LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-CONF-754816

Investigation of Spectral Clustering for Signed Graph Matrix Representations

A. Fox, G. Sanders, A. Knyazev

July 17, 2018

IEEE HPEC conference
Waltham, MA, United States
September 25, 2018 through September 27, 2018

Disclaimer

Investigation of Spectral Clustering for Signed Graph Matrix Representations

1st Alyson Fox
Lawrence Livermore National Lab
Center for Applied Scientific Computing
Livermore, CA
fox33@llnl.gov

2nd Geoffrey Sanders
Lawrence Livermore National Lab
Center for Applied Scientific Computing
Livermore, CA
sanders29@llnl.gov

3rd Andrew Knyazev
University of Colorado Denver
Department of Mathematical & Statistical Sciences
Denver, CO
andrew.knyazev@ucdenver.edu

***Abstract*—Signed graphs [11], which allow for both favorable and adverse relationships, are becoming a common model choice for various data analysis applications, e.g., correlation clustering [1] and spectral clustering [9]. Unlike for unsigned graphs, there is no collective agreement of a matrix representation that relates to the unsigned graph Laplacian. There currently exists three proposed matrix representations: [9] proposes a zero row-sum preserving Laplacian (signed Laplacian), [8] proposes a physics preserving Laplacian (Physics Laplacian), and [4] proposes an expansion of the signed Laplacian into an unsigned Laplacian with twice the number of degrees-of-freedom (Gremban's expansion.) We investigate these three proposed matrix representations with respect to the quality of traditional (unsigned graph) spectral clustering concepts. We provide three numerical tests that use a stochastic block model with negative edge weights. We observe that the best matrix representation for spectral clustering depends on the underlying structure of a signed graph, which may be unavailable. However, since the Gremban's expansion matrix provides higher quality clusters for more combinations of inner and outer-connection probabilities for positive and negative valued edges, we conclude that the Gremban's expansion is the most robust representation to use for spectral clustering.**

***Index Terms*—signed graphs, Laplacian, spectral embedding, spectral clustering, Gremban's expansion**

I. Introduction

Spectral clustering [12], a form of graph partitioning, groups or clusters the vertices using spectral information of matrices associated with the graph. An unsigned graph assumes that the connection between vertices are positive (or favorable) and the spectrum of the associated adjacency matrix and Laplacian are well known. A classic spectral clustering approach for unsigned graphs is to use the Feidler vector [3], the second smallest eigenvector associated with the unsigned graph Laplacian. The values of the Fiedler vector identify two partitions for the graph vertices.

Intuitively, a positive edge in a graph indicates a similarity, and a negative edge indicates dissimilarity. A simple example of a signed graph are friend-enemy social networks [11], where positive edges represent "friends" and negative edges represent "foes." The negative edge weights provide extremely valuable information, that if ignored in a spectral clustering algorithm, may result in a poor clustering ([5], [9], [10]). Currently in the literature, there have been three proposed signed graph matrix representations that relate to the unsigned graph Laplacian, the signed Laplacian, the Physics Laplacian, and the Gremban expansion matrix.

The signed Laplacian, introduced in [9], is a sign-variant of the unsigned Laplacian. Each diagonal elements of the unsigned Laplacian is the sum of the off-diagonal elements in each row. The signed Laplacian, instead, sums the absolute value of each off-diagonal element, retaining a positive-semidefinite matrix where the eigenvalues remain non-negative. The Physics Laplacian, introduced in [8], argues that the original definition of the graph Laplacian (the unsigned Laplacian) is a more natural choice based on spectral clustering techniques for a classical model of a mass-spring system with some springs having negative stiffness. Based on the numerical results with respect to spectral clustering, the eigenvectors associated with the smallest, possibly negative, eigenvalues produce clusters of a higher quality than the signed Laplacian. Often, for unsigned graphs the normalized Laplacian is used for spectral clustering, however, since the degree matrix associated with the Physics Laplacian maybe singular, the normalized signed variants of each representation will not be discussed. Lastly, the Gremban expansion matrix, introduced in [6] and studied specifically for signed graphs in [4], expands the signed Laplacian into an unsigned Laplacian with twice the number of degrees-of-freedom. It is argued, that even though the representation is double the size, all the traditional concepts of an unsigned graph can easily be generalized to the signed case. The signed Laplacian is the most commonly used signed graph representation, however, [8] and [4] make compelling arguments for alternative representations. Thus, depending on the application, a different representation may be more beneficial. In this paper we empirically investigate each of these matrix representations with respect to spectral clustering.

The remainder of the paper is organized as follows. Section II briefly reviews basic spectral clustering concepts and the proposed signed graph matrix representations. Section III presents the numerical results. Lastly, Section IV, concludes that the Gremban's expansion results in the most robust representation for spectral clustering on a block stochastic model with negative edge weights.

II. Background

A. Short discussion of spectral clustering

An unsigned, undirected graph, $\mathcal{G}(\mathcal{V};\mathcal{E};w)$, relates a set of n vertices, $\mathcal{V}$, by m connections or edges in the set $\mathcal{E}$ with strictly positive weights, $w > 0$. An edge $(i,j) \in \mathcal{E}$ between two vertices i and j is undirected, meaning (j,i) is also in $\mathcal{E}$, such that $w_{ij} = w_{ji}$. The unsigned Laplacian is represented as $L_u = D - A$, where the adjacency matrix, A, and degree matrix, D, are defined as

$$A_{ij} = \begin{cases} w_{ij} > 0 & (i,j) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}$$

and $D_{ii} = \sum_j A_{ij}$.

