Data Mining, ID2222 HT201

Homework Assignment 4

Frano Rajic, Ivan Stresec

November 30, 2020

1 Objective

Our objective in this homework was to implement a spectral clustering algorithm proposed by Andrew Y. Ng, Michael I. Jordan and Yair Weiss in On Spectral Clustering, Analysis and an Algorithm.

2 Implementation

Andrew Ng et al. proposed an algorithm for clustering that uses the top k eigenvectors of a matrix derived from the distance matrix. These k eigenvectors are then used to cluster the data. The best value of k can be guessed by looking at the eigengap of the Laplacian matrix.

The implementation of the algorithm was made in Matlab, as suggested in the assignment. The algorithm can be divided into 6 steps, as written in the original paper:

1. Create affinity/adjacency matrix $\mathbf{A}$
2. Create Laplacian $\mathbf{L} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$
   - $\mathbf{D}$ diagonal matrix, $[\mathbf{D}]_{ii} = \sum_{j=1}^{n} \mathbf{A}_{ij}$
3. Create matrix $\mathbf{X}$ by stacking k largest eigenvectors (determine k by eigengap)
4. Create matrix $\mathbf{Y}$ by normalizing rows of X
5. Use k-means on $\mathbf{Y}$ (same k as before)
6. Appoint clusters to nodes according to k-means clustering of $\mathbf{Y}$

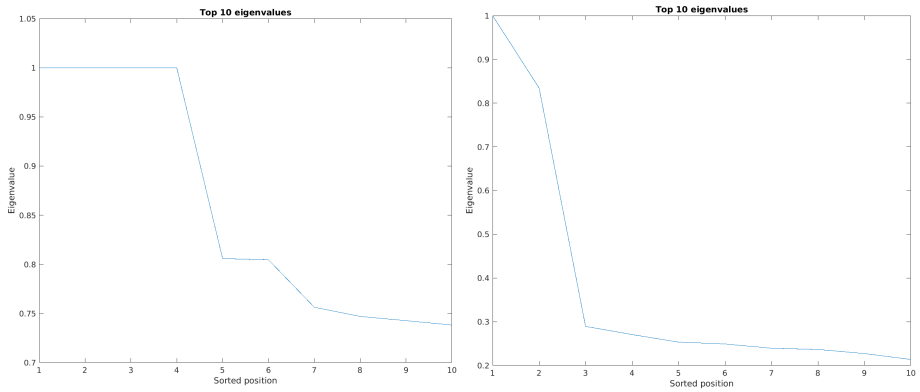
The algorithm proposed by Andrew Ng et al. proposed to create the affinity matrix A given a set of points $S = \{s_1, \ldots, s_n\}$ using the formula $A_{ij} =$

$\exp\left(-\|s_i - s_j\|^2 / 2\sigma^2\right)$. In our implementation, we use the number of (undirected) edges between nodes to initialize the proximity matrix, as the datasets we worked with did not have data points in a coordinate space, but the edges between graph vertices.

Each of the steps of the algorithm were simply implemented using Matlab's matrix operations, eigen decomposition and k-means functions. The clusters are visualized by plotting instances of Matlab's graph classes.

3 Testing and evaluation

A real graph dug by Ron Burt from the 1966 data collected by Coleman, Katz and Menzel on medical innovation and a synthetic graph given in the assignment were the two datasets used to evaluate the performance of the algorithm. The value of k was picked according to the eigengap, which can be noted in Figure 1 where the first 10 largest eigenvalues of both graphs are shown. The results for the first dataset are shown in Figure 2, whereas for the second in Figure 3.



