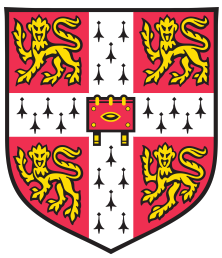
UNIVERSITY OF CAMBRIDGE



L42: MACHINE LEARNING AND ALGORITHMS FOR DATA MINING

MPHIL IN ADVANCED COMPUTER SCIENCE

Research report

Clustering Signed Networks with the Geometric Mean of Laplacians

Pedro Mercado, Francesco Tudisco, Matthias Hein

Author: Ioana Bica (ib354)

Word count: 2468

May 3, 2018

Research report

Clustering Signed Networks with the Geometric Mean of Laplacians [1]

Abstract

Signed graphs offer the possibility of incorporating both positive and negative relationships between vertices, thus having interesting applications in social balance. However, applying spectral clustering to identify communities in such networks represents a challenge, since there is not a well-defined Laplacian operator that can be reliably used in this sense. This research report focuses on analysing the approach proposed by Mercado et al. [1] that uses the geometric mean of Laplacians for cluster signed networks, thus claiming to overcome the limitations of previous spectral methods. The techniques presented in the paper will be discussed and evaluated using both the Stochastic Block Model and real-world datasets.

1 Introduction

Developing powerful data mining algorithms for extracting information from signed networks would have important applications, as many real-world relations can be illustrated in this way. In particular, clustering signed networks, such that vertices in a cluster are linked by positive edges and vertices between clusters are linked by negative edges, would be very useful for mining social networks [2] or for identifying social balance [3, 4].

While many different approaches have been proposed for extending spectral clustering methods for unsigned graphs to signed graphs, there still does not exist a well-established Laplacian operator that is guaranteed to give the best results. The paper 'Clustering Signed Networks with the Geometric Mean of Laplacians' by Mercado *et al.* [1] proposes a novel such Laplacian operator, which the authors claim to overcome the limitations of Laplacians based on some fort of arithmetic mean.

This report will describe and investigate the methods and results obtained in the paper mentioned above. To begin with, the main contributions of Mercado *et al.* involve:

- identifying a Laplacian operator and proposing a spectral clustering algorithm that can find communities in signed networks.
- developing an algorithm that efficiently computes the first (eigenvalue, eigenvector) pairs of the proposed Laplacian operator.

- building mathematical theory based on the stochastic block models that justifies why the proposed Laplacian operator performs better and why previous approaches have severe limitations.

The following sections will go through the main definitions (§2) and mathematical theories developed by the paper (§3). Then, the experiments performed will be described and their outcomes interpreted (§4). Finally, some limitations of the paper (§5) and other related work for finding communities in signed networks (§6) are discussed.

2 Signed graphs and spectral clustering

A weighted signed graph $G^{\pm}$ is defined as a pair $G^{\pm} = (G^+, G^-)$, where $G^+ = (V, W^+)$ describes the positive relations between the vertices while $G^- = (V, W^-)$ expresses the negative relations. W^+ and W^- are symmetric non-negative weight matrices and $V = \{v_1, v_2, \dots v_n\}$ denotes the set of vertices in the graph.

The Laplacian operator and its normalized version, defined on W^+, are given by:

$$L^+ = D^+ - W^+ \qquad\qquad L^+_{\text{sym}} = (D^+)^{-1/2} L^+ (D^+)^{-1/2} \tag{1}$$

where $D^+_{ii} = \sum_{j=1}^{n} w^+_{ij}$. Spectral clustering based on this Laplacian identifies clusters were there are more edges between elements within a cluster than between different clusters (assortative case).

The signless Laplacian, and its normalized version, defined on W^-, is obtained as:

$$Q^- = D^- + W^- \qquad\qquad Q^-_{\text{sym}} = (D^-)^{-1/2} Q^- (D^-)^{-1/2} \tag{2}$$

where $D^-_{ii} = \sum_{j=1}^{n} w^-_{ij}$. Spectral clustering based on the signless Laplacian finds clusters where there are more edges between clusters than between elements within a cluster. (disassortative case).

