DataMining ID2222 - Homework 4
Graph Spectra

Sherly Sherly and Anna Martignano

December 1, 2019

1 Introduction

In this project, we will implement and test the algorithm for spectral clustering described in the paper "On Spectral Clustering: Analysis and an algorithm" by Andrew Y. Ng, Michael I. Jordan, Yair Weiss. [1]
The algorithm includes the following steps:

1. Form the affinity matrix A, $A_{ij} = exp((-||s_i - s_j||^2)/2(\sigma)^2)$ for i not j and Aii = 0.
2. Define D to be the diagonal matrix whose (i,i)-element is the sum of A's i-th row and construct the matrix $L = D^{-1/2}AD^{-1/2}$.
3. Find x1, x2, .., xk the k largest eigenvectors of L (chosen to be orthogonal to each other in the case of repeated eigenvalues), and form the matrix X = [x1 x2 ...xk] by stacking the eigenvectors in columns.
4. Form the matrix Y from X by renormalizing each of X's rows to have unit length (i.e. $Y_{ij} = X_{ij}/(\sum_j (X_{ij})^2)^(1/2))$.
5. Treating each row of Y as a point in Rk, cluster them into k clusters via K-means or any other algorithm.
6. Finally, assign the original point si to the cluster j if and only if row i of the matrix Y was assigned to cluster j.

There are two parameters to tune in this scenario, k the number of clusters and σ the scaling parameter.

The project is organised in the following settings: **Secton 2** describes the implementation, **Section 3** describes the datasets used to implement and test the algorithms, **Section 4** describes the results obtained by running the algorithm over the given datasets, and **Section 5** describes how to run the codes.

2 Implementation Strategy

The algorithm has two parameters that need to be tuned in order to achieve good performance namely the number of cluster k and the scaling parameter σ.

The first parameter is the number of cluster k. A naive approach to determine the cluster size is to have a look at the original graph and try to look at the network pattern. If the cluster are well distinguished among each other, it will be possible to determine the number by simply by looking at it. A formal alternative is to check if the difference between two eigenvalues, sorting in descending order, changes radically. This techniques aim to maximize the **eigengap**, i.e. the difference between two consecutive eigenvalues. Indeed, most stable clustering is generally given by the value k that maximizes eigengap, $\Delta_k = |\lambda_k - \lambda_{(k-1)}|$.

The second parameter is σ. In order to find the best σ, it is important to assess after the clustering of Y which values gives the tightest clusters, i.e. the one with the smallest distortion.

Based on the algorithm proposed in the paper, we would need to compute the affinity matrix A given the set of points as the first step. However, in this project, the datasets provided are edge lists. In this case, instead of generating the matrix A with their proposed calculation, we will first convert the edge list into a graph and use the graph to generate the adjacency matrix. This adjacency matrix will be used as the affinity matrix A. Due to this, we will not need to optimize for the parameter σ in our computation anymore.

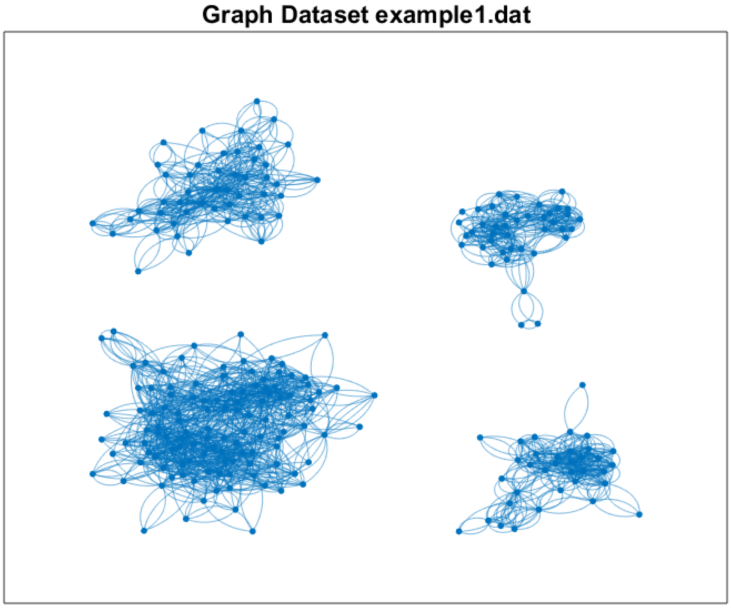
3 Graph Datasets

There are two graph datasets used in this project. The sections below describes the dataset and its graph plot.

3.1 Dataset 1

This dataset is a real graph. It was prepared by Ron Burt. He dug out the 1966 data collected by Coleman, Katz and Menzel on medical innovation. They had collected data from physicians in four towns in Illinois, Peoria, Bloomington, Quincy and Galesburg.

