Regularized spectral methods for clustering signed networks

Mihai Cucuringu   MIHAI.CUCURINGU@STATS.OX.AC.UK
Department of Statistics and Mathematical Institute
University of Oxford
The Alan Turing Institute, London, UK

Apoorv Vikram Singh   APOORV.SINGH@NYU.EDU
Department of Computer Science and Engineering
New York University

Déborah Sulem   DEBORAH.SULEM@STATS.OX.AC.UK
Department of Statistics
University of Oxford

Hemant Tyagi   HEMANT.TYAGI@INRIA.FR
UMR 8524 - Laboratoire Paul Painlevé, F-59000
Inria, Univ. Lille, CNRS *

Abstract

We study the problem of k-way clustering in signed graphs. Considerable attention in recent years has been devoted to analyzing and modeling signed graphs, where the affinity measure between nodes takes either positive or negative values. Recently, Cucuringu et al. (2019) proposed a spectral method, namely SPONGE (*Signed Positive over Negative Generalized Eigenproblem*), which casts the clustering task as a generalized eigenvalue problem optimizing a suitably defined objective function. This approach is motivated by social balance theory, where the clustering task aims to decompose a given 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 mainly connected by negative edges. Through extensive numerical experiments, SPONGE was shown to achieve state-of-the-art empirical performance. On the theoretical front, Cucuringu et al. (2019) analyzed SPONGE, as well as the popular Signed Laplacian based spectral method under the setting of a Signed Stochastic Block Model, for $k = 2$ equal-sized clusters, in the regime where the graph is moderately dense.

In this work, we build on the results in Cucuringu et al. (2019) on two fronts for the normalized versions of SPONGE and the Signed Laplacian. Firstly, for both algorithms, we extend the theoretical analysis in Cucuringu et al. (2019) to the general setting of $k \geq 2$ unequal-sized clusters in the moderately dense regime. Secondly, we introduce regularized versions of both methods to handle sparse graphs – a regime where standard spectral methods are known to underperform – and provide theoretical guarantees under the same setting of a Signed Stochastic Block Model. To the best of our knowledge, regularized spectral methods have so far not been considered in the setting of clustering signed graphs. We complement our theoretical results with an extensive set of numerical experiments on synthetic data, and three real world data sets standard in the signed networks literature.

---

* Authors are listed in alphabetical order.

1. Introduction

Signed graphs. The recent years have seen a significant increase in interest for analysis of signed graphs, for tasks such as clustering (Chiang et al., 2014; Cucuringu et al., 2019), link prediction (Leskovec et al., 2010; Kumar et al., 2016) and visualization (Kunegis et al., 2010). Signed graphs are an increasingly popular family of undirected graphs, for which the edge weights may take both positive and negative values, thus encoding a measure of similarity or dissimilarity between the nodes. Signed social graphs have also received considerable attention to model trust relationships between entities, with positive (respectively, negative) edges encoding trust (respectively, distrust) relationships.

Clustering is arguably one of the most popular tasks in unsupervised machine learning, aiming at partitioning the node set such that the average connectivity or similarity between pairs of nodes within the same cluster is larger than that of pairs of nodes spanning different clusters. While the problem of clustering undirected unsigned graphs has been thoroughly studied for the past two decades (and to some extent, also that of clustering directed graphs in recent years), a lot less research has been undertaken on studying signed graphs.

Spectral clustering and regularization. Spectral clustering methods have become a fundamental tool with a broad range of applications in areas including network science, machine learning and data mining (von Luxburg, 2007). The attractivity of spectral clustering methods stems, on one hand, from its computational scalability by leveraging state-of-the-art eigensolvers, and on the other hand, from the fact that such algorithms are amenable to a theoretical analysis under suitably defined stochastic block models that quantify robustness to noise and sparsity of the measurement graph. Furthermore, on the theoretical side, understanding the spectrum of the adjacency matrix and its Laplacians, is crucial for the development of efficient algorithms with performance guarantees, and leads to a very mathematically rich set of problems. One such example from the latter class is that of Cheeger inequalities for general graphs, which relate the dominant eigenvalues of the Laplacian to edge expansion on graphs (Chung, 1996), extended to the setup of directed graphs (Chung, 2005), and more recently, to the graph Connection Laplacian arising in the context of the group synchronization problem (Bandeira et al., 2013), and higher-order Cheeger inequalities for multiway spectral clustering (Lee et al., 2014). There has been significant recent advances in theoretically analyzing spectral clustering methods in the context of stochastic block models; for a detailed survey, we refer the reader to the comprehensive recent survey of Abbe (2017).

In general, spectral clustering algorithms for unsigned and signed graphs typically have a common pipeline, where a suitable graph operator is considered (e.g., the graph Laplacian), its (usually k) extremal eigenvectors are computed, and the resulting point cloud in $\mathbb{R}^k$ is clustered using a variation of the popular k-means algorithm (Rohe et al., 2011). The main motivation for our current work stems from the lack of statistical guarantees in the above literature for the signed clustering problem, in the context of sparse graphs and large number of clusters $k \geq 3$. The problem of k-way clustering in signed graphs aims to find a partition of the node set into k disjoint clusters, such that most edges within clusters are positive,

while most edges across clusters are negative, thus altogether maximizing the number of *satisfied* edges in the graph. Another potential formulation to consider is to minimize the number of (*unsatisfied*) edges violating the partitions, i.e, the number of negative edges within clusters and positive edges across clusters.

A regularization step has been introduced in the recent literature motivated by the observation that properly regularizing the adjacency matrix A of a graph can significantly improve performance of spectral algorithms in the sparse regime. It was well known beforehand that standard spectral clustering often fails to produce meaningful results for sparse networks that exhibit strong degree heterogeneity (Amini et al., 2013; Jin, 2015). To this end, Chaudhuri et al. (2012) proposed the regularized graph Laplacian $L^\tau = D_\tau^{-1/2} A D_\tau^{-1/2}$, where $D_\tau = D + \tau I$, for $\tau \geq 0$. The spectral algorithm introduced and analyzed in Chaudhuri et al. (2012) splits the nodes into two random subsets and only relies on the subgraph induced by only one of the subsets to compute the spectral decomposition. Qin and Rohe (2013) studied the more traditional formulation of a spectral clustering algorithm that uses the spectral decomposition on the entire matrix (Ng et al., 2001), and proposed a regularized spectral clustering which they analyze. Subsequently, Joseph and Yu (2016) provided a theoretical justification for the regularization $A_\tau = A + \tau J$, where J denotes the all ones matrix, partly explaining the empirical findings of Amini et al. (2013) that the performance of regularized spectral clustering becomes insensitive for larger values of regularization parameters, and show that such large values can lead to better results. It is this latter form of regularization that we would be leveraging in our present work, in the context of clustering signed graphs. Additional references and discussion on the regularization literature are provided in Section 1.2.

Motivation & Applications. The recent surge of interest in analyzing signed graphs has been fueled by a very wide range of real-world applications, in the context of clustering, link prediction, and node rankings. Such social signed networks model trust relationships between users with positive (trust) and negative (distrust) edges. A number of online social services such as Epinions and Slashdot that allow users to express their opinions are naturally represented as signed social networks (Leskovec et al., 2010). Banerjee et al. (2012) considered shopping bipartite networks that encode like and dislike preferences between users and products. Other domain specific applications include personalized rankings via signed random walks (Jung et al., 2016), node rankings and centrality measures (Li et al., 2019), node classification (Tang et al., 2016), community detection (Yang et al., 2007; Chu et al., 2016), and anomaly detection, as in Kumar et al. (2014) which classifies users of an online signed social network as malicious or benign. In the very active research area of synthetic data generation, generative models for signed networks inspired by Structural Balance Theory have been proposed in Derr et al. (2018). Learning low-dimensional representations of graphs (network embeddings) have received tremendous attention in the recent machine learning literature, and graph convolutional networks-based methods have also been proposed for the setting of signed graphs, including Derr et al. (2018); Li et al. (2020), which provide network embeddings to facilitate subsequent downstream tasks, including clustering and link prediction.

A key motivation for our line of work stems from time series clustering (Aghabozorgi et al., 2015), an ubiquitous task arising in many applications that consider biological gene

3

expression data (Fujita et al., 2012), economic time series that capture macroeconomic variables (Focardi, 2005), and financial time series corresponding to large baskets of instruments in the stock market (Ziegler et al., 2010; Pavlidis et al., 2006). Driven by the clustering task, a popular approach in the literature is to consider similarity measures based on the Pearson correlation coefficient that captures linear dependence between variables and takes values in $[-1, 1]$. By construing 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, indicating the probability of observing a correlation at least as large as the measured sample correlation, assuming the null hypothesis is true. Such a thresholding step on the p-value associated to each individual sample correlation (Ha et al., 2015), renders the correlation network as a **sparse** matrix, which is one of the main motivations of our current work which proposes and analyzes algorithms for handling such sparse signed networks. We refer the reader to the popular work of Smith et al. (2011) for a detailed survey and comparison of various methodologies for turning time series data into networks, where the authors explore the interplay between fMRI time series and the network generation process. Importantly, they conclude that, in general, correlation-based approaches can be quite successful at estimating the connectivity of brain networks from fMRI time series.

Paper outline. This paper is structured as follows. The remainder of this Section 1 establishes the notation used throughout the paper, followed by a brief survey of related works in the signed clustering literature and graph regularization techniques for general graphs, along by a brief summary of our main contributions. Section 2 lays out the problem setup leading to our proposed algorithms in the context of the signed stochastic block model we subsequently analyze. Section 3 is a high-level summary of our main results across the two algorithms we consider. Section 4 contains the analysis of the proposed SPONGE$_{sym}$ algorithm, for both the sparse and dense regimes, for general number of clusters. Similarly, Section 5 contains the main theoretical results for the symmetric Signed Laplacian, under both sparsity regimes as well. Section 6 contains detailed numerical experiments on various synthetic and real world data sets, showcasing the performance of our proposed algorithms, as we vary the number of clusters, the relative cluster sizes, the sparsity regimes, and the regularization parameters. Finally, Section 7 is a summary and discussion of our main findings, with an outlook towards potential future directions. We defer to the Appendix additional proof details and a summary of the main technical tools used throughout.

1.1 Notation

We denote by $G = (V, E)$ a signed graph with vertex set V, edge set E, and adjacency matrix $A \in \{0, \pm 1\}^{n \times n}$. We will also refer to the unsigned subgraphs of positive (resp. negative) edges $G^+ = (V, E^+)$ (resp. $G^- = (V, E^-)$) with adjacency matrices A^+ (resp. A^-), such that $A = A^+ - A^-$. More precisely, $A_{ij}^+ = \max\{A_{ij}, 0\}$ and $A_{ij}^- = \max\{-A_{ij}, 0\}$, with $E^+ \cap E^- = \emptyset$, and $E^+ \cup E^- = E$. We denote by $\overline{D} = D^+ + D^-$ the signed degree matrix, with the unsigned versions given by $D^+ := A^+ \mathbb{1}$ and $D^- := A^- \mathbb{1}$. For a subset of nodes $C \subset V$, we denote its complement by $\overline{C} = V \setminus C$.

For a matrix $M \in \mathbb{R}^{m \times n}$, $\|M\|$ denotes its spectral norm $\|M\|_2$, i.e., its largest singular value, and $\|M\|_F$ denotes its Frobenius norm. When M is a $n \times n$ symmetric matrix, we denote $V_k(M)$ be the $n \times k$ matrix whose columns are given by the eigenvectors corresponding to the k smallest eigenvalues, and let $\mathcal{R}(V_k(M))$ denote the range space of these eigenvectors. We denote the eigenvalues of M by $(\lambda_j(M))_{j=1}^n$, with the ordering

$$\lambda_n(M) \leq \lambda_{n-1}(M) \leq \cdots \leq \lambda_1(M).$$

We also denote M_{i*} to be the i-th row of M. We denote $\mathbb{1} = (1, \ldots, 1)$ (resp. $\mathbb{1}_k$) the all ones column vector of size n (resp. k) and $\chi_1 = \frac{1}{\sqrt{k}} \mathbb{1}_k$. I_m denotes the square identity matrix of size m and is shortened to I when $m = n$. J_{mn} is the $m \times n$ matrix of all ones. Finally, for $a, b \geq 0$, we write $a \lesssim b$ if there exists a universal constant $C > 0$ such that $a \leq b$. If $a \lesssim b$ and $b \lesssim a$, then we write $a \asymp b$.

1.2 Related literature on signed clustering and graph regularization techniques

Signed clustering. There exists a very rich literature on algorithms developed to solve the k-way clustering problem, with spectral methods playing a central role in the developments of the last two decades. Such spectral techniques optimize an objective function via the eigen-decomposition of a suitably chosen graph operator (typically a graph Laplacian) built directly from the data, in order to obtain a low-dimensional embedding (most often of dimension k or $k - 1$). A clustering algorithm such as k-means or k-means++ is subsequently applied in order to extract the final partition.

Kunegis et al. (2010) introduced the combinatorial Signed Laplacian $\overline{L} = \overline{D} - A$ for the 2-way clustering problem. For heterogeneous degree distributions, normalized extensions are generally preferred, such as the random-walk Signed Laplacian $\overline{L_{rw}} = I - \overline{D}^{-1}A$, and the symmetric Signed Laplacian $\overline{L_{sym}} = I - \overline{D}^{-1/2}A\overline{D}^{-1/2}$. Chiang et al. (2012) pointed out a weakness in the Signed Laplacian objective for k-way clustering with $k > 2$, and proposed instead a Balanced Normalized Cut (BNC) objective based on the operator $\overline{L_{BNC}} = \overline{D}^{-1/2}(D^+ - A)\overline{D}^{-1/2}$. Mercado et al. (2016) based their clustering algorithm on a new operator called the *Geometric Mean of Laplacians*, and later extended this method in (Mercado et al., 2019) to a family of operators called the *Matrix Power Mean of Laplacians*. Previous work (Cucuringu et al., 2019) by a subset of the authors of the present paper introduced the symmetric SPONGE objective using the matrix operator $T = (L_{sym}^- + \tau^+I)^{-1/2}(L_{sym}^+ + \tau^-I)(L_{sym}^- + \tau^+I)^{-1/2}$, using the unsigned normalized Laplacians $L_{sym}^{\pm} = I - (D^{\pm})^{-1/2}A^{\pm}(D^{\pm})^{-1/2}$ and regularization parameters $\tau^+, \tau^- > 0$. This work also provides theoretical guarantees for the SPONGE and Signed Laplacian algorithms, in the setting of a Signed Stochastic Block Model.

Mercado et al. (2016) and Mercado et al. (2019) study the eigenspaces - in expectations and in probability - of several graph operators in a certain Signed Stochastic Block Model. However, this generative model differs from the one proposed in Cucuringu et al. (2019) that we analyze in this work. In the former, the positive and negative adjacency matrices do not have disjoint support, contrary to the latter. Moreover, their analysis is performed in the case of equal-size clusters. We will later show in our analysis that their result for the symmetric Signed Laplacian is not applicable in our setting.

Hsieh et al. (2012) proposed to perform low-rank matrix completion as a preprocessing step, before clustering using the top k eigenvectors of the completed matrix. For $k = 2$, Cucuringu (2015) showed that signed clustering can be cast as an instance of the group synchronization (Singer, 2011) problem over $\mathbb{Z}_2$, potentially with constraints given by available side information, for which spectral, semidefinite programming relaxations, and message passing algorithms have been considered. In recent work, Cucuringu et al. (2021) proposed a formulation for the signed clustering problem that relates to graph-based diffuse interface models utilizing the Ginzburg-Landau functionals, based on an adaptation of the classic numerical Merriman-Bence-Osher (MBO) scheme for minimizing such graph-based functionals (Merkurjev et al., 2014). We refer the reader to Gallier (2013) for a recent survey on clustering signed and unsigned graphs.

In a different line of work, known as *correlation clustering*, Bansal et al. (2004) considered the problem of clustering signed complete graphs, proved that it is NP-complete, and proposed two approximation algorithms with theoretical guarantees on their performance. On a related note, Demaine et al. (2006) studied the same problem but for arbitrary weighted graphs, and proposed an $O(\log n)$ approximation algorithm based on linear programming. For correlation clustering, in contrast to k-way clustering, the number of clusters is not given in advance, and there is no normalization with respect to size or volume.

Regularization in the sparse regime. In many applications, real-world networks are sparse. In this context, regularization methods have increased the performance of traditional spectral clustering techniques, both for synthetic Stochastic Block Models and real data sets (Chaudhuri et al., 2012; Amini et al., 2013; Joseph and Yu, 2016; Le et al., 2015).

Chaudhuri et al. (2012) regularize the Laplacian matrix by adding a (typically small) weight τ to the diagonal entries of the degree matrix $L_\tau = I - D_\tau^{-1/2} A D_\tau^{-1/2}$ with $D_\tau = D + \tau I$. Amini et al. (2013) regularize the graph by adding a weight τ/n to every edge, leading to the Laplacian $\widetilde{L}_\tau = I - D_\tau^{-1/2} A_\tau D_\tau^{-1/2}$ with $A_\tau = A + \tau/n \mathbb{1}\mathbb{1}^T$ and $D_\tau = A_\tau \mathbb{1}$. Le et al. (2017) prove that this technique makes the adjacency and Laplacian matrices concentrate for inhomogeneous Erdős-Rényi graphs. Zhang and Rohe (2018) show that this technique prevents spectral clustering from overfitting through the analysis of dangling sets. In (Le et al., 2017), Le et al. propose a graph trimming method in order to reduce the degree of certain nodes. This is achieved by reducing the entries of the adjacency matrix that lead to high-degree vertices. Zhou and Amini (2018) add a spectral truncation step after this regularization method, and prove consistency results in the bipartite Stochastic Block Model.

Very recently, regularization methods using powers of the adjacency matrix have been introduced. Abbe et al. (2020) transform the adjacency matrix into the operator $A_r = \mathbb{1}\left\{(I + A)^r \geq 1\right\}$, where the indicator function is applied entrywise. With this method, spectral clustering achieves the fundamental limit for weak recovery in the sparse setting. Very similarly, Stephan and Massoulié (2019) transform the adjacency matrix into a distance matrix of outreach l, which links pairs of nodes that are l far apart w.r.t the graph distance.

