SPONGE: A generalized eigenproblem for clustering signed networks

Mihai Cucuringu
University of Oxford
The Alan Turing Institute
mihai.cucuringu@stats.ox.ac.uk

Peter Davies
University of Warwick
P.W.Davies@warwick.ac.uk

Aldo Glielmo
King's College London
aldo.glielmo@kcl.ac.uk

Hemant Tyagi
INRIA Lille - Nord Europe
hemant.tyagi@inria.fr

Abstract

We introduce a principled and theoretically sound spectral method for k-way clustering in signed graphs, where the affinity measure between nodes takes either positive or negative values. Our approach is motivated by social balance theory, where the task of clustering aims to decompose the network into disjoint groups, such that individuals within the same group are connected by as many positive edges as possible, while individuals from different groups are connected by as many negative edges as possible. Our algorithm relies on a generalized eigenproblem formulation inspired by recent work on constrained clustering. We provide theoretical guarantees for our approach in the setting of a signed stochastic block model, by leveraging tools from matrix perturbation theory and random matrix theory. An extensive set of numerical experiments on both synthetic and real data shows that our approach compares favorably with state-of-the-art methods for signed clustering, especially for large number of clusters and sparse measurement graphs.

1 Introduction

Clustering is a popular unsupervised learning task aimed at extracting groups of nodes in a weighted graph in such a way that the average connectivity or similarity between pairs of nodes within the same group is larger than that of pairs of nodes from different groups. While most of the literature has focused on clustering graphs where the edge weights are non-negative, the task of clustering signed graphs (whose edge weights can take negative values as well) remained relatively unexplored, and has recently become an increasingly important research topic [34].

The motivation for recent studies arose from a variety of examples from social networks, where users express relationships of trust-distrust or friendship-enmity, online news and review websites such as Epinions [1] and Slashdot [2] that allow users to approve or denounce others [33], and shopping bipartite networks encoding like-dislike preferences between users and products [5].

Another application stems from time series analysis, in particular clustering time series [3], a task broadly used for analyzing gene expression data in biology [20], economic time series that capture macroeconomic variables [19], and financial time series corresponding to large baskets of instruments in the stock market [53, 40]. In such contexts, a popular similarity measure in the literature is given by the Pearson correlation coefficient that measures linear dependence between variables and takes values in $[-1, 1]$. By interpreting the correlation matrix as a weighted network whose (signed) edge weights capture the pairwise correlations, we cluster the multivariate time series by clustering the underlying signed network. To increase robustness, tests of statistical significance are often applied to individual pairwise correlations, leading to sparse networks after thresholding on the p-value associated to each individual sample correlation [24]. We refer the reader to the popular work of Smith et al. [47] for a detailed survey and comparison of various methods for turning time series into networks. Importantly, they conclude that in general correlation-based approaches can be quite successful at estimating the connectivity of brain networks from fMRI time series.

Contributions. Our contributions are as follows.

• We propose a regularized spectral algorithm for clustering signed graphs that is based on solving a generalized eigenproblem. Our approach is scalable and compares favorably to state-of-the-art methods.

• We provide a detailed theoretical analysis of our algorithm with respect to its robustness against sampling sparsity and noise level, under a Signed Stochas-

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

tic Block Model (SSBM).

• To our knowledge, we provide the first theoretical guarantees – in the SSBM framework – for the Signed Laplacian method introduced in the popular work of Kunegis et al. [32] for clustering signed graphs.

• Finally, we provide extensive numerical experiments on both synthetic and real data, showing that our algorithm compares favourably to state-of-art methods. In particular, it is able to recover partitions in the regime where the graph is sparse and the number of clusters k is large, where existing methods completely fail.

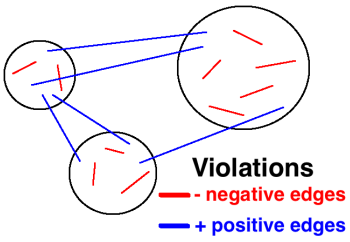
Paper outline. The remainder of this paper is organized as follows. Section 2 is a summary of related work from the signed clustering literature. Section 3 formulates our SPONGE (Signed Positive Over Negative Generalized Eigenproblem) algorithm for clustering signed graphs. Section 4 introduces the Signed Stochastic Block Model (SSBM) and contains our theoretical analysis of SPONGE in the SSBM. Section 5 contains a similar theoretical analysis for signed spectral clustering via the Signed Laplacian. Section 6 contains numerical experiments on various synthetic and real data sets. Finally, Section 7 summarizes our findings along with future research directions.

Notation. For a matrix $A \in \mathbb{R}^{n\times n}$, we denote its eigenvalues and eigenvectors by $\lambda_i(A)$ and $v_i(A)$ respectively, $\forall i = 1,\dots,n$. For symmetric A, we assume the ordering $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$. For $A \in \mathbb{R}^{m\times n}$, $\|A\|_2$ denotes its spectral norm, i.e., the largest singular value of A. We denote $\mathbf{1}$ to be the all one's column vector. For a matrix U, $\mathcal{R}(U)$ denotes the range space of its columns. Throughout, $G = (V, E)$ denotes the signed graph with vertex set V, edge set E, and adjacency matrix $A \in \{0, \pm 1\}^{n\times n}$. We let $G^+ = (V, E^+)$ (resp. $G^- = (V, E^-)$) denote the unsigned subgraphs of positive (resp. negative) edges with adjacency matrices A^+ (resp. A^-), such that $A = A^+ - A^-$. Here, $A^+_{ij} = \max\{A_{ij}, 0\}$ and $A^-_{ij} = \max\{-A_{ij}, 0\}$. Moreover $E^+ \cap E^- = \emptyset$, and $E^+ \cup E^- = E$.

2 Related literature

The problem of clustering signed graphs can be traced back to the work of Cartwright and Harary from the 1950s on social balance theory [25, 7], who explored the concept of *balance* in signed graphs. A signed graph is said to be balanced iff (i) all the edges are positive, or (ii) the nodes can be partitioned into two disjoint sets such that positive edges exist only within clusters, and negative edges are only present across clusters. The "weak balance theory" of Davis [16] relaxed the balanced relationship – a signed graph is weakly balanced iff (i) all the edges are positive, or (ii) the nodes can be partitioned into $k \in \mathbb{N}$ disjoint sets such that positive edges exist only within clusters, and negative edges are only present across clusters.

Motivated by this theory, the k-way clustering problem in signed graphs amounts to finding a partition into k clusters such that most edges within clusters are positive, and most edges across clusters are negative. Alternatively, one may seek a partition such that the number of **violations** is minimized, i.e., negative edges within the cluster and positive edges across clusters, as depicted in the figure above. In order to avoid partitions where clusters contain only a few nodes, one often wishes to also incentivize clusters of large size or volume.



A number of algorithms have been proposed for clustering signed graphs. Doieran and Mrvar [18] proposed a local search approach in the spirit of the Kernighan-Lin algorithm [28]. Yang et al. [52] introduced an agent-based approach by considering a certain random walk on the graph. In recent years, several efforts for the analysis of signed graphs have lead to novel extensions for various tasks, including edge prediction [31, 33], node classification [6, 48], node embeddings [9, 17, 29, 50], node ranking [11, 45], and clustering [8, 32, 36]. We refer the reader to [49] for a recent survey on the topic.