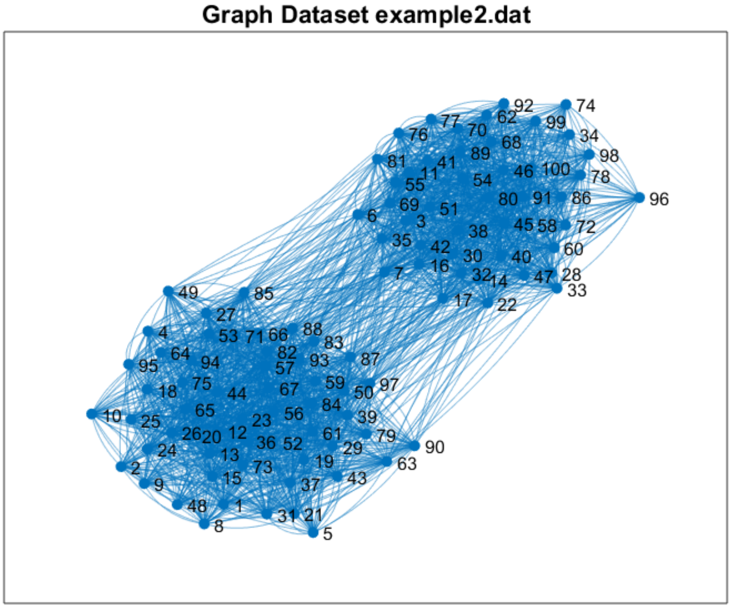
Details	
Number of Nodes	241
Number of Edges	2196



3.2 Dataset 2

This dataset is a synthetic graph.

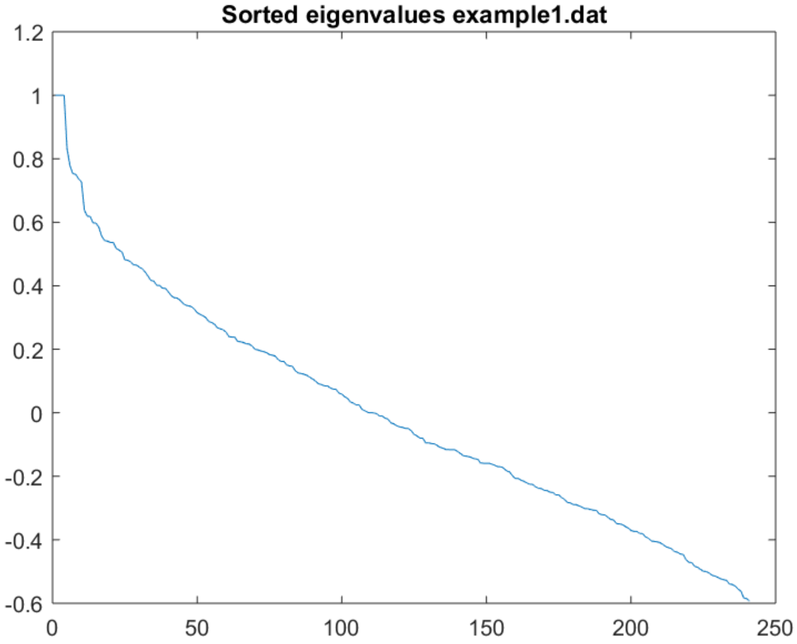
Details	
Number of Nodes	100
Number of Edges	2418



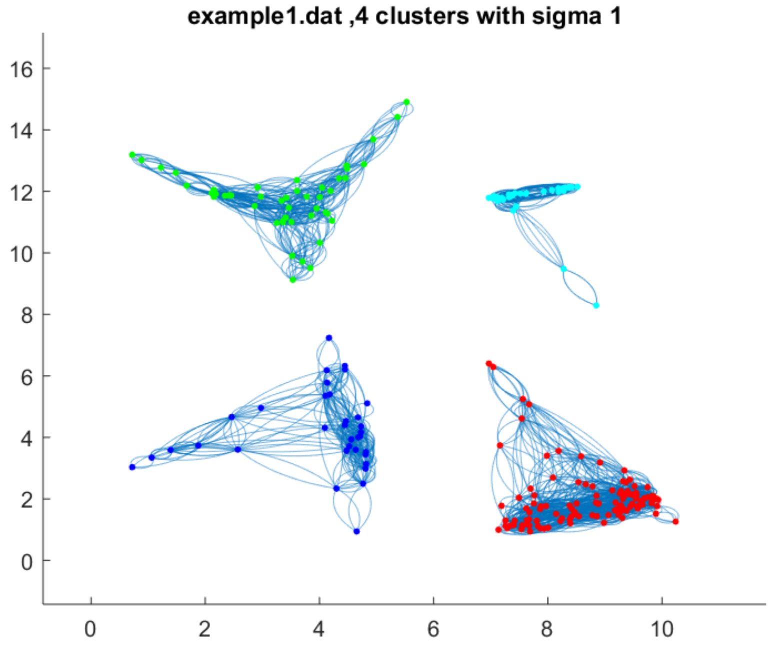
4 Results

4.1 Dataset 1

The number of clusters k based on **eigengap** is 4 in this dataset. The figure below shows the plot of the eigenvalues for this dataset.