1.3 Summary of our main contributions

This work extends the results obtained in Cucuringu et al. (2019) by a subset of the authors of our present paper. This previous work introduced the SPONGE algorithm, a princi-

pled and scalable spectral method for the signed clustering task that amounts to solving a generalized eigenvalue problem. Cucuringu et al. (2019) provided a theoretical analysis of both the newly introduced SPONGE algorithm and the popular Signed Laplacian-based method (Kunegis et al., 2010), quantifying their robustness against the sampling sparsity and noise level, under the setting of a Signed Stochastic Block Model (SSBM). These were the first such theoretical guarantees for the signed clustering problem under a suitably defined stochastic graph model. However, the analysis in Cucuringu et al. (2019) was restricted to the setting of two equally-sized clusters, which is less realistic in light of most real world applications. Furthermore, the same previous line of work considered the moderately dense regime in terms of the edge sampling probability p, in particular, it operated in the setting where $\mathbb{E}[\overline{D}_{jj}] \gtrsim \ln n$, i.e., $p \gtrsim \frac{\ln n}{n}$. Many real world applications involve large but very sparse graphs, with $p = \Theta\left(\frac{1}{n}\right)$, which provides motivation for our present work.

We summarize below our main contributions, and start with the remark that the theoretical analysis in the present paper pertains to the normalized version of SPONGE (denoted as SPONGE_{sym}) and the symmetric Signed Laplacian, while Cucuringu et al. (2019) analyzed only the un-normalized versions of these signed operators. The experiments reported in Cucuringu et al. (2019) also consider such normalized matrix operators, and show their superior performance over their respective un-normalized versions, further providing motivational ground for our current work.

(i) Our first main contribution is to analyze the two above-mentioned signed operators, namely SPONGE_{sym} and the symmetric Signed Laplacian, in the general SSBM model with $k \geq 2$ and unequal-cluster sizes, in the moderately dense regime. In particular, we evaluate the accuracy of both signed clustering algorithms by bounding the mis-clustering rate of the entire pipelines, as achieved by the popular k-means algorithm.

(ii) Our second contribution is to introduce and analyze new regularized versions of both SPONGE_{sym} and the symmetric Signed Laplacian, under the same general SSBM model, but in the sparse graph regime $\mathbb{E}[\overline{D}_{jj}] \gtrsim 1$, a setting where standard spectral methods are known to underperform. To the best of our knowledge, this sparsity regime has not been previously considered in the literature of signed networks; such regularized spectral methods have so far not been considered in the setting of clustering signed networks, or more broadly in the signed networks literature, where such regularization could prove useful for other related downstream tasks. One important aspect of regularization techniques is the choice of the regularization parameters. We show that our proposed algorithms can benefit from careful regularization and attain a higher level of accuracy in the sparse regime, provided that the regularization parameters scale as an adequate power of the average degree in the graph. These findings are supported by our experiments on real-world datasets.

2. Problem setup

This section details the two algorithms for the signed clustering problem that we will analyze subsequently, namely, SPONGE_{sym} (Symmetric Signed Positive Over Negative Generalized

Eigenproblem) and the symmetric Signed Laplacian, along with their respective regularized versions.

2.1 Clustering via the SPONGE$_{sym}$ algorithm

The symmetric SPONGE method, denoted as SPONGE$_{sym}$, aims at jointly minimizing two measures of badness in a signed clustering problem. For an unsigned graph G and $X, Y \subset V$, we define the cut function $\text{Cut}_G(X, Y) := \sum_{i \in X, j \in Y} A_{ij}$, and denote the volume of X by $\text{Vol}_G(X) := \sum_{i \in X} \sum_{j=1}^n A_{ij}$.

For a given cluster set $C \subset V$, $\text{Cut}_G(C, \overline{C})$ is the total weight of edges crossing from C to $\overline{C}$ and $\text{Vol}_G(C)$ is the sum of (weighted) degrees of nodes in C. With this notation in mind and motivated by the approach of Cucuringu et al. (2016) in the context of constrained clustering, the symmetric SPONGE algorithm for signed clustering aims at minimizing the following two measures of *badness* given by $\frac{\text{Cut}_{G^+}(C, \overline{C})}{\text{Vol}_{G^+}(C)}$ and $\left(\frac{\text{Cut}_{G^-}(C, \overline{C})}{\text{Vol}_{G^-}(C)} \right)^{-1} = \frac{\text{Vol}_{G^-}(C)}{\text{Cut}_{G^-}(C, \overline{C})}$. To this end, we consider "merging" the objectives, and aim to solve

$$\min_{C \subset V} \frac{\frac{\text{Cut}_{G^+}(C, \overline{C})}{\text{Vol}_{G^+}(C)} + \tau^-}{\frac{\text{Cut}_{G^-}(C, \overline{C})}{\text{Vol}_{G^-}(C)} + \tau^+},$$

where $\tau^+ > 0, \tau^- \geq 0$ denote trade-off parameters. For k-way signed clustering into disjoint clusters $C_1, \ldots, C_k$, we arrive at the combinatorial optimization problem

$$\min_{C_1, \ldots, C_k} \sum_{i=1}^k \left(\frac{\frac{\text{Cut}_{G^+}(C_i, \overline{C_i})}{\text{Vol}_{G^+}(C_i)} + \tau^-}{\frac{\text{Cut}_{G^-}(C_i, \overline{C_i})}{\text{Vol}_{G^-}(C_i)} + \tau^+} \right). \tag{1}$$

Let D^+, L^+ denote respectively the degree matrix and un-normalized Laplacian associated with G^+, and $L_{sym}^+ = (D^+)^{-1/2} L^+ (D^+)^{-1/2}$ denote the symmetric Laplacian matrix for G^+ (similarly for L_{sym}^-, D^-, L^-). For a subset $C_i \subset V$, denote $\mathbb{1}_{C_i}$ to be the indicator vector for C_i so that $(\mathbb{1}_{C_i})_j$ equals 1 if $j \in C_i$, and is 0 otherwise. Now define the normalized indicator vector $x_{C_i} \in \mathbb{R}^n$ where

$$x_{C_i} = \left(\frac{\text{Cut}_{G^-}(C_i, \overline{C_i})}{\text{Vol}_{G^-}(C_i)} + \tau^+ \right)^{-1/2} \frac{1}{\sqrt{\text{Vol}_{G^+}(C_i)}} (D^+)^{1/2} \mathbb{1}_{C_i}.$$

In light on this, one can verify that

$$x_{C_i}^\top x_{C_i} = \left(\frac{\text{Cut}_{G^-}(C_i, \overline{C_i})}{\text{Vol}_{G^-}(C_i)} + \tau^+ \right)^{-1} \frac{\mathbb{1}_{C_i}^\top D^+ \mathbb{1}_{C_i}}{\text{Vol}_{G^+}(C_i)} = \left(\frac{\text{Cut}_{G^-}(C_i, \overline{C_i})}{\text{Vol}_{G^-}(C_i)} + \tau^+ \right)^{-1},$$

$$x_{C_i}^\top L_{sym}^+ x_{C_i} = \left(\frac{\text{Cut}_{G^-}(C_i, \overline{C_i})}{\text{Vol}_{G^-}(C_i)} + \tau^+ \right)^{-1} \frac{\mathbb{1}_{C_i}^\top L^+ \mathbb{1}_{C_i}}{\text{Vol}_{G^+}(C_i)}$$

$$= \left(\frac{\text{Cut}_{G^-}(C_i, \overline{C_i})}{\text{Vol}_{G^-}(C_i)} + \tau^+ \right)^{-1} \frac{\text{Cut}_{G^+}(C_i, \overline{C_i})}{\text{Vol}_{G^+}(C_i)}.$$

Hence (1) is equivalent to the following discrete optimization problem

$$\min_{C_1,\ldots,C_k} \sum_{i=1}^{k} x_{C_i}^{\top}(L_{sym}^+ + \tau^- I)x_{C_i} \tag{2}$$

which is NP-Hard. A common approach to solve this problem is to drop the discreteness constraints, and allow x_{C_i} to take values in $\mathbb{R}^n$. To this end, we introduce a new set of vectors $z_1,\ldots,z_k \in \mathbb{R}^n$ such that they are orthonormal with respect to the matrix $L_{sym}^- + \tau^+ I$, i.e., $z_i^{\top}(L_{sym}^- + \tau^+ I)z_{i'} = \delta_{ii'}$. This leads to the continuous optimization problem

$$\min_{z_i^{\top}(L_{sym}^-+\tau^+I)z_{i'}=\delta_{ii'}} \sum_{i=1}^{k} z_i^{\top}(L_{sym}^+ + \tau^- I)z_i. \tag{3}$$

Note that the above choice of vectors $z_1,\ldots,z_k$ is not really a relaxation of (2) since $x_{C_1},\ldots,x_{C_k}$ are not necessarily $(L_{sym}^- + \tau^+ I)$-orthonormal, but (3) can be conveniently formulated as a suitable generalized eigenvalue problem, similar to the approach in Cucuringu et al. (2016). Indeed, denoting $y_i = (L_{sym}^- + \tau^+ I)^{1/2}z_i$, and $Y = [y_1,\ldots,y_k] \in \mathbb{R}^{n\times k}$, (3) can be rewritten as

$$\min_{Y^{\top}Y=I} \text{Tr}\Big(Y^{\top}(L_{sym}^- + \tau^+ I)^{-1/2}(L_{sym}^+ + \tau^- I)(L_{sym}^- + \tau^+ I)^{-1/2}Y\Big),$$

the solution to which is well known to be given by the smallest k eigenvectors of

$$T = (L_{sym}^- + \tau^+ I)^{-1/2}(L_{sym}^+ + \tau^- I)(L_{sym}^- + \tau^+ I)^{-1/2},$$

see for e.g. (Sameh and Tong, 2000, Theorem 2.1). However this is not practically viable for large scale problems, since computing T itself is already expensive. To circumvent this issue, one can instead consider the embedding in $\mathbb{R}^k$ corresponding to the smallest k generalized eigenvectors of the symmetric definite pair $(L_{sym}^+ + \tau^- I, L_{sym}^- + \tau^+ I)$. There exist many efficient solvers for solving large scale generalized eigenproblems for symmetric definite matrix pairs. In our experiments, we use the LOBPCG (Locally Optimal Block Preconditioned Conjugate Gradient method) solver introduced in Knyazev (2001).

One can verify that (λ, v) is an eigenpair[1] of T iff $(\lambda, (L_{sym}^- + \tau^+ I)^{-1/2}v)$ is a generalized eigenpair of $(L_{sym}^+ + \tau^- I, L_{sym}^- + \tau^+ I)$. Indeed, for symmetric matrices A, B with $A \succ 0$, it holds true for $w = A^{-1/2}v$ that

$$A^{-1/2}BA^{-1/2}v = \lambda v \iff Bw = \lambda A w.$$

Therefore, denoting $V_k(T) \in \mathbb{R}^{n\times k}$ to be the matrix consisting of the smallest k eigenvectors of T, and $G_k(T) \in \mathbb{R}^{n\times k}$ to be the matrix of the smallest k generalized eigenvectors of $(L_{sym}^+ + \tau^- I, L_{sym}^- + \tau^+ I)$, it follows that

$$G_k(T) = (L_{sym}^- + \tau^+ I)^{-1/2}V_k(T). \tag{4}$$

Hence upon computing $G_k(T)$, we will apply a suitable clustering algorithm on the rows of $G_k(T)$ such as the popular k-means++ (Arthur and Vassilvitskii, 2007), to arrive at the final partition.

---

[1] With λ denoting its eigenvalue, and v the corresponding eigenvector.

Remark 1 *In Cucuringu et al. (2019), similar arguments as above were shown for the SPONGE algorithm which led to computing the k smallest generalized eigenvectors of the matrix pair $(L^+ + \tau^- D^-, L^- + \tau^+ D^+)$. SPONGE$_{sym}$ was proposed in Cucuringu et al. (2019) but no theoretical results were provided.*

Clustering in the sparse regime. We also provide a version of SPONGE$_{sym}$ for the case where G is sparse, i.e., the graph has very few edges and is typically disconnected. In this setting, we consider a regularized version of SPONGE$_{sym}$ wherein a weight is added to each edge (including self-loops) of the positive and negative subgraphs, respectively. Formally, for regularization parameters $\gamma^+, \gamma^- \geq 0$, let us define $A^{\pm}_{\gamma^{\pm}} := A^{\pm} + \frac{\gamma^{\pm}}{n}\mathbb{1}\mathbb{1}^{\top}$ to be the regularized adjacency matrices for the unsigned graphs G^+, G^- respectively. Denoting $D^{\pm}_{\gamma^{\pm}}$ to be the degree matrix of $A^{\pm}_{\gamma^{\pm}}$, the normalized Laplacians corresponding to $A^{\pm}_{\gamma^{\pm}}$ are given by

$$L^{\pm}_{sym,\gamma^{\pm}} = I - (D^{\pm}_{\gamma^{\pm}})^{-1/2}A^{\pm}_{\gamma^{\pm}}(D^{\pm}_{\gamma^{\pm}})^{-1/2}.$$

Given the above modifications, let $V_k(T_{\gamma^+,\gamma^-}) \in \mathbb{R}^{n \times k}$ denote the matrix consisting of the smallest k eigenvectors of

$$T_{\gamma^+,\gamma^-} = (L^-_{sym,\gamma^-} + \tau^+ I)^{-1/2}(L^+_{sym,\gamma^+} + \tau^- I)(L^-_{sym,\gamma^-} + \tau^+ I)^{-1/2}.$$

For the same reasons discussed earlier, we will consider the embedding given by the smallest k generalized eigenvectors of the matrix pencil $(L^+_{sym,\gamma^+} + \tau^- I, L^-_{sym,\gamma^-} + \tau^+ I)$, namely $G_k(T_{\gamma^+,\gamma^-})$ where

$$G_k(T_{\gamma^+,\gamma^-}) = (L^-_{sym,\gamma^-} + \tau^+ I)^{-1/2}V_k(T_{\gamma^+,\gamma^-}),$$

as in (44). The rows of $G_k(T_{\gamma^+,\gamma^-})$ can then be clustered using an appropriate clustering procedure, such as k-means++.

Remark 2 *Regularized spectral clustering for unsigned graphs involves adding $\frac{\gamma}{n}\mathbb{1}\mathbb{1}^{\top}$ to the adjacency matrix, followed by clustering the embedding given by the smallest k eigenvectors of the normalized Laplacian (of the regularized adjacency), see for e.g. Amini et al. (2013); Le et al. (2017). To the best of our knowledge, regularized spectral clustering methods have not been explored thus far in the context of sparse signed graphs.*

2.2 Clustering via the symmetric Signed Laplacian

The rationale behind the use of the (un-normalized) Signed Laplacian $\overline{L}$ for clustering is justified by Kunegis et al. in Kunegis et al. (2010) using the signed ratio cut function. For $C \subset V$,

$$sRCut(C,\overline{C}) = \left(2\text{Cut}_{G+}(C,\overline{C}) + \text{Cut}_{G-}(C,C) + \text{Cut}_{G-}(\overline{C},\overline{C})\right)\left(\frac{1}{|C|} + \frac{1}{|\overline{C}|}\right). \quad (5)$$

For 2-way clustering, minimizing this objective corresponds to minimizing the number of positive edges between the two classes and the number of negative edges inside each class. Moreover, (5) is equivalent to the following optimization problem

$$\min_{u \in \mathcal{U}} u^{\top}\overline{L}u,$$

where $\mathcal{U} \in \mathbb{R}^n$ is the set of vectors of the form $\forall i \in [n], u_i = \pm \frac{1}{2}\left(\sqrt{\frac{|C|}{|\overline{C}|}} + \sqrt{\frac{|\overline{C}|}{|C|}}\right)$.

However, Gallier (2016) noted that this equivalence does not generalize to $k > 2$, and defined a new notion of signed cut, called the signed normalized cut function. For a partition $C_1, \ldots, C_k$ with membership matrix $X \in \{0,1\}^{n \times k}$,

$$sNCut(C_1, \ldots, C_k) = \sum_{i=1}^{k} \frac{\text{Cut}_G(C_i, \overline{C_i})}{\text{Vol}_G(C_i)} + 2\frac{\text{Cut}_{G^-}(C_i, C_i)}{\text{Vol}_G(C_i)} = \sum_{i=1}^{k} \frac{(X^i)^\top \overline{L} X^i}{(X^i)^\top \overline{D} X^i},$$

with X^i the i-th column of X. Compared to (5), this objective also penalizes the number of negative edges across two subsets, which may not be a desirable feature for signed clustering. Minimizing this function with a relaxation of the constraint that $X^i \in \{0,1\}^n$ leads to the following problem

$$\min_{Y^\top Y = I} \text{Tr}\left(Y^\top \overline{L_{sym}} Y\right).$$

The minimum of this problem is obtained by stacking column-wise the k eigenvectors of $\overline{L_{sym}}$ corresponding to the smallest eigenvalues, i.e. $V_k(\overline{L_{sym}})$. Therefore, one can apply a clustering algorithm to the rows of the matrix $V_k(\overline{L_{sym}})$ to find a partition of the set of nodes V.

In fact, we will consider using only the $k-1$ smallest eigenvectors of $\overline{L_{sym}}$ and applying the k-means++ algorithm on the rows of $V_{k-1}(\overline{L_{sym}})$. This will be justified in our analysis via a stochastic generative model, namely the Signed Stochastic Block Model (SSBM), introduced in the next subsection. Under this model assumption, we will see later that the embedding given by the $k-1$ smallest eigenvectors of the symmetric Signed Laplacian of the expected graph has k distinct rows (with two rows being equal if and only if the corresponding nodes belong to the same cluster).