On a graph, $\mathcal{G}$, the graph clustering problem is defined by grouping the vertex set $\mathcal{V}$ into non-overlapping subgroups. It is well known that spectral clustering is a form graph partitioning. A common optimization problem for graph partitioning is

$$\min Ratiocut(\mathcal{V}_1, \cdot, \mathcal{V}_k),$$

where

$$Ratiocut(\mathcal{V}_1, \cdots, \mathcal{V}_k) = \sum_i \frac{cut(\mathcal{V}_i, \bar{\mathcal{V}}_i)}{|\mathcal{V}_i|},$$

$cut(\mathcal{V}_l, \mathcal{V}_h) = \sum_{i \in \mathcal{V}_l, j \in j \in \mathcal{V}_h} w_{ij}$, and $\mathcal{V}_i$ are subsets of $\mathcal{V}$. The goal of the minimization problem is to find a partition of the graph such that the edges between different groups have a very low weight. For $k = 2$ the minimization problem can be relaxed to

$$\min_{\boldsymbol{v}} \boldsymbol{v}^t L_u \boldsymbol{v} \text{ subject to } \boldsymbol{v}^t \mathbf{1} = 0, \text{ where } v_i \in \{-1, 1\} \qquad (1)$$

which can be further relaxed to an eigenvalue problem [12]. More formally, spectral clustering uses spectral information to perform dimensionality reduction, then the vertex set may be partitioned in fewer dimensions. The typical strategy, for a well-connected unsigned graph, involves computing k eigenvectors associated with the k smallest non-negative eigenvalues of the Laplacian, i.e.,

$$LV_k = V_k \Lambda_k$$

where the eigenvectors are the columns of V_k and the associated k smallest eigenvalues in the diagonal entries of Λ_k. Then a clustering algorithm is applied to a spectral embedding of the nodes defined by the rows of V_k, i.e., $V_k^t \boldsymbol{e}_i \quad \forall \quad i$, where $\boldsymbol{e}_i$ is the canonical basis vector.

For the purposes of this paper we will use a clustering algorithm developed in [2] that is based on a column-pivoted QR factorization. The algorithm is quite simple, the QR factorization with column-pivoting of V_k is found, such that

$$V_k \Pi = QR,$$

where Π is a the permutation matrix. The number of eigenvectors used in the spectral embedding, k, is typically chosen by the eigengap, i.e., the largest difference between two successive eigenvalues. Then each vertex is assigned to a cluster based on the index of the largest value in each row, i.e.,

$$c_j = \underset{i}{\operatorname{argmax}}(Q_{ji}).$$

The actual algorithm used in for this paper is slightly modified, using a polar factorization on the first k columns found from the column pivoting so that there is no preferential treatment to various vertices based on the order of computations of the QR factorization. For more details, please see [2].

B. Signed Graph Matrix Representations

For a signed, undirected graph, $\mathcal{G}(\mathcal{V}, \mathcal{E} = \mathcal{E}_+ \cup \mathcal{E}_+)$, the edge weights are signed meaning $w_{ij} \neq 0$, where positive edges in the positive edge set ($\mathcal{E}_+$), $w_{ij} > 0$, represent "friends" and negative edges in the negative edge set ($\mathcal{E}_-$), $w_{ij} < 0$, represent "foes." Define the negative-valued adjacency matrix as

$$(A_-)_{ij} = \begin{cases} w_{ij} & (i,j) \in \mathcal{E} \text{ and } w_{ij} < 0 \\ 0 & \text{otherwise} \end{cases}$$

Similarly define the positive-valued adjacency matrix as

$$(A_+)_{ij} = \begin{cases} w_{ij} & (i,j) \in \mathcal{E} \text{ and } w_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

The signed adjacency matrix can then be represented as the sum of the positive and negative-valued adjacency matrix, $A = A_+ + A_-$. Define the diagonal degree matrices $(D_p)_{ii} = \sum_j A_{ij}$ and $(D_s)_{ii} = \sum_j |A_{ij}|$. We can then define the *Physics Laplacian*, *Signed Laplacian*, and the *Unsigned Laplacian* as

$$L_p = D_p - A, \quad L_s = D_s - A, \text{ and } L_u = D_s - A_+ + A_-,$$

respectively. Lastly, define the Gremban's expansion matrix as

$$G = \begin{bmatrix} D_s - A_+ & A_- \\ A_- & D_s - A_+ \end{bmatrix}.$$

The signed Laplacian is the most studied of the three representations and has many nice theoretical properties. It is well known that if the signed graph is balanced, meaning the vertex set $\mathcal{V}$ can be partitioned into two groups such that the edges connecting the two groups are strictly negative edges, then the spectrum of the signed Laplacian maps directly to the spectrum of the unsigned Laplacian ([13], [9]). This can easily be seen if we let $\boldsymbol{y}$ be a bipartition that partitions the vertices of a balanced signed graph into two groups, then $L_u = diag(\boldsymbol{y}) L_s diag(\boldsymbol{y})$, where $diag(\boldsymbol{y})$ is a diagonal matrix with the components of the vector $\boldsymbol{y}$ along the diagonal. The eigenvector associated with the signed Laplacian are thus a signed variant of the eigenvectors of the unsigned Laplacian. The signed Laplacian also remains positive semi-definite as one can see by the quadratic form

$$\boldsymbol{x}^t L_s \boldsymbol{x} = \sum_{\substack{(i,j) \in \mathcal{E} \\ w_{ij} > 0}} w_{ij} (\boldsymbol{x}_i - \boldsymbol{x}_j)^2 + \sum_{\substack{(p,q) \in \mathcal{E} \\ w_{pq} < 0}} |w_{pq}| (\boldsymbol{x}_p + \boldsymbol{x}_q)^2,$$

ensuring that the spectrum of the signed Laplacian remains non-negative. If the graph is unbalanced, then the signed Laplacian is strictly positive-definite [9].

