Report

Shadman Ahmed, Oguzhan Ugur

Problem Formulation

In this homework we were instructed to study and implement and to test spectral graph clustering. The paper "On Spectral Clustering: Analysis and an algorithm" by Andrew Y. Ng, Michael I. Jordan, Yair Weiss was analyzed in order to implement the spectral graph clustering using the K-eigenvector algorithm. The problems we had to solve in order to create this algorithm were,

- First, we had to form a Adjency matrix which shows the relation between the nodes, in other words the edges.
- Second, the Laplacian matrix had to be determined
- Third, find [x1,x2,…,x_k] k largest eigenvectors of L and form matrix X = [$x_1 x_2$ . . . x_k] by stacking the eigenvectors in columns
- Fourth, each row of X had to be renormalized and clustered using K-means.

Dataset

The datasets used in this homework were "example1.dat" and "example2.dat". The first dataset is a real graph and consists of edges and the second one is a synthetic graph and consists of weighted edges.

Result

Example1.dat dataset:
The result we obtained is depicted on the figures below Figure 1. The graph to the left represent the the non clustered data. The graph to the right is the spectral clustered nodes where we chose to have 4 clusters. It is possible to see that the algorithm clustered the graphs correctly since each community got clustered.

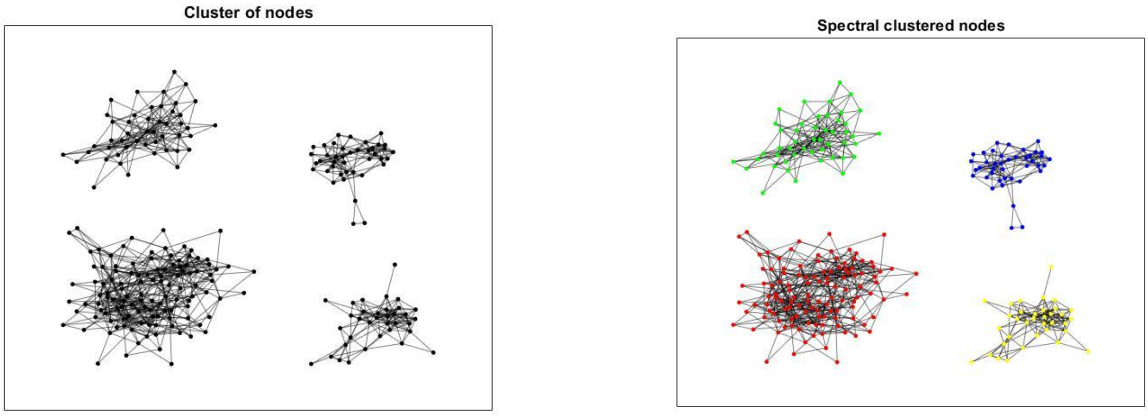

Figure 1

The sparsity pattern depicted on Figure 2(left picture) shows the edges between the nodes. It is possible to see that we have 4 communities, one big and three smaller communities. This plot shows the communities we want to cluster. The conductivity of this plot is smaller than the sparsity pattern of example2.dat dataset since it is less dense.

The second plot on figure 2 correspond to the Fiedler vector (eigenvector corresponding to the smallest eigenvalue). The Fiedler vector shows the partitioning of the graph. The signs of the Fiedler vector can therefore be used in order partition this graph into four components in our case the negative signs the positive and those in middle. Seer figure below.