Clustering in the sparse regime. When G is sparse, we propose a spectral clustering method based on a regularization of the signed graph, leading to a regularized Signed Laplacian. To this end, for $\gamma^+, \gamma^- \geq 0$, recall the regularized adjacency matrices $A_{\gamma^\pm}^\pm$, with degree matrices $D_{\gamma^\pm}^\pm$, for the unsigned graphs G^+, G^- respectively. In light of this, the regularized signed adjacency and degree matrices are defined as follows

$$A_\gamma := A_{\gamma^+}^+ - A_{\gamma^-}^- = A + \frac{\gamma^+ - \gamma^-}{n} \mathbb{1}\mathbb{1}^\top,$$

$$\overline{D}_\gamma := D_{\gamma^+}^+ + D_{\gamma^-}^- = D^+ + \gamma^+ I + D^- + \gamma^- I = \overline{D} + (\gamma^+ + \gamma^-)I = \overline{D} + \gamma I,$$

with $\gamma := \gamma^+ + \gamma^-$. Our regularized Signed Laplacian is the symmetric Signed Laplacian on this regularized signed graph, i.e.

$$L_\gamma := I - (\overline{D}_\gamma)^{-1/2} A_\gamma (\overline{D}_\gamma)^{-1/2}. \tag{6}$$

Similarly to the symmetric Signed Laplacian, our clustering algorithm in the sparse case finds the $k-1$ smallest eigenvectors of L_γ and applies the k-means algorithm on the rows of $V_{k-1}(L_\gamma)$.

Remark 3 *For the choice $\gamma^+ = \gamma^-$, the regularized Laplacian becomes*

$$L_\gamma := I - (\overline{D}_\gamma)^{-1/2} A (\overline{D}_\gamma)^{-1/2},$$

with $\overline{D}_\gamma = \overline{D} + (\gamma^+ + \gamma^-) I$. This regularization scheme is very similar to the degree-corrected normalized Laplacian defined in Chaudhuri et al. (2012).

2.3 Signed Stochastic Block Model (SSBM)

Our work theoretically analyzes the clustering performance of SPONGE_{sym} and the symmetric Signed Laplacian algorithms under a signed random graph model, also considered previously in (Cucuringu et al., 2019; Cucuringu et al., 2021). We recall here its definition and parameters.

- n: the number of nodes in network;

- k: the number of planted communities;

- p: the probability of an edge to be present;

- η: the probability of flipping the sign of an edge;

- $C_1, \ldots, C_k$: an arbitrary partition of the vertices with sizes $n_1, \ldots, n_k$.

We first partition the vertices (arbitrarily) into clusters $C_1, \ldots, C_k$ where $|C_i| = n_i$. Next, we generate a *noiseless* measurement graph from the Erdős-Rényi 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 $\eta \in [0, 1/2)$. This results in the realization of a signed graph instance G from the SSBM ensemble.

Let $A \in \{0, \pm1\}^{n \times n}$ denote the adjacency matrix of G, and note that $(A_{jj'})_{j \leq j'}$ 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, $(A_{jj'}^+)_{j \leq j'}$ are independent, and similarly $(A_{jj'}^-)_{j \leq j'}$ are also independent. But for given $j, j' \in [n]$ with $j \neq j'$, $A_{jj'}^+$ and $A_{jj'}^-$ are dependent. Let $d_i^{\pm}$ denote the degree of a node in cluster i, for $i \in [k]$ in the graph $\mathbb{E}[A^{\pm}]$. Moreover, under this model, the expected signed degree matrix is the scaled identity matrix $\mathbb{E}\overline{D} = \overline{d}I$, with $\overline{d} = p(n-1)$.

Remark 4 *Contrary to stochastic block models for unsigned graphs, we do not require (for the purpose of detecting clusters) that the intra-cluster edge probabilities to be different from those of inter-cluster edges, since the sign of the edges already achieves this purpose implicitly. In fact, it is the noise parameter η that is crucial for identifying the underlying latent cluster structure.*

To formulate our theoretical results we will also need the following notations. Let $s_i = n_i/n$ denote the fraction of nodes in cluster i, with l (resp. s) denoting the fraction for the largest (resp. smallest) cluster. Hence, the size of the largest (resp. smallest) cluster is nl (resp. ns). Following the notation in Lei and Rinaldo (2015), we will denote $\mathbb{M}_{n,k}$

to be the class of "membership" matrices of size $n \times k$, and denote $\hat{\Theta} \in \mathbb{M}_{n,k}$ to be the ground-truth membership matrix containing k distinct indicator row-vectors (one for each cluster), i.e., for $i \in [k]$ and $j \in [n]$,

$$\hat{\Theta}_{ji} = \begin{cases} 1 & \text{if node } j \in \text{ cluster } C_i, \\ 0 & \text{otherwise.} \end{cases}$$

We also define the normalized membership matrix Θ corresponding to $\hat{\Theta}$, where for $i \in [k]$ and $j \in [n]$,

$$\Theta_{ji} = \begin{cases} 1/\sqrt{n_i} & \text{if node } j \in \text{ cluster } C_i, \\ 0 & \text{otherwise.} \end{cases}$$

3. Summary of main results

We now summarize our theoretical results for SPONGE_{sym} and the symmetric Signed Laplacian methods, when the graph is generated from the SSBM ensemble.

3.1 Symmetric SPONGE

We begin by describing conditions under which the rows of the matrix $G_k(T)$ approximately preserve the ground truth clustering structure. Before explaining our results, let us denote the matrix $\overline{T}$ to be the analogue of T for the expected graph, i.e.,

$$\overline{T} = (\overline{L_{sym}^-} + \tau^+ I)^{-1/2}(\overline{L_{sym}^+} + \tau^- I)(\overline{L_{sym}^-} + \tau^+ I)^{-1/2},$$

where $\overline{L_{sym}^\pm} = I - (\mathbb{E}[D^\pm])^{-1/2} \mathbb{E}[A^\pm](\mathbb{E}[D^\pm])^{-1/2}$. We first show that for suitable values of $\tau^+ > 0, \tau^- \geq 0$ (with n large enough), the smallest k eigenvectors of $\overline{T}$, denoted by $V_k(\overline{T})$, are given by $V_k(\overline{T}) = \Theta R$, for some $k \times k$ rotation matrix R. Hence, the rows of $V_k(\overline{T})$ have the same clustering structure as that of Θ. Denoting $G_k(\overline{T}) \in \mathbb{R}^{n \times k}$ to be the matrix consisting of the k smallest generalized eigenvectors of $(L_{sym}^+ + \tau^- I, \overline{L_{sym}^-} + \tau^+ I)$, and recalling (4), we can relate $G_k(\overline{T})$ and $V_k(\overline{T})$ via

$$G_k(\overline{T}) = (\overline{L_{sym}^-} + \tau^+ I)^{-1/2} V_k(\overline{T}). \tag{7}$$

It turns out that when $V_k(\overline{T}) = \Theta R$, and in light of the expression for $\overline{L_{sym}^-} + \tau^+ I$ from (24), we arrive at $G_k(\overline{T}) = \Theta(C^-)^{-1/2}R$, where $C^- \succ 0$ is as in (18). Since $(C^-)^{-1/2}R$ is invertible, it follows that $G_k(\overline{T})$ has k distinct rows, with the rows that belong to the same cluster being identical. The remaining arguments revolve around deriving concentration bounds on $\|T - \overline{T}\|$, which imply (for p large enough) that the distance between the column spans of $V_k(T)$ and $V_k(\overline{T})$ is small, i.e., there exists an orthonormal matrix O such that $\|V_k(T) - V_k(\overline{T})O\|$ is small. Finally, the expressions in (4) and (7) altogether imply that $\|G_k(T) - G_k(\overline{T})O\|$ is small, which is an indication that the rows of $G_k(T)$ approximately preserve the clustering structure encoded in Θ.

The above discussion is summarized in the following theorem, which is our first main result for SPONGE_{sym} in the moderately dense regime.

Theorem 5 (Restating Theorem 29) (Eigenspace alignment of SPONGE$_{sym}$ in the dense case) *Assuming $n \geq \max\left\{\frac{2(1-\eta)}{s(1-2\eta)}, \frac{2\eta}{(1-l)(1-\eta)}\right\}$, suppose that $\tau^+ > 0, \tau^- \geq 0$ are chosen to satisfy*

$$\tau^+ > \frac{16\eta}{\beta s(1-2\eta)}, \qquad \tau^- < \frac{\beta}{2}\left(\frac{s(1-2\eta)}{s(1-2\eta)+2\eta}\right)\min\left\{\frac{1}{4(1-\beta)}, \frac{\tau^+}{8}\right\}$$

where β, η satisfy one of the following conditions

1. *$\beta = \frac{4\eta}{s(1-2\eta)+4\eta}$ and $0 < \eta < \frac{1}{2}$, or*

2. *$\beta = \frac{1}{2}$ and $\eta \leq \frac{s}{2s+4}$.*

Then $V_k(\overline{T}) = \Theta R$ and $G_k(\overline{T}) = \Theta(C^-)^{-1/2}R$, where R is a rotation matrix, and $C^- \succ 0$ is as defined in (18). Moreover, for any $\varepsilon, \delta \in (0,1)$, there exists a constant $\widetilde{c}_\varepsilon > 0$ such that the following is true. If p satisfies

$$p \geq \max\left\{\widetilde{c}_\varepsilon C_2(s,\eta,l), \frac{256 C_1^4(\tau^+,\tau^-)(2+\tau^+)^4}{\delta^4(1+\tau^-)^4(1-\beta)^4}C_2(s,\eta,l), \frac{81}{(1-l)\delta^4}\right\}\frac{\ln(4n/\varepsilon)}{n}$$

with $C_1(\cdot), C_2(\cdot)$ as in (45), then with probability at least $1 - 2\varepsilon$, there exists an orthogonal matrix $O \in \mathbb{R}^{k \times k}$ such that

$$\left\|V_k(T) - V_k(\overline{T})O\right\| \leq \delta, \qquad and \qquad \left\|G_k(T) - G_k(\overline{T})O\right\| \leq \frac{\delta}{\sqrt{\tau^+}} + \frac{\delta}{(\tau^+)^2}.$$

Let us now interpret the scaling of the terms n, p, τ^+ and τ^- in Theorem 5, and provide some intuition.

1. In general, when no assumption is made on the noise level η, we have $\beta = \frac{4\eta}{s(1-2\eta)+4\eta}$ and the requirement on n is $n \gtrsim \max\left\{\frac{1}{s(1-2\eta)}, \frac{\eta}{1-l}\right\}$. Then a sufficient set of conditions on $\tau^+ > 0, \tau^- \geq 0$ are

$$\tau^+ \gtrsim 1 + \frac{\eta}{s(1-2\eta)}, \quad \tau^- \lesssim \frac{\eta}{s(1-2\eta)+2\eta}. \tag{8}$$

Moreover, we see from (45) that $C_1(\tau^+, \tau^-) \lesssim 1/\tau^+$, and thus $\frac{(2+\tau^+)C_1(\tau^+,\tau^-)}{1+\tau^-} \lesssim 1$. Hence, a sufficient condition on p is

$$p \gtrsim \frac{1}{\delta^4}\left(1 + \frac{\eta}{s(1-2\eta)}\right)^4 C_2(s,\eta,l)\frac{\ln n}{n}.$$

2. In the "low-noise" regime where $\eta \leq \frac{s}{2s+4}$, the condition on τ^- in (8) becomes strict, especially as $\eta \to 0$. In this regime, the second condition in Theorem 5 allows for a wider range of values for τ^-; in particular, the following set of conditions suffice

$$\tau^+ \gtrsim 1, \quad \tau^- \lesssim \frac{s(1-2\eta)}{s(1-2\eta)+2\eta}.$$

Moreover, we then obtain that the condition $p \gtrsim \frac{1}{\delta^4}C_2(s,\eta,l)\frac{\ln n}{n}$ is sufficient.

14

3. When $\tau^+ \to \infty$, then $\left\|G_k(T) - G_k(\overline{T})O\right\| \to 0$, which might lead one to believe that the clustering performance improves accordingly. This is not the case however, since when τ^+ is large, then $G_k(T) \approx \frac{1}{\sqrt{\tau^+}}V_k(T)$ and $G_k(\overline{T}) \approx \frac{1}{\sqrt{\tau^+}}V_k(\overline{T})$, which means that clustering the rows of $G_k(T)$ (resp. $G_k(\overline{T})$) is roughly equivalent to clustering the rows of $V_k(T)$ (resp. $V_k(\overline{T})$). Moreover, note that for large τ^+, we have $T \approx \frac{1}{\tau^+}(L^+_{sym}+\tau^- I)$ and $\overline{T} \approx \frac{1}{\tau^+}(\overline{L^+_{sym}}+\tau^- I)$ and thus the negative subgraph has no effect on the clustering performance.

SPONGE$_{sym}$ in the sparse regime.   Notice that the above theorem required the sparsity parameter $p = \Omega(\ln n/n)$, when n is large enough. This condition on p is essentially required to show concentration bounds on $\left\|L^\pm_{sym} - \overline{L^\pm_{sym}}\right\|$ in Lemma 27, which in turn implies a concentration bound on $\left\|T - \overline{T}\right\|$ (see Lemma 28). However, in the sparse regime p is of the order $o(\ln n)/n$, and thus Lemma 27 does not apply in this setting. In fact, it is not difficult to see that the matrices $L^\pm_{sym}$ will not concentrate[2] around $\overline{L^\pm_{sym}}$ in the sparse regime. On the other hand, by relying on a recent result in (Le et al., 2017, Theorem 4.1) on the concentration of the normalized Laplacian of regularized adjacency matrices of inhomogeneous Erdős-Rényi graphs in the sparse regime (see Theorem 31), we show concentration bounds on $\left\|L^+_{sym,\gamma^+} - \overline{L^+_{sym}}\right\|$ and $\left\|L^-_{sym,\gamma^-} - \overline{L^-_{sym}}\right\|$, which hold when $p \gtrsim 1/n$ and $\gamma^+,\gamma^- \asymp (np)^{6/7}$ (see Lemma 32). As before, these concentration bounds can then be shown to imply a concentration bound on $\left\|T_{\gamma^+,\gamma^-} - \overline{T}\right\|$ (see Lemma 33). Other than these technical differences, the remainder of the arguments follow the same structure as in the proof of Theorem 5, thus leading to the following result in the sparse regime.

Theorem 6 (Restating Theorem 34) *Assuming* $n \geq \max\left\{\frac{2(1-\eta)}{s(1-2\eta)}, \frac{2\eta}{(1-\eta)(1-l)}\right\}$, *suppose* $\tau^+ > 0, \tau^- \geq 0$ *are chosen to satisfy*

$$\tau^+ > \frac{16\eta}{\beta s(1-2\eta)}, \qquad \tau^- < \frac{\beta}{2}\left(\frac{s(1-2\eta)}{s(1-2\eta)+2\eta}\right)\min\left\{\frac{1}{4(1-\beta)}, \frac{\tau^+}{8}\right\}$$

where β,η *satisfy one of the following conditions*

1. $\beta = \frac{4\eta}{s(1-2\eta)+4\eta}$ *and* $0 < \eta < \frac{1}{2}$, *or*

2. $\beta = \frac{1}{2}$ *and* $\eta \leq \frac{s}{2s+4}$.

Then $V_k(\overline{T}) = \Theta R$ *and* $G_k(\overline{T}) = \Theta(C^-)^{-1/2}R$, *where* R *is a rotation matrix, and* $C^- \succ 0$ *is as defined in (18). Moreover, there exists a constant* $C > 0$ *such that for* $r \geq 1$ *and* $\delta \in (0,1)$, *if* p *satisfies*

$$p \geq \max\left\{1, \left(\frac{4C_1(\tau^+,\tau^-)(2+\tau^+)}{3(\tau^+)^2(1-\beta)(1+\tau^-)}\right)^{28}\right\}\frac{C_4^{14}(r,s,\eta,l)}{\delta^{28}(1-\eta)n},$$

and $\gamma^+,\gamma^- = [np(1-\eta)]^{6/7}$, *then with probability at least* $1 - 2e^{-r}$, *there exists a rotation* $O \in \mathbb{R}^{k\times k}$ *so that*

$$\left\|V_k(T_{\gamma^+,\gamma^-}) - V_k(\overline{T})O\right\| \leq \delta, \qquad and \qquad \left\|G_k(T_{\gamma^+,\gamma^-}) - G_k(\overline{T})O\right\| \leq \frac{\delta}{\sqrt{\tau^+}} + \frac{\delta}{(\tau^+)^2}.$$

---

[2]See for e.g., Le et al. (2017).

Here, $C_4(r, s, \eta, l) := 2^{5/2} C r^2 + 3\sqrt{2 C_2(s, \eta, l)}$, with $C_2(s, \eta, l)$ as defined in (45).

The following remarks are in order.

1. It is clear that γ^+, γ^- can neither be too small (since this would imply lack of concentration), nor too large (since this would destroy the latent geometries of G^+, G^-). The choice $\gamma^+, \gamma^- \asymp (np)^{6/7}$ provides a trade-off, and leads to the bounds $\left\|L^+_{sym,\gamma^+} - \overline{L^+_{sym}}\right\|$, $\left\|L^-_{sym,\gamma^-} - \overline{L^-_{sym}}\right\| = O((np)^{-1/14})$ when $p \gtrsim 1/n$ (see Lemma 32).

2. In general, for $\eta \in (0, 1/2)$, it suffices that τ^+, τ^- satisfy (8) and $n \gtrsim \max\left\{\frac{1}{s(1-2\eta)}, \frac{\eta}{1-l}\right\}$. As discussed earlier, $\frac{(2+\tau^+)C_1(\tau^+,\tau^-)}{1+\tau^-} \lesssim 1$, and hence it suffices that $p \gtrsim \frac{C_4^{14}(r,s,\eta,l)}{\delta^{28} n}$.

Mis-clustering error bounds. Thus far, our analysis has shown that under suitable conditions on n, p, τ^+ and τ^-, the matrix $G_k(T)$ (or $G_k(T_{\gamma^+,\gamma^-})$ in the sparse regime) is close to $G_k(\overline{T})O$ for some rotation O, with the rows of $G_k(\overline{T})$ preserving the ground truth clustering. This suggests that by applying the k-means clustering algorithm on the rows of $G_k(T)$ (or $G_k(T_{\gamma^+,\gamma^-})$) one should be able to approximately recover the underlying communities. However, the k-means problem for clustering points in $\mathbb{R}^d$ is known to be NP-Hard in general, even for $k = 2$ or $d = 2$ (Aloise et al., 2009; Dasgupta, 2008; Mahajan et al., 2012). On the other hand, there exist efficient $(1 + \xi)$-approximation algorithms (for $\xi > 0$), such as, for e.g., the algorithm of Kumar et al. (2004) which has a running time of $O(2^{(k/\xi)^{O(1)}} nd)$.