Using simple signed graph examples, [8] demonstrates that for spectral clustering there is an advantage to using the Physics Laplacian over the signed Laplacian. For example, when using spectral information associated with the signed Laplacian for a "dumbbell" graph - a graph with two fully connected cliques each with six vertices, connected internally with positive edges and connected together by two positive and two negative edges - the smallest nontrivial eigenvector associated with the signed Laplacian is unable to fully capture one of the cliques. From the definition of the Physics Laplacian, L_p, it might be possible to have zero diagonal entries in the degree matrix, breaking the traditional concept of the Laplacian of having non-negative eigenvalues. One can see, using the quadratic form of the Physics Laplacian,

$$\boldsymbol{x}^t L_p \boldsymbol{x} = \sum_{\substack{(i,j) \in \mathcal{E} \\ w_{ij} > 0}} w_{ij} (\boldsymbol{x}_i - \boldsymbol{x}_j)^2 + \sum_{\substack{(p,q) \in \mathcal{E} \\ w_{pq} < 0}} w_{pq} (\boldsymbol{x}_p - \boldsymbol{x}_q)^2,$$

that the Physics Laplacian is no longer guaranteed to be positive semi-definite. Based on the numerical results and the mass-spring model, [8] argues that one should use the eigenvectors associated with the k smallest, possibly negative, eigenvalues. For the "dumbbell" graph, the Physics Laplacian, using the eigenvector associated with the most negative eigenvalue, perfectly clusters the two cliques.

The Gremban's expansion, first introduced in [6], was first applied to signed graphs in [4] for solving linear systems involving the signed Laplacian, i.e., $L_s \boldsymbol{x} = \boldsymbol{b}$. The expansion decomposes any diagonally dominant matrix, into a diagonally dominant Z-matrix and a nonnegative matrix. One could then solve $L_s \boldsymbol{x} = \boldsymbol{b}$ by solving the larger system

$$G \boldsymbol{w} = \begin{bmatrix} D_s - A_+ & A_- \\ A_- & D_s - A_+ \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ -\boldsymbol{x} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ -\boldsymbol{b} \end{bmatrix} = \boldsymbol{z},$$

where G is an unsigned graph Laplacian and is amendable to traditional graph Laplacian linear system solvers. Although, not intentionally studied for its spectral properties, the Gremban's expansion of the signed Laplacian has many useful properties that relate to the unsigned Laplacian. The graph of the Gremban's expansion matrix can be seen as an unsigned graph with twice the number of vertices. The vertices form two distinct groups: those connected internally by edges that relate to the positive-valued edges of the original signed graph and those connected to each other by edges that relate to the original negative-valued edges, as seen in Figure 1.

A well known random walk theorem for unsigned graphs was generalized to signed graphs via the Gremban's expansion matrix. For an unsigned graph the kth power of the associated adjacency matrix defines the number of k-length walks connecting the vertices i and j. The k powers of the associated binary adjacency matrix of the Gremban's expansion graph defines an even or odd number of negative-valued edges in a walk of length k between node i and j

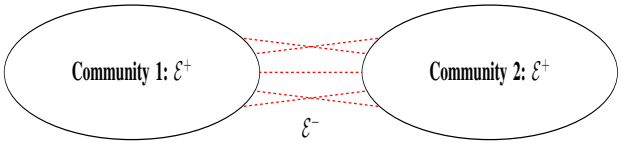


Fig. 1: The graph of Gremban's expansion matrix is an unsigned graph with twice the number of vertices. Each community is connected internally by edges that are associated with the positive edges of the original signed graph. The two communities are connected to each other by edges that are associated with the negative edges of the original signed graph.

[4]. This theorem is simply an extension of the classical phrase "an enemy of my enemy is my friend" and relates to the minimization problem in Equation (1) since

$$\boldsymbol{e}_j L_s^\dagger \boldsymbol{e}_i = \begin{bmatrix}\boldsymbol{e}_j \mathbf{0}\end{bmatrix} G^\dagger \begin{bmatrix}\boldsymbol{e}_i \\ \mathbf{0}\end{bmatrix} - \begin{bmatrix}\boldsymbol{e}_j \mathbf{0}\end{bmatrix} G^\dagger \begin{bmatrix}\mathbf{0} \\ \boldsymbol{e}_i\end{bmatrix}$$

can be seen as a difference between a friendly score and an unfriendly score. Another useful property of the Gremban's expansion matrix, as shown in [4], is that the spectrum of G is the union of the spectrum of L_s and L_u, i.e.,

$$\sigma(G) = \sigma(L_u) \cup \sigma(L_s).$$

If $(\lambda, \boldsymbol{v})$ and $(\mu, \boldsymbol{u})$ are eigenpairs of L_s and L_u, respectively, then $(\lambda, \begin{bmatrix}\boldsymbol{v}, -\boldsymbol{v}\end{bmatrix}^t)$ and $(\mu, \begin{bmatrix}\boldsymbol{u}, \boldsymbol{u}\end{bmatrix}^t)$ are eigenpairs of G. The union spectrum theorem is useful if the signed graph has mostly unsigned structure, then the spectral embedding will use the eigenvector associated with unsigned Laplacian. The eigenvectors of the Gremban's expansion has twice the number of elements than necessary for the spectral embedding. If only the first n elements of each eigenvector are used for the QR clustering algorithm, then from the spectral union theorem proved in [4] we know that the first n elements of the eigenvectors are the eigenvectors associated with the signed or unsigned Laplacian. The eigengap for G should be twice the number of clusters however, this is not guaranteed. Thus, we will instead use the true eigengap associated with the Gremban's expansion matrix. The QR clustering algorithm will cluster the larger graph associated with the Gremban's expansion matrix that has $2n$ vertices. The result that is used to assess the clustering will only use the clusters that are associated with the first n vertices.

