, sklearn, networkx and SciPy are installed
- Go into the Graph Spectra map
- Run instruction

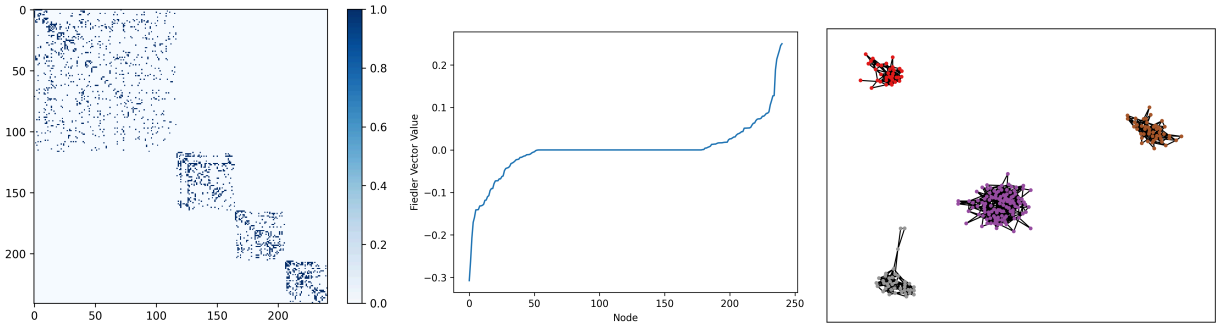
```
python3 spectral_clustering.py
```

Solution

1. The sample graphs were loaded in using the networkx functions "read_edgelist" and "read_weighted_edgelist" depending on the graph
2. The adjacency matrix A was obtained using the networkx function "to_numpy_array" and was used as the affinity matrix described in the paper
3. The numpy function "diagflat" was used to create a diagonal matrix D with the flattened input as the diagonal. The inverted square root of the D matrix and the A matrix was then used to calculate the L matrix (normalized Laplacian matrix).
4. The L matrix was used to calculate its eigenvalues and eigenvectors
5. The eigengap was then used to calculate k (comparing the amount of difference between sets of continuous eigenvalues. K is the approximation of the number of clusters in the graph.
6. The Fiedler vector (2nd smallest eigenvector) was saved along with k and the k largest eigenvectors
7. The k largest eigenvectors were then normalized
8. Using the K-means algorithm, a representation of the graphs with k clusters was created

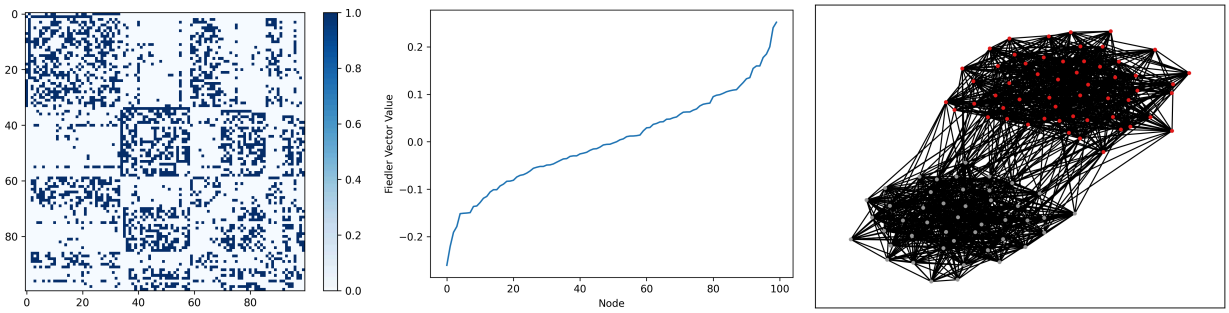
Result

Graph 1



From the adjacency matrix plot to the left, we can see 4 distinct clusters which is then confirmed by looking at the clustered graph to the right. The graph in the middle shows the sorted Fiedler vector which has three distinct sets of values, the extremes and the middle point at zero. So, just by looking at the middle graph, it would be very hard to guess that there are 4 clusters.

Graph 2



The adjacency matrix plot to the left shows a very cluttered pattern which indicates that the clusters are not distinct such as in graph 1 but are instead connected to each other as supposed to only having connections strictly within the cluster. This is confirmed by the right plot where we can distinguish 2 clusters also having connections to each other. The middle plot of the Fiedler Vector shows the existence of two clusters but with a slow transition. This can be an indication that the clusters are connected butw not separated as in the first graph.