Spectral methods on signed networks began with Anchuri et al. [4]; they proposed optimizing modularity and other objective functions in signed graphs. Kunegis et al. [32] proposed solving a 2-way "signed" ratio-cut problem via the (combinatorial) Signed Laplacian [26] $\overline{L} = \overline{D} - A$, where $\overline{D}$ is a diagonal matrix with $\overline{D}_{ii} = \sum_{i=1}^n |A_{ij}|$. Similar signed extensions also exist for the random-walk Laplacian $\overline{L}_{\text{rw}} = I - \overline{D}^{-1}A$, and the symmetric graph Laplacian $\overline{L}_{\text{sym}} = I - \overline{D}^{-1/2}A\overline{D}^{-1/2}$, the latter of which is particularly suitable for skewed degree distributions. Chiang et al. [10] put forth the claim that the Signed Laplacian $\overline{L}$ faces a fundamental weakness when directly extended to k-way clustering for $k > 2$. They proposed a formulation based on the *Balanced Normalized Cut* (BNC) objective

$$\min_{\{x_1,\dots,x_k\}\in\mathcal{I}} \left(\sum_{c=1}^{k} \frac{x_c^T(D^+ - A)x_c}{x_c^T\overline{D}x_c} \right).$$

Here, D^+ denotes the diagonal matrix with degrees $D^+_{ii} = \sum_{j=1}^n A^+_{ij}$; $C_1,\dots,C_k$ denote the k clusters, and $\mathcal{I}$ denotes a k-cluster indicator set, where $(x_t)_i = 1$, if node $i \in C_t$, and 0 otherwise. The same authors also consider the closely related *Balanced Ratio Cut*, which replaces $\overline{D}$ in the denominator by I. We remark that spectral clustering algorithms

(for signed/unsigned graphs) typically have a common pipeline, wherein a suitable graph operator is considered (for eg. Laplacian), its k (or $k-1$) extremal eigenvectors are found, and the resulting points in $\mathbb{R}^k$ (or $\mathbb{R}^{k-1}$) are then clustered using k-means.

Hsieh et al. [27] propose performing matrix completion as a preprocessing step before clustering using the top k eigenvectors of the completed matrix. Mercado et al. [36] present an extended spectral method based on the geometric mean of Laplacians. For $k=2$, Cucuringu [12] showed that signed clustering can be cast as an instance of the group synchronization [46] problem over $\mathbb{Z}_2$. Finally, we refer the reader to [21] for a recent survey on clustering signed and unsigned graphs.

3 SPONGE: a signed generalized eigenproblem formulation

Given an unsigned graph H with adjacency matrix W with non-negative entries, for any cluster $C \subset V$ define $\mathrm{cut}_H(C,\overline{C}) := \sum_{i\in C, j\in\overline{C}} W_{ij}$ as the total weight of edges crossing from C to $\overline{C}$. Also define the volume of C, $\mathrm{vol}_H(C) := \sum_{i\in C}\sum_{j=1}^n W_{ij}$ as the sum of degrees of nodes in C. Motivated by the approach of [13] in the context of constrained clustering, we aim to minimize the following two measures of *badness*

$$\frac{\mathrm{cut}_{G^+}(C,\overline{C})}{\mathrm{vol}_{G^+}(C)}, \tag{3.1}$$

$$\left(\frac{\mathrm{cut}_{G^-}(C,\overline{C})}{\mathrm{vol}_{G^-}(C)}\right)^{-1} = \frac{\mathrm{vol}_{G^-}(C)}{\mathrm{cut}_{G^-}(C,\overline{C})}. \tag{3.2}$$

Ideally, C is such that both (3.1) and (3.2) are small. To this end, we first consider "merging" the objectives (3.1) and (3.2), and would like to solve

$$\min_{C\subset V} \frac{\mathrm{cut}_{G^+}(C,\overline{C}) + \tau^- \mathrm{vol}_{G^-}(C)}{\mathrm{cut}_{G^-}(C,\overline{C}) + \tau^+ \mathrm{vol}_{G^+}(C)},$$

with $\tau^+, \tau^- > 0$ denoting trade-off or regularization parameters. While at first sight this may seem rather ad-hoc in nature, we provide a sound theoretical justification for our approach in later sections. A natural extension to $k>2$ disjoint clusters $C_1,\dots,C_k$ leads to the following discrete optimization problem

$$\min_{C_1,\dots,C_k} \sum_{i=1}^k \frac{\mathrm{cut}_{G^+}(C_i,\overline{C_i}) + \tau^- \mathrm{vol}_{G^-}(C_i)}{\mathrm{cut}_{G^-}(C_i,\overline{C_i}) + \tau^+ \mathrm{vol}_{G^+}(C_i)}. \tag{3.3}$$

For a subset $C_i \subset V$, the normalized indicator vector

$$(x_{C_i})_j = \begin{cases} (\mathrm{cut}_{G^-}(C_i,\overline{C_i}) + \tau^+ \mathrm{vol}_{G^+}(C_i))^{-1/2}; & j \in C_i \\ 0; & j \notin C_i \end{cases}$$

renders (3.3) as the discrete optimization problem

$$\min_{C_1,\dots,C_k} \sum_{i=1}^k \frac{x_{C_i}^T (L^+ + \tau^- D^-) x_{C_i}}{x_{C_i}^T (L^- + \tau^+ D^+) x_{C_i}}, \tag{3.4}$$

which is NP-hard. Here L^+ (resp. L^-) denotes the Laplacian of G^+ (resp. G^-), and D^+ (resp. D^-) denotes a diagonal matrix with the degrees of G^+ (resp. G^-). A common approach in this situation is to drop the discreteness constraint and allow each x_{C_i} to take values in $\mathbb{R}^n$. To this end, we introduce a new set of vectors $z_1,\dots,z_k \in \mathbb{R}^n$, such that they are orthonormal with respect to $L^- + \tau^+ D^+$, i.e.,

- $z_i^T (L^- + \tau^+ D^+) z_i = 1$, and
- $z_i^T (L^- + \tau^+ D^+) z_j = 0$, for $i \neq j$.

This leads to the following modified version of (3.4)

$$\min_{z_i^T (L^- + D^+) z_j = \delta_{ij}} \sum_{i=1}^k \frac{z_i^T (L^+ + \tau^- D^-) z_i}{z_i^T (L^- + \tau^+ D^+) z_i}. \tag{3.5}$$

The above choice of $(L^- + \tau^+ D^+)$-orthonormality of vectors $z_1,\dots,z_k$ is not – strictly speaking – a relaxation of (3.4). But it leads to a suitable eigenvalue problem. Indeed, assuming $L^- + \tau^+ D^+$ is full rank, consider the change of variables $y_i = (L^- + \tau^+ D^+)^{1/2} z_i$ which changes the orthonormality constraints of (3.4) to $y_i^T y_j = \delta_{ij}$. Furthermore, denoting $Y = [y_1,\dots,y_k] \in \mathbb{R}^{n\times k}$, one can rewrite (3.5) as

$$\min_{Y^T Y = I} \mathrm{Tr}\Big(Y^T (L^- + \tau^+ D^+)^{-1/2} (L^+ + \tau^- D^-)(L^- + \tau^+ D^+)^{-1/2} Y\Big). \tag{3.6}$$

The solution to (3.6) is given by the eigenvectors corresponding to the k-smallest eigenvalues of $(L^- + \tau^+ D^+)^{-1/2}(L^+ + \tau^- D^-)(L^- + \tau^+ D^+)^{-1/2}$ (see for eg. [44, Theorem 2.1]). One can also verify[1] that (λ, v) is an eigenpair of the previous matrix if and only if $(\lambda, (L^- + \tau^+ D^+)^{-1/2} v)$ is a *generalized* eigenpair of $(L^+ + \tau^- D^-, L^- + \tau^+ D^+)$.