Using standard tools (Lei and Rinaldo, 2015, Lemma 5.1), we can bound the mis-clustering error when a $(1 + \xi)$-approximate k-means algorithm is applied on the rows of $G_k(T)$ (or $G_k(T_{\gamma^+,\gamma^-})$), provided the estimation error bound δ is small enough. In the following theorem, the sets S_i, $i = 1, \ldots, k$ contain those vertices in C_i for which we cannot guarantee correct clustering.

Theorem 7 (Re-Stating Theorem 36) *Under the notation and assumptions of Theorem 5, let $(\widetilde{\Theta}, \widetilde{X}) \in \mathbb{M}_{n \times k} \times \mathbb{R}^{k \times k}$ be a $(1 + \xi)$-approximate solution to the k-means problem $\min_{\Theta \in \mathbb{M}_{n \times k}, X \in \mathbb{R}^{k \times k}} \|\Theta X - G_k(T)\|_F^2$. Denoting*

$$S_i = \left\{ j \in C_i \; : \; \left\|(\widetilde{\Theta}\widetilde{X})_{j*} - (\Theta(C^-)^{-1/2}RO)_{j*}\right\| \geq \frac{1}{2\sqrt{n_i(\tau^+ + \frac{2}{1-l})}} \right\}$$

it holds with probability at least $1 - 2\varepsilon$ that

$$\sum_{i=1}^{k} \frac{|S_i|}{n_i} \leq \delta^2 (64 + 32\xi)k \left(\tau^+ + \frac{2}{1-l}\right) \left(\frac{(\tau^+)^3 + 1}{(\tau^+)^4}\right). \tag{9}$$

In particular, if δ satisfies

$$\delta < \frac{(\tau^+)^2}{\sqrt{(64 + 32\xi)k(\tau^+ + \frac{2}{1-l})((\tau^+)^3 + 1)}}$$

then there exists a $k \times k$ permutation matrix π such that $\widetilde{\Theta}_G = \hat{\Theta}_G \pi$, where $G = \cup_{i=1}^{k}(C_i \backslash S_i)$.

In the sparse regime, the above statement holds under the notation and assumptions of Theorem 6 with $G_k(T)$ replaced with $G_k(T_{\gamma^+, \gamma^-})$, and with probability at least $1 - 2e^{-r}$.

We remark that when $\tau^+ \to \infty$, the bound on δ becomes independent of τ^+ and is of the form $\delta \lesssim \frac{1}{\sqrt{k}}$. This is also true for the mis-clustering bound in (9), which is of the form $\sum_{i=1}^{k} \frac{|S_i|}{n_i} \lesssim \delta^2 k$.

3.2 Symmetric Signed Laplacian

We now describe our results for the symmetric Signed Laplacian. We recall that $\mathbb{E}[A] = \mathbb{E}[A^+] - \mathbb{E}[A^-]$ and $\mathbb{E}[\overline{D}]$ denote the adjacency and degree matrices of the expected graph, under the SSBM ensemble. We define

$$\mathcal{L}_{sym} = I_n - (\mathbb{E}[\overline{D}])^{-1/2} \mathbb{E}[A](\mathbb{E}[\overline{D}])^{-1/2}, \tag{10}$$

to be the normalized Signed Laplacian of the expected graph. Moreover, $\rho = \frac{s}{l} \leq 1$ denotes the *aspect ratio*, measuring the discrepancy between the smallest and largest cluster sizes in the SSBM.

We will first show that for ρ large enough, the smallest $k-1$ eigenvectors of $\mathcal{L}_{sym}$, denoted by $V_{k-1}(\mathcal{L}_{sym})$, are given by $V_{k-1}(\mathcal{L}_{sym}) = \Theta R_{k-1}$, with $R_{k-1} \in \mathbb{R}^{k \times (k-1)}$ a matrix whose columns are the $k-1$ smallest eigenvectors of a $k \times k$ matrix $\overline{C}$ defined in Lemma 37. We will then prove that the rows of $V_{k-1}(\mathcal{L}_{sym})$ impart the same clustering structure as that of Θ. The remaining arguments revolve around deriving concentration bounds on $\left\| \overline{L_{sym}} - \mathcal{L}_{sym} \right\|$, which imply, for n, p and ρ large enough, that the distance between the column spans of $V_{k-1}(\overline{L_{sym}})$ and $V_{k-1}(\mathcal{L}_{sym})$ is small, i.e. there exists a unitary matrix O such that $\left\| V_{k-1}(\overline{L_{sym}}) - V_{k-1}(\mathcal{L}_{sym})O \right\|$ is small. Altogether, this allows us to conclude that the rows of $V_{k-1}(\overline{L_{sym}})$ approximately encode the clustering structure of Θ. The above discussion is summarized in the following theorem, which is our first main result for the symmetric Signed Laplacian, in the moderately dense regime.

Theorem 8 (Eigenspace alignment in the dense case) *Assuming $\eta \in [0, 1/2)$, $k \geq 2$, $n \geq 10$, suppose the aspect ratio satisfies*

$$\sqrt{\rho} > 1 - \frac{1}{4k(2 + \sqrt{k})}, \tag{11}$$

and suppose that, for $\delta \in (0, \frac{1}{2})$, it holds true that

$$p > C(k, \eta, \delta)\frac{\ln n}{n} \qquad with \quad C(k, \eta, \delta) = \left(\frac{2Ck}{\delta(1 - 2\eta)}\right)^2 \quad and \, C < 43, \tag{12}$$

Then there exists a universal constant $c > 0$, such that with probability at least $1 - \frac{2}{n} - n\exp\left(\frac{-np}{c}\right)$, there exists an orthogonal matrix $O \in \mathbb{R}^{(k-1) \times (k-1)}$ such that

$$\|V_{k-1}(\overline{L_{sym}}) - \Theta R_{k-1}O\| \leq 2\delta,$$

where $R_{k-1} \in \mathbb{R}^{k \times (k-1)}$ is a matrix whose columns are the $(k-1)$ smallest eigenvectors of the matrix $\overline{C}$ defined in Lemma 37.

Remark 9 *(Related work) As previously explained, for the special case where $k = 2$ and with equal-size clusters, a similar result was proved in (Cucuringu et al., 2019, Theorem 3). Under a different SSBM model, the Signed Laplacian clustering algorithm was analyzed by Mercado et al. (2019) for general k. Although their generative model is more general than our SSBM, their results on the symmetric Signed Laplacian do not apply here. More precisely, one assumption of Theorem 3 of Mercado et al. (2019) translates into our model as $p(k-2)(1-2\eta) < 0$, which does not hold for $\eta < \frac{1}{2}$ and $k \geq 2$.*

Remark 10 *(Assumptions) The condition on the aspect ratio (11) is essential to apply a perturbation technique, where the reference is the setting with equal-size clusters, i.e. $n_i = \frac{n}{k}, \forall i \in [k]$ (see Lemma 39). In the sparsity condition (12), we note that the constant $C(k, \eta, \delta)$ scales quadratically with the number of classes k and as δ^{-2} with $\delta > 0$ the error on the eigenspace. However, we conjecture that this assumption is only an artefact of the proof technique, and that the result could hold for more general graphs with very unbalanced cluster sizes.*

Regularized Signed Laplacian. We now consider the sparse regime $p = o(\ln n)/n$ and show that we can recover the ground-truth clustering structure up to some small error using the regularized Signed Laplacian L_γ, provided that n, p and ρ are large enough, and that the regularization parameters γ^+, γ^- are well-chosen. We denote $\mathcal{L}_\gamma$ to be the equivalent of the regularized Laplacian for the expected graph in our SSBM, *i.e.*

$$\mathcal{L}_\gamma = I - (\mathbb{E}[\overline{D}_\gamma])^{-1/2}\mathbb{E}[A_\gamma](\mathbb{E}[\overline{D}_\gamma])^{-1/2},$$

with $\mathbb{E}[A_\gamma]$, resp. $\mathbb{E}[\overline{D}_\gamma]$, denoting the adjacency matrix, resp. the degree matrix, of the expected regularized graph. The next theorem is an intermediate result, which provides a high probability bound on $\|L_\gamma - \mathcal{L}_\gamma\|$ and $\|L_\gamma - \mathcal{L}_{sym}\|$.

Theorem 11 *(Error bound for the regularized Signed Laplacian) Assuming $\eta \in [0, 1/2)$, $k \geq 2$, and regularization parameters $\gamma^+, \gamma^- \geq 0$, $\gamma := \gamma^+ + \gamma^-$, it holds true that for any $r \geq 1$, with probability at least $1 - 7e^{-2r}$, we have*

$$\|L_\gamma - \mathcal{L}_\gamma\| \leq \frac{Cr^2}{\sqrt{\gamma}}\left(1 + \frac{\overline{d}}{\gamma}\right)^{5/2} + \frac{32\sqrt{2r}}{\sqrt{\gamma}} + \frac{8}{\sqrt{\overline{d}}}, \tag{13}$$

with $C > 1$ an absolute constant. Moreover, it also holds true that

$$\|L_\gamma - \mathcal{L}_{sym}\| \leq \frac{Cr^2}{\sqrt{\gamma}}\left(1 + \frac{\overline{d}}{\gamma}\right)^{5/2} + \frac{32\sqrt{2r}}{\sqrt{\gamma}} + \frac{8}{\sqrt{\overline{d}}} + \frac{\gamma}{\overline{d} + \gamma}. \tag{14}$$

In particular, for the choice $\gamma = \overline{d}^{7/8}$, if $p \geq 2/n$, we obtain

$$\|L_\gamma - \mathcal{L}_{sym}\| \leq \left(128Cr^2 + 1\right)(\overline{d})^{-\frac{1}{8}}.$$

Remark 12 *The above theorem shows the concentration of our regularized Laplacian L_γ towards the regularized Laplacian (13) and the Signed Laplacian (14) of the expected graph. More precisely, if for some well-chosen parameters $\gamma^+, \gamma^- \geq 0$, these upper bounds are small, e.g $\|L_\gamma - \mathcal{L}_{sym}\| << 1$, then we have $\|L_\gamma - \mathcal{L}_{sym}\| << \|\mathcal{L}_{sym}\|$ since $\|\mathcal{L}_{sym}\| = 2$ (see Appendix E).*

Using this concentration bound, we can show that the eigenspaces $V_{k-1}(L_\gamma)$ and $V_{k-1}(\mathcal{L}_{sym})$ are "close", provided that $p = \Omega(1/n)$, ρ is close enough to 1, and γ is well-chosen. This is stated in the next theorem.

Theorem 13 (Eigenspace alignment in the sparse case) *Assuming $\eta \in [0, 1/2)$, $k \geq 2$, and $n \geq 10$, suppose that (11) holds true, and for $\delta \in (0, \frac{1}{2})$ and $r \geq 1$, the sparsity p satisfies*

$$p > \left(\frac{2kC_4}{\delta(1 - 2\eta)} \right)^8 \frac{2}{n} \qquad with \qquad C_4 = 128Cr^2 + 1 \qquad (15)$$

and $C > 1$ the constant defined in (13). If the regularization parameters $\gamma^+, \gamma^- \geq 0$ are chosen so that $\gamma = \overline{d}^{7/8}$, then with probability at least $1 - 7e^{-2r} - \frac{2}{n} - ne^{-np/c}$, there exists an orthogonal matrix $O \in \mathbb{R}^{(k-1)\times(k-1)}$ so that

$$\|V_{k-1}(L_\gamma) - \Theta R_{k-1}O\| \leq 2\delta.$$

Remark 14 *In the sparse setting, the constant before the factor $\frac{1}{n}$ in the sparsity condition (15) scales as $\left(\frac{k}{\delta}\right)^8$. However for k fixed, it would hold if $p = \omega(1/n)$ as $n \to \infty$.*

Remark 15 *In practice, one can choose the regularization parameters by first estimating the sparsity parameter p, e.g. from the fraction of connected pairs of nodes*

$$p = \frac{2}{n(n-1)} \sum_{i<j} |A_{ij}|,$$

then choosing $\gamma \geq 0$ so that $\gamma = (\hat{p}(n-1))^{7/8}$. However, from this analysis, it is not clear how one would suitably choose γ^+ and γ^-.

Mis-clustering error bounds. Since $V_{k-1}(\overline{L_{sym}})$ and $V_{k-1}(L_\gamma)$ are "close" to $V_{k-1}(\mathcal{L}_{sym})$, we recover the ground-truth clustering structure up to some error, which we quantify in the following theorem, where we bound the mis-clustering rate when using a $(1+\xi)$-approximate k-means error on the rows of $V_{k-1}(\overline{L_{sym}})$ (resp. $V_{k-1}(L_\gamma)$).

Theorem 16 *(Number of mis-clustered nodes) Let $\xi > 0$ and $\delta \in \left(0, \sqrt{\frac{1}{12(16+8\xi)(k-1)}}\right)$, and suppose that ρ and p satisfy the assumptions of Theorem 8 (resp. Theorem 13 and $r \geq 1$). Let $(\widetilde{\Theta}, \widetilde{R}_{k-1})$ be the $(1+\xi)$-approximation of the k-means problem*

$$\min_{\Theta \in \mathbb{M}_{n,k}, R \in \mathbb{R}^{k\times(k-1)}} \left\|\Theta R - V_{k-1}(\overline{L_{sym}})\right\|_F \qquad (resp. \quad \min_{\Theta \in \mathbb{M}_{n,k}, R \in \mathbb{R}^{k\times(k-1)}} \left\|\Theta R - V_{k-1}(L_\gamma)\right\|_F \ ).$$

Let $S_i = \left\{ j \in C_i; \left\|(\widetilde{\Theta}\widetilde{R}_{k-1})_{j*} - (\Theta R_{k-1}O)_{j*}\right\|^2 \geq \frac{2}{3n_i} \right\}$ and $\widetilde{V} = \cup_{i=1}^k C_i \backslash S_i$. Then with probability at least $1 - \frac{2}{n} - n\exp(\frac{-np}{c})$ (resp. $1 - 7e^{-2r} - \frac{2}{n} - ne^{-np/c}$), there exists a permutation $\pi \in \mathbb{R}^{k\times k}$ such that $\widetilde{\Theta}_{\widetilde{V}*} = \hat{\Theta}_{\widetilde{V}*}\pi$ and

$$\sum_{i=1}^k \frac{|S_i|}{n_i} \leq 96(2+\xi)(k-1)\delta^2.$$

In particular, the set of mis-clustered nodes is a subset of $\cup_{i=1}^k S_i$.

19

4. Analysis of SPONGE Symmetric

This section contains the proof of our main results for SPONGE_{sym}, divided over the following subsections. Section 4.1 describes the eigen-decomposition of the matrix $\overline{T}$, thus revealing that a subset of its eigenvectors contain relevant information about Θ. Section 4.2 provides conditions on τ^+, τ^- which ensure that $V_k(\Theta) = \Theta R$ (for some rotation matrix R), along with a lower bound on the eigengap $\lambda_{n-k+1}(\overline{T}) - \lambda_{n-k}(\overline{T})$. Section 4.3 then derives concentration bounds on $\|T - \overline{T}\|$ using standard tools from the random matrix literature. These results are combined in Section 4.4 to derive error bounds for estimating $V_k(\overline{T})$ and $G_k(\overline{T})$ up to a rotation (using the Davis-Kahan theorem). The results summarized thus far pertain to the "dense" regime, where we require $p = \Omega(\ln n/n)$ when n is large. Section 4.5 extends these results to the sparse regime where $p = o(\ln n)/n$, for the regularized version of SPONGE_{sym}. Finally, we conclude in Section 4.6 by translating our results from Sections 4.4 and 4.5 to obtain mis-clustering error bounds for a $(1 + \xi)$-approximate k-means algorithm, by leveraging previous tools from the literature (Lei and Rinaldo, 2015).

4.1 Eigen-decomposition of $\overline{T}$

The following lemma shows that a subset of the eigenvectors of $\overline{T}$ indeed contain information about Θ, i.e., the ground-truth clustering.

Lemma 17 (Spectrum of $\overline{T}$) *Let*

$$d_i^+ = p\left(n(s_i(1 - 2\eta) + \eta) - (1 - \eta)\right),$$

$$d_i^- = p\left(n(-s_i(1 - 2\eta) + (1 - \eta)) - \eta\right),$$

denote the expected degree of a node in cluster C_i, $i \in [k]$. Let $u^+ = \left(\sqrt{\frac{n_1}{d_1^+}}, \ldots, \sqrt{\frac{n_k}{d_k^+}}\right)^\top$, $u^- = \left(\sqrt{\frac{n_1}{d_1^-}}, \ldots, \sqrt{\frac{n_k}{d_k^-}}\right)^\top$, $\alpha_i^+ = 1 + \tau^- + p(1 - \eta)/d_i^+$, and $\alpha_i^- = 1 + \tau^+ + p\eta/d_i^-$, for $i \in [k]$, for some $\tau^+ > 0, \tau^- \geq 0$. Let the columns of $V^\perp$ contain eigenvectors of $\mathbb{E}[D^+]$ which are orthogonal to the column span of Θ. It holds true that

$$\overline{T} = \begin{bmatrix} \Theta R & V^\perp \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda} & & & \\ & \frac{\alpha_1^+}{\alpha_1^-} I_{n_1 - 1} & & \\ & & \ddots & \\ & & & \frac{\alpha_k^+}{\alpha_k^-} I_{n_k - 1} \end{bmatrix} \begin{bmatrix} (\Theta R)^\top \\ V^{\perp \top} \end{bmatrix}, \tag{16}$$

where R is a $k \times k$ rotation matrix, and Λ is a diagonal matrix, such that $(C^-)^{-1/2} C^+ (C^-)^{-1/2} = R\Lambda R^T$, where

$$C^+ = -p\eta u^+(u^+)^\top + diag\left(1 + \tau^- + \frac{p}{d_i^+}(1 - \eta - n_i(1 - 2\eta))\right), \tag{17}$$

$$C^- = -p(1 - \eta)u^-(u^-)^\top + diag\left(1 + \tau^+ + \frac{p}{d_i^-}(\eta + n_i(1 - 2\eta))\right). \tag{18}$$

Proof We first consider the spectrum of D^+, D^-, A^+, A^-, followed by that of $(\overline{L^+_{sym}} + \tau^- I)$ and $(\overline{L^-_{sym}} + \tau^+ I)$, which altogether will reveal the spectral decomposition of $\overline{T}$.