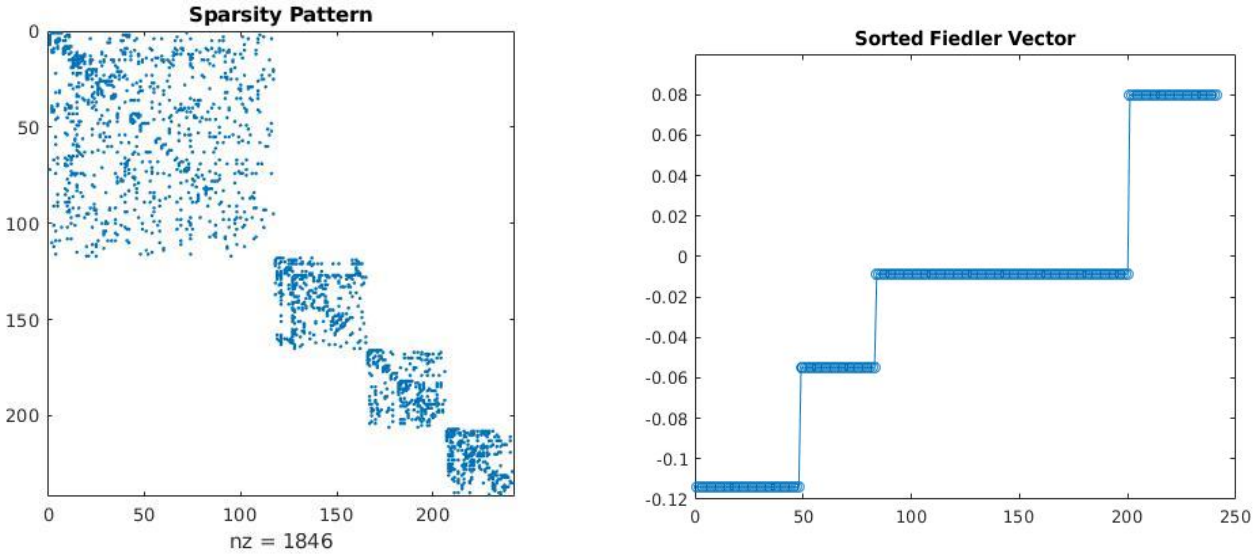

Figure 2

Laplacian Matrix

We used two different Laplacian matrices, one was ratio cut and the other was the normalized Laplacian matrix. The Fiedler vector depicted on figure 2 shows the Fiedler vector for partitioning with Ratio cut L = D – A.

The other Laplacian matrix we used was the normalized Laplacian matrix $\mathbf{L = D^{1/2}AD^{1/2}}$ **,** the Fiedler vector obtained for this Laplacian matrix is depicted on Figure 3. This result is not what we were seeking that's why we tried the ratio cut and it worked well. As it's possible to see on Figure 2, the Fiedler vector has four partitions while figure 3 has two which is not the case for this dataset.

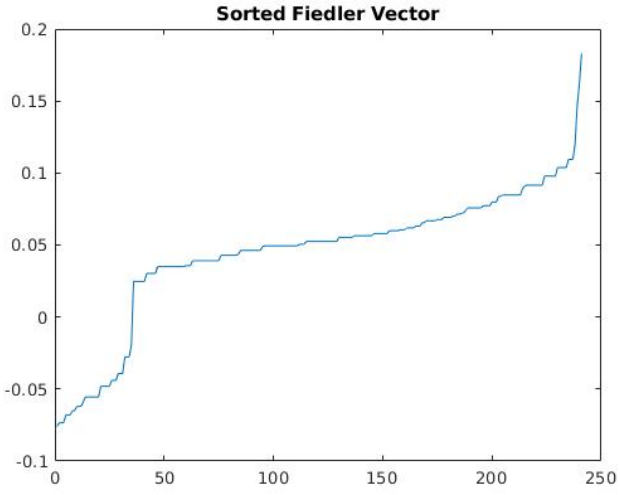

Figure 3

Example2.dat Dataset

Spectral graph portioning of the second data, Example 2, consisting of edges and weights from a synthetic graph.

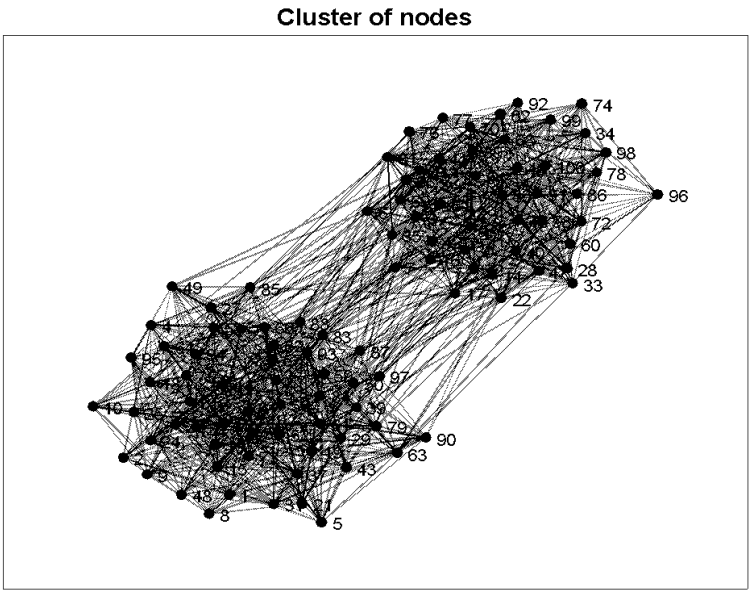

Figure 4. Graph model of Example 2 dataset.