In the next section, we will investigate using the QR spectral clustering algorithm discussed in Section II-A with each of the three proposed representations eigenvectors. Using classical block stochastic graphs we hope to gain an understanding of which representation provides eigenvectors that are robust to the underlying sign structure of the signed graph.

III. Numerical Results

We measure the accuracy of a proposed clustering by *pairwise-recall* and *pairwise-precision* [7]. Both precision and recall consider every pair of vertices. Precision measures the fraction of the predicted pairs that match the ground truth while recall measures the fraction of ground truth pairs that are correctly represented in the predictions, i.e.,

$$\text{precison} = \frac{\#\text{true-positives}}{\#\text{true-positives} + \#\text{false-positives}}$$

and

$$\text{recall} = \frac{\#\text{true-positives}}{\#\text{true-positives} + \#\text{false-negatives}}.$$

As discussed in Section II-A, we chose to use a QR clustering method described in [2], using the largest eigengap of each representation to determine the dimension of the embedding. For the Physics Laplacian and the signed Laplacian, the number of eigenvalues found to determine the eigngap is the number of ground truth clusters plus ten extra eigenvalues. For the Gremban's expansion we found twice the number of ground truth clusters plus ten extra eigenvalues. Let k be the index of the eigengap, then for all signed graph representations, the eigenvectors associated with the k smallest eigenvalues are used in the QR clustering method. Since the Gremban's expansion matrix has twice the degrees of freedom, only the clusters associated with the first n vertices are used to evaluate the performance.

A. Test 1:

The first numerical example is designed to investigate how the distribution of positive and negative edges effect the clustering for various proposed matrix representations. Using a stochastic block model with k communities, each with n vertices, we let p be the probability of an inner-block connection. If there exists a connection then there is a probability p_{neg} that the connection has a negative weight. The outer-block connections are formed in a similar way, q is the probability of an outer-block connection with probability q_{neg} that the connection has a negative weight. For the sake simplicity, the graph weights are chosen as -1 or 1, respectively. Let $k = 2$ and $n = 100$. For each combination of (p, q, p_{neg}, q_{neg}) we performed ten trials and displayed the average precision and recall. Since $k = 2$, the eigenvector associated with the smallest nontrivial eigenvalue is used to distinguish the clusters. This is a relativity simple problem that depends on the probabilities of the positive and negative connections. Typically, for an unsigned graph, if $p \gg q$ then the communities are easy to distinguish with the eigenvector associated with the smallest nontrivial eigenvalue. We would expect, if $p > q$ and $q_{neq} > p_{neg}$, any of the signed matrix representations to be able to distinguish the clusters using the eigenvector associated their smallest nontrivial eigenvalue. We would also expect that for any p and q if there are a large enough number of negative-value connections between two communities, i.e., $q_{neg} \gg p_{neg}$, the clustering algorithm should be able to provide high quality clusters.

Figure 2 displays the difference in recall and precision using the eigenvectors associated with the Gremban's expansion matrix and the Physics Laplacian. For various pairs of (p, q), the probability p_{neg} and q_{neg} are varied. The figures in the left column display the difference in recall, while the figures in the right column display the difference in precision. The value in each block is the precision (or recall) with respect to the clustering using the eigenvalues of the Gremban's expansion matrix. For each sub-figure, the x-axis and y-axis vary the values of p_{neq} and q_{neg}, respectively.

The top row is when $(p, q) = (0.1, 0.1)$, the inner and outer-block connection probabilities are the same and relatively low. For these particular values, if only positive edges exists in the graph, then the classic spectral clustering techniques would be unable to distinguish between the two communities. Rationally, if there exists proportionally more negative edges in the outer-blocks than the inner-blocks then we should be able to distinguish the two communities, thus, only the values were $q_{neg} \geq p_{neg}$ are considered. Otherwise, the ground-truth of the block-stochastic matrix is no longer valid. In most combinations, the Gremban's expansion resulted in a higher recall. If $q_{neg} \gg p_{neg}$ then the Gremban's expansion resulted in better precision. When $q_{neg} \approx p_{neg}$ both representations had poor precision, with the Physics Laplacian having slightly better precision values. However, one could argue that we should not be able to distinguish the clusters when $q_{neg} \approx p_{neg}$.

For the second row $(p, q) = (0.1, 0.9)$, the inner-block connection probability is low while the outer-block connection probability is large. In this case, only values where $q_{neg} \geq 0.5$ are considered, otherwise the ground truth is invalid. In this case the clustering that used the Gremban's expansion eigenvector resulted in near perfect precision and recall, while the Physics Laplacian struggled to produce high precision and high recall for all combinations.