• **Analysis in expectation of the spectra of D^+, D^-, A^+, A^-.** Without loss of generality, we may assume that cluster C_1 contains the first n_1 vertices, cluster C_2 the next n_2 vertices and similarly for the remaining clusters. Note that $\mathbb{E}[D^\pm] = \text{diag}\left(d_1^\pm I_{n_1}, \ldots, d_k^\pm I_{n_k}\right)$, where for $i \in [k]$, straightforward calculations reveal that $d_i^+ = p\left(n(s_i(1 - 2\eta) + \eta) - (1 - \eta)\right)$, and $d_i^- = p\left(n(-s_i(1 - 2\eta) + (1 - \eta)) - \eta\right)$. One can rewrite the matrices $(\mathbb{E}[D^\pm])^{-1}$ in the more convenient form

$$(\mathbb{E}[D^\pm])^{-1} = [\Theta \ V^\perp] \, \text{diag}\left(\frac{1}{d_1^\pm}, \ldots, \frac{1}{d_k^\pm}, \frac{1}{d_1^\pm} I_{n_1-1}, \ldots, \frac{1}{d_k^\pm} I_{n_k-1}\right) \, [\Theta \ V^\perp]^\top \quad (19)$$

since the column vectors of Θ are eigenvectors of $(\mathbb{E}[D^\pm])^{-1}$, and the eigenvalues of $(\mathbb{E}[D^\pm])^{-1}$ are apparent because $\mathbb{E}[D^\pm]$ is a diagonal matrix. Note that (19) is true in general, and does not make any assumption on the placement of the vertices into their respective C_i cluster. Furthermore, one can verify that $\mathbb{E}[A^+]$ admits the eigen-decomposition

$$\mathbb{E}[A^+] = \Theta_{n\times k} \begin{bmatrix} n_1 p(1-\eta) & \sqrt{n_1 n_2} p\eta & \cdots & \sqrt{n_1 n_k} p\eta \\ \sqrt{n_2 n_1} p\eta & n_2 p(1-\eta) & \cdots & \sqrt{n_2 n_k} p\eta \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{n_k n_1} p\eta & \sqrt{n_k n_2} p\eta & \cdots & n_k p(1-\eta) \end{bmatrix}_{k\times k} \Theta^\top_{k\times n} - p(1-\eta) I_{n\times n} \quad (20)$$

and similarly, $\mathbb{E}[A^-]$ can be decomposed as

$$\mathbb{E}[A^-] = \Theta_{n\times k} \begin{bmatrix} n_1 p\eta & \sqrt{n_1 n_2} p(1-\eta) & \cdots & \sqrt{n_1 n_k} p(1-\eta) \\ \sqrt{n_2 n_1} p(1-\eta) & n_2 p\eta & \cdots & \sqrt{n_2 n_k} p(1-\eta) \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{n_k n_1} p(1-\eta) & \sqrt{n_k n_2} p(1-\eta) & \cdots & n_k p\eta \end{bmatrix}_{k\times k} \Theta^\top_{k\times n} - p\eta I_{n\times n} .$$

• **Analysis of the spectra of $(\overline{L^+_{sym}} + \tau^- I)$ and $(\overline{L^-_{sym}} + \tau^+ I)$.** We start by observing that

$$\overline{L^\pm_{sym}} + \tau^\mp I = I - (\mathbb{E}[D^\pm])^{-1/2}(\mathbb{E}[A^\pm])(\mathbb{E}[D^\pm])^{-1/2} + \tau^\mp I$$
$$= (1 + \tau^\mp)I - (\mathbb{E}[D^\pm])^{-1/2}(\mathbb{E}[A^\pm])(\mathbb{E}[D^\pm])^{-1/2} . \quad (21)$$

In light of (20), one can write $(\mathbb{E}[D^+])^{-1/2}(\mathbb{E}[A^+])(\mathbb{E}[D^+])^{-1/2}$ as

$$(\mathbb{E}[D^+])^{-1/2}(\mathbb{E}[A^+])(\mathbb{E}[D^+])^{-1/2} = -p(1-\eta)(\mathbb{E}[D^+])^{-1}$$

$$+ [\Theta \ \ V^\perp] \begin{bmatrix} \overbrace{\begin{bmatrix} \frac{n_1}{d_1^+}p(1-\eta) & \sqrt{\frac{n_1 n_2}{d_1^+ d_2^+}}p\eta & \cdots & \sqrt{\frac{n_1 n_k}{d_1^+ d_k^+}}p\eta \\ \sqrt{\frac{n_2 n_1}{d_2^+ d_1^+}}p\eta & \frac{n_2}{d_2^+}p(1-\eta) & \cdots & \sqrt{\frac{n_2 n_k}{d_2^+ d_k^+}}p\eta \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\frac{n_k n_1}{d_k^+ d_1^+}}p\eta & \sqrt{\frac{n_k n_2}{d_k^+ d_2^+}}p\eta & \cdots & \frac{n_k}{d_k^+}p(1-\eta) \end{bmatrix}_{k\times k}}^{\stackrel{\text{def}}{=} B^+} & \mathbf{0}_{k\times(n-k)} \\ \mathbf{0}_{(n-k)\times k} & \mathbf{0}_{(n-k)\times(n-k)} \end{bmatrix} \begin{bmatrix} \Theta^\top \\ V^{\perp \top} \end{bmatrix} .$$

$$(22)$$

Similarly, using the expression for $\mathbb{E}[A^-]$, the expression for $(\mathbb{E}[D^-])^{-1/2}(\mathbb{E}[A^-])(\mathbb{E}[D^-])^{-1/2}$ can be written as

$$(\mathbb{E}[D^-])^{-1/2}(\mathbb{E}[A^-])(\mathbb{E}[D^-])^{-1/2} = -p\eta(\mathbb{E}[D^-])^{-1}+$$

$$\begin{bmatrix} \Theta & V^\perp \end{bmatrix} \begin{bmatrix} \overbrace{\begin{bmatrix} \frac{n_1}{d_1^-}p\eta & \sqrt{\frac{n_1 n_2}{d_1^- d_2^-}}p(1-\eta) & \cdots & \sqrt{\frac{n_1 n_k}{d_1^- d_k^-}}p(1-\eta) \\ \sqrt{\frac{n_2 n_1}{d_2^- d_1^-}}p(1-\eta) & \frac{n_2}{d_2^-}p\eta & \cdots & \sqrt{\frac{n_2 n_k}{d_2^- d_k^-}}p(1-\eta) \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\frac{n_k n_1}{d_k^- d_1^-}}p(1-\eta) & \sqrt{\frac{n_k n_2}{d_k^- d_2^-}}p(1-\eta) & \cdots & \frac{n_k}{d_k^-}p\eta \end{bmatrix}_{k\times k}}^{\stackrel{\text{def}}{=}B^-} & \mathbf{0}_{k\times(n-k)} \\ \mathbf{0}_{(n-k)\times k} & \mathbf{0}_{(n-k)\times(n-k)} \end{bmatrix} \begin{bmatrix} \Theta^\top \\ V^{\perp\top} \end{bmatrix}.$$

$$(23)$$

Combining (19), (22), and (23) into (21), we readily arrive at

$$(\overline{L_{sym}^{\pm}}+\tau^{\mp}I) = \begin{bmatrix} \Theta & V^\perp \end{bmatrix} \begin{bmatrix} \underbrace{[\mathrm{diag}(\alpha_i^{\pm}) - B^{\pm}]_{k\times k}}_{\stackrel{\text{def}}{=}C^{\pm}} & \mathbf{0}_{k\times(n-k)} \\ & \alpha_1^{\pm}I_{n_1-1} \\ & & \alpha_2^{\pm}I_{n_k-1} \\ & & & \ddots \\ & & & & \alpha_k^{\pm}I_{n_k-1,} \end{bmatrix} \begin{bmatrix} \Theta^\top \\ V^{\perp\top} \end{bmatrix}$$

$$(24)$$

where $\alpha_i^{\pm}$ and C^+, C^- are defined as in the statement of the lemma. The spectral decomposition of $\overline{T}$ now follows trivially using (24), along with the spectral decomposition $(C^-)^{-1/2}C^+(C^-)^{-1/2} = R\Lambda R^T$. ∎

Lemma 17 reveals that we need to extract the k-informative eigenvectors ΘR from the n-eigenvectors $\begin{bmatrix} \Theta R & V^\perp \end{bmatrix}$ of $\overline{T}$. Clearly, it suffices to recover any orthonormal basis for the column span of Θ, since the rows of any such corresponding matrix (one instance of which is ΘR) will exhibit the same clustering structure as Θ.

4.2 Ensuring $V_k(\overline{T}) = \Theta R$ and bounding the spectral gap

In this section, our aim is to show that, for suitable values of $\tau^+ > 0, \tau^- \geq 0$, the eigenvectors corresponding to the smallest k eigenvalues of $\overline{T}$ are given by ΘR, i.e., $V_k(\overline{T}) = \Theta R$. This is equivalent to ensuring (recall Lemma 17) that

$$\lambda_{n-k+1}(\overline{T}) = \left\| (C^-)^{-1/2}C^+(C^-)^{-1/2} \right\| < \min_{i\in[k]} \frac{\alpha_i^+}{\alpha_i^-} = \lambda_{n-k}(\overline{T}). \qquad (25)$$

Moreover, we will need to find a strictly positive lower-bound on the spectral gap $\lambda_{n-k}(\overline{T}) - \lambda_{n-k+1}(\overline{T})$, as it will be used later on, in order to show that the column span of $V_k(T)$ is close to that of $V_k(\overline{T})$. We first consider the equal-sized clusters case, and then proceed to the general-sized clusters case.

4.2.1 SPECTRAL GAP FOR EQUAL-SIZED CLUSTERS

When the cluster sizes are equal, the analysis is considerably cleaner than the general setting. Let us first establish notation specific to the equal-sized clusters case.

Remark 18 (Notation for the equal-sized clusters) *For clusters of equal size, we have that $n_1 = ... = n_k = n/k$, $d^+ := d_1^+ = ... = d_k^+$, $d^- := d_1^- = ... = d_k^-$, $\alpha^+ := \alpha_1^+ = ... = \alpha_k^+$, and $\alpha^- := \alpha_1^- = ... = \alpha_k^-$. Let C_e^+, C_e^-, and $\overline{T}_e$ denote the respective counterparts of C^+, C^-, and $\overline{T}$, for the equal-sized case. In light of (17) and (18), one can verify that C_e^+ and C_e^- are simultaneously diagonalizable, which we show in Lemma 60.*

In the following lemma, we show the exact value of $\|\Lambda\| = \left\| (C_e^-)^{-1/2} C_e^+ (C_e^-)^{-1/2} \right\|$.

Lemma 19 (Bounding the spectral norm of $(C_e^-)^{-1/2} C_e^+ (C_e^-)^{-1/2}$) *For equal-sized clusters, the following holds true*

$$\left\| (C_e^-)^{-1/2} C_e^+ (C_e^-)^{-1/2} \right\| = \max \left\{ \frac{\tau^-}{\tau^+}, \frac{\tau^- + \frac{pn\eta}{d^+}}{\tau^+ + \frac{pn(1-\eta)}{d^-}} \right\} .$$

Proof The lemma follows directly from Lemma 60. ∎

Next, we derive conditions on $\tau^+ > 0, \tau^- \geq 0$ which ensure $V_k(\overline{T}) = \Theta R$.

Lemma 20 (Conditions on τ^- and τ^+) *Suppose $n \geq \frac{2k(1-\eta)}{1-2\eta}$, and $\tau^- \geq 0$, $\tau^+ > 0$. If τ^-, τ^+ satisfy*

1.

$$\tau^- \left(1 + \frac{p\eta}{d^-} \right) < \tau^+ \left(1 + \frac{p(1-\eta)}{d^+} \right) ,$$

2.

$$\tau^- \left[\frac{(1-2\eta)/k}{(1-\eta) - \frac{1-2\eta}{k}} \right] + \tau^+ \left[\frac{(1-2\eta)/k}{\eta + \frac{1-2\eta}{k}} \right] + 1 > \frac{2\eta}{\eta + \frac{1-2\eta}{2k}} .$$

Then it holds true that $V_k(\overline{T}) = \Theta R$, i.e., $\lambda_{n-k+1}(\overline{T}) = \left\| (C_e^-)^{-1/2} C_e^+ (C_e^-)^{-1/2} \right\| < \frac{\alpha^+}{\alpha^-} = \lambda_{n-k}(\overline{T})$.

Proof Recalling the expression for $\left\| (C_e^-)^{-1/2} C_e^+ (C_e^-)^{-1/2} \right\|$ from Lemma 19, we will ensure that each term inside the max is less than α^+/α^-. To derive the first condition of the lemma, we simply ensure that

$$\frac{\tau^-}{\tau^+} < \frac{1 + \tau^- + p(1-\eta)/d^+}{1 + \tau^+ + p\eta/d^-} \Leftrightarrow \tau^- \left(1 + \frac{p\eta}{d^-} \right) < \tau^+ \left(1 + \frac{p(1-\eta)}{d^+} \right) .$$

Before deriving the second condition, let us note additional useful bounds on $\frac{np}{d^-}, \frac{np}{d^+}$ which will be needed later.

1. $d^-/np = 1 - \eta - (1-2\eta)/k - \eta/n \leq 1 - \eta$.

2. Since $n \geq k \geq 2$, we obtain that $d^-/np \geq (1 - \eta) - (1 - 3\eta)/k \geq \frac{1-\eta}{2}$. This also implies that $p\eta/d^- \leq 1$.

Therefore, combining the above two bounds, we arrive at

$$\frac{1}{1 - \eta} \leq \frac{np}{d^-} \leq \frac{2}{1 - \eta} \ .$$

3. $d^+/np = (1 - 2\eta)/k + \eta - (1 - \eta)/n \leq \eta + (1 - 2\eta)/k$.

4. Since $n \geq \frac{2k(1-\eta)}{1-2\eta}$, it holds that $d^+/np = (1-2\eta)/k + \eta - (1-\eta)/n \geq \eta + (1 - 2\eta)/2k$.

5. Therefore, combining the above two conditions yields

$$\frac{1}{\eta + \frac{1-2\eta}{k}} \leq \frac{np}{d^+} \leq \frac{1}{\eta + \frac{1-2\eta}{2k}} \ .$$

To derive the second condition, we need to ensure $\frac{\tau^- + \frac{pn\eta}{d^+}}{\tau^+ + \frac{pn(1-\eta)}{d^-}} < \frac{1 + \tau^- + p(1-\eta)/d^+}{1 + \tau^+ + p\eta/d^-}$, which is equivalent to

$$\tau^- \left[1 - \frac{np}{d^-} \left((1 - \eta) - \frac{\eta}{n} \right) \right] < \tau^+ \left[1 - \frac{np}{d^+} \left(\eta - \frac{1 - \eta}{n} \right) \right]$$

$$+ \underbrace{\left[\frac{np(1 - \eta)}{d^-} \left(1 + \frac{p(1 - \eta)}{d^+} \right) - \frac{np\eta}{d^+} \left(1 + \frac{p\eta}{d^-} \right) \right]}_{\text{term 2}} \ .$$

Now, we can lower bound "term 2" in the above equation as

$$\frac{np(1 - \eta)}{d^-} \left(1 + \frac{p(1 - \eta)}{d^+} \right) - \frac{np\eta}{d^+} \left(1 + \frac{p\eta}{d^-} \right) \geq 1 - \frac{2\eta}{\eta + \frac{(1-2\eta)}{k}} \ .$$

Hence from the above two equations, we observe that it suffices that τ^+, τ^- satisfy

$$\tau^- \left[\frac{(1 - 2\eta)/k}{(1 - \eta) - \frac{1-2\eta}{k}} \right] + \tau^+ \left[\frac{(1 - 2\eta)/k}{\eta + \frac{1-2\eta}{k}} \right] + 1 > \frac{2\eta}{\eta + \frac{1-2\eta}{2k}} \ .$$

∎

Next, we derive sufficient conditions on τ^+, τ^- which ensure a lower bound on the *spectral gap*

$$\lambda_{n-k}(\overline{T}) - \lambda_{n-k+1}(\overline{T}) = \frac{\alpha^+}{\alpha^-} - \left\| (C_e^-)^{-1/2} C_e^+ (C_e^-)^{-1/2} \right\| \ .$$

Lemma 21 (Conditions on τ^+, τ^-, and lower-bound on spectral gap) *Suppose $n \geq \frac{2k(1-\eta)}{1-2\eta}$, then the following holds.*

24

1. If $\tau^+ > 0, \tau^- \geq 0$ satisfy

$$\tau^+ > \frac{32\eta k}{3(1-2\eta)}, \quad \tau^- < \min\left\{\frac{3}{2}, \frac{3}{16}\tau^+, \frac{3(1-\eta)}{8(\eta + \frac{1-2\eta}{k})}\right\},$$

then $V_k(\overline{T}) = \Theta R$, and $\left\|(C_e^-)^{-1/2}C_e^+(C_e^-)^{-1/2}\right\| < \left(1 - \frac{(1-2\eta)}{2k(1-\eta)}\right)\frac{\alpha^+}{\alpha^-}$, i.e., $\lambda_{n-k}(\overline{T}) - \lambda_{n-k+1}(\overline{T}) > \left(\frac{(1-2\eta)}{2k(1-\eta)}\right)\frac{\alpha^+}{\alpha^-}$.

