Cluster Ranking using Spectral clustering and Eigenvector centrality

Keerthiraj Nagaraj
Department of Electrical and Computer Engineering
University of Florida
UFID: 36726114

Overview

- Abstract
- Motivation
- Introduction
- Problem formulation
- Methodology
- Dataset generation
- Results and observations
- Conclusions
- References

Abstract

- Clustering technique is an important technique in analysis of large datasets.
- In large networks, it is better to analyze small groups of network nodes to understand the behavior of network rather than working on the whole network at once.
- It is important to understand and quantify the effectiveness of each cluster in a network.
- I have used Spectral clustering and Eigenvector centrality to divide any given network into optimal clusters and rank them.

Motivation

- In most cases the networks are divided into different clusters for easy operation and management.
- It is beneficial to provide more resources to clusters of nodes which are more important than others in order to improve overall efficiency of network.
- Cluster ranking information can be used to distribute resources optimally among the different clusters of the network.
- Cluster ranks can be used to understand the vulnerability of the network i.e. to understand strength and weakness of a network.

Introduction

- Spectral clustering is one among the rich family of clustering algorithms.
- Spectral clustering uses the spectral properties i.e. Eigenvalues and Eigenvectors of Graph Laplacian matrix and K-means technique to divide the given network into ‘k’ number of clusters.
- Eigenvector centrality measure specifies the importance of a given node in a network.
- Higher the eigenvector centrality of a node, more is the number of connections it has with other important nodes.

Problem Formulation

- Consider a network N of 'n' nodes.
- Eigenvector centrality 'e' of each node in N is calculated.
- N is divided into 'k' number of clusters using Spectral clustering technique.
- Mean eigen score for each cluster is calculated by averaging the eigenvector measures for the nodes that belongs to each cluster using the below equation.

$$m_j = \frac{1}{|C_j|} \sum_{e_a \epsilon C_j} e_a \quad \text{j=1,2,3… k}$$

Where 'mj' represent mean eigenscore for each cluster and |Cj| represents the number of nodes in cluster 'j'

Methodology

1. Spectral Clustering

- Spectral clustering can be implemented by using standard linear algebra techniques
- Graph laplacian matrix is defined in terms of weighted adjacency matrix and degree diagonal matrix as following, [1]

$$L = D^{-1/2} A D^{1/2}$$

Where L represent Graph Laplacian matrix, D is degree diagonal matrix and A is weighted adjacency matrix for the considered network of nodes.

Methodology

1. Spectral Clustering (Contd)

- Spectral clustering requires the number of clusters 'k' to be given as input. I have used 'Eigengap heuristic' to decide the optimal number of clusters for the given network.

- Evaluate the set of all eignevalues of graph laplacian matrix and arrange them in ascending order and find the difference between consecutive sorted eigenvalues and form a new set $\{l_1, l_2, \ldots l_n\}$ using these values

$$k = \max_i(l_{i+1} - l_i)$$

Methodology

1. Spectral Clustering (Contd)

Algorithm 1 : Steps of Spectral clustering [1] [2]

- Find D, A and then L for the given network.
- Decide the optimal number of clusters ‘k’ using Eigengap heuristic.
- Find first ‘E’ eigenvectors of L and form a matrix E of size ‘n x k’
- Normalize the rows of ‘k’ to make norm of each row equal to 1.
- Apply K-means to normalized E to find the ‘k’ clusters of the given network.
- K-means will return the cluster ID of each node which represent the number of cluster each node belongs to.

Methodology

2. Eigenvector centrality

Algorithm 2 : Steps for evaluating eigenvector centrality of each node. [3] [4]

- Find D, A and L for the given network.
- Find the maximum of all the 'n' eigenvalues of L, let it be denoted by 'lmax'
- Find the degree of all nodes in network and form matrix 'C' of size 'n x 1' using these values.
- The eigenvector centrality matrix is obtained by using the following equation.

$$E = \frac{1}{l_{max}}(A \times C)$$

Methodology

3. Ranking of clusters

Algorithm 3 : Steps for evaluating the rank of each cluster.

- Calculate the Mean eigen score of each cluster using the equation,

$$m_j = \frac{1}{|C_j|} \sum_{e_a \epsilon C_j} e_a \quad \text{j=1,2,3… k}$$

Where ‘ea’ is the eigenscore of node ‘a’ and a=1,2,3,…. n

- Assign ranks to the ‘k’ clusters based on the value ‘m’ of each cluster i.e. higher the value of ‘m’, better is the rank.

Dataset generation

- I have used programs and algorithms designed by Andrea Lancichinetti and Santo Fortunato to generate different datasets of weighted graphs.

Parameter flag	Parameter explanation
N	Number of nodes in the graph
k	Average degree of nodes
maxk	Maximum degree of nodes
minC	Minimum for community sizes
maxC	Maximum for community sizes
C	Clustering coefficient

Dataset Generation

- I generated 3 datasets with different values for parameter flags and used them to test the proposes techniques.