For the third row $(p, q) = (0.9, 0.1)$, the inner-block connection probability is large while the outer-block connection probability is

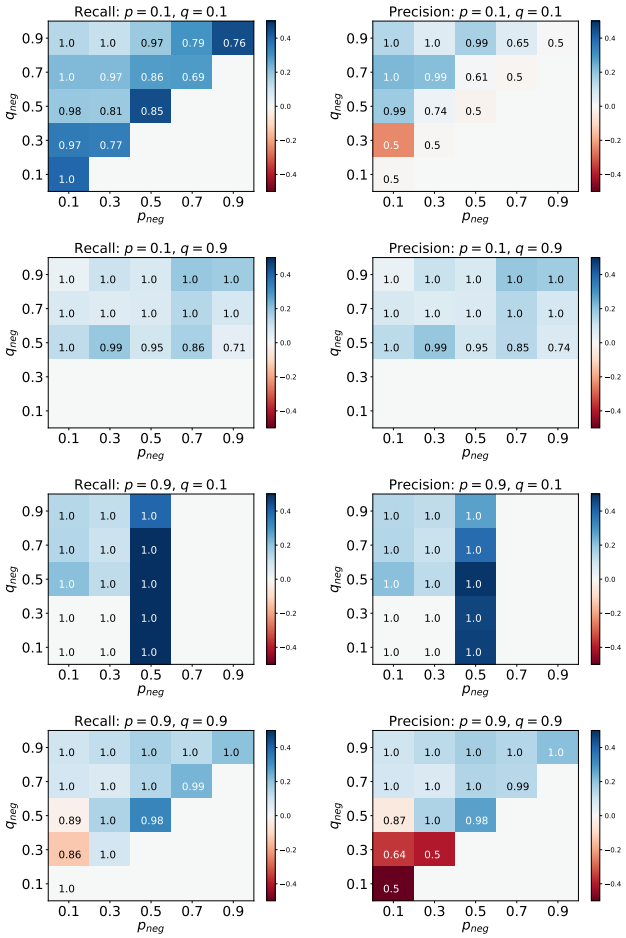


Fig. 2: Difference in recall (left column) and precision (right column) for spectral clustering using the eigen-embedding of eigenvectors associated with the Gremban's expansion matrix and the Physics Laplacian. The printed value in each block is the precision (or recall) with respect to the clustering using the Gremban's expansion matrix. The color displays by how much the precision (or recall) of the Gremban's expansion clustering is larger than that of the Physics Laplacian clustering, e.g., blue indicates that the Gremban's expansion clustering outperforms Physics Laplacian clustering, and red otherwise.

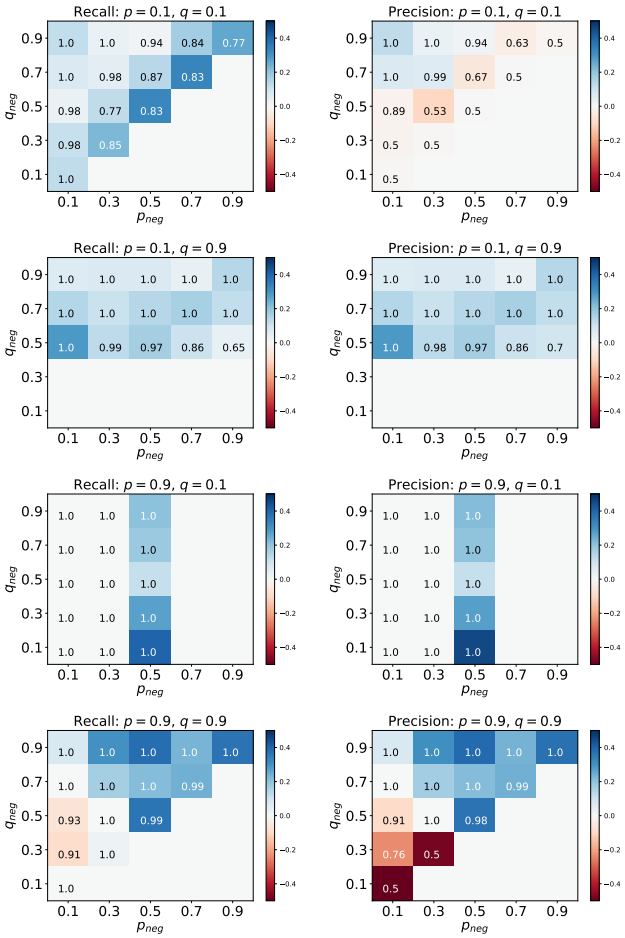


Fig. 3: Difference in recall (left column) and precision (right column) for spectral clustering using the eigen-embedding of eigenvectors associated with the Gremban's expansion matrix and the signed Laplacian. The printed value in each block is the precision (or recall) with respect to the clustering using the Gremban's expansion matrix.

small, thus only $p_{neg} \leq 0.5$ is considered. In this case, both representations had near perfect recall when p_{neg} was small. For high p_{neg}, the Gremban's expansion clearly out performed the Physics Laplacian. However, again for the case when $p_{neg} \approx 0.5$, one could argue that the ground truth is no longer valid.

Lastly, for the last row $(p, q) = (0.9, 0.9)$, the inner and outer-block connection probability are both large, thus only $q_{neg} \geq p_{neg}$ is considered. Similar to previous results, the Physics Laplacian resulted in better recall and precision for when q_{neg} and p_{neg} is small. Otherwise, the Gremban's expansion produced higher quality clusters.

The same test is presented in Figure 3, but displaying the difference between recall and precision when using eigenvectors associated with the Gremban's expansion matrix and the signed Laplacian. Similar conclusions can be drawn. When $p \approx q$ and p_{neg} and q_{neg} are both low, the Gremban's expansion results is lower precision and recall than if the signed Laplacian is used. When the inner and outer-connection probability is equal, both representations have poor precision when $q_{neg} \approx p_{neg}$. Otherwise, the Gremban's expansion better captures the two clusters than the signed Laplacian and has near perfect precision and recall for most combinations of p, q, p_{neg}, and q_{neg}. Thus, we can conclude for this small block stochastic test, that the eignevalue associated with the Gremban's expansion matrix resulted in a robust method. However, it should be noted that there are combinations of p, q, p_{neg}, and q_{neg}, where the signed or Physics Laplacian resulted in higher quality clusters.