2. If $\eta < \frac{1}{3k+2}$ and $\tau^+ > 0, \tau^- \geq 0$ satisfy

$$\tau^- < \min\left\{\left(\frac{\frac{1-2\eta}{k} - \eta}{\frac{1-2\eta}{k} + \eta}\right), \frac{1}{2}, \frac{\tau^+}{8}\right\},$$

then $V_k(\overline{T}) = \Theta R$, and $\left\|(C_e^-)^{-1/2}C_e^+(C_e^-)^{-1/2}\right\| < \frac{\alpha^+}{2\alpha^-}$, i.e., $\lambda_{n-k}(\overline{T}) - \lambda_{n-k+1}(\overline{T}) > \frac{\alpha^+}{2\alpha^-}$.

Proof We need to ensure the following two conditions for a suitably chosen $\beta \in (0, 1]$.

$$\frac{\tau^- + \frac{pn\eta}{d^+}}{\tau^+ + \frac{pn(1-\eta)}{d^-}} < \beta\left(\frac{1 + \tau^- + p(1-\eta)/d^+}{1 + \tau^+ + p\eta/d^-}\right), \tag{26}$$

$$\frac{\tau^-}{\tau^+} < \beta\left(\frac{1 + \tau^- + p(1-\eta)/d^+}{1 + \tau^+ + p\eta/d^-}\right). \tag{27}$$

1. Ensuring (26) We can rewrite (26) as

$$\tau^-\left(1 + \frac{p\eta}{d^-} - \beta\frac{pn(1-\eta)}{d^-}\right) + \tau^+\left(\frac{pn\eta}{d^+} - \beta\left(1 + \frac{p(1-\eta)}{d^+}\right)\right) + \tau^+\tau^-(1-\beta)$$
$$< \beta\frac{pn(1-\eta)}{d^-}\left(1 + \frac{p(1-\eta)}{d^+}\right) - \frac{pn\eta}{d^+}\left(1 + \frac{p\eta}{d^-}\right). \tag{28}$$

Using the expressions for d^+, d^-, we can write the coefficients of the terms τ^+, τ^- as follows.

$$1 + \frac{p\eta}{d^-} - \beta\frac{pn(1-\eta)}{d^-} = \frac{-(\frac{1-2\eta}{k}) + (1-\eta)(1-\beta)}{-(\frac{1-2\eta}{k}) + (1-\eta) - \frac{\eta}{n}},$$

$$\frac{pn\eta}{d^+} - \beta\left(1 + \frac{p(1-\eta)}{d^+}\right) = \frac{np}{d^+}(\eta - \beta\frac{1-\eta}{n}) - \beta = \frac{\eta(1-\beta) - \beta(\frac{1-2\eta}{k})}{\frac{1-2\eta}{k} + \eta - \frac{1-\eta}{n}}.$$

Moreover, using the bounds on $\frac{d^-}{np}, \frac{d^+}{np}$ derived in Lemma 20, we can lower bound the RHS term in (28) as

$$\beta\frac{pn(1-\eta)}{d^-}\left(1 + \frac{p(1-\eta)}{d^+}\right) - \frac{pn\eta}{d^+}\left(1 + \frac{p\eta}{d^-}\right) > \beta - \frac{2\eta}{\eta + \frac{1-2\eta}{k}}.$$

From the above considerations, we see that (28) is ensured provided

$$\tau^-\left[\frac{(\frac{1-2\eta}{k}) - (1-\eta)(1-\beta)}{-(\frac{1-2\eta}{k}) + (1-\eta) - \frac{\eta}{n}}\right] + \tau^+\left[\frac{-\eta(1-\beta) + \beta(\frac{1-2\eta}{k})}{\frac{1-2\eta}{k} + \eta - \frac{1-\eta}{n}}\right] + \beta > \frac{2\eta}{\eta + \frac{1-2\eta}{k}} + \tau^+\tau^-(1-\beta). \tag{29}$$

We outline two possible ways in which (29) is ensured.

- Note that the denominators of the coefficients of τ^+, τ^- in (29) are positive, while the numerators are non-negative provided $1 - \beta \leq \frac{(1-2\eta)}{2k(1-\eta)}$. Therefore, choosing

$$\beta = 1 - \frac{(1-2\eta)}{2k(1-\eta)} \quad \left(\geq \frac{3}{4} \right),$$

note that (29) is ensured provided

$$\tau^- \left[\frac{(1-2\eta)}{2k(1-\eta)} \right] + \tau^+ \left[\frac{3(1-2\eta)}{8k\left(\eta + \frac{1-2\eta}{k}\right)} \right] + \frac{3}{4} > \frac{2\eta}{\eta + \frac{1-2\eta}{k}} + \tau^+ \tau^- \left[\frac{(1-2\eta)}{2k(1-\eta)} \right]. \quad (30)$$

Finally, we observe that in order for (30) to hold, it suffices that

$$\tau^+ \tau^- \left[\frac{(1-2\eta)}{2k(1-\eta)} \right] < \frac{\tau^+}{2} \left[\frac{3(1-2\eta)}{8k\left(\eta + \frac{1-2\eta}{k}\right)} \right] \iff \tau^- < \frac{3(1-\eta)}{8\left(\eta + \frac{1-2\eta}{k}\right)}, \text{ and}$$

$$\frac{2\eta}{\eta + \frac{1-2\eta}{k}} < \frac{\tau^+}{2} \left[\frac{3(1-2\eta)}{8k\left(\eta + \frac{1-2\eta}{k}\right)} \right] \iff \tau^+ > \frac{32\eta k}{3(1-2\eta)}.$$

- Alternatively, by setting $\beta = 1/2$, (29) can be rewritten as

$$\tau^+ \left[\frac{-\frac{\eta}{2} + \frac{1-2\eta}{2k}}{\frac{1-2\eta}{k} + \eta - \frac{1-\eta}{n}} \right] + \frac{1}{2} > \frac{2\eta}{\eta + \frac{1-2\eta}{k}} + \tau^- \left[\frac{-(\frac{1-2\eta}{k}) + \frac{1-\eta}{2}}{-(\frac{1-2\eta}{k}) + (1-\eta) - \frac{\eta}{n}} \right] + \frac{\tau^+ \tau^-}{2}. \quad (31)$$

Clearly, it holds true that

$$\frac{1}{2} > \frac{2\eta}{\eta + \frac{1-2\eta}{k}} \iff \eta < \frac{1}{3k+2},$$

which also ensures that the numerator of the coefficient of τ^+ is positive. Therefore, if $\eta < \frac{1}{3k+2}$, then in order for (31) to hold, it suffices that

$$\tau^- < \left[\frac{-\eta + \frac{1-2\eta}{k}}{\frac{1-2\eta}{k} + \eta} \right] \implies \tau^+ \left[\frac{-\frac{\eta}{2} + \frac{1-2\eta}{2k}}{\frac{1-2\eta}{k} + \eta - \frac{1-\eta}{n}} \right] > \frac{\tau^+ \tau^-}{2}.$$

2. Ensuring (27) Note that one can rewrite (27) as

$$\tau^- \tau^+ (1-\beta) + \tau^- \left(1 + \frac{p\eta}{d^-} \right) < \beta \tau^+ \left(1 + \frac{p(1-\eta)}{d^+} \right). \quad (32)$$

Since $\frac{p\eta}{d^-} \leq 1$, (32) is ensured provided

$$\tau^- \tau^+ (1-\beta) + 2\tau^- < \beta \tau^+$$

which in turn holds if each LHS term is respectively less than half of the RHS term. This leads to the condition

$$\tau^- < \min\left\{\frac{\beta}{2(1-\beta)}, \frac{\beta}{4}\tau^+\right\}.$$

Finally, plugging the choices $\beta = 1 - \frac{(1-2\eta)}{2k(1-\eta)} (\geq 3/4)$ and $\beta = \frac{1}{2}$ in the above equation, and combining it with the conditions derived for ensuring (26), we readily arrive (after minor simplifications) at the statements in the Lemma. ∎

4.2.2 Spectral gap for the general case

For the general-sized clusters case, it is difficult to find the exact value of $\left\|(C^-)^{-1/2}C^+(C^-)^{-1/2}\right\|$. Therefore, in the following lemma, we show an upper bound on this quantity by bounding the spectral norms of C^+ and $(C^-)^{-1}$.

Lemma 22 (Bounding the spectral norm of $(C^-)^{-1}$ and C^+) *Recall $s := \min_{i \in [k]} n_i/n$. Then it holds true that*

$$\lambda_{\max}(C^+) \leq \tau^- + \frac{n\eta}{n(s(1-2\eta)+\eta)-(1-\eta)}, \tag{33}$$

$$\lambda_{\min}(C^-) \geq \tau^+. \tag{34}$$

From the above two inequalities, it follows that

$$\left\|(C^-)^{-1/2}C^+(C^-)^{-1/2}\right\| \leq \frac{\lambda_{max}(C^+)}{\lambda_{min}(C^-)} \leq \frac{\tau^- + \frac{n\eta}{n(s(1-2\eta)+\eta)-(1-\eta)}}{\tau^+}.$$

The proof of the above lemma is deferred to Appendix D.

Remark 23 *It is difficult to obtain more precise bounds on $\lambda_{\max}(C^+)$ and $\lambda_{\min}(C^-)$, given the expressions for C^+ in (17), and C^- in (18). Clearly, a tighter bound on $\left\|(C^-)^{-1/2}C^+(C^-)^{-1/2}\right\|$ would yield a tighter analysis in the general case.*

Recall $l := \max_{i \in [k]} n_i/n$; with a slight abuse of notation, let $d_l^{\pm}$ denote the degree of the largest cluster (of size nl). As before, we now derive conditions on $\tau^+ > 0, \tau^- \geq 0$ which ensure $V_k(\overline{T}) = \Theta R$, or equivalently,

$$\lambda_{n-k+1}(\overline{T}) = \left\|(C^-)^{-1/2}C^+(C^-)^{-1/2}\right\| < \min_{i \in [k]} \frac{\alpha_i^+}{\alpha_i^-} = \frac{1 + \tau^- + p(1-\eta)/d_l^+}{1 + \tau^+ + p\eta/d_l^-} = \frac{\alpha_l^+}{\alpha_l^-} = \lambda_{n-k}(\overline{T}). \tag{35}$$

Additionally, we find sufficient conditions on $\tau^+ > 0, \tau^- \geq 0$ which ensure a lower bound on the *spectral gap* $\lambda_{n-k}(\overline{T}) - \lambda_{n-k+1}(\overline{T}) = \min_{i \in [k]} \frac{\alpha_i^+}{\alpha_i^-} - \left\|(C^-)^{-1/2}C^+(C^-)^{-1/2}\right\|$. These are shown in the following lemma.

Lemma 24 (Conditions on τ^+, τ^-, and Lower-Bound on Spectral Gap) *Suppose $n \geq \max\left\{\frac{2(1-\eta)}{s(1-2\eta)}, \frac{2\eta}{(1-l)(1-\eta)}\right\}$, then the following is true.*

1. If $\tau^+ > 0, \tau^- \geq 0$ *satisfy*

$$2\tau^- + \frac{4\eta}{s(1-2\eta)+2\eta} < \frac{s(1-2\eta)}{s(1-2\eta)+2\eta}\tau^+ \tag{36}$$

then $V_k(\overline{T}) = \Theta R$, *i.e.,* $\lambda_{n-k+1}(\overline{T}) = \left\|(C^-)^{-1/2}C^+(C^-)^{-1/2}\right\| < \frac{\alpha_l^+}{\alpha_l^-} = \lambda_{n-k}(\overline{T})$.

2. *For* $\beta = \frac{4\eta}{s(1-2\eta)+4\eta}$ *with* $0 < \eta < \frac{1}{2}$, *if* $\tau^+ > 0, \tau^- \geq 0$ *satisfy*

$$(1-\beta)\tau^-\tau^+ + 2\tau^- + \frac{4\eta}{s(1-2\eta)+2\eta} < \frac{\beta}{2}\left(\frac{s(1-2\eta)}{s(1-2\eta)+2\eta}\right)\tau^+ \tag{37}$$

then $V_k(\overline{T}) = \Theta R$, *and* $\left\|(C^-)^{-1/2}C^+(C^-)^{-1/2}\right\| < \beta\frac{\alpha_l^+}{\alpha_l^-}$, *i.e.,* $\lambda_{n-k}(\overline{T}) - \lambda_{n-k+1}(\overline{T}) > (1-\beta)\frac{\alpha_l^+}{\alpha_l^-}$. *Moreover, for* (37) *to hold, it suffices that*

$$\tau^+ > \frac{16\eta}{\beta s(1-2\eta)}, \quad \tau^- < \frac{\beta}{2}\left(\frac{s(1-2\eta)}{s(1-2\eta)+2\eta}\right)\min\left\{\frac{1}{4(1-\beta)}, \frac{\tau^+}{8}\right\}.$$

3. *The statement in part* (2) *also holds for the choice* $\beta = \frac{1}{2}$, *and provided* $\eta \leq \frac{s}{2s+4}$.

Proof From (35) and Lemma 22, it suffices to show for $\beta \in (0,1]$ that

$$\frac{\tau^- + \frac{\eta}{s(1-2\eta)+\eta-\frac{(1-\eta)}{n}}}{\tau^+} < \beta\left(\frac{1+\tau^-+p(1-\eta)/d_l^+}{1+\tau^++p\eta/d_l^-}\right). \tag{38}$$

For the stated condition on n, it is easy to verify that

$$n \geq \frac{2(1-\eta)}{s(1-2\eta)} \implies s(1-2\eta)+\eta-\frac{(1-\eta)}{n} \geq \frac{s(1-2\eta)}{2}+\eta,$$

$$n \geq \frac{2\eta}{(1-l)(1-\eta)} \implies \frac{p\eta}{d_l^-} \leq \frac{2\eta}{n(1-\eta)(1-l)} \leq 1.$$

Using these bounds in (38), observe that it suffices that τ^+, τ^- satisfy

$$\frac{\tau^- + \frac{2\eta}{s(1-2\eta)+2\eta}}{\tau^+} < \beta\left(\frac{1+\tau^-}{2+\tau^+}\right). \tag{39}$$

Then for $\beta = 1$, we readily see that (39) is equivalent to (36).

To establish the second part of the Lemma, we begin by rewriting (39) as

$$(1-\beta)\tau^+\tau^- + 2\tau^- + \frac{4\eta}{s(1-2\eta)+2\eta} < \left(\beta - \frac{2\eta}{s(1-2\eta)+2\eta}\right)\tau^+$$

$$= \left[\frac{\beta s(1-2\eta)-2\eta(1-\beta)}{s(1-2\eta)+2\eta}\right]\tau^+, \tag{40}$$

28

and observe that

$$\beta s(1 - 2\eta) \geq 4\eta(1 - \beta) \iff \beta \geq \frac{4\eta}{s(1 - 2\eta) + 4\eta} \tag{41}$$

This verifies (37) in the statement of the Lemma. The "moreover" part is established by ensuring that each term on the LHS of (37) is a sufficiently small fraction of the RHS term. In particular, it is enough to choose this fraction to be $1/4$ for the first two terms, and $1/2$ for the third term.

Finally, the third part of the Lemma can be shown in the same manner as the second part. The starting point is to ensure (40), and we simply observe that for $\beta = 1/2$, (41) is equivalent to $\eta \leq \frac{s}{2s+4}$. The rest follows identically. ∎

4.3 Concentration bound for $\left\| T - \overline{T} \right\|$

In this section, we bound the "distance" between T and $\overline{T}$, i.e., $\left\| T - \overline{T} \right\|$. This is shown via individually bounding the terms $\left\| L_{sym}^+ - \overline{L_{sym}^+} \right\|$, and $\left\| L_{sym}^- - \overline{L_{sym}^-} \right\|$. To this end, we first recall the following Theorem from Chung and Radcliffe (2011).

Theorem 25 (Bounding $\left\| L_{sym} - \overline{L_{sym}} \right\|$, (Chung and Radcliffe, 2011)) *Let L_{sym} denote the normalized Laplacian of a random graph, and $\overline{L_{sym}}$ the normalized Laplacian of the expected graph. Let δ be the minimum expected degree of the graph. Choose $\varepsilon > 0$. Then there exists a constant c_ε such that, if $\delta \geq c_\varepsilon \ln n$, then with probability at least $1 - \varepsilon$, it holds true that*

$$\left\| L_{sym} - \overline{L_{sym}} \right\| \leq 2\sqrt{\frac{3\ln(4n/\varepsilon)}{\delta}} \ .$$

Remark 26 *A similar result appears in Imbuzeiro Oliveira (2009) for the (unsigned) inhomogeneous Erdős-Rényi model, where $\left\| L_{sym} - \overline{L_{sym}} \right\| = O(\sqrt{\ln n/d_0})$, with d_0 the smallest expected degree of the graph.*

Using Theorem 25, we readily obtain the following concentration bounds for $\left\| L_{sym}^+ - \overline{L_{sym}^+} \right\|$ and $\left\| L_{sym}^- - \overline{L_{sym}^-} \right\|$.

Lemma 27 (Bounding $\left\| L_{sym}^\pm - \overline{L_{sym}^\pm} \right\|$) *Assuming $n \geq \max\left\{ \frac{2(1-\eta)}{s(1-2\eta)}, \frac{2\eta}{(1-l)(1-\eta)} \right\}$, there exists a constant $c_\varepsilon > 0$ such that if $p \geq \frac{c_\varepsilon \ln n}{n} \max\left\{ \frac{1}{s(1-2\eta)+2\eta}, \frac{2}{1-l} \right\}$, then with probability at least $1 - 2\varepsilon$,*

$$\left\| L_{sym}^+ - \overline{L_{sym}^+} \right\| \leq 2\sqrt{\frac{6\ln(4n/\varepsilon)}{np[s(1 - 2\eta) + 2\eta]}}, \qquad and \qquad \left\| L_{sym}^- - \overline{L_{sym}^-} \right\| \leq 2\sqrt{\frac{12\ln(4n/\varepsilon)}{np(1 - l)}}.$$

Proof Note that the minimum expected degrees of the positive and negative subgraphs are given by d_s^+, d_l^-, respectively. For the stated condition on n, it is easily seen that

$$d_s^+ \geq \frac{np}{2}\left[s(1 - 2\eta) + 2\eta\right], \quad d_l^- \geq \frac{np}{2}(1 - l)(1 - \eta) \geq \frac{np(1 - l)}{4}. \tag{42}$$

29

Invoking Theorem 25, and observing that $d_s^+, d_l^- \geq \frac{c_\varepsilon}{2} \ln n$ are ensured for the stated condition on p, the statement follows via the union bound. ∎

Next, using the above lemma, we can upper bound $\|T - \overline{T}\|$. This will help us show that $V_k(T)$ and $V_k(\overline{T})$ are "close".