Our complete algorithm SPONGE first finds the smallest k generalized eigenvectors of $(L^+ + \tau^- D^-, L^- + \tau^+ D^+)$ for suitably chosen $\tau^+, \tau^- > 0$. We then cluster the resulting embedding of the vertices in $\mathbb{R}^k$ using k-means++. We also consider a variant of SPONGE, namely SPONGE$_{sym}$, where the embedding is generated using the smallest k generalized eigenvectors of $(L^+_{sym} + \tau^- I, L^-_{sym} + \tau^+ I)$, wherein $L^+_{sym} = (D^+)^{-1/2} L^+ (D^+)^{-1/2}$ is the so-called symmetric Laplacian of G^+ (similarly for L^-_{sym}).

Remark 1. *Solving (3.6) is computationally expensive in practice as it involves computing a matrix-inverse. This is not the case if we solve the generalized eigenproblem version of (3.6). In our experiments, we use LOBPCG [30], a preconditioned eigensolver[2] for solving large positive definite generalized eigenproblems.*

[1] Let A, B be symmetric matrices with $A \succ 0$. Then (λ, v) is an eigenpair of $A^{-1/2} B A^{-1/2}$ iff $(\lambda, A^{-1/2} v)$ is a generalized eigenpair of (B, A). Indeed, for $w = A^{-1/2} v$, $A^{-1/2} B A^{-1/2} v = \lambda v \Leftrightarrow Bw = \lambda A w$.

[2] Locally Optimal Block Preconditioned Conjugate Gradient method.

4 Analysis of SPONGE under SSBM

We begin by introducing the signed stochastic block model (SSBM) in Section 4.1 and then theoretically analyze the performance of SPONGE in Section 4.2.

4.1 Signed stochastic block model

For ease of exposition, we assume n is a multiple of k, and partition the vertices of G into k-equally sized clusters $C_1, \dots, C_k$. In particular, we assume w.l.o.g that $C_l = \left\{ \frac{(l-1)n}{k} + 1, \dots, \frac{ln}{k} \right\}$ for $l = 1, \dots, k$. The graph G follows the Erdős-Rényi random graph model $G(n,p)$ wherein each edge takes value $+1$ if both its endpoints are contained in the same cluster, and -1 otherwise. To model noise, we flip the sign of each edge independently with probability $\eta \in [0, 1/2)$.

Let $A \in \{0, \pm 1\}^{n \times n}$ denote the adjacency matrix of G, then $(A_{ij})_{i \le j}$ are independent random variables. Recall that $A = A^+ - A^-$, where $A^+, A^- \in \{0,1\}^{n \times n}$ are the adjacency matrices of the unsigned graphs G^+, G^- respectively. Then, $(A^+_{ij})_{i \le j}$ are independent, and similarly $(A^-_{ij})_{i \le j}$ are also independent. But clearly, for given $i, j \in [n]$ with $i \neq j$, A^+_{ij} and A^-_{ij} are dependent.

Remark 2. *Contrary to stochastic block models for unsigned graphs, we do not require the intra-cluster edge probabilities to be different from those of inter-cluster edges. While this is necessary in the unsigned case for detecting clusters (eg. [38, 39]), it is not the case for signed networks since the sign of the edge already achieves this purpose implicitly. In fact, as one would expect, it is the parameter η that is crucial for identifiability, as shown formally in our analysis.*

4.2 Theoretical results for SPONGE

We now theoretically analyze the performance of SPONGE under the SSBM. In particular, we analyze the embedding given by the smallest k eigenvectors of

$$T = (L^- + \tau^+ D^+)^{-1/2} (L^+ + \tau^- D^-)(L^- + \tau^+ D^+)^{-1/2},$$

for parameters $\tau^-, \tau^+ > 0$. Recall that (λ, v) is an eigenpair of T if and only if $(\lambda, (L^- + \tau^+ D^+)^{-1/2} v)$ is a generalized eigenpair for the matrix pencil $(L^+ + \tau^- D^-, L^- + \tau^+ D^+)$. We assume throughout that both $L^+ + \tau^- D^-$ and $L^- + \tau^+ D^+$ are full rank. For ease of exposition, we focus on the case $k = 2$ but the results can be extended to the general $k \ge 2$ setting (work in progress) using the same proof outline. Denote

$$\overline{T} = (\mathbb{E}[L^-] + \tau^+ \mathbb{E}[D^+])^{-1/2} (\mathbb{E}[L^+] + \tau^- \mathbb{E}[D^-])(\mathbb{E}[L^-] + \tau^+ \mathbb{E}[D^+])^{-1/2},$$

and also denote

$$V_2(T) = [v_n(T)\ v_{n-1}(T)], \quad V_2(\overline{T}) = [v_n(\overline{T})\ v_{n-1}(\overline{T})],$$

to be $n \times 2$ matrices consisting of the smallest two (unit ℓ_2 norm) eigenvectors of $T, \overline{T}$ respectively. Let

$$w = \frac{1}{\sqrt{n}} (\underbrace{1, \dots, 1}_{n/2}, \underbrace{-1, \dots, -1}_{n/2})^T \in \mathbb{R}^n \tag{4.1}$$

correspond to the "ground truth" or "planted clusters" we seek to recover. Our main result is the following.

Theorem 1. *For $\eta \in [0, 1/2)$ let $\tau^+, \tau^- > 0$ satisfy*

$$\tau^- < \tau^+ \left(\frac{\frac{n}{2} - 1 + \eta}{\frac{n}{2} - \eta} \right). \tag{4.2}$$

Then it holds that $\{v_{n-1}(\overline{T}), v_n(\overline{T})\} = \left\{ \frac{1}{\sqrt{n}} \mathbf{1}, w \right\}$ where w is defined in (4.1). Moreover, assuming $n \ge 6$, for given $0 < \varepsilon \le 1/2$, $\epsilon \in (0,1)$ and $\varepsilon_\tau \in (0,1)$ let $\tau^- \le \varepsilon_\tau \tau^+ \left(\frac{\frac{n}{2} - 1 + \eta}{\frac{n}{2} - \eta} \right)$ and $p \ge c_1'(\varepsilon, \tau^+, \tau^-, \varepsilon_\tau, \eta, \epsilon) \frac{\log n}{n}$ where $c_1'(\varepsilon, \tau^+, \tau^-, \varepsilon_\tau, \eta, \epsilon) > 0$ depends only on $\varepsilon, \tau^+, \tau^-, \varepsilon_\tau, \eta, \epsilon$. Then there exists a constant $c_\varepsilon > 0$ depending only on ε such that with probability at least $1 - \frac{4}{n} - 2n \exp\left(\frac{-pn}{c_\varepsilon}\right)$, it holds that

$$\left\| (I - V_2(T) V_2(T)^T) V_2(\overline{T}) \right\|_2 \le \frac{\epsilon}{1 - \epsilon}.$$