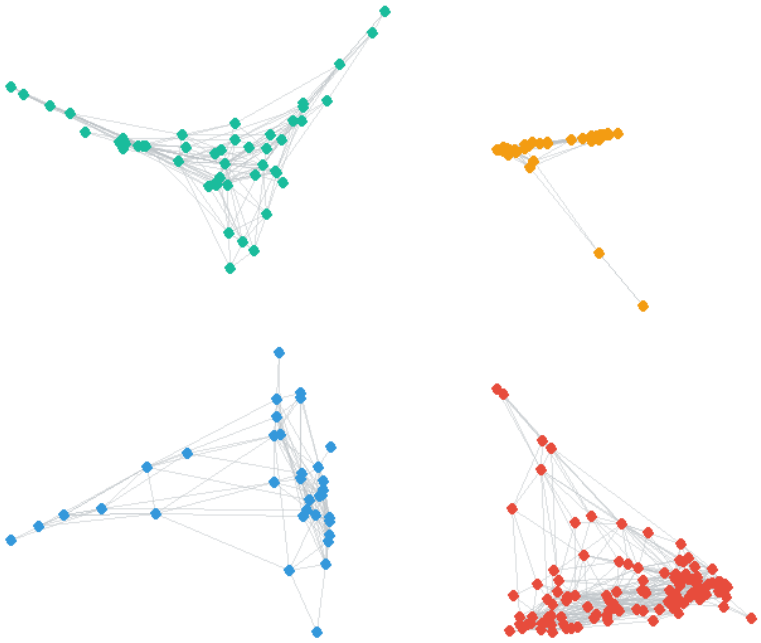
(a) Graph dug by Ron Burt

(b) The synthetic graph

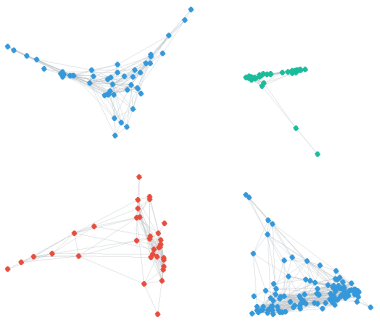
Figure 1: Eigengaps for the two graphs used during evaluation and testing. Eigengap for the first occurs after 4^{th} largest eigenvalue and after 2^{nd} for the second graph.

4 Final remarks

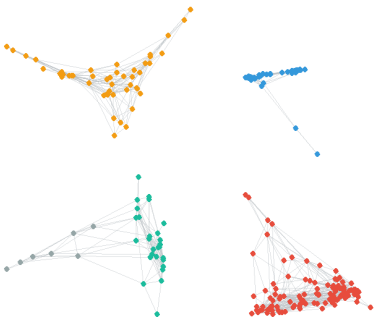
This assignment was a good way to see the practical side of spectral graph theory and showing its versitility. We saw how easily the necessary functionality could be implemented in Matlab and checked. We have verified the validity of the k-eigenvector clustering method and our implementation on the two datasets.



(a) k=4



(b) k=3



(c) k=5

Figure 2: K-eigenvectors spectral clustering on a real graph dug by Ron Burt from the 1966 data collected by Coleman, Katz and Menzel on medical innovation. Results shown for three different values of parameter k. $k = 4$ was matched with the eigengap and gave the most intuitive result, as the four graph components can easily be recognized.

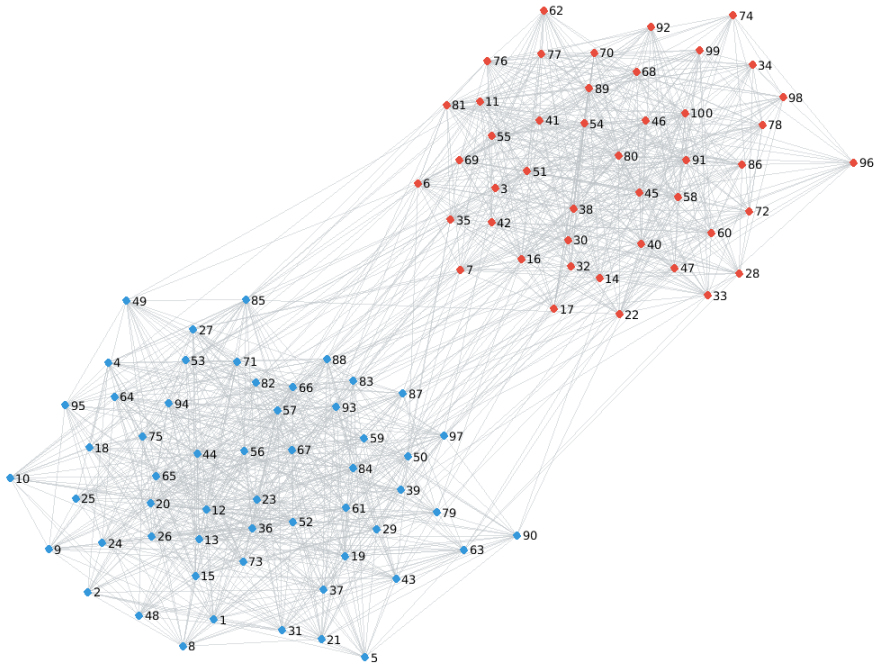


Figure 3: Results on the synthetic graph given in the assignment. $k = 2$ gave the expected, intuitive result.