Lemma 28 (Bounding $\|T - \overline{T}\|$) *Let $P = (L_{sym}^- + \tau^+ I)$, $\overline{P} = (\overline{L_{sym}^-} + \tau^+ I)$, $Q = (L_{sym}^+ + \tau^- I)$, and $\overline{Q} = (\overline{L_{sym}^+} + \tau^- I)$. Assume that $\|P - \overline{P}\| \leq \Delta_P$, and $\|Q - \overline{Q}\| \leq \Delta_Q$. Then it holds true that*

$$\|T - \overline{T}\| \leq \frac{(\alpha_s^+ + \Delta_Q)}{\tau^+} \left(\frac{\Delta_P}{\tau^+} + 2\sqrt{\frac{\Delta_P}{\tau^+}} \right) + \frac{\Delta_Q}{\tau^+}$$

where $\alpha_s^+ = 1 + \tau^- + \frac{p(1-\eta)}{d_s^+}$ (see Lemma 17).

Proof Since $P, \overline{P}, Q, \overline{Q}$ are positive definite, therefore using Proposition 59, we obtain the bound

$$\|T - \overline{T}\| \leq \|P^{-1}\| \|Q\| \left(\|(\overline{P})^{-1}\| \|\overline{P} - P\| + 2 \|(\overline{P})^{-1/2}\| \|\overline{P} - P\|^{1/2} \right) + \|(\overline{P})^{-1}\| \|Q - \overline{Q}\|. \tag{43}$$

We know that $\|P^{-1}\| = 1/\tau^+ = \|\overline{P}^{-1}\|$ and $\|(\overline{P})^{-1/2}\| = 1/\sqrt{\tau^+}$. Moreover, $\|Q\| \leq \|\overline{Q}\| + \Delta_Q$ by Weyl's inequality (Weyl, 1912) (see Appendix B). Hence (43) simplifies to

$$\|T - \overline{T}\| \leq \frac{(\|\overline{Q}\| + \Delta_Q)}{\tau^+} \left(\frac{\Delta_P}{\tau^+} + 2\sqrt{\frac{\Delta_P}{\tau^+}} \right) + \frac{\Delta_Q}{\tau^+} \leq \frac{(\alpha_s^+ + \Delta_Q)}{\tau^+} \left(\frac{\Delta_P}{\tau^+} + 2\sqrt{\frac{\Delta_P}{\tau^+}} \right) + \frac{\Delta_Q}{\tau^+},$$

where the last inequality can be verified by examining the expression of $\overline{Q}$ in (24), and noting from the definition of C^+ that $\|C^+\| < \max\{\alpha_1^+, ..., \alpha_k^+\} = \alpha_s^+$ holds (via Weyl's inequality). ∎

4.4 Estimating $V_k(\overline{T})$ and $G_k(\overline{T})$ up to a rotation

We are now ready to combine the results of the previous sections to show that if n, p are large enough, then the distance between the subspaces spanned by $V_k(T)$ and $V_k(\overline{T})$ is small, i.e., there exists an orthonormal matrix O such that $V_k(T)$ is close to $V_k(\overline{T})O$. For τ^+, τ^- chosen suitably, we have seen in Lemma 24 that $V_k(\overline{T}) = \Theta R$ for a rotation R, hence this suggests that the rows of $V_k(T)$ will then also approximately preserve the clustering structure of $V_k(\overline{T})$.

With $P, \overline{P}, Q, \overline{Q}$ as defined in Lemma 28 recall from (4), (7) that $G_k(T), G_k(\overline{T})$ can be written as

$$G_k(\overline{T}) = \overline{P}^{-1/2} V_k(\overline{T}), \quad G_k(T) = P^{-1/2} V_k(T). \tag{44}$$

Therefore if $V_k(\overline{T}) = \Theta R$, then using the expression for $\overline{P}$ from (24) we see that $G_k(\overline{T}) = \Theta(C^-)^{-1/2}R$, and thus the rows of $G_k(\overline{T})$ also preserve the ground truth clustering structure. Moreover, if $\left\|V_k(T) - V_k(\overline{T})O\right\|$ is small, then it can be shown to imply a bound on $\left\|G_k(T) - G_k(\overline{T})O\right\|$. Hence the rows of $G_k(T)$ will approximately preserve the clustering structure of $G_k(\overline{T})$.

Before stating the theorem, let us define the terms

$$C_1(\tau^+, \tau^-) = 3\left(\frac{(3+\tau^-)(2\sqrt{\tau^+}+1)+\tau^+}{(\tau^+)^2}\right), \quad C_2(s, \eta, l) = \max\left\{\frac{1}{s(1-2\eta)+2\eta}, \frac{2}{1-l}\right\}. \tag{45}$$

Theorem 29 *Assuming $n \geq \max\left\{\frac{2(1-\eta)}{s(1-2\eta)}, \frac{2\eta}{(1-l)(1-\eta)}\right\}$, suppose $\tau^+ > 0, \tau^- \geq 0$ are chosen to satisfy*

$$\tau^+ > \frac{16\eta}{\beta s(1-2\eta)}, \quad \tau^- < \frac{\beta}{2}\left(\frac{s(1-2\eta)}{s(1-2\eta)+2\eta}\right)\min\left\{\frac{1}{4(1-\beta)}, \frac{\tau^+}{8}\right\}$$

where β, η satisfy one of the following conditions.

1. $\beta = \frac{4\eta}{s(1-2\eta)+4\eta}$ and $0 < \eta < \frac{1}{2}$, or

2. $\beta = \frac{1}{2}$ and $\eta \leq \frac{s}{2s+4}$.

Then $V_k(\overline{T}) = \Theta R$ and $G_k(\overline{T}) = \Theta(C^-)^{-1/2}R$ where R is a rotation matrix, and $C^- \succ 0$ is as defined in (18). Moreover, for any $\varepsilon, \delta \in (0,1)$, there exists a constant $\widetilde{c}_\varepsilon > 0$ such that the following is true. If p satisfies

$$p \geq \max\left\{\widetilde{c}_\varepsilon C_2(s, \eta, l), \frac{256C_1^4(\tau^+, \tau^-)(2+\tau^+)^4}{\delta^4(1+\tau^-)^4(1-\beta)^4}C_2(s, \eta, l), \frac{81}{(1-l)\delta^4}\right\}\frac{\ln(4n/\varepsilon)}{n},$$

with $C_1(\cdot), C_2(\cdot)$ as in (45), then with probability at least $1 - 2\varepsilon$, there exists an orthogonal matrix $O \in \mathbb{R}^{k \times k}$ such that

$$\left\|V_k(T) - V_k(\overline{T})O\right\| \leq \delta, \quad \text{and} \quad \left\|G_k(T) - G_k(\overline{T})O\right\| \leq \frac{\delta}{\sqrt{\tau^+}} + \frac{\delta}{(\tau^+)^2}.$$

Proof We will first simplify the upper bound on $\left\|T - \overline{T}\right\|$ in Lemma 28, starting by bounding α_s^+. If $n \geq \frac{2(1-\eta)}{s(1-2\eta)}$, it is easy to verify that $\frac{(1-\eta)p}{d_s^+} \leq 1$ which implies $\alpha_s^+ \leq 2+\tau^-$. Moreover, we observe from Lemma 27 that $\Delta_P, \Delta_Q \leq 1$ is ensured if $p \geq \widetilde{c}_\varepsilon C_2(s, \eta, l)\frac{\ln(4n/\varepsilon)}{n}$ where $\widetilde{c}_\varepsilon = \max\{24, c_\varepsilon\}$. These considerations altogether imply

$$\left\|T - \overline{T}\right\| \leq \frac{(3+\tau^-)(2\sqrt{\tau^+}+1)}{(\tau^+)^2}\sqrt{\Delta_P} + \frac{\Delta_Q}{\tau^+}$$

$$\leq \frac{(3+\tau^-)(2\sqrt{\tau^+}+1)+\tau^+}{(\tau^+)^2}\max\left\{\sqrt{\Delta_P}, \sqrt{\Delta_Q}\right\}$$

$$\leq C_1(\tau^+, \tau^-)C_2^{1/4}(s, \eta, l)\left(\frac{\ln(4n/\varepsilon)}{np}\right)^{1/4} \tag{46}$$

31

where in the penultimate inequality we used $\Delta_Q \leq \sqrt{\Delta_Q}$, and the last inequality uses Lemma 27.

Next, we will use the Davis-Kahan theorem (Davis and Kahan, 1970) (see Appendix B) for bounding the distance $\left\| (I - V_k(\overline{T}) V_k(\overline{T})^T) V_k(T) \right\|$. Applied to our setup, it yields

$$\left\| (I - V_k(\overline{T}) V_k(\overline{T})^T) V_k(T) \right\| \leq \frac{\left\| T - \overline{T} \right\|}{\lambda_{n-k+1}(T) - \lambda_{n-k}(\overline{T})}, \tag{47}$$

provided $\lambda_{n-k+1}(T) - \lambda_{n-k}(\overline{T}) > 0$. From Weyl's inequality, we know that $\lambda_{n-k+1}(T) \geq \lambda_{n-k+1}(\overline{T}) - \left\| T - \overline{T} \right\|$. Moreover, under the stated conditions on τ^+, τ^-, we obtain from Lemma 24 the bound

$$\lambda_{n-k+1}(\overline{T}) - \lambda_{n-k}(\overline{T}) \geq (1 - \beta) \frac{\alpha_l^+}{\alpha_l^-} \geq (1 - \beta) \left(\frac{1 + \tau^-}{2 + \tau^+} \right),$$

where in the last inequality we used the simplifications $p(1 - \eta)/d_l^+ \geq 0$ and $p\eta/d_l^- \leq 1$ in the expressions for α_l^+, α_l^-. Hence using (46), we observe that if

$$C_1(\tau^+, \tau^-) C_2^{1/4}(s, \eta, l) \left(\frac{\ln(4n/\varepsilon)}{np} \right)^{1/4} \leq \left(\frac{1 - \beta}{2} \right) \left(\frac{1 + \tau^-}{2 + \tau^+} \right)$$

$$\iff p \geq \left(\frac{16 C_1^4(\tau^+, \tau^-) C_2(s, \eta, l)(2 + \tau^+)^4}{(1 + \tau^-)^4 (1 - \beta)^4} \right) \frac{\ln(4n/\varepsilon)}{n},$$

then the RHS of (47) can be bounded as

$$\left\| (I - V_k(\overline{T}) V_k(\overline{T})^T) V_k(T) \right\| \leq \frac{2(2 + \tau^+)}{(1 + \tau^-)(1 - \beta)} C_1(\tau^+, \tau^-) C_2^{1/4}(s, \eta, l) \left(\frac{\ln(4n/\varepsilon)}{np} \right)^{1/4}.$$

It follows that there exists an orthogonal matrix $O \in \mathbb{R}^{k \times k}$ so that

$$\left\| V_k(T) - V_k(\overline{T}) O \right\| \leq 2 \left\| (I - V_k(\overline{T}) V_k(\overline{T})^T) V_k(T) \right\| \quad \text{(using Proposition 57)}$$

$$\leq \frac{4(2 + \tau^+)}{(1 + \tau^-)(1 - \beta)} C_1(\tau^+, \tau^-) C_2^{1/4}(s, \eta, l) \left(\frac{\ln(4n/\varepsilon)}{np} \right)^{1/4}$$

$$\leq \delta$$

for the stated bound on p. This establishes the first part of the Theorem.

In order to bound $\left\| G_k(T) - G_k(\overline{T}) O \right\|$, we obtain from (44) that

$$\left\| G_k(T) - G_k(\overline{T}) O \right\| = \left\| P^{-1/2}(V_k(T) - V_k(\overline{T}) O) + (P^{-1/2} - \overline{P}^{-1/2}) V_k(\overline{T}) O \right\|$$

$$\leq \underbrace{\left\| P^{-1/2} \right\|}_{(\tau^+)^{-1/2}} \underbrace{\left\| V_k(T) - V_k(\overline{T}) O \right\|}_{\leq \delta} + \left\| P^{-1/2} - \overline{P}^{-1/2} \right\| \underbrace{\left\| V_k(\overline{T}) \right\|}_{=1}$$

$$\leq \frac{\delta}{\sqrt{\tau^+}} + \left\| P^{-1/2} - \overline{P}^{-1/2} \right\|. \tag{48}$$

The term $\left\| P^{-1/2} - \overline{P}^{-1/2} \right\|$ can be bounded as

$$
\begin{aligned}
\left\| P^{-1/2} - \overline{P}^{-1/2} \right\| = \left\| P^{-1}(P^{1/2} - \overline{P}^{1/2})\overline{P}^{-1} \right\| &\leq \frac{\left\| P^{1/2} - \overline{P}^{1/2} \right\|}{(\tau^+)^2} \\
&\leq \frac{\left\| P - \overline{P} \right\|^{1/2}}{(\tau^+)^2} \\
&\leq \frac{3}{(\tau^+)^2} \left[\frac{\ln(4n/\varepsilon)}{np(1-l)} \right]^{1/4},
\end{aligned} \tag{49}
$$

where the penultimate inequality uses Proposition 58, and the last inequality follows from Lemma 27 with a minor simplification of the constant. Plugging (49) in (48) leads to the stated bound for $p \geq \frac{81}{(1-l)\delta^4} \frac{\ln(4n/\varepsilon)}{n}$. ∎

4.5 Clustering sparse graphs

We now turn our attention to the sparse regime where $p = o(\ln n)/n$. In this regime, Lemma 27 is no longer applicable since it requires $p = \Omega\left(\frac{\ln n}{n}\right)$. In fact, it is not difficult to see that the matrices $L_{sym}^{\pm}$ will not concentrate around $\overline{L_{sym}^{\pm}}$ in this sparsity regime. To circumvent this issue, we will aim to show that the normalized Laplacian $L_{sym,\gamma^\pm}^{\pm}$ corresponding to the regularized adjacencies $A_{\gamma^\pm}^{\pm} := A^{\pm} + \frac{\gamma^\pm}{n} \mathbb{1}\mathbb{1}^\top$ concentrate around $\overline{L_{sym}^{\pm}}$, for carefully chosen values of γ^+, γ^-.

To show this, we rely on the following theorem from Le et al. (2017), which states that the symmetric Laplacian $L_{sym,\gamma}$ of the regularized adjacency matrix $A_\gamma := A + \frac{\gamma}{n}\mathbb{1}\mathbb{1}^\top$ is close to the symmetric Laplacian $\overline{L_{sym,\gamma}}$ of the expected regularized adjacency matrix, for inhomogeneous Erdős-Rényi graphs.

Theorem 30 (Theorem 4.1 of Le et al. (2017)) *Consider a random graph from the inhomogeneous Erdős-Rényi model ($G = (n, p_{ij})$), and let $d = \max_{p_{ij}} np_{ij}$. Choose a number $\gamma > 0$. Then, for any $r \geq 1$, C being an absolute constant, with probability at least $1 - e^{-r}$*

$$
\left\| L_{sym,\gamma} - \overline{L_{sym,\gamma}} \right\| \leq \frac{Cr^2}{\sqrt{\gamma}} \left(1 + \frac{d}{\gamma} \right)^{5/2}. \tag{50}
$$

The above result leads to a bound on the distance between $L_{sym,\gamma}$ and the normalized Laplacian $\overline{L_{sym}}$ of the expected (un-regularized) adjacency matrix.