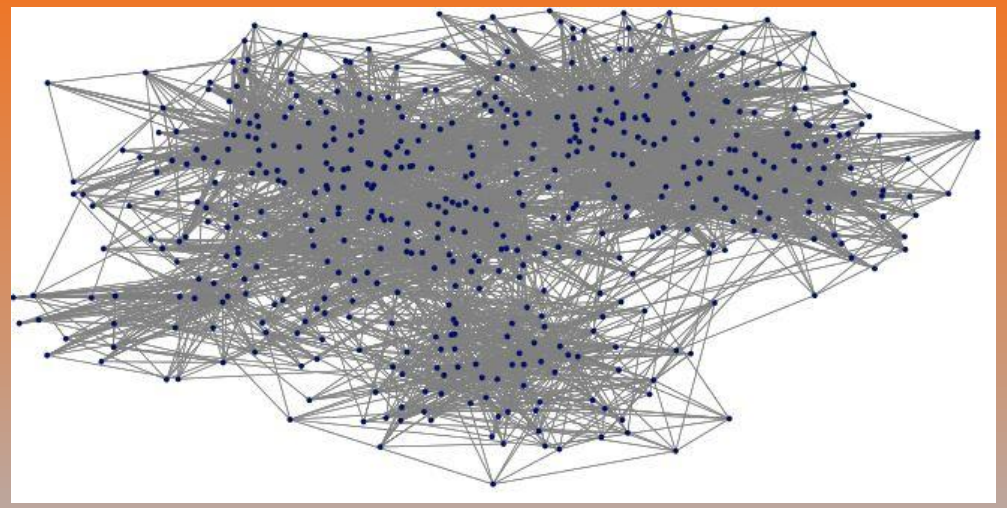
Parameter flag	Dataset 1	Dataset 2	Dataset 3
N	500	500	500
k	15	10	10
maxk	50	15	20
minC	20	10	30
maxC	50	30	60
C	0.5	0.1	0.6

Results and Observations

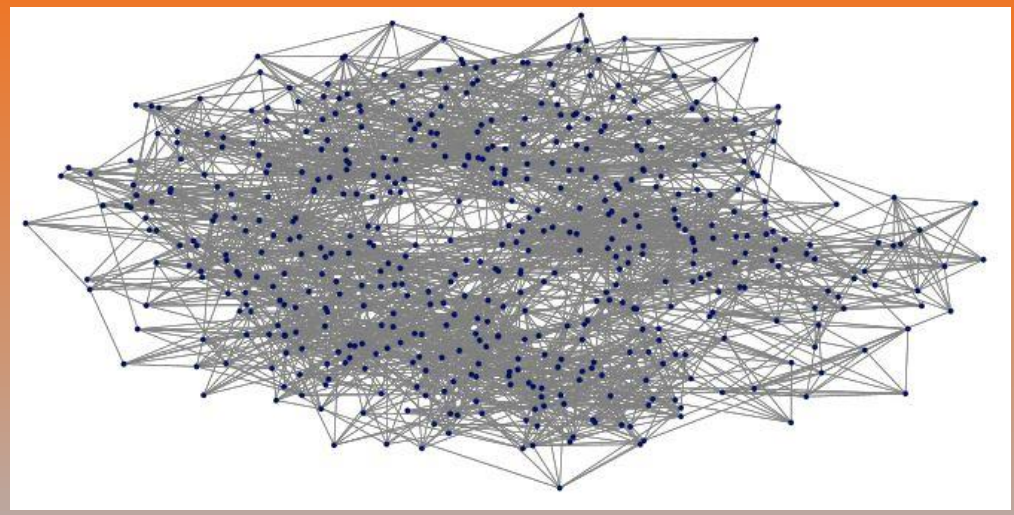
- I implemented the three algorithms proposed using MATLAB software and used the generated 3 datasets to test the technique proposed.
- Number of optimal clusters found using Eigengap heuristic for different datasets.