B. Test 2:

The following tests consider graphs from the IEEE HPEC Streaming Graph Challenge Stochastic Block Partition data sets [7]. Each graph using a stochastic block model with unsigned weights with a known ground truth. They are static graphs and have non-overlapping communities. Since these graphs were created for only positive edge weights, the inner and outer-block probabilities are known to be $p > q$. Since the inner-block connection probability is higher than the outer-block connection probability, only values of $p_{neg} \leq 0.5$ are considered, otherwise, the standard clustering concepts do not apply. To create a signed graph, we will deliberately switch the sign of some of the edges at random. For a particular value of p_{neg} and q_{neg}, if there exists an edge, then the sign of the edge is switched based on the respective negative edge probability. We again will present the difference in precision and recall for spectral clustering using the

eigenvectors of the various matrix representation as the probabilities p_{neg} and q_{neg} vary. For each combination of p_{neg} and q_{neg}, ten random trials are performed and the average recall and precision values are presented.

Figure 4 presents the results with respect to the 50 node graph. The left column presents the difference with respect to recall and the right column presents the difference with respect to precision. The top row is the difference in recall (or precision) of the clustering using the eigenvalues of the Gremban's expansion matrix and the Physics Laplacian. The value displayed in each block is the the recall (or precision) of the clustering with respect to the Gremban's expansion. Again we see similar results as the first test. Both the Physics Laplacian and the Gremban's expansion do not have high precision when both q_{neg} and p_{neg} are low. If p_{neg} is low, the performance of both matrix representations improves as q_{neg} is increased. However, as p_{neg} increases, the Gremban's expansion has a much higher recall and precision than the Physics Laplacian. It should again be noted that as p_{neg} increases the validity of the ground truth becomes arguable.

The middle row is the difference in recall (or precision) of the clustering using the eigenvalues of the signed Laplacian and the Physics Laplacian. The value displayed in each block is the recall (or precision) of the clustering with respect to the signed Laplacian. For small p_{neg}, meaning not many of the inner-block connections are negative-valued, both representations have high recall and precision, with the signed Laplacian performing better than the Physics Laplacian. However, as p_{neg} increase we again see the performance of both representations degrade.

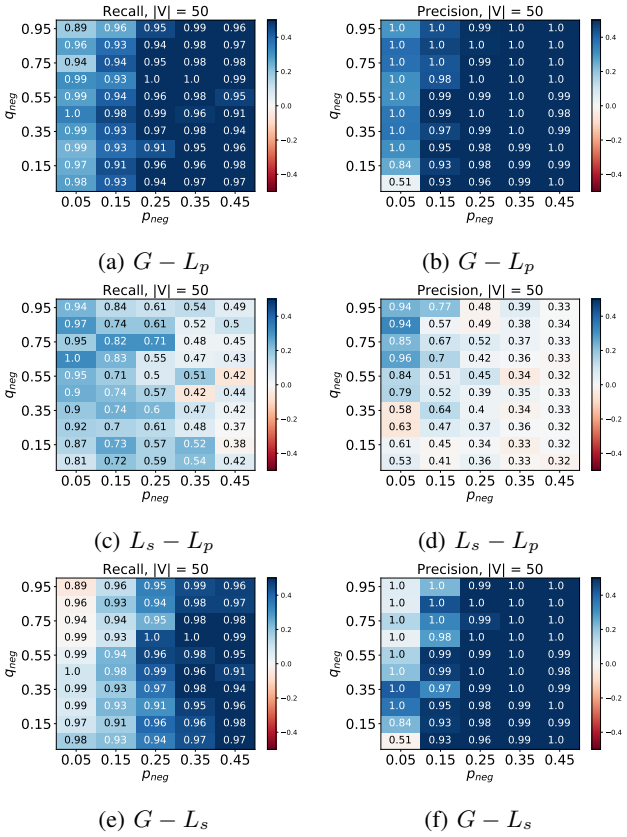
The last row is the difference in recall (or precision) of the clustering using the eigenvalues of the Gremban's expansion matrix and the signed Laplacian. The value displayed in each block is the recall (or precision) of the clustering with respect to the Gremban's expansion. For low p_{neg} and q_{neg} the clustering using the signed Laplacian is much better than using the Gremban's expansion. However, as q_{neg} and p_{neg} increases the Gremban's expansion produces higher quality clusters than the signed Laplacian.

Figure 5 displays the same test but for the 1000 vertex block stochastic model. Similar results to the 50 vertex model can be concluded. For low p_{neg}, both the clustering using the signed Laplacian and the Physics Laplacian had higher precision (or recall) than using the Gremban's expansion. As p_{neg} increases the clustering using the Gremban's expansion {out performed the signed and Physics Laplacian.}.

C. Test 3

Using the static block stochastic graphs with non-overlapping ground-truth provided by the Graph Challenge [7], we investigate how the spectral clustering performs using the eigenvectors of each signed matrix representations as the graph size increases with $n = [50, 100, 500, 1000, 5000]$. With predetermined values of p_{neg} and q_{neg}, if there exists an edge, then based on the negative edge probability, the sign of the edge is switched. We again will present the difference in precision and recall for the various matrix representation as the probabilities p_{neg} and q_{neg} vary. For each combination ten trails were computed and the average value is displayed. From Test 1 and Test 2, if p_{neg} and q_{neg} were small the clustering using the Gremban's expansion matrix typically performed worse than the signed or Physics Laplacian. Thus, we consider $p_{neg} = 0.1$ and study how increasing q_{neg} effects the performance.