The graph in figure 1 shows two clusters, but not as separated compared to fig 1. The conductance will result in higher value due to many endpoints from both sets, are linked across.

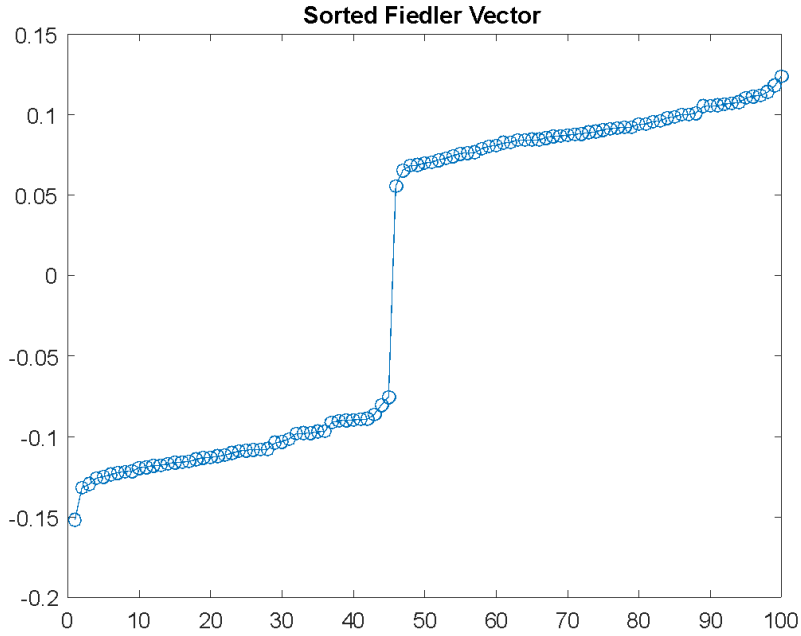

Figure 5. Fiddler vectors from Example 2 dataset. Y-axis represent nodes value from the eigenvector. And the X-axis is the node index.

After having created the Laplacian matrix representing G, the spectral portioning for the second largest eigenvector namely the fiddle vector is shown in figure 2. One can clearly see the separation of the two clusters, where the nodes from the first set have values below 0 and the other set have values above 0.

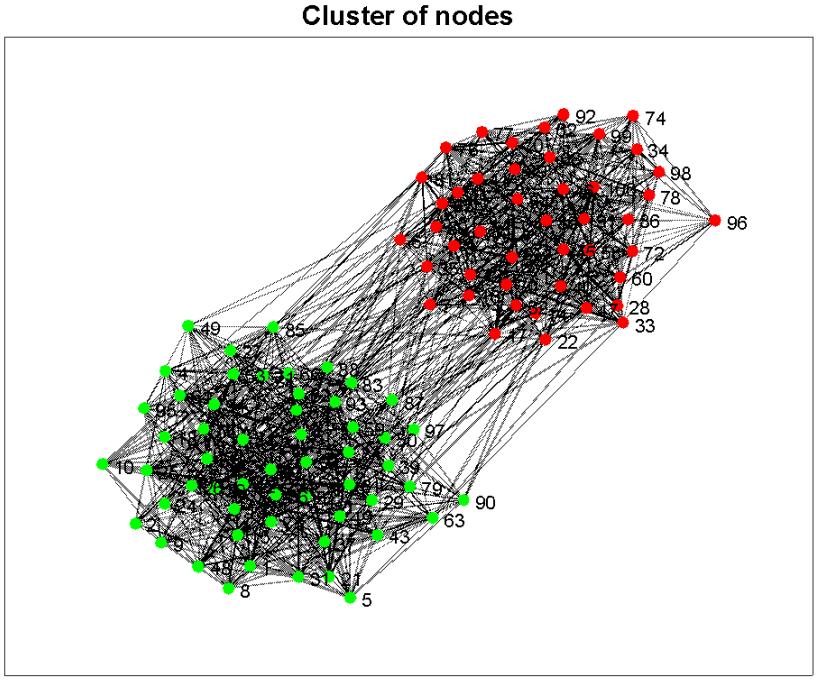

Figure 6. Spectral clustered nodes of example 2 dataset.

After the pre-processing and decomposition of the graph, the grouping was done by using k-means of the eigenvector matrix extracted from the Laplacian matrix. As figure 2 indicated, we got a good spectral clustering of the graph with two different colors indicating which cluster the nodes belong.