Dataset number	'k' - clusters
1	14
2	32
3	27

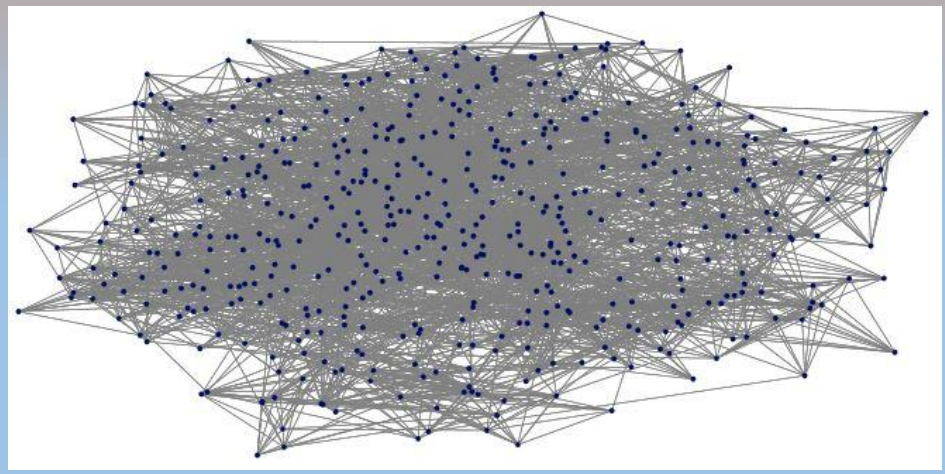
Results and Observations



Edges and nodes of network for dataset 1



Edges and nodes of network for dataset 2

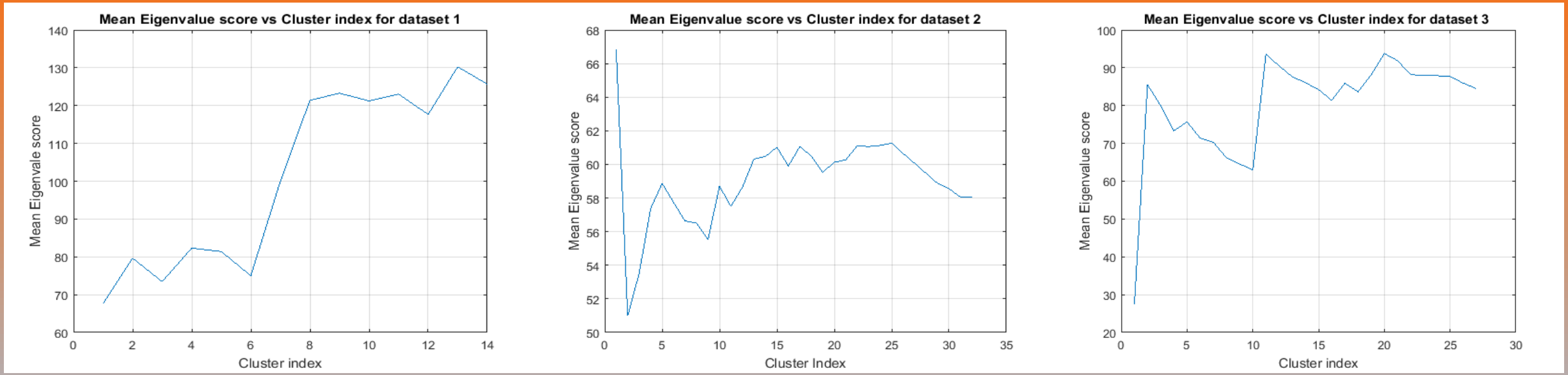


Edges and nodes of network for dataset 3

Results and Observations

- Mean Eigenvalue scores for different clusters of networks for each dataset were calculated and observed that they were different from each other.
- The variation of Mean Eigenvalue scores for different clusters in each network shows that, the importance of each cluster is different from others in a network.
- Hence, the need to find the rank of different clusters of the network so that the higher rank networks can get better resources, which in turn improves the efficiency of the whole network.

Results and Observations



Variation of Mean Eigenvalue score for dataset 1, 2 and 3 respectively

Conclusions and Scope for Future work

- In this work, I have worked on identifying and ranking the clusters of any given network using Spectral clustering and Eigenvector centrality.
- Mean eigenvalue score for each cluster was calculated and was used to categorize the importance of the network and also rank the clusters.
- Higher the Mean eigenvalue score of a cluster is, better is the connectivity and more important nodes of the whole network are interconnected to nodes belonging to that particular cluster.

Conclusions and Scope for Future work

- The proposed technique is useful when we have to find the vulnerable groups of a network, which can be helpful when the resources available for the network are limited.

- The limited resources can be distributed more to the clusters of nodes which have higher rank.

References

[1] Von Luxburg, U., A tutorial on spectral clustering, Statistics and computing, 17(4), 395-416 (2007)

[2] Ng, A. Y., Jordan, M. I., & Weiss, Y., On spectral clustering: Analysis and an algorithm, Advances in neural information processing systems, 2, 849-856 (2002)

[3] White, S., & Smyth, P., A Spectral Clustering Approach To Finding Communities in Graph, In SDM (Vol. 5, pp. 76-84) (2005)

References

[4] Ruhnau, B., Eigenvector-centralitya node-centrality?, Social networks, 22(4), 357-365 (2000)

[5] Bonacich, P., Some unique properties of eigenvector centrality, Social Networks, 29(4), 555-564 (2007)

[6] Lancichinetti, A., Fortunato, S., & Radicchi, F., Benchmark graphs for testing community detection algorithms, Physical review E, 78(4), 046110 (2008)

References

[7] Lancichinetti, A., & Fortunato, S., Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities, Physical Review E, 80(1), 016118 (2009)

[8] Bar-Yossef, Z., Guy, I., Lempel, R., Maarek, Y. S., & Soroka, V., Cluster ranking with an application to mining mailbox networks, Knowledge and Information Systems, 14(1), 101-139 (2008)

THANK YOU