The theorem states that $\mathcal{R}(V_2(T))$ is close to $\mathcal{R}(V_2(\overline{T}))$ with high probability provided n, p are suitably large, and τ^- is sufficiently small compared to τ^+. The latter condition is required to ensure that the smallest two eigenvectors of $\overline{T}$ are $\left\{ \frac{1}{\sqrt{n}} \mathbf{1}, w \right\}$. Also note that the embedding generated by any orthonormal basis for $\mathcal{R}(V_2(T))$ leads to the same clustering performance[3]. Since the embedding corresponding to $V_2(\overline{T})$ leads to perfectly separated (ground truth) clusters, hence the closer $\mathcal{R}(V_2(T))$ is to $\mathcal{R}(V_2(\overline{T}))$, the better is the clustering performance. Using standard tools, one can actually use bounds on subspace recovery to bound the misclustering rate of k-means (see for eg. [42]).

Proof sketch. The proof is deferred to the appendix, but the main steps involved are as follows. We first compute the spectra of $\mathbb{E}[L^-], \mathbb{E}[L^+], \mathbb{E}[D^-]$, and $\mathbb{E}[D^+]$ by finding the eigenvalues and the corresponding *relevant* eigenvectors (i.e., associated to the smallest two eigenvalues). We then identify conditions on the parameters τ^+, τ^- under which the smallest two eigenvectors of $\overline{T}$ are $\left\{ w, \frac{1}{\sqrt{n}} \mathbf{1} \right\}$. As shown in the proof, w is always one of the smallest two eigenvectors (since $\tau^+, \tau^- > 0$). The condition (4.2) leads to $\frac{1}{\sqrt{n}} \mathbf{1} \in \{v_{n-1}(\overline{T}), v_n(\overline{T})\}$. Next, we derive concentration bounds using tools from random matrix theory for A^-, A^+, D^-, D^+ holding with high probability. This in turn leads to a bound on $||T - \overline{T}||_2$. Combining the above results and by controlling the perturbation term $||T - \overline{T}||_2$, we obtain via the Davis-Kahan theorem [15], a bound on $\left\| \sin(\Theta(\mathcal{R}(V_2(T)), \mathcal{R}(V_2(\overline{T})))) \right\|_2$ which equals $\left\| (I - V_2(T) V_2(T)^T) V_2(\overline{T}) \right\|_2$. Here,

[3] For a 2×2 orthogonal matrix O, the rows of the matrix $V_2(T)O$ are obtained via the same orthogonal transformation on the corresponding rows of $V_2(T)$.

$\Theta(\mathcal{R}(V_2(T)), \mathcal{R}(V_2(\overline{T})))$ is the diagonal matrix of canonical angles between $\mathcal{R}(V_2(T))$ and $\mathcal{R}(V_2(\overline{T}))$.

Selecting only $v_n(T)$. Alternately, one could consider taking just the smallest eigenvector of T, i.e. $v_n(T)$, leading to a one-dimensional embedding. The following theorem states that, provided τ^- is sufficiently larger than τ^+, and if n, p are suitably large, then $\mathcal{R}(v_n(T))$ is close to $\mathcal{R}(w)$ with high probability.

Theorem 2. *For $\eta \in [0, 1/2)$ let $\tau^+, \tau^- > 0$ satisfy*

$$\tau^- > \left(\frac{\eta}{1-\eta}\right)\left(\frac{\frac{n}{2}-1+\eta}{\frac{n}{2}-\eta}\right)\tau^+. \tag{4.3}$$

Then it holds that $v_n(\overline{T}) = w$ with w defined in (4.1).

Moreover, assuming $n \geq 6$, for given $0 < \varepsilon \leq 1/2, \epsilon \in (0,1)$ and $\varepsilon_\tau \in (0,1)$ let $\tau^- \geq \frac{1}{\varepsilon_\tau}\left(\frac{\eta}{1-\eta}\right)\left(\frac{\frac{n}{2}-1+\eta}{\frac{n}{2}-\eta}\right)\tau^+$,

$$p \geq c_2'(\varepsilon, \tau^+, \tau^-, \varepsilon_\tau, \eta, \epsilon)\frac{\log n}{n},$$

where $c_2'(\varepsilon, \tau^+, \tau^-, \varepsilon_\tau, \eta, \epsilon) > 0$ depends only on the indicated parameters. Then there exists a constant $c_\varepsilon > 0$ depending only on ε such that with probability at least $\left(1 - \frac{4}{n} - 2n\exp\left(\frac{-pn}{c_\varepsilon}\right)\right)$, it holds that

$$\left\|(I - v_n(T)v_n(T)^T)w\right\|_2 \leq \frac{\epsilon}{1-\epsilon}.$$

The proof is deferred to the appendix, being similar to that of Theorem 1. The main difference is in the conditions on τ^-, τ^+ for ensuring that the smallest (two) eigenvector(s) of $\overline{T}$ correspond to the ground truth clustering; these are clearly weaker in Theorem 2 compared to Theorem 1. For eg. if $\eta = 0$, then any $\tau^+, \tau^- > 0$ imply $v_n(\overline{T}) = w$ by Theorem 2, while the analogous statement is not true in Theorem 1.

Figure 1 (left) compares the 40 smallest eigenvalues of SPONGE, SPONGE$_{sym}$, and $\overline{L}_{sym}$ in the scenario $n = 10,000$, $p = 0.01$, $\eta = 0.1$, and $k = 10$. SPONGE$_{sym}$ clearly exhibits the largest spectral gap between the 9^{th} and 10^{th} eigenvalue. Figure 1 (right) also compares the spectral densities of SPONGE$_{sym}$ for several $\eta \in \{0, 0.1, 0.2\}$. As expected, the spectral gap decreases as the noise level increases. Figure 2 compares heatmaps of recovery rates for SPONGE and SPONGE$_{sym}$ for $k = 2$ clusters and varying $\tau^+, \tau^- > 0$. Observe that (4.3) shows up when we consider only the smallest eigenvector for SPONGE. Figure 3 shows similar plots for $k = 8$, where we observe that SPONGE$_{sym}$ allows for a wider choice of $\tau^+, \tau^- > 0$ for successful clustering.

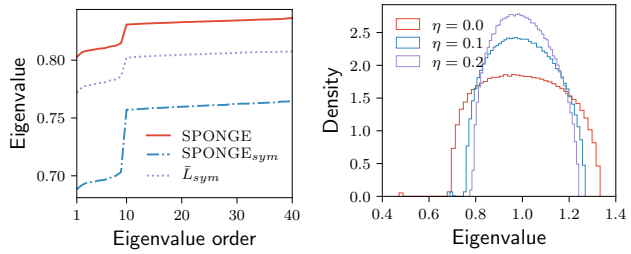


Figure 1: Left: Bottom spectrum of SPONGE$_{sym}$, SPONGE, and $\overline{L}_{sym}$. Right: Spectrum of SPONGE$_{sym}$ for three values of noise η, ($n = 10000$, $p = 0.01$, and $k = 10$).