Mercado *et al.* propose to perform spectral clustering on signed networks using the normalized geometric mean Laplacian of $G^{\pm}$, defined as:

$$L_{GM} = L^+_{\text{sym}} \# Q^-_{\text{sym}} \tag{3}$$

where $A \# B = A^{1/2}(A^{-1/2} B A^{-1/2})^{1/2} A^{1/2}$ is the geometric mean of positive definite matrices A and B.

The algorithm used in the paper for finding k clusters in a signed network involves firstly computing the eigenvectors $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_k$ corresponding to the k smallest eigenvalues of L_{GM}. Then, k-means clustering is used on the rows of $U = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_k)$ to identify which cluster each vertex is part of.

The following Laplacian operators for clustering signed graphs that have been previously proposed [5, 6] and compared against L_{GM}:

$$L_{BR} = D^+ - W^+ + W^- \qquad L_{BN} = \bar{D}^{-1} L_{BR} \tag{4}$$

$$L_{SR} = \bar{D} - W^+ + W^- \qquad L_{SN} = \bar{D}^{-1/2} L_{SR} \bar{D}^{-1/2} \tag{5}$$

where $\bar{D}_i i = \sum_{j=1}^{n} w_{ij}^+ + w_{ij}^-$. L_{SR} is the arithmetic mean of L^+ and Q^-.

The paper also uses for comparison during evaluation the arithmetic mean of the normalized Laplacians:

$$L_{AM} = L_{\text{sym}}^+ + Q_{\text{sym}}^- \tag{6}$$

3 Mathematical ideas

Mercado *et al.* prove how the Laplacians using the arithmetic mean are based on the assumption that there exists a k-balance in $G^{\pm}$, namely that the set of vertices can be partitioned into k sets where there are only positive edges within a set and only negative edges between the sets [7]. A k-balance implies the presence of assortative connectivity in G^+ and disassortative connectivity in G^-. However, this assumption is unrealistic for real-world datasets, thus causing poor performance when using the arithmetic mean of Laplacians.

The proof is based on the following theorem: "Let $\mathbf{u}$ be an eigenvector of A and B with eigenvalues λ and μ, respectively. Then, $\mathbf{u}$ is an eigenvector of $A + B$ and $A\#B$ with eigenvalue $\lambda + \mu$ and $\sqrt{\lambda\mu}$, respectively."

The eigenvectors corresponding to small eigenvalues in L_{sym}^+ are related to assortative clusters in $G+$, while the eigenvectors corresponding to small eigenvalues in Q_{sym}^- are related to disassortative clusters in G^-. Therefore, based on the previously mentioned theorem, the ordering of the small eigenvalues in the arithmetic mean of Laplacians is influenced by having assortative clusters in G^+ and disassortative clusters in G^- which is unrealistic. However, the ordering of the small eigenvalues in the geometric mean only requires for one of the conditions to be satisfied, which means that there need to be either assortative clusters in G^+ or disassortative clusters in G^-. This model is less restrictive, thus making it more adequate for real-world datasets which may only exhibit clustering in one of G^+ or G^-.

Additionally, by using the Stochastic Block Model, the authors prove how L_{BN} and L_{SN} are based on properties of the graph structure that are not related to clustering. One such property is that the expected volume in the negative graph (G^-) needs to be smaller than the expected volume in the positive one (G^+). This does not need to be the case for L_{GM}.

Moreover, the paper also proposes an algorithm for efficiently computing the small eigenvalues and the corresponding eigenvectors for $L_{\text{sym}}^+ \# Q_{\text{sym}}^+$ that does not require the actual computation

of the geometric mean which involves inverting a matrix. The algorithm uses the Krylov-based inverse power method [8] but it shall not be further discussed as it is outside the scope of this report.

4 Evaluation

A theoretical evaluation of the algorithm proposed by Mercado *et al.* involves using Stochastic Block Models. The results of these models show how different probability for having positive/negative edges in W^+/W^- affects the clustering performance. This report explores some more parameter settings for the Stochastic Block Models and it also performs experiments are on two real-world datasets.