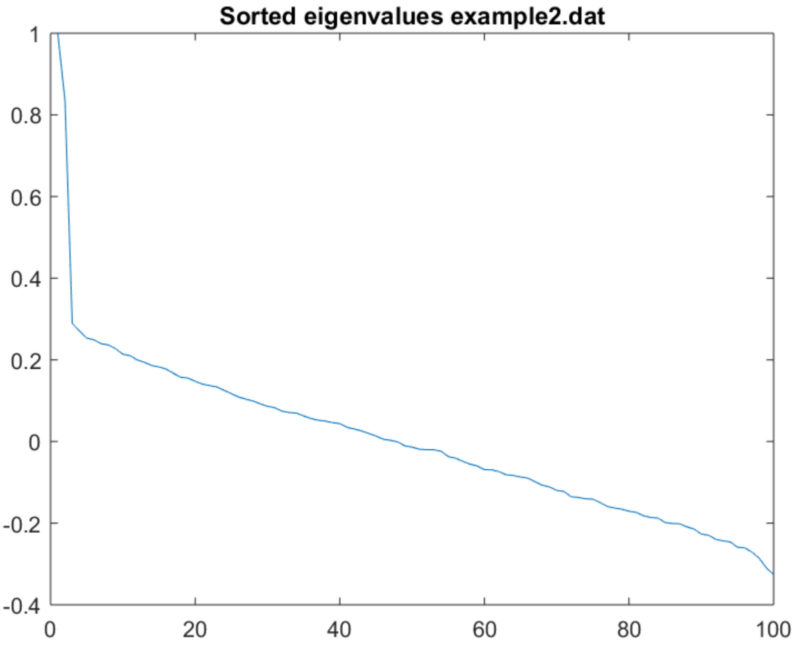


Based on the clustering with K-Means and $k = 4$ we obtain the plot below for the clusters generated.

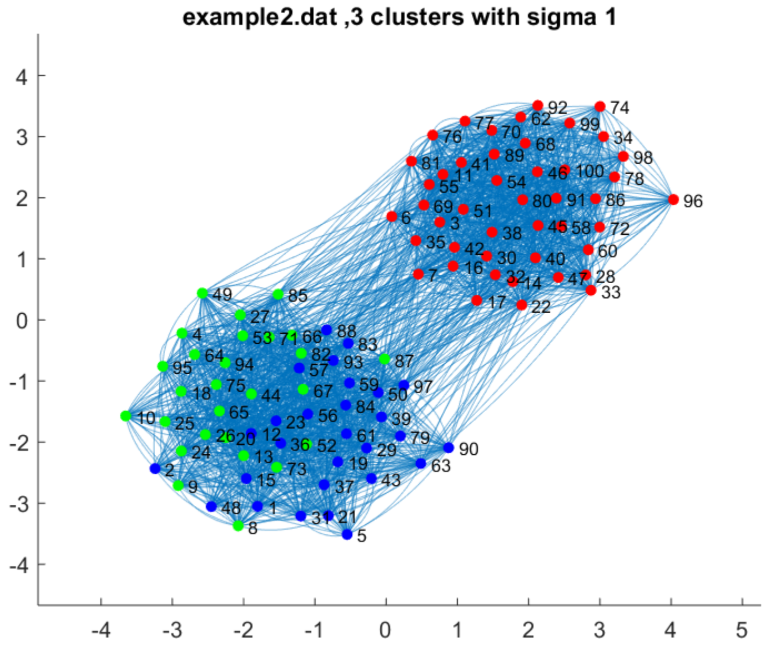


4.2 Dataset 2

The number of clusters k based on **eigengap** is 3 in this dataset. The figure below shows the plot of the eigenvalues for this dataset.



Based on the clustering with K-Means and $k = 3$ we obtain the plot below for the clusters generated.



In this case, visually, we observe that there is 2 main clusters. However, there could be nested communities which resulted in best $k = 3$.

5 How to run the code

The code is written in Matlab. To run the code, load the code and the dataset into the same path on Matlab application and *Run* it.

References

[1] Andrew Y. Ng, Michael I. Jordan, Yair Weiss, er "On Spectral Clustering: Analysis and an algorithm".