At a high level, our proof technique, using tools from matrix perturbation and random matrices, has been used before for analyzing spectral methods for clustering unsigned graphs [42]. In the sparse regime where $p \to 0$ as $n \to \infty$, Theorems 1, 2 state that $p \gtrsim \frac{\log n}{n}$ ensures that the success probability tends to one. Similar scalings are known for unsigned graphs, however there, the intra-cluster and inter-cluster edge probabilities necessarily must be different (see Remark 2).

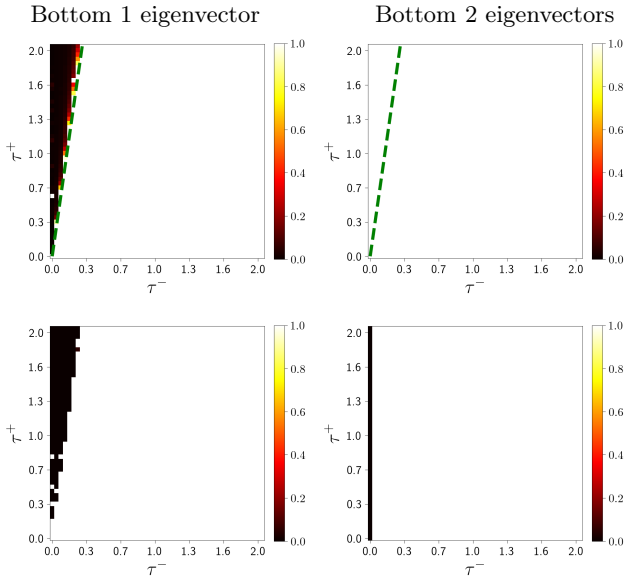


Figure 2: Heatmap of recovery rates for $k = 2$ clusters for SPONGE (top) and SPONGE$_{sym}$ (bottom), with $n = 5000, p = 0.012$ and $\eta = 0.125$, via the bottom one or two eigenvectors. The green dotted line is the condition (4.3).

5 Theoretical analysis of the Signed Laplacian $\overline{L}$ under SSBM

In this section, we theoretically analyze the popular Signed Laplacian based method of Kunegis et al. [32] for clustering signed graphs under the SSBM. This method is particularly appealing due to its simplicity, but to our knowledge, there do not exist any theoretical guarantees on the performance of this approach. We fill this gap by providing a detailed analysis for the $k = 2$ case. This choice is for ease of exposition, but the proof outline clearly extends[4] to any $k \geq 2$.

Recall that for a signed graph G with adjacency matrix $A \in \{0, \pm 1\}^{n \times n}$, and with the diagonal matrix $\overline{D}$ consisting of the degree terms defined as $\overline{D}_{ii} :=$

[4] This is part of work currently in progress.

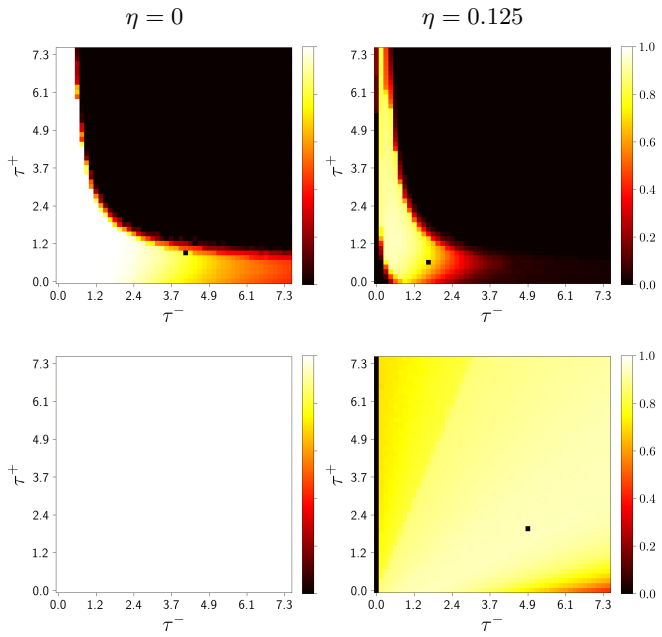


Figure 3: Heatmap of recovery rates for SPONGE (TOP) and SPONGE$_{sym}$ (bottom), using the bottom $k-1$ eigenvectors, as we vary τ^+, τ^-, with $n = 5000$, $p = 0.012$, $k = 8$ and $\eta \in \{0, 0.125\}$.

$\sum_{j=1}^{n} |A_{ij}|$, the Signed Laplacian of G, denoted by $\overline{L} \in \mathbb{R}^{n\times n}$, is given by $\overline{L} = \overline{D} - A$. Kunegis et al. [32] showed that $\overline{L}$ is positive semi-definite for any graph (see [32, Theorem 4.1]). Moreover, they also showed that $\overline{L}$ is positive definite iff the graph is unbalanced [32, Theorem 4.4]. Kunegis et al. proposed using $\overline{L}$ to first compute a lower dimensional embedding of the graph – obtained from the smallest k eigenvectors of $\overline{L}$ (in fact, as we will see, taking only $k-1$ is sufficient and more effective in signed graphs) and then clustering the obtained points in $\mathbb{R}^k$ (or $\mathbb{R}^{k-1}$) using any standard clustering method (e.g. k-means).

Our main result for the Signed Laplacian based clustering approach of Kunegis et al. [32] is stated below, and the proof is deferred to the appendix.

Theorem 3. *Assuming $0 \le \eta < 1/2$, it holds that $v_n(\mathbb{E}[\overline{L}]) = w$, where w is defined in (4.1). Moreover, let $n \ge 2$ and for given $0 < \epsilon < 1$, $0 < \varepsilon \le 1/2$ let*

$$p \ge \frac{4((1+\varepsilon)2\sqrt{2}+1)^2}{\epsilon^2(1-2\eta)^2}\frac{\log n}{n}.$$

Then there exists a constant $c_\varepsilon > 0$ depending only on ε such that with probability at least $1 - \frac{2}{n} - n\exp(-\frac{pn}{4c_\varepsilon})$ it holds that $\left\|(I - ww^T)v_n(\overline{L})\right\|_2 \le \frac{\epsilon}{1-\epsilon}$.

Theorem 3 states that for n, p suitably large, $\mathcal{R}(v_n(\overline{L})) \approx \mathcal{R}(w)$ with high probability. In particular, if η is bounded away from $1/2$, then in the sparse regime where $p \to 0$ as $n \to \infty$, the success probability approaches one if $p \gtrsim \frac{\log n}{n}$. As seen in the proof, $\mathbb{E}[\overline{L}]$ is positive definite if $\eta \neq 0$, and positive semi-definite otherwise. This makes sense since for $\eta = 0$, the generated graph (under the SSBM) is balanced by construction and thus is positive semi-definite [32, Theorem 4.4]. The fact that $\mathbb{E}[\overline{L}]$ is positive definite for $\eta \neq 0$ tells us that the resulting graph will be unbalanced with high probability. Finally, we note that as η approaches $1/2$, the condition on p becomes stricter since the expected number of intra-cluster positive edges is almost the same as the number of inter-cluster positive edges (similarly for negative edges). Hence, to get a non-trivial lower bound on p, we require n to be sufficiently large.

