Data Mining - Homework 4

Robert-Andrei Damian and Alice De Schutter

Graph Spectra

3 december 2021

Task

Study, implement and test the spectral graph clustering algorithm as described in the paper "On Spectral Clustering: Analysis and an algorithm" by Andrew Y. Ng, Michael I. Jordan, Yair Weiss. Using our implementation of the K-eigenvector algorithm, we are to analyse two sample graphs: a real graph and a synthetic graph.

1 Detailed Information

Python was used to implement the spectral graph clustering algorithm presented in the paper. The algorithm is summarized in **Figure 1**. For each data-set, the **k** value (**k** = number of clusters) was determined by the eigengap of the Laplacian matrix **L**, meaning that **k** is given by the value that maximizes the eigengap (difference between consecutive eigenvalues).

Given a set of points $S = \{s_1, \ldots, s_n\}$ in $\mathbb{R}^l$ that we want to cluster into k subsets:

1. Form the affinity matrix $A \in \mathbb{R}^{n \times n}$ defined by $A_{ij} = \exp(-||s_i - s_j||^2 / 2\sigma^2)$ if $i \neq j$, and $A_{ii} = 0$.
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 $L = D^{-1/2} A D^{-1/2}$.[1]
3. Find $x_1, x_2, \ldots, x_k$, the k largest eigenvectors of L (chosen to be orthogonal to each other in the case of repeated eigenvalues), and form the matrix $X = [x_1 x_2 \ldots x_k] \in \mathbb{R}^{n \times k}$ 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 $\mathbb{R}^k$, cluster them into k clusters via K-means or any other algorithm (that attempts to minimize distortion).
6. Finally, assign the original point s_i to cluster j if and only if row i of the matrix Y was assigned to cluster j.

Figur 1: Summary of algorithm

Instructions on how to build and run the program

- python3 graph-spectra.py

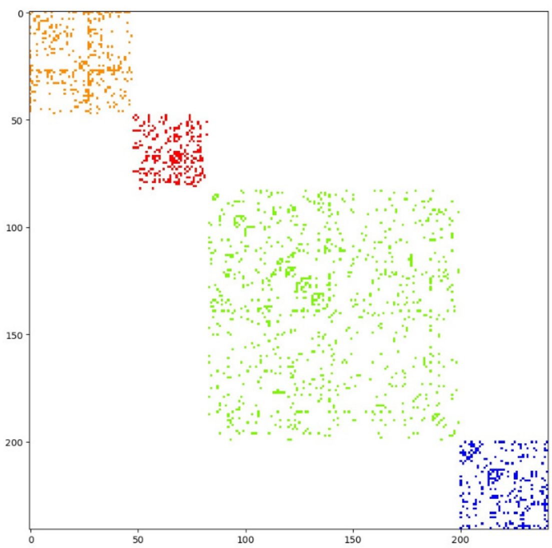
2 Results

Real Graph:

Here, **k** was determined to be equal to 4.

The clustering labels associated with the data is: [2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]

The adjacency matrix was rearranged so that nodes belonging to the same cluster were grouped together. This is visualised below.



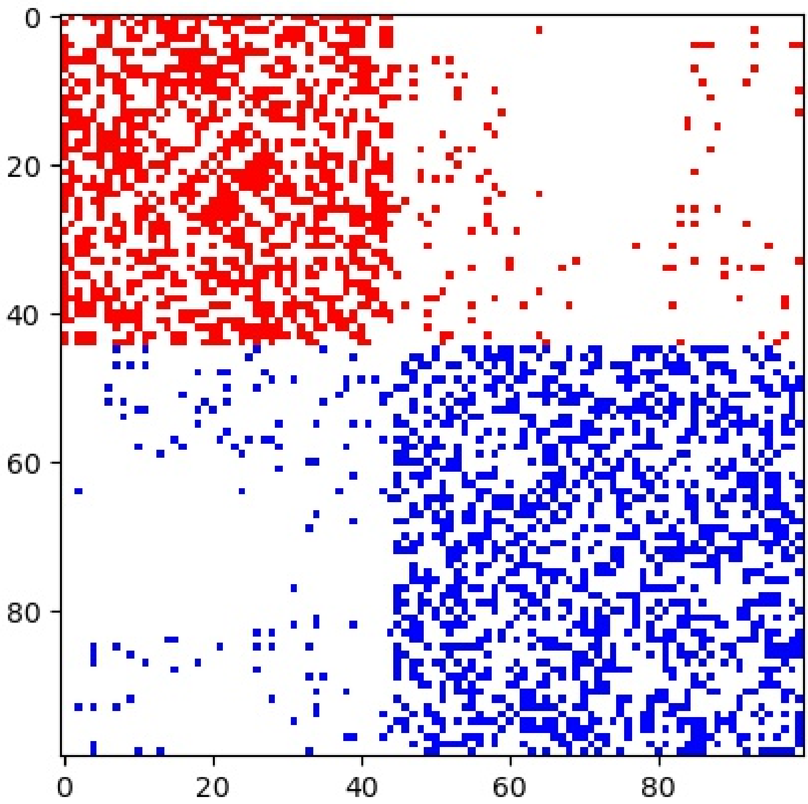
Figur 2: Clusters for real graph data-set

Synthetic Graph:

Here, **k** was determined to be equal to 2.

The clustering labels associated with the data is: [1 1 0 1 1 0 0 1 1 1 0 1 1 0 1 0 0 1 1 1 1 0 1 1 1 1 1 0 1 0 1 0 0 0 0 1 1 0 1 0 0 0 1 1 0 0 0 1 1 1 0 1 1 0 0 1 1 0 1 0 1 0 1 1 1 1 1 0 0 0 1 0 1 0 1 0 0 0 1 0 0 1 1 1 1 0 1 1 0 1 0 0 1 1 1 0 1 0 0 0]

Again, the adjacency matrix was rearranged so that nodes belonging to the same cluster were grouped together. This is visualised below.



Figur 3: Clusters for synthetic graph data-set