4.1 Experimental set-up

The spectral clustering algorithm proposed in the paper will be compared with variants involving swapping L_{GM} with the other possible Laplacian operators, namely L_{SN}, L_{BN}, L_{AM}, L^+_{sym} and Q^-_{sym}. State-of-the-art algorithms for clustering L_{SN} and L_{BN} also rely on using the smallest eigenvalues of the Laplacian operator and their corresponding eigenvectors. Therefore, the comparison will be meaningful for showing the shortcomings of these approaches.

Moreover, the authors use a small diagonal shift for L^+_{sym} and Q^-_{sym}, namely $L^+_{\text{sym}} + \epsilon_1 I$ and $Q^-_{\text{sym}} + \epsilon_2 I$, to ensure that these matrices are positive definite when computing the geometric mean. The paper gives very little details about the values of ϵ_1 and ϵ_2 used for experiments. It was interesting to notice that varying these parameters had a significant influence of the performance. The experiments in this report will use $\epsilon_1 = 0.1$ and $\epsilon_2 = 0.1$.

The evaluation metric used to assess the clustering performance in all of the situations is the normalised mutual information between the clustering labels obtained from spectral clustering and the true labels for the datasets used. This is possible because in all situations evaluated, the true labels will be known.

Given two cluster assignment $\mathcal{X}$ and $\mathcal{Y}$ of n data points, the mutual information ($\mathcal{MI}$) is obtained as follows:

$$\mathcal{MI}(\mathcal{X}, \mathcal{Y}) = \sum_{i=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{Y}|} P(i,j) \log \frac{P(i,j)}{P(i)P'(j)} \tag{7}$$

where $P(i) = |\mathcal{X}_i|/n$ is the probability that a data point picked at random from $\mathcal{X}$ is part of cluster $\mathcal{X}_i$, $P'(j) = |\mathcal{Y}_i|/n$ is the probability that a data point picked at random from $\mathcal{Y}$ is part of cluster $\mathcal{Y}_j$ and $P(i,j) = |\mathcal{X}_i \cap \mathcal{Y}_j|/n$ is the probability that a data point picked at random is part of both clusters $\mathcal{X}_i$ and $\mathcal{Y}_j$.

The normalized mutual information is given by:

$$\mathcal{NMI}(\mathcal{X}, \mathcal{Y}) = \frac{\mathcal{MI}(\mathcal{X}, \mathcal{Y})}{\sqrt{H(\mathcal{X})H(\mathcal{Y})}} \tag{8}$$

where $H(\mathcal{X})$ and $H(\mathcal{Y})$ are the entropies of each clustering assignment respectively.

The main advantages of using the normalized mutual information is that it makes no assumption on the clustering structure. This clustering score is bounded by $[0, 1]$, with 1 indicating agreement between the clustering labels and the true labels and with 0 indicating that the labellings compared are independent.

4.2 Stochastic Block Models

4.2.1 Signed graph creation

The Stochastic Block Model (SBM) is a benchmark generative model that can be used to study network clustering [9]. This method was used to generate signed graphs as follows: assume the network has k clusters, each with the same number of vertices. k and the number of vertices are parameters that can be modified. The condition used for determining if vertices v_i and v_j are part of the same cluster is whether $(i/k == j/k)$.

Then, the weight matrices were generated as follows:

- If vertices v_i and v_j are in the same cluster:

$$W^+_{i,j} = W^+_{j,i} = \begin{cases} 1, & \text{with probability } p^+_{\text{in}} \\ 0, & \text{with probability } 1 - p^+_{\text{in}} \end{cases} \tag{9}$$

- Otherwise,

$$W^+_{i,j} = W^+_{j,i} = \begin{cases} 1, & \text{with probability } p^+_{\text{out}} \\ 0, & \text{with probability } 1 - p^+_{\text{out}} \end{cases} \tag{10}$$

The negative weight matrix W^- is built similarly, by replacing p^+_{in} with p^-_{in} and p^+_{out} with p^-_{out}. By generating positive and negative weights in this way, the weight matrices in expectation $\mathcal{W}^+$ and $\mathcal{W}^-$ have the property that $\mathcal{W}^+_{i,j} = p^+_{\text{in}}$ and $\mathcal{W}^-_{i,j} = p^-_{\text{in}}$ if v_i and v_j belong to the same cluster, while $\mathcal{W}^+_{i,j} = p^+_{\text{out}}$ and $\mathcal{W}^-_{i,j} = p^-_{\text{out}}$ if v_i and v_j belong to different clusters.