6 Numerical experiments

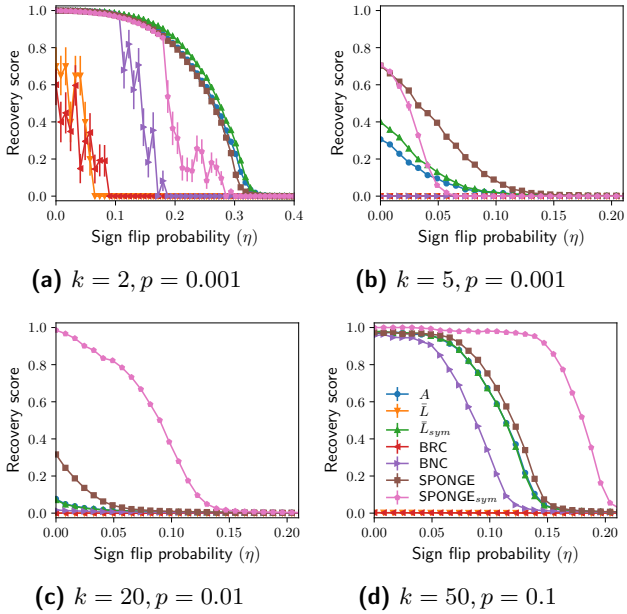
This section contains numerical experiments comparing our SPONGE and SPONGE$_{sym}$ algorithms[5] (setting $\tau^+ = \tau^- = 1$), with several existing spectral signed clustering techniques based on: the adjacency matrix A, the Signed Laplacian matrix $\overline{L}$, its symmetrically normalized version $\overline{L}_{sym}$ [32], and the two algorithms introduced in [10] that optimize the Balanced Ratio Cut and the Balanced Normalized Cut objectives. In all cases, the bottom $k-1$ (or top $k-1$ in the case of adjacency matrix A) eigenvectors of the relevant matrix or generalized eigenvalue problem are considered as an embedding, and kmeans++ is applied to obtain a k-clustering. Section 6.1 contains numerical experiments on synthetic graphs generated under the SSBM, while Section 6.2 details the results obtained on four different real-world data sets. Additional experiments are available in the appendix.

6.1 Signed stochastic block model

This section compares all algorithms on a variety of synthetic graphs generated from the SSBM. Since the ground truth partition is available, we measure accuracy by the Adjusted Rand Index (ARI) [22], an improved version of the popular Rand Index [43]. Both measures indicate how well the recovered partition matches ground truth, with a value close to 1, resp. 0, indicating an almost perfect recovery, resp. an almost random assignment of the nodes into clusters.

The SSBM considered here and introduced in Section 4.1 has four parameters: n, k, p and η. In our experiments, we fix $n = 10000$, and let $k \in \{2, 5, 10, 20, 50\}$ with clusters chosen of equal size n/k. We analyze the performance of all algorithms by plotting mean and standard error, over 20 repetitions, of the ARI as a function of η for $p \in \{0.001, 0.01, 0.1\}$. The results are reported in Figure 4. When $k = 2$ (Figure 4 (a)), $\overline{L}_{sym}$ performs slightly better than all other algorithms. As k increases, the SPONGE algorithms start to significantly outperform all other methods. In particular, while for intermediate values of k (Figure

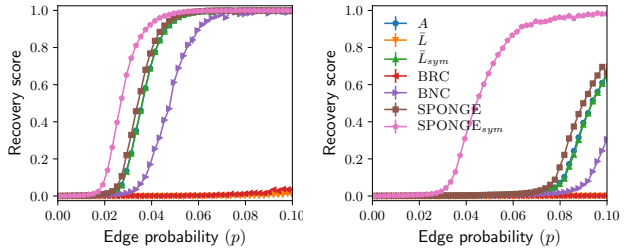
[5] Our current Python implementations are available at https://github.com/alan-turing-institute/signet



(a) $k = 2, p = 0.001$ **(b)** $k = 5, p = 0.001$

(c) $k = 20, p = 0.01$ **(d)** $k = 50, p = 0.1$

Figure 4: ARI recovery scores versus η for increasing k, with communities of equal size and $n = 10000$.

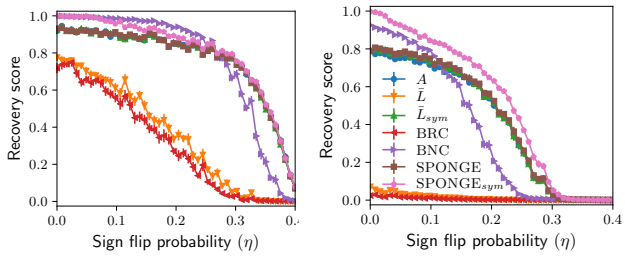


(a) $k = 20, \eta = 0.2$ **(b)** $k = 50, \eta = 0.1$

Figure 5: ARI recovery scores as a function of the edge probability p, for $k = 20$ and $k = 50$ at two different noise levels. The communities are of equal size, and $n = 10000$.

4 (b)) SPONGE was the best performer, once $k = 20$ or $k = 50$ (Figure 4 (c) and (d)) SPONGE$_{sym}$ was greatly superior, being able to perfectly recover the cluster structure (ARI $= 1$) when all other methods completely fail (ARI ≈ 0). We remark that similar results, showing excellent recovery for large k via SPONGE$_{sym}$, hold true over a wider range of values of the sparsity p, and are reported in the appendix.

We also tested the algorithms on SSBM graphs with clusters of unequal sizes, with the probability of each node to be part of a given cluster being uniformly sampled in $[0, 1]$, and subsequently normalized, which typically lead to widely different cluster sizes. Under this setting (see Figure 6), SPONGE$_{sym}$ was still the best performer, although the extent of the performance gap was less pronounced. Interestingly, the performance of BNC often matched (but rarely overcame) that of SPONGE$_{sym}$. Overall, we find that for large enough



(a) $k = 20, p = 0.1$ **(b)** $k = 50, p = 0.1$

Figure 6: ARI recovery scores of all algorithms, as a function of the noise η, for $k \in \{20, 50\}$ clusters of randomly chosen sizes, and fixed edge density $p = 0.1$.

$k \geq 5$, SPONGE and especially SPONGE$_{sym}$, outperform all state-of-art algorithms across a broad range of values for p, η, and for n sufficiently large for a clustering to be recoverable.

6.2 Real data

This section details the outcomes of experiments on a variety of real-world signed network data sets. Due to space constraints, we show results for the four algorithms that performed best on the synthetic experiments: SPONGE, SPONGE$_{sym}$, BNC and $\overline{L}_{sym}$. Since we no longer have ground truth, we compare the output of the algorithms by plotting the network adjacency matrix sorted by membership of the clusters produced. For our time series data applications, we also demonstrate visually that our algorithms have recovered meaningful information in their clusterings.