Figure 6 displays the average recall and precision using the eigenvectors associated with each signed graph matrix representation. The x-axis displays the size of the graph and the y-axis represents the precision (or recall). The blue markers depict the precision (or recall) using the eigenvectors associated with the Gremban's expansion, the green markers using the signed Laplacian, and the orange markers using the Physics Laplacian. The top row is when $q_{neg} = 0.2$ and the bottom row is when $q_{neg} = 1.0$, with $q_{neg} = 0.6$ displayed in between. For all values of q_{neg} the clustering using the Gremban's expansion improves in both precision and recall as q_{neg} increases. This is unsurprising as we saw a similar result in the previous tests. Also, as the graph increases in size the Gremban's expansion average performs better, however, there are a few cases where either and/or both the signed Laplacian and Physics Laplacian produces higher quality clusters. Figure 7 displays the same test but for $p_{neg} = 0.2$. With a slight increase in the probability for negative connections in the inner-block community, the Gremban's expansion is able to capture the clusters well as the size of the graph increases. The performance of Physics Laplacian and the signed Laplacian are erratic and do not scale with the size of the graph. Thus, it can be concluded that performance of each method depends on the underlying sign structure. However, using the eigengap and the eigenvectors associated with the Gremban's expansion matrix consistently provided high quality clusters than the signed and Physics Laplacian for varying combinations of (p, q, p_{neg}, q_{neg}).



(a) $G - L_p$ (b) $G - L_p$

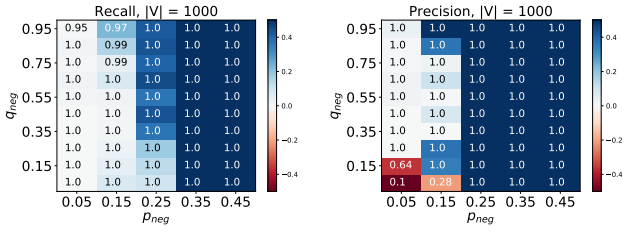
(c) $L_s - L_p$ (d) $L_s - L_p$

(e) $G - L_s$ (f) $G - L_s$

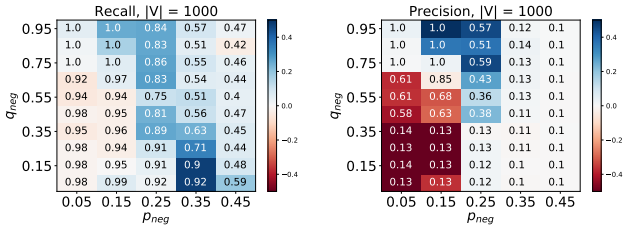
Fig. 4: Difference in recall (left column) and precision (right column) for the 50 vertex ground truth model. (a)-(b) Difference using the Gremban's expansion verse the Physics Laplacian ($G - L_p$). The printed values are with respect to the Gremban's expansion matrix. (c)-(d) Difference using the signed Laplacian verse the Physics Laplacian ($L_s - L_p$). The printed values are with respect to the signed Laplacian. (e)-(f) Difference using the Gremban's expansion verse the signed Laplacian ($G - L_s$). The printed values are with respect to the Gremban's expansion matrix.

IV. Conclusion

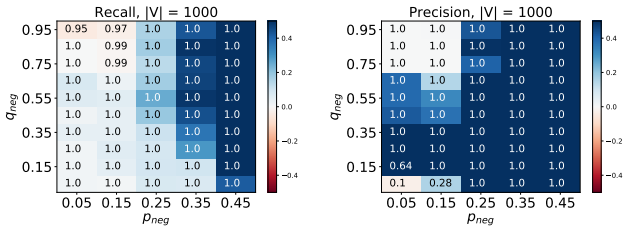
In this paper we empirically studied classical spectral clustering concepts of signed graphs using three different signed graph matrix representations: the signed Laplacian, the Physics Laplacian, and the Gremban's expansion. No one representation was able to produce the highest quality clusters for all combinations of graph size, and the



(a) $G - L_p$ (b) $G - L_p$



(c) $L_s - L_p$ (d) $L_s - L_p$



(e) $G - L_s$ (f) $G - L_s$

Fig. 5: Difference in recall (left column) and precision (right column) for the 1000 vertex ground truth model. (a)-(b) Difference using the Gremban's expansion verse the Physics Laplacian ($G - L_p$). The printed values are with respect to the Gremban's expansion matrix. (c)-(d) Difference using the signed Laplacian verse the Physics Laplacian ($L_s - L_p$). The printed values are with respect to the signed Laplacian. (e)-(f) Difference using the Gremban's expansion verse the signed Laplacian ($G - L_s$). The printed values are with respect to the Gremban's expansion matrix.