4.2.2 Experiments using sthocastic block methods.

The experiments performed illustrate how the geometric mean of Laplacians can find clusters when either G^+ is assortative or G^- is disassortative, properties indicated by the weight matrices

in expectation $\mathcal{W}^+$ and $\mathcal{W}^-$. Conversely, the Laplacians based on the arithmetic means will only be able to find clusters when both conditions are satisfied.

In the first experiment, the probability p^+_{out} is varied such that G^+ changes from being assortative to having no clustering structure to being disassortative. The other probabilities are set to: $p^+_{\text{in}} = 0.4$, $p^-_{\text{in}} = 0.05$, $p^-_{\text{out}} = 0.4$. The results are illustrated in Figures 1. Since G^- is disassortative, Q^-_{sym} gives the best clustering score. While the change in G^+ severely affects the performance of L_{SN}, L_{BN} and L_{AM}, L_{GM} is capable to overcome this problem by identifying the disassortative clustering structure in G^-.

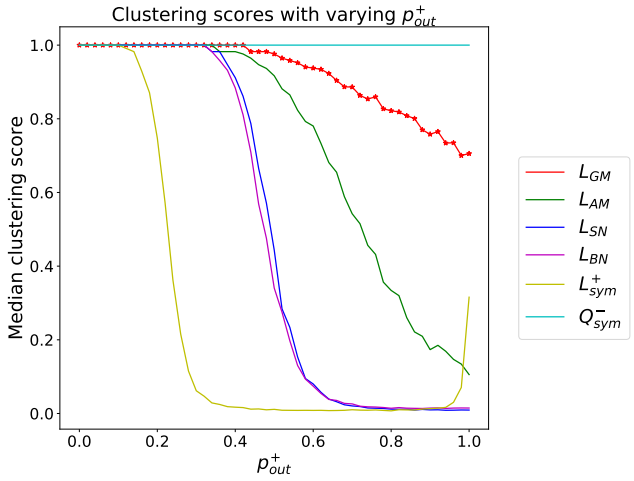


Figure 1: Median clustering scores for using each Laplacian operator under the stochastic block model after 50 runs.

Moreover, Figure 2 illustrates the structure of the weight matrices in expectation, namely $\mathcal{W}^+$ and $\mathcal{W}^+$ for two settings of p^+_{out}. The visualization of the weight matrices together with the results in Figure 1 can help us understand under which conditions each Laplacian can identify clustering structure.

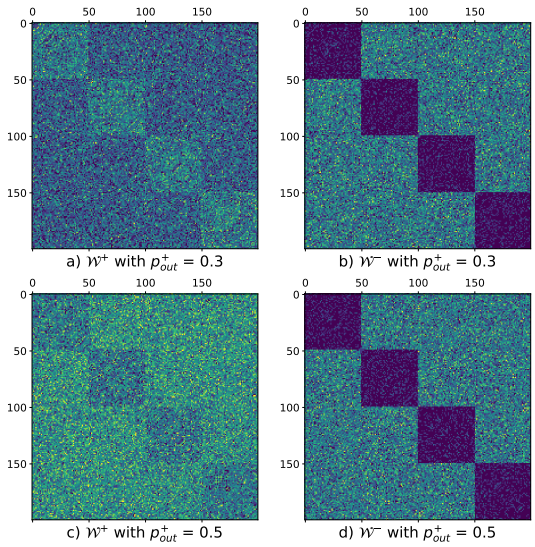


Figure 2: Weight matrices in expectation under the stochastic block model after 50 runs for two different settings of p^+_{out}. The darker colours indicate the absence of edges, while lighter colours indicate the presence of edges.