Wikipedia elections. We consider the classic data set of Wikipedia Requests for Adminship [51] from SNAP [35]; a network of positive, neutral, and negative votes between Wikipedia editors running in adminship elections. We construct a signed, undirected, weighted graph using the sums of edge weights for each pair of nodes. We then discard 0-weighted edges and consider only the largest connected component of the resulting graph. Thus, we obtain a graph on $n = 11,259$ nodes with $132,412$ (resp. $37,423$) positive (resp. negative) edges. Figure 7 shows the resulting adjacency matrix sorted by cluster membership with $k = 6$, where blue (resp. red) denotes positive (resp. negative) edges. Previous work on signed networks [36], also succeeded in finding clustering structure in this data. However, the majority of the nodes are placed in a single large cluster which is very sparse and does not exhibit discernible associations. A major advantage of the clustering in Figure 7 is that all clusters demonstrate a significantly higher ratio of positive to negative internal edges, compared to that of the graph as a whole.

Correlations of financial market returns. We consider daily prices for $n = 1500$ stocks in the S&P 1500 Index, during 2003-2015, and build correlation

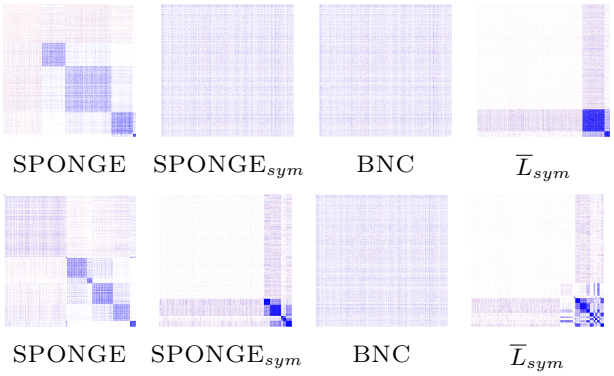


Figure 7: Sorted adjacency matrix of the Wikipedia graph for $k = 6$ (top row) and $k = 50$ (bottom row).

matrices from market excess returns. We refer the reader to the appendix, for a detailed overview of our steps. Figure 8 shows that, for $k \in \{10, 30\}$, we are able to find a meaningful segmentation of the market. In Figure 9, we interpret our results in light of the popular GICS sector decomposition [41].

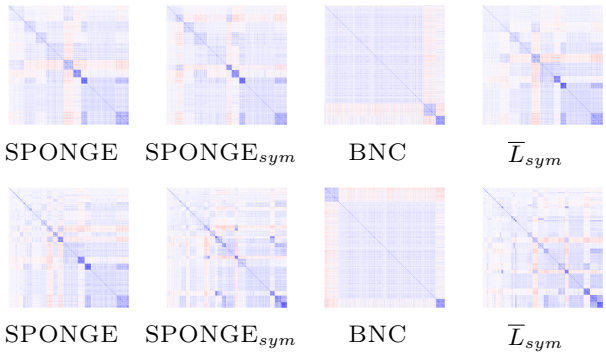


Figure 8: Adjacency matrix of the S&P 1500 data, sorted by cluster membership; $k = 10$ (top) and $k = 30$ (bottom).

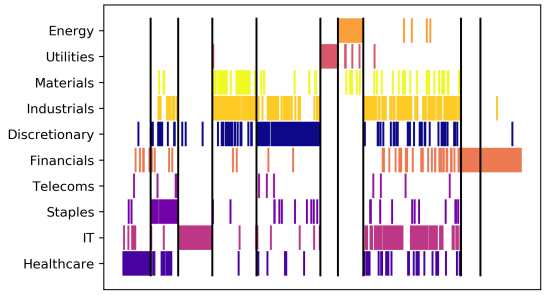


Figure 9: GICS decomposition for SPONGE$_{sym}$ clusters.

Correlations of Australian rainfalls. We also consider time series of historical rainfalls in locations throughout Australia. Edge weights are obtained from the pairwise Pearson correlation, leading to a complete signed graph on $n = 306$ nodes.

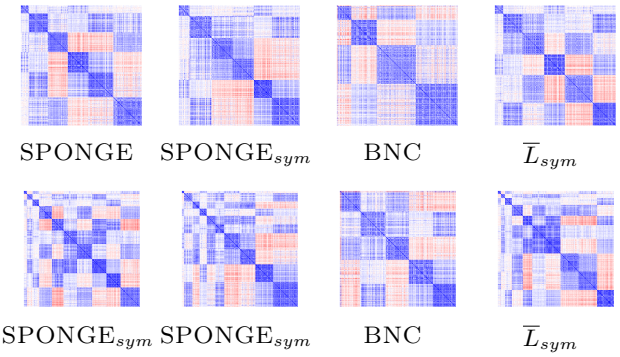


Figure 11: Sorted adjacency matrix of the Australian rainfall data set, with $k = 6$ (top) and $k = 10$ (bottom).

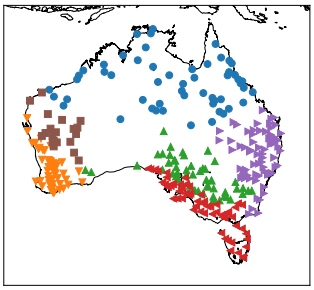


Figure 10: SPONGE: $k = 6$, Australian rainfalls data.

Figure 11 shows a clear clustering structure, for $k = \{6, 10\}$ clusters, and Figure 10 plots the points onto the corresponding geographic locations. SPONGE has very effectively identified geographic regions with similar climate, based only on the correlations of the rainfall measurements.

7 Discussion and future directions

We introduced a principled spectral algorithm (SPONGE) for clustering signed graphs, that amounts to solving a generalized eigenvalue problem, and provided a theoretical analysis for $k = 2$ clusters. Extensive numerical experiments demonstrate its robustness to noise and sampling sparsity. In particular, for very sparse graphs and large number of clusters k, we are able to recover clusterings when all state-of-the-art methods completely fail.

There are several directions for future work such as: (i) considering a more general SSBM that allows for different edge sampling probabilities and noise levels; (ii) handling the challenging setting of very sparse graphs, where $p = \Theta(\frac{1}{n})$; (iii) exploring the usefulness of the SPONGE embedding as a dimensionality reduction tool in multivariate time series analysis; (iv) exploring semidefinite programming based approaches, inspired by recent work on community detection [23]; and (v) investigating graph-based diffuse interface models utilizing the Ginzburg-Landau functionals, based on the MBO scheme [14, 37].

Acknowledgments. We thank Pedro Mercado for useful discussions. This work was funded by EPSRC grant EP/N510129/1 at The Alan Turing Institute.

References

[1] Epinions data set. http://www.epinions.com. Accessed: 2010-09-30.

[2] Slashdot data set. http://www.slashdot.com. Accessed: 2010-09-30.

[3] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering–a decade review. *Information Systems*, 53:16–38, 2015.

[4] Pranay Anchuri and Malik Magdon-Ismail. Communities and balance in signed networks: A spectral approach. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 235–242. IEEE, 2012.

[5] Sujogya Banerjee, Kaushik Sarkar, Sedat Gokalp, Arunabha Sen, and Hasan Davulcu. Partitioning signed bipartite graphs for classification of individuals and organizations. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 196–204. Springer, 2012.