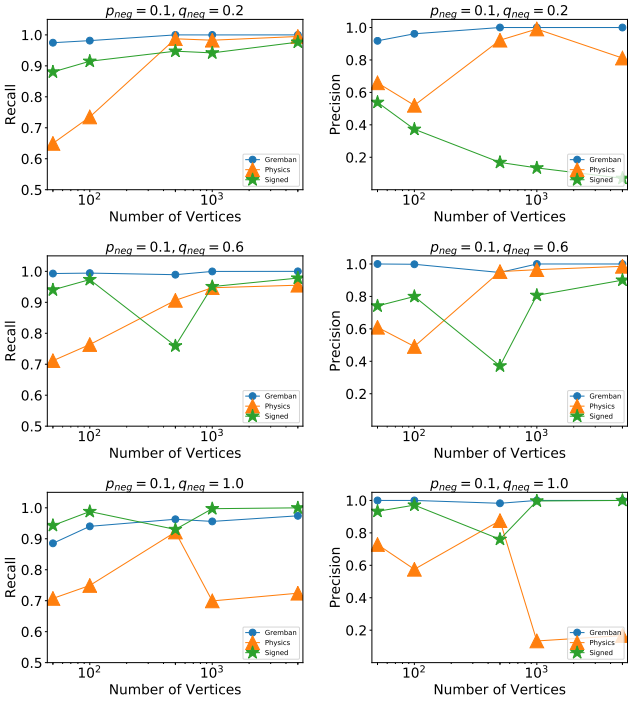


Fig. 6: Recall (left column) and precision (right column) for $n = [50, 100, 500, 1000, 5000]$ node ground truth model for the Gremban's expansion matrix, the Physics Laplacian, and the signed Laplacian with $p_{neg} = 0.1$ held constant.

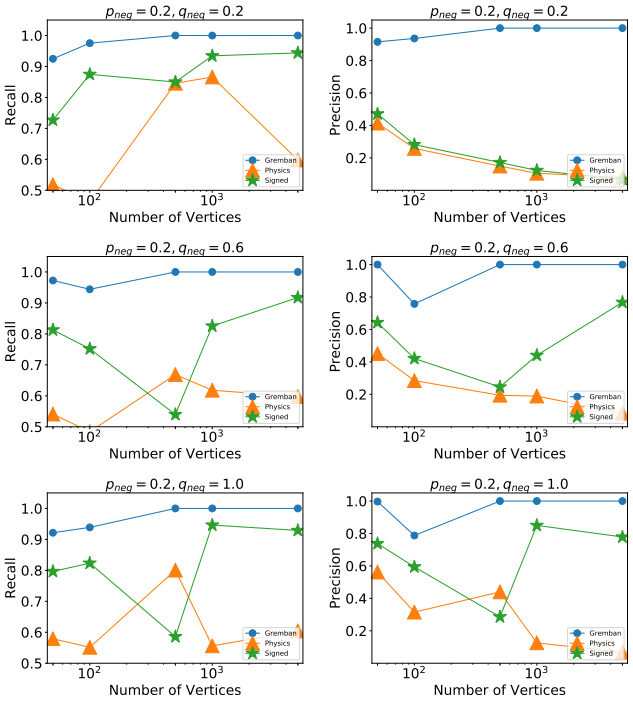


Fig. 7: Recall (left column) and precision (right column) for $n = [50, 100, 500, 1000, 5000]$ node ground truth model for the Gremban's expansion matrix, the Physics Laplacian, and the signed Laplacian with $p_{neg} = 0.2$ held constant.

connection probabilities for positive and negative edge weights. The signed Laplacian and the Physics Laplacian usually produced higher quality clusters when both the inner and outer-block connection negative-valued probabilities were relatively small. Otherwise, the Gremban's expansion produced a higher quality result. Thus, depending on the sign structure and the edge densities one could argue for any one of the representations. However, since the internal sign structure will be unknown in real-world signed graphs, we conclude, that the Gremban's expansion is the most robust representation for spectral clustering.

Acknowledgments

This work was funded by LLNL Laboratory Directed Research and Development as Project 17-SI-004: *Variable Precision Computing* and was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Release Number: LLNL-CONF-754816.

References

[1] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation Clustering. *Machine Learning*, 56(1-3):89–113, July 2004.

[2] Anil Damle, Victor Minden, and Lexing Ying. Robust and efficient multi-way spectral clustering. *arXiv:1609.08251 [physics]*, September 2016. arXiv: 1609.08251.

[3] Miroslav Fiedler. Algebraic Connectivity of Graphs. *Czechoslovak Mathematical Journal*, 23, January.

[4] A. Fox, T. Manteuffel, and G. Sanders. Numerical Methods for Gremban's Expansion of Signed Graphs. *SIAM Journal on Scientific Computing*, 39(5):S945–S968, January 2017.

[5] Jean Gallier. Spectral theory of unsigned and signed graphs. applications to graph clustering: a survey. *CoRR*, abs/1601.04692, 2016.

[6] Keith D. Gremban. *Combinatorial Preconditioners for Sparse, Symmetric, Diagonally Dominant Linear Systems*. PhD thesis, Carnegie Mellon University, Pittsburgh, October 1996. CMU CS Tech Report CMU-CS-96-123.

[7] Edward Kao, Vijay Gadepally, Michael Hurley, Michael Jones, Jeremy Kepner, Sanjeev Mohindra, Paul Monticciolo, Albert Reuther, Siddharth Samsi, William Song, Diane Staheli, and Steven Smith. Streaming Graph Challenge: Stochastic Block Partition. August 2017.

[8] Andrew V. Knyazev. On spectral partitioning of signed graphs. *arXiv:1701.01394 [cs, math, stat]*, January 2017. arXiv: 1701.01394.

[9] J. Kunegis, S. Schmidt, A. Lommatzsch, J. Lerner, E. 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*, Proceedings, pages 559–570. Society for Industrial and Applied Mathematics, April 2010.

[10] Jérôme Kunegis, Andreas Lommatzsch, and Christian Bauckhage. The slashdot zoo: Mining a social network with negative edges. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 741–750, New York, NY, USA, 2009. ACM.

[11] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. page 1361. ACM Press, 2010.

[12] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December 2007.

[13] Leting Wu, Xiaowei Ying, Xintao Wu, Aidong Lu, and Zhi-Hua Zhou. Spectral analysis of k-balanced signed graphs. In Joshua Zhexue Huang, Longbing Cao, and Jaideep Srivastava, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–12, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.