The second experiment is very similar to the first one and it involves varying p^-_{in} such that G^- changes from being disassortative to having no clustering structure to being assortative. The other probabilities are set to: $p^+_{\text{in}} = 0.3$, $p^+_{out} = 0.05$, $p^-_{\text{out}} = 0.2$.

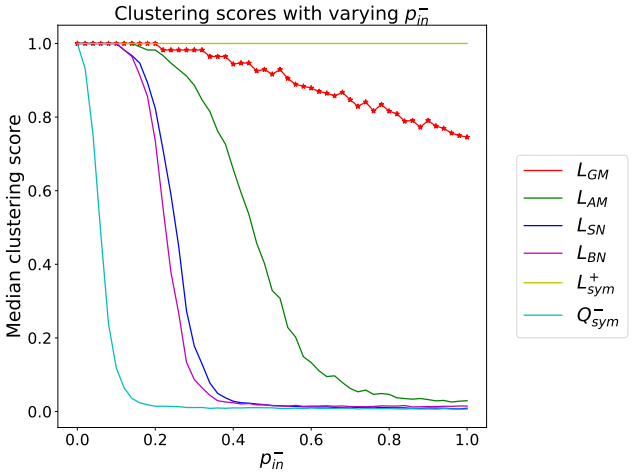


Figure 3: Median clustering scores for using each Laplacian operator under the stochastic block model after 50 runs.

The clustering scores are illustrated in 3 and, as before, they indicate that the Laplacian operators based on the arithmetic mean require G^- to have a disassortative structure in order to be able to identify the clusters. Figure 4 help us visualize the structure of the weight matrices under different parameter settings for p^-_{out}.

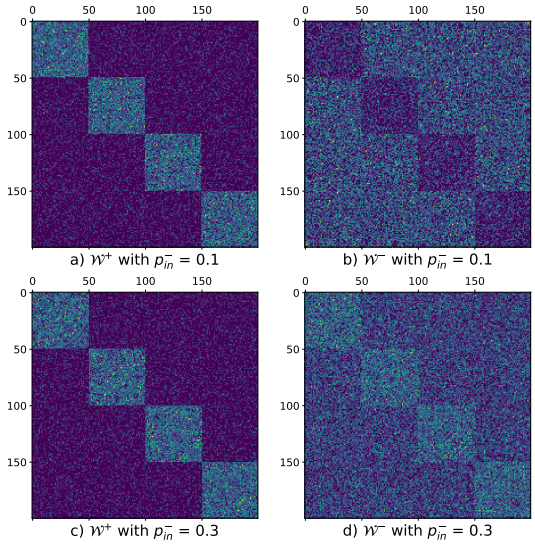


Figure 4: Weight matrices in expectation under the stochastic block model after 50 runs for two different settings of p^-_{in}. The darker colours indicate the absence of edges, while lighter colours indicate the presence of edges.

An additional experiment was set-up to determine whether the number of clusters affects the clustering score. In this situation, the parameters were set as follows: $p^+_{\text{in}} = 0.4$, $p^-_{\text{in}} = 0.1$, $p^+_{\text{out}} = 0.05$ and $p^-_{\text{out}} = 0.2$. Since $p^+_{out} < p^+_{\text{in}}$ and $p^-_{\text{in}} < p^-_{\text{out}}$ there are both assortative and disassortative clusters in expectation in G^+ and G^- respectively.

The results in Figure 5 indicate that, while all Laplacians are affected by the increase in the number of clusters, the clustering score of L_{SN} and L_{BN} starts decreasing the most as the number of clusters increases.

This behaviour can be explained by the fact that spectral clustering using L_{SN} and L_{BN} is based on the assumption that the expected volume in the negative graph (G^-) needs to be smaller than the expected volume in the positive one (G^+). More formally, this can be written as

$$p^-_{\text{in}} + (k-1)p^-_{\text{out}} < p^+_{\text{in}} + (k-1)p^+_{\text{out}} \tag{11}$$

For our setting of the probabilities, this condition holds only for a small values of k.