[6] Jessica Bosch, Pedro Mercado, and Martin Stoll. Node classification for signed networks using diffuse interface methods. *CoRR*, abs/1809.06432, 2018.

[7] Dorwin Cartwright and Frank Harary. Structural balance: a generalization of Heider's theory. *Psychological review*, 63(5):277, 1956.

[8] K. Chiang, J. Whang, and I. Dhillon. Scalable clustering of signed networks using balance normalized cut. CIKM, 2012.

[9] Kai-Yang Chiang, Nagarajan Natarajan, Ambuj Tewari, and Inderjit S. Dhillon. Exploiting longer cycles for link prediction in signed networks. CIKM, 2011.

[10] Kai-Yang Chiang, Joyce Whang, and Inderjit S. Dhillon. Scalable Clustering of Signed Networks using Balance Normalized Cut. In *ACM Conference on Information and Knowledge Management (CIKM)*, oct 2012.

[11] Fan Chung, Alexander Tsiatas, and Wensong Xu. Dirichlet pagerank and ranking algorithms based on trust and distrust. *Internet Mathematics*, 9(1):113–134, 2013.

[12] M. Cucuringu. Synchronization over Z_2 and community detection in multiplex networks with constraints. *Journal of Complex Networks*, 3:469–506, 2015.

[13] M. Cucuringu, I. Koutis, S. Chawla, G. Miller, and R. Peng. Scalable Constrained Clustering: A Generalized Spectral Method. *Artificial Intelligence and Statistics Conference (AISTATS) 2016*, 2016.

[14] M. Cucuringu, A. Pizzoferrato, and Y. van Gennip. An MBO scheme for clustering and semi-supervised clustering of signed networks. *CoRR*, abs/1901.03091, 2019.

[15] Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[16] J. A. Davis. Clustering and structural balance in graphs. *Human Relations*, 20(2):181–187, 1967.

[17] T. Derr, Y. Ma, and J. Tang. Signed Graph Convolutional Network. *ArXiv e-prints*, August 2018.

[18] Patrick Doreian and Andrej Mrvar. Partitioning signed social networks. *Social Networks*, 31(1):1–11, 2009.

[19] Sergio M Focardi. Clustering economic and financial time series: Exploring the existence of stable correlation conditions. *The Intertek Group*, 2005.

[20] André Fujita, Patricia Severino, Kaname Kojima, João Ricardo Sato, Alexandre Galvão Patriota, and Satoru Miyano. Functional clustering of time series gene expression data by Granger causality. *BMC systems biology*, 6(1):137, 2012.

[21] Jean H. Gallier. Notes on elementary spectral graph theory. applications to graph clustering using normalized cuts. *CoRR*, abs/1311.2492, 2013.

[22] Alexander J. Gates and Yong-Yeol Ahn. The impact of random models on clustering similarity. *Journal of Machine Learning Research*, 18(87):1–28, 2017.

[23] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via Grothendieck's inequality. *Probability Theory and Related Fields*, 165(3-4):1025–1049, 2016.

[24] Gyeong-Gyun Ha, Jae Woo Lee, and Ashadun Nobi. Threshold network of a financial market using the p-value of correlation coefficients. *Journal of the Korean Physical Society*, 66(12):1802–1808, Jun 2015.

[25] Frank Harary. On the notion of balance of a signed graph. *Michigan Math. J.*, 2(2):143–146, 1953.

[26] Jao Ping Hou. Bounds for the least Laplacian eigenvalue of a signed graph. *Acta Mathematica Sinica*, 21(4):955–960, 2005.

[27] Cho-Jui Hsieh, Kai-Yang Chiang, and Inderjit S. Dhillon. Low-Rank Modeling of Signed Networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.

[28] B. W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49(2):291–307, 1970.

[29] Junghwan Kim, Haekyu Park, Ji-Eun Lee, and U Kang. Side: Representation learning in signed directed networks. In *WWW*, 2018.

[30] A. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM Journal on Scientific Computing*, 23(2):517–541, 2001.

[31] Srijan Kumar, Francesca Spezzano, V.S. Subrahmanian, and Christos Faloutsos. Edge weight prediction in weighted signed networks. In *ICDM*, 2016.

[32] Jérôme Kunegis, Stephan Schmidt, Andreas Lommatzsch, Jürgen Lerner, Ernesto William De Luca, and Sahin Albayrak. Spectral analysis of signed graphs for clustering, prediction and visualization. *SDM*, 10:559–570, 2010.

[33] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *WWW*, pages 641–650, 2010.

[34] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed Networks in Social Media. In *CHI*, pages 1361–1370, 2010.

[35] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data, June 2014.

[36] Pedro Mercado, Francesco Tudisco, and Matthias Hein. Clustering signed networks with the geometric mean of Laplacians. In *NIPS*. 2016.

[37] Ekaterina Merkurjev, Tijana Kostic, and Andrea L Bertozzi. An MBO scheme on graphs for classification and image processing. *SIAM Journal on Imaging Sciences*, 6(4):1903–1930, 2013.

[38] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pages 69–75, 2015.

[39] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3):431–461, Aug 2015.

[40] Nicos G Pavlidis, Vassilis P Plagianakos, Dimitris K Tasoulis, and Michael N Vrahatis. Financial forecasting through unsupervised clustering and neural networks. *Operational Research*, 6(2):103–127, 2006.

[41] Ryan L. Phillips and Rita Ormsby. Industry classification schemes: An analysis and review. *Journal of Business & Finance Librarianship*, 21(1):1–25, 2016.

[42] Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3120–3128, 2013.

[43] W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[44] Ahmed Sameh and Zhanye Tong. The trace minimization method for the symmetric generalized eigenvalue problem. *Journal of Computational and Applied Mathematics*, 123(1):155 – 175, 2000.

[45] Moshen Shahriari and Mahdi Jalili. Ranking nodes in signed social networks. *Social Network Analysis and Mining*, 4(1):172, Jan 2014.

[46] A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.*, 30(1):20–36, 2011.

[47] Stephen M. Smith, Karla L. Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F. Beckmann, Thomas E. Nichols, Joseph D. Ramsey, and Mark W. Woolrich. Network modelling methods for FMRI. *NeuroImage*, 54(2):875 – 891, 2011.

[48] Jiliang Tang, Charu Aggarwal, and Huan Liu. Node classification in signed social networks. In *SDM*, 2016.

[49] Jiliang Tang, Yi Chang, Charu Aggarwal, and Huan Liu. A survey of signed network mining in social media. *ACM Comput. Surv.*, 49(3):42:1–42:37, 2016.

[50] Suhang Wang, Jiliang Tang, Charu Aggarwal, Yi Chang, and Huan Liu. *Signed Network Embedding in Social Media*. 2017.

[51] Robert West, Hristo S. Paskov, Jure Leskovec, and Christopher Potts. Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association for Computational Linguistics*, 2:297–310, 2014.

[52] B. Yang, W. K. Cheung, and J. Liu. Community mining from signed social networks. *IEEE Trans Knowl Data Eng*, 19(10):1333–1348, 2007.

[53] Hartmut Ziegler, Marco Jenny, Tino Gruse, and Daniel A Keim. Visual market sector analysis for financial time series data. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 83–90. IEEE, 2010.