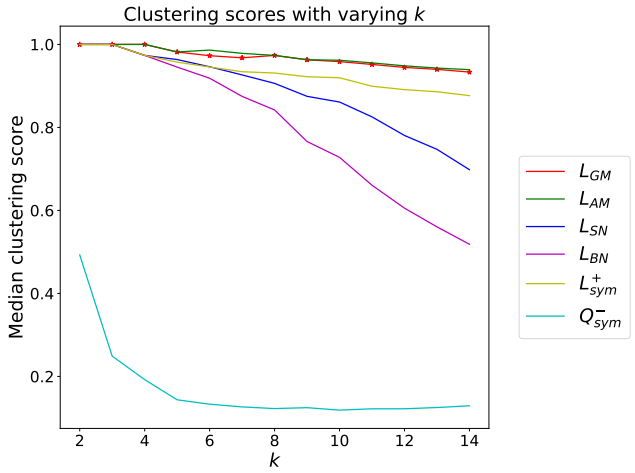


Figure 5: Median clustering scores for using each Laplacian operator under the stochastic block model after 50 runs.

4.3 Real-world datasets

Signed network clustering can be performed on real-world datasets by building the signed graph with the data points as the vertices and with positive edges added between similar points and negative edges added between dissimilar points. The approach presented in the paper is used in this sense. Thus, W^+ is obtained from the k^+-nearest neighbour graph, while W^+ is given by the k^--furthest neighbour. The parameters k^+ and k^- influence the clustering score and various settings for them will be analysed.

In order to construct both G^+ and G^- , k- nearest neighbours is performed, but with k and the distance metric depending on the situation. G^+ is obtained by setting $k = k^+$ and using as the distance between data points $\mathbf{x}_i$ and $\mathbf{x}_j$ the Euclidean distance $\text{euclidean}(\mathbf{x}_i, \mathbf{x}_j)$. G^- is built by setting $k = k^-$ taking as the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ the inverse Euclidean distance $1/(\text{euclidean}(\mathbf{x}_i, \mathbf{x}_j) + \epsilon)$, where ϵ was set to 0.001 to avoid division by zero.

An important element to notice is that the weight matrices obtained this way are not necessarily symmetric. Since the paper does does not account for this, the problem was addressed in this report by adding extra edges to ensure that the graph is undirected.

The real-world datasets for which signed network clustering is evaluated are the iris flowers dataset and the Olivetti faces dataset. The iris dataset consists 150 points, each having several features for three different types of iris flowers.The Olivetti faces dataset, consists of 10 face images, of size 64×64 for 40 different people.

4.3.1 Experiments using real-world datasets

For both datasets we computed clustering scores, by selecting the parameters

$$k^+, k^- \in \{3, 5, 8, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120\} \tag{12}$$

thus evaluating 225 settings. Table 1 shows the results where using each Laplacian achieves the best clustering error. While using the L_{GM} gives best clustering scores in more cases than other Laplacians do, there does not seem to be a significant difference.

	Iris flowers	**Olivetti faces**
# vertices	150	400
# classes	3	40
L_{BN}	0.04	0.13
L_{SN}	0.30	0.23
L_{AM}	0.31	0.33
L_{GM}	0.39	0.37

Table 1: Fraction of cases where each Laplacian operator achieves best clustering score.

Moreover, a grid search was performed to identify which parameters k^+ and k^- give the best clustering score. For both the iris and the Olivetti datasets, $k^+ = 10$ works best with the Laplacian operators compared. In this sense, k^- was varied to observe how the clustering error is affected.

For the iris dataset, the performance for using the different Laplacians does not seem to be considerably different (Figure 6). This behaviour could be explained by the fact that, in this situation G^+ has assortative structure, while G^- has disassortative structure, which makes all of the Laplacians suitable for identifying clusters.

Conversely, for the Olivetti faces dataset (Figure 7), L_{GM}, L_{AM} and L^+_{sym} give the best clustering score, while the other Laplacians have consistently lower performance. The fact that the performance of L_{GM} and L_{AM} is better when there is a large number of clusters, as it is the case of the Olivetti dataset, was also confirmed by the Stochastic Block Models.

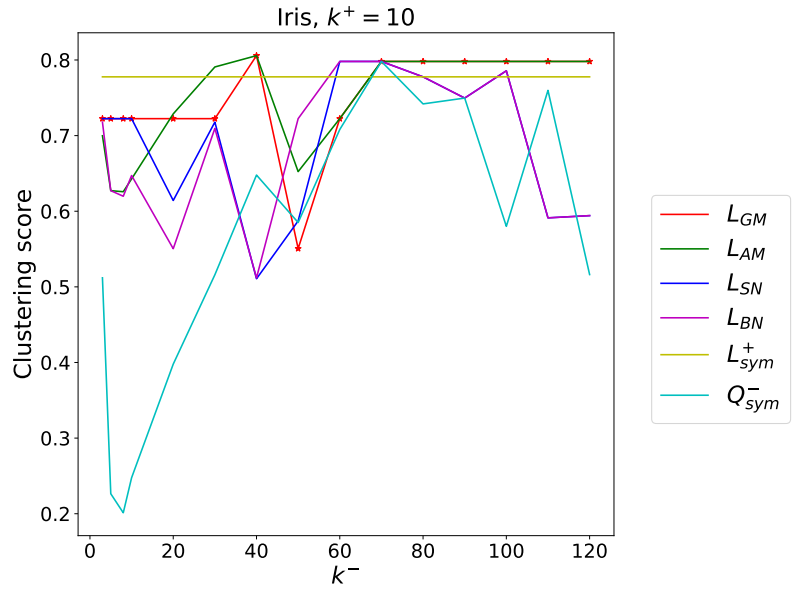


Figure 6: Clustering scores for using each Laplacian operator on the iris dataset.

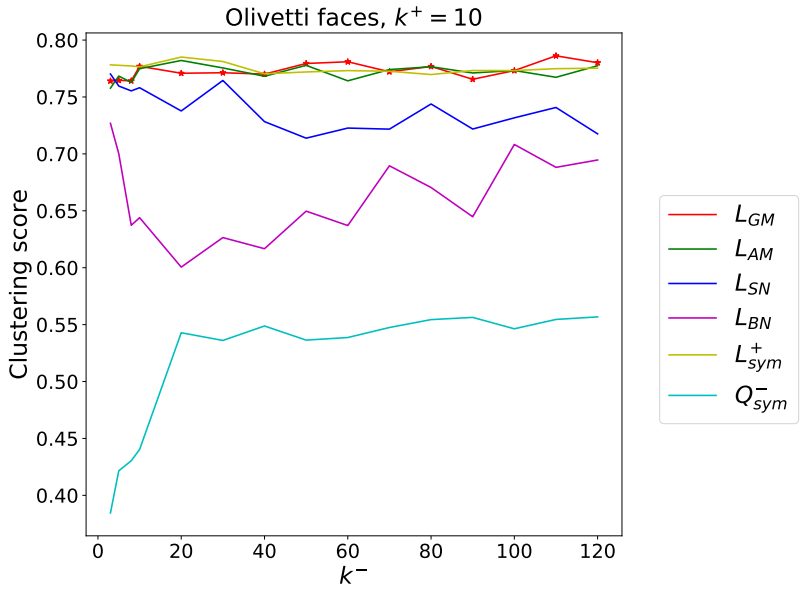


Figure 7: Clustering scores for using each Laplacian operator on the Olivetti faces dataset.

5 Limitations

The results indicate that L_{GM} can find clusters in many situations where other Laplacians fail. However, the scores achieved in the situations when both G^+ is assortative and G^- is disassortative does not seem to be much better than the ones obtained by using L_{AM} or L_{SN}.

Another important factor that should be taken into account is the computational complexity of the different Laplacians. While the paper proposes an efficient algorithm for computing the smallest eigenvalues and the associated eigenvectors for L_{GM}, the execution time for performing the same computation for L_{SN} and L_{BN} is still shorter.

6 Related work

Developing algorithms for clustering signed networks represents an important research topic. Chiang *et al.* [10] proposed a method that involves initially performing low-rank matrix completion on the adjacency (weight) matrix and then cluster the eigenvectors corresponding to the smallest eigenvalues. Interestingly enough, the authors showed how this model worked better than using the signed Laplacian developed by Kunegis *et al.* [6].

Some other methods that do not involve spectral clustering have also been used for finding communities in signed networks. For example, Bansal *et al.* [11] modelled this task as a correlation clustering problem where a positive edge indicates that the nodes it connects are similar, while a negative edge indicates that the nodes it connects are different.

An important extension to the work presented in this report involves clustering signed directed networks [12]. In this sense, Li *et al.* [13] have developed a spectral clustering method for directed graphs that is based on a graph partitioning algorithm.

7 Conclusion

The task of finding clusters in signed networks has received significant importance due to its application in social balance and social mining. This report has analysed how the geometric mean of Laplacians, proposed by Mercado *et al.*, can be used to identify community in signed networks, especially in situations where other proposed Laplacians fail to do so. Despite some of the limitations of using the geometric mean of Laplacians, including that it only works on undirected graphs and that it is still slower to use than the Laplacians it was compared against, it represents a good stepping stone for developing powerful data mining techniques for signed networks.

References

[1] P. Mercado, F. Tudisco, and M. Hein, "Clustering signed networks with the geometric mean of laplacians," in *Advances in Neural Information Processing Systems*, pp. 4421–4429, 2016.

[2] J. Tang, Y. Chang, C. Aggarwal, and H. Liu, "A survey of signed network mining in social media," *ACM Computing Surveys (CSUR)*, vol. 49, no. 3, p. 42, 2016.

[3] F. Harary *et al.*, "On the notion of balance of a signed graph.," *The Michigan Mathematical Journal*, vol. 2, no. 2, pp. 143–146, 1953.

[4] D. Cartwright and F. Harary, "Structural balance: a generalization of heider's theory.," *Psychological review*, vol. 63, no. 5, p. 277, 1956.

[5] K.-Y. Chiang, J. J. Whang, and I. S. Dhillon, "Scalable clustering of signed networks using balance normalized cut," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 615–624, ACM, 2012.

[6] J. Kunegis, S. Schmidt, A. Lommatzsch, J. Lerner, E. W. De Luca, and S. Albayrak, "Spectral analysis of signed graphs for clustering, prediction and visualization," in *Proceedings of the 2010 SIAM International Conference on Data Mining*, pp. 559–570, SIAM, 2010.

[7] J. A. Davis, "Clustering and structural balance in graphs," *Human relations*, vol. 20, no. 2, pp. 181–187, 1967.

[8] V. Druskin and L. Knizhnerman, "Extended krylov subspaces: approximation of the matrix square root and related functions," *SIAM Journal on Matrix Analysis and Applications*, vol. 19, no. 3, pp. 755–771, 1998.

[9] K. Rohe, S. Chatterjee, B. Yu, *et al.*, "Spectral clustering and the high-dimensional stochastic blockmodel," *The Annals of Statistics*, vol. 39, no. 4, pp. 1878–1915, 2011.

[10] K.-Y. Chiang, C.-J. Hsieh, N. Natarajan, I. S. Dhillon, and A. Tewari, "Prediction and clustering in signed networks: a local to global perspective," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1177–1213, 2014.

[11] N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," *Machine Learning*, vol. 56, no. 1-3, pp. 89–113, 2004.

[12] F. D. Malliaros and M. Vazirgiannis, "Clustering and community detection in directed networks: A survey," *Physics Reports*, vol. 533, no. 4, pp. 95–142, 2013.

[13] Y. Li, X. Wu, and A. Lu, "On spectral analysis of directed signed graphs," *arXiv preprint arXiv:1612.08102*, 2016.