Theorem 31 (Concentration of Regularized Laplacians) *Consider a random graph from the inhomogeneous Erdős-Rényi model ($G = (n, p_{ij})$), and let $d = \max_{p_{ij}} np_{ij}$, $d_{\min} = \min_i \sum_j p_{ij}$. Choose a number $\gamma > 0$. Then, for any $r \geq 1$, C being an absolute constant, with probability at least $1 - e^{-r}$*

$$
\left\| L_{sym,\gamma} - \overline{L_{sym}} \right\| \leq \frac{Cr^2}{\sqrt{\gamma}} \left(1 + \frac{d}{\gamma} \right)^{5/2} + 3\sqrt{\frac{\gamma}{d_{\min} + \gamma}}. \tag{51}
$$

Proof To establish the above lemma we make use of triangle inequality, where we use the fact that $\left\|L_{sym,\gamma} - \overline{L_{sym}}\right\| \leq \left\|L_{sym,\gamma} - \overline{L_{sym,\gamma}}\right\| + \left\|\overline{L_{sym,\gamma}} - \overline{L_{sym}}\right\|$. We know the bound on the first term on the RHS from Lemma 30 (which holds with probability $1 - e^{-r}$). To bound the second term on the RHS, note that

$$\left\|\overline{L_{sym,\gamma}} - \overline{L_{sym}}\right\| = \left\|\overline{D}^{-1/2}\overline{A}\,\overline{D}^{-1/2} - \overline{D}_\gamma^{-1/2}\overline{A}_\gamma\overline{D}_\gamma^{-1/2}\right\|$$

$$= \left\|\overline{D}^{-1/2}\overline{A}\,\overline{D}^{-1/2} - \overline{D}_\gamma^{-1/2}\overline{A}\,\overline{D}_\gamma^{-1/2} + \overline{D}_\gamma^{-1/2}\overline{A}\,\overline{D}_\gamma^{-1/2} - \overline{D}_\gamma^{-1/2}\overline{A}_\gamma\overline{D}_\gamma^{-1/2}\right\|$$

$$\leq \left\|\overline{D}^{-1/2}\overline{A}\,\overline{D}^{-1/2} - \overline{D}_\gamma^{-1/2}\overline{A}\,\overline{D}_\gamma^{-1/2}\right\| + \left\|\overline{D}_\gamma^{-1/2}\overline{A}\,\overline{D}_\gamma^{-1/2} - \overline{D}_\gamma^{-1/2}\overline{A}_\gamma\overline{D}_\gamma^{-1/2}\right\|.$$

The second term of the inequality can be easily bounded as follows.

$$\left\|\overline{D}_\gamma^{-1/2}\overline{A}\,\overline{D}_\gamma^{-1/2} - \overline{D}_\gamma^{-1/2}\overline{A}_\gamma\overline{D}_\gamma^{-1/2}\right\| \leq \left\|\overline{D}_\gamma^{-1/2}\right\|^2 \left\|\overline{A} - \overline{A}_\gamma\right\| \leq \frac{\gamma}{d_{\min} + \gamma} \leq \sqrt{\frac{\gamma}{d_{\min} + \gamma}}.$$

To analyse the first term, we observe that

$$\left\|\overline{D}^{-1/2}\overline{A}\,\overline{D}^{-1/2} - \overline{D}_\gamma^{-1/2}\overline{A}\,\overline{D}_\gamma^{-1/2}\right\| =$$

$$\left\|\overline{D}^{-1/2}\overline{A}\,\overline{D}^{-1/2} - \overline{D}_\gamma^{-1/2}\overline{D}^{1/2}\overline{D}^{-1/2}\overline{A}\,\overline{D}^{-1/2}\overline{D}^{1/2}\overline{D}_\gamma^{-1/2}\right\|$$

$$= \left\|(I - \overline{L_{sym}})(I - \overline{D}^{1/2}\overline{D}_\gamma^{-1/2}) + (I - \overline{D}_\gamma^{-1/2}\overline{D}^{1/2})(I - \overline{L_{sym}})\overline{D}^{1/2}\overline{D}_\gamma^{-1/2}\right\|$$

$$\leq \left\|I - \overline{D}^{1/2}\overline{D}_\gamma^{-1/2}\right\| + \left\|I - \overline{D}_\gamma^{-1/2}\overline{D}^{1/2}\right\| \left\|\overline{D}^{1/2}\overline{D}_\gamma^{-1/2}\right\|$$

$$\leq \left(1 - \sqrt{\frac{d_{\min}}{d_{\min} + \gamma}}\right) + \left(1 - \sqrt{\frac{d_{\min}}{d_{\min} + \gamma}}\right)$$

$$\leq 2\sqrt{\frac{\gamma}{d_{\min} + \gamma}},$$

where in the first inequality we use the fact that $\left\|I - \overline{L_{sym}}\right\| \leq 1$, and in the last inequality we use the fact that for two numbers $a, b > 0$ if $a > b$ then $\sqrt{a} - \sqrt{b} \leq \sqrt{a - b}$. We have all the components to plug into the triangle inequality, which yields the desired statement of the theorem. ∎

We now translate Theorem 31 to our setting for G^+, G^- and show that if $p = \Omega(1/n)$ for n large enough, then for the choices $\gamma^+, \gamma^- \asymp (np)^{6/7}$, the bounds $\left\|L_{sym,\gamma^\pm}^\pm - \overline{L_{sym}^\pm}\right\| = O\left(\frac{1}{(np)^{1/14}}\right)$ hold with sufficiently high probability.

Lemma 32 *Let $n \geq \max\left\{\frac{2(1-\eta)}{s(1-2\eta)}, \frac{2\eta}{(1-\eta)(1-l)}\right\}$ and $p \geq \frac{1}{n(1-\eta)}$. Then for the choices $\gamma^+, \gamma^- = [np(1-\eta)]^{6/7}$, and any $r \geq 1$, there exists a constant $C > 0$ such that with probability at least $1 - 2e^r$, it holds true that*

$$\left\|L_{sym,\gamma^+}^+ - \overline{L_{sym}^+}\right\| \leq \left(2^{5/2}Cr^2 + \frac{3\sqrt{2}}{\sqrt{s(1-2\eta) + 2\eta}}\right)\frac{1}{[np(1-\eta)]^{1/14}}, \tag{52}$$

$$\left\|L_{sym,\gamma^-}^- - \overline{L_{sym}^-}\right\| \leq \left(2^{5/2}Cr^2 + \frac{6}{\sqrt{1-l}}\right)\frac{1}{[np(1-\eta)]^{1/14}}. \tag{53}$$

Proof We will apply Theorem 31 to the subgraphs G^+, G^-. Let us denote $d^{\pm}$ to be the quantity $\max_{ij} np_{ij}$, and $d^{\pm}_{min}$ to be the minimum expected degree for the positive and negative subgraphs, respectively. From the SSBM model, it can be verified that $d^{\pm} = np(1 - \eta)$. We also know that $d^+_{\min} = d^+_s$ and $d^-_{\min} = d^-_l$, where for the stated condition on n, d^+_s, d^-_l satisfy the bounds in (42). The latter can be written as

$$d^+_{\min} \geq \frac{d^+}{2}[s(1 - 2\eta) + 2\eta], \qquad d^-_{\min} \geq \frac{d^-(1 - l)}{4}.$$

Let us denote $C_3(s, \eta) = s(1 - 2\eta) + 2\eta$ for convenience. In order to show (52), we obtain from Theorem 31 that, with probability at least $1 - e^{-r}$,

$$\left\| L^+_{sym,\gamma^+} - \overline{L^+_{sym}} \right\| \leq \frac{Cr^2}{\sqrt{\gamma^+}} \left(1 + \frac{d^+}{\gamma^+} \right)^{5/2} + 3\sqrt{\frac{\gamma^+}{d^+_{\min} + \gamma^+}}$$

$$\leq \frac{Cr^2}{\sqrt{\gamma^+}} \left(1 + \frac{d^+}{\gamma^+} \right)^{5/2} + 3\sqrt{\frac{\gamma^+}{C_3(s, \eta)d^+}},$$

where the last inequality uses $d^+_s + \gamma^+ \geq d^+_s$. Now note that if $\gamma^+ \leq d^+$, then the above bound simplifies to

$$\left\| L^+_{sym,\gamma^+} - \overline{L^+_{sym}} \right\| \leq \frac{2^{5/2} Cr^2 (d^+)^{5/2}}{(\gamma^+)^3} + \frac{3\sqrt{2}}{\sqrt{C_3(s, \eta)}} \sqrt{\frac{\gamma^+}{d^+}}. \tag{54}$$

Choosing γ^+ such that $\frac{(d^+)^{5/2}}{(\gamma^+)^3} = \sqrt{\frac{\gamma^+}{d^+}}$, or equivalently, $\gamma^+ = (d^+)^{6/7}$, and plugging this in (54), we arrive at (52). Clearly, $\gamma^+ \leq d^+$ is equivalent to the stated condition on p. The bound in (53) follows in an identical manner and is omitted. ∎

We are now in a position to write the bound on $\left\| T_{\gamma^+, \gamma^-} - \overline{T} \right\|$ in terms of $\left\| L^{\pm}_{sym,\gamma^{\pm}} - \overline{L^{\pm}_{sym}} \right\|$, in a completely analogous manner to Lemma 28.

Lemma 33 (Adapting Lemma 28 for the sparse regime) *Let* $P_{\gamma^-} = (L^-_{sym,\gamma^-} + \tau^+ I)$, $\overline{P} = (\overline{L^-_{sym}} + \tau^+ I)$, $Q_{\gamma^+} = (L^+_{sym,\gamma^+} + \tau^- I)$, *and* $\overline{Q} = (\overline{L^+_{sym}} + \tau^- I)$. *Assume that* $\left\| P_{\gamma^-} - \overline{P} \right\| \leq \Delta_{P_{\gamma^-}}$, $\left\| Q_{\gamma^+} - \overline{Q} \right\| \leq \Delta_{Q_{\gamma^+}}$. *Then it holds true that*

$$\left\| T_{\gamma^+, \gamma^-} - \overline{T} \right\| \leq \frac{(\alpha^+_s + \Delta_{Q_{\gamma^+}})}{\tau^+} \left(\frac{\Delta_{P_{\gamma^-}}}{\tau^+} + 2\sqrt{\frac{\Delta_{P_{\gamma^-}}}{\tau^+}} \right) + \frac{\Delta_{Q_{\gamma^+}}}{\tau^+},$$

where $\alpha^+_s = 1 + \tau^- + \frac{p(1 - \eta)}{d^+_s}$ *(see Lemma 17).*

Next, we derive the main theorem for SPONGE_{sym} in the sparse regime, which is the analogue of Theorem 29. The first part of the Theorem remains unchanged, i.e., for n large enough and τ^+, τ^- chosen suitably, we have $V_k(\overline{T}) = \Theta R$ and $G_k(\overline{T}) = \Theta(C^-)^{-1/2} R$ for a $k \times k$ rotation R, and $C^- \succ 0$. The remaining arguments follow the same outline of Theorem 29, i.e., (a) using Lemma 33 and Lemma 32 to obtain a concentration bound on

$\left\|T_{\gamma^+,\gamma^-} - \overline{T}\right\|$ (when $p = \Omega(1/n)$), and (b) using the Davis-Kahan theorem to show that the column span of $V_k(T_{\gamma^+,\gamma^-})$ is close to $V_k(\overline{T})$. The latter bound then implies that $G_k(T_{\gamma^+,\gamma^-})$ is close (up to a rotation) to $G_k(\overline{T})$, where we recall

$$G_k(\overline{T}) = \overline{P}^{-1/2}V_k(\overline{T}), \quad G_k(T_{\gamma^+,\gamma^-}) = P_{\gamma^-}^{-1/2}V_k(T_{\gamma^+,\gamma^-}) \tag{55}$$

with $P_{\gamma^-}, \overline{P}$ as defined in Lemma 33.

Theorem 34 *Assuming* $n \geq \max\left\{\frac{2(1-\eta)}{s(1-2\eta)}, \frac{2\eta}{(1-l)(1-\eta)}\right\}$, *suppose* $\tau^+ > 0, \tau^- \geq 0$ *are chosen to satisfy*

$$\tau^+ > \frac{16\eta}{\beta s(1-2\eta)}, \quad \tau^- < \frac{\beta}{2}\left(\frac{s(1-2\eta)}{s(1-2\eta)+2\eta}\right)\min\left\{\frac{1}{4(1-\beta)}, \frac{\tau^+}{8}\right\}$$

where β, η *satisfy one of the following conditions.*

1. $\beta = \frac{4\eta}{s(1-2\eta)+4\eta}$ *and* $0 < \eta < \frac{1}{2}$, *or*

2. $\beta = \frac{1}{2}$ *and* $\eta \leq \frac{s}{2s+4}$.

Then $V_k(\overline{T}) = \Theta R$ *and* $G_k(\overline{T}) = \Theta(C^-)^{-1/2}R$ *where* R *is a rotation matrix, and* $C^- \succ 0$ *is as defined in* (18). *Moreover, there exists a constant* $C > 0$ *such that for* $r \geq 1$ *and* $\delta \in (0,1)$, *if* p *satisfies*

$$p \geq \max\left\{1, \left(\frac{4C_1(\tau^+,\tau^-)(2+\tau^+)}{3(\tau^+)^2(1-\beta)(1+\tau^-)}\right)^{28}\right\}\frac{C_4^{14}(r,s,\eta,l)}{\delta^{28}(1-\eta)n},$$

and $\gamma^+, \gamma^- = [np(1-\eta)]^{6/7}$, *then with probability at least* $1 - 2e^{-r}$, *there exists a rotation* $O \in \mathbb{R}^{k\times k}$ *so that*

$$\left\|V_k(T_{\gamma^+,\gamma^-}) - V_k(\overline{T})O\right\| \leq \delta, \qquad \left\|G_k(T_{\gamma^+,\gamma^-}) - G_k(\overline{T})O\right\| \leq \frac{\delta}{\sqrt{\tau^+}} + \frac{\delta}{(\tau^+)^2}.$$

Here, $C_4(r,s,\eta,l) := 2^{5/2}Cr^2 + 3\sqrt{2C_2(s,\eta,l)}$ *with* $C_2(s,\eta,l)$ *as defined in* (45).

Proof We will first simplify the upper bound on $\left\|T_{\gamma^+,\gamma^-} - \overline{T}\right\|$ in Lemma 33. Note that $n \geq \frac{2(1-\eta)}{s(1-2\eta)}$ implies $\alpha_s^+ \leq 2+\tau^-$, and moreover, we can bound $\left\|L_{sym,\gamma^\pm}^\pm - \overline{L_{sym}^\pm}\right\|$ uniformly (from (52), (53)) as

$$\left\|L_{sym,\gamma^\pm}^\pm - \overline{L_{sym}^\pm}\right\| \leq \frac{2^{5/2}Cr^2 + 3\sqrt{2C_2(s,\eta,l)}}{[np(1-\eta)]^{1/14}} \leq \frac{C_4(r,s,\eta,l)}{[np(1-\eta)]^{1/14}} \ (= \Delta_{P_{\gamma^-}}, \Delta_{Q_{\gamma^+}}). \tag{56}$$

Note that $\Delta_{P_{\gamma^-}}, \Delta_{Q_{\gamma^+}} \leq 1$ if $p \geq \frac{C_4^{14}(r,s,\eta,l)}{n(1-\eta)}$. Under these considerations, the bound in Lemma 33 simplifies to

$$\left\|T_{\gamma^+,\gamma^-} - \overline{T}\right\| \leq \frac{(3+\tau^-)(2\sqrt{\tau^+}+1)+\tau^+}{(\tau^+)^2}\max\left\{\sqrt{\Delta_{P_{\gamma^-}}}, \sqrt{\Delta_{Q_{\gamma^+}}}\right\}$$

$$= \frac{C_1(\tau^+,\tau^-)\sqrt{C_4(r,s,\eta,l)}}{3(\tau^+)^2[np(1-\eta)]^{1/28}}.$$

36

Following the steps in the proof of Theorem 29, we observe that

$$\left\| T_{\gamma^+,\gamma^-} - \overline{T} \right\| \leq \frac{1}{2}(\lambda_{n-k+1}(T_{\gamma^+,\gamma^-}) - \lambda_{n-k}(\overline{T})),$$

is guaranteed to hold, provided

$$\frac{C_1(\tau^+,\tau^-)\sqrt{C_4(r,s,\eta,l)}}{3(\tau^+)^2[np(1-\eta)]^{1/28}} \leq (\frac{1-\beta}{2})\left(\frac{1+\tau^-}{2+\tau^+}\right)$$

$$\Longleftrightarrow p \geq \left(\frac{2C_1(\tau^+,\tau^-)(2+\tau^+)}{3(\tau^+)^2(1-\beta)(1+\tau^-)}\right)^{28}\frac{C_4^{14}(r,s,\eta,l)}{n(1-\eta)}.$$

Then, we obtain via the Davis-Kahan theorem that there exists an orthogonal matrix $O \in \mathbb{R}^{k \times k}$ such that

$$\left\| V_k(T_{\gamma^+,\gamma^-}) - V_k(\overline{T})O \right\| \leq \frac{4\left\| T_{\gamma^+,\gamma^-} - \overline{T} \right\|}{\lambda_{n-k+1}(\overline{T}) - \lambda_{n-k}(\overline{T})}$$

$$\leq \frac{4C_1(\tau^+,\tau^-)\sqrt{C_4(r,s,\eta,l)}(2+\tau^+)}{3(\tau^+)^2[np(1-\eta)]^{1/28}(1-\beta)(1+\tau^-)} \leq \delta,$$

for the stated bound on p in the theorem. This establishes the first part of the theorem.

In order to bound $\left\| G_k(T_{\gamma^+,\gamma^-}) - G_k(\overline{T})O \right\|$, first observe that

$$\left\| G_k(T_{\gamma^+,\gamma^-}) - G_k(\overline{T})O \right\| = \left\| P_{\gamma^-}^{-1/2}(V_k(T_{\gamma^+,\gamma^-}) - V_k(\overline{T})O) + (P_{\gamma^-}^{-1/2} - \overline{P}^{-1/2})V_k(\overline{T})O \right\|$$

$$\leq \underbrace{\left\| P_{\gamma^-}^{-1/2} \right\|}_{\leq (\tau^+)^{-1/2}}\underbrace{\left\| V_k(T_{\gamma^+,\gamma^-}) - V_k(\overline{T})O \right\|}_{\leq \delta} + \left\| P^{-1/2} - \overline{P}^{-1/2} \right\|\underbrace{\left\| V_k(\overline{T}) \right\|}_{=1}$$

$$\leq \frac{\delta}{\sqrt{\tau^+}} + \left\| P_{\gamma^-}^{-1/2} - \overline{P}^{-1/2} \right\|. \tag{57}$$

The second term $\left\| P_{\gamma^-}^{-1/2} - \overline{P}^{-1/2} \right\|$ can be bounded as

$$\left\| P_{\gamma^-}^{-1/2} - \overline{P}^{-1/2} \right\| = \left\| P_{\gamma^-}^{-1}(P_{\gamma^-}^{1/2} - \overline{P}^{1/2})\overline{P}^{-1} \right\| \tag{58}$$

$$\leq \frac{\left\| P_{\gamma^-}^{1/2} - \overline{P}^{1/2} \right\|}{(\tau^+)^2} \leq \frac{\left\| P_{\gamma^-} - \overline{P} \right\|^{1/2}}{(\tau^+)^2} \leq \frac{\sqrt{C_4(r,s,\eta,l)}}{(\tau^+)^2[np(1-\eta)]^{1/28}}, \tag{59}$$

where the penultimate inequality uses Proposition 58, and the last inequality uses (56). Plugging (58) into (57) leads to the stated bound for $p \geq \frac{C_4^{14}(r,s,\eta,l)}{n(1-\eta)\delta^{28}}$. ∎

4.6 Mis-clustering rate from k-means

We now analyze the mis-clustering error rate when we apply a $(1+\xi)$-approximate k-means algorithm (e.g., (Kumar et al., 2004)) on the rows of $G_k(T)$ (respectively, $G_k(T_{\gamma^+,\gamma^-})$ in the sparse regime). To this end, we rely on the following result from Lei and Rinaldo (2015),

which when applied to our setting, yields that the mis-clustering error is bounded by the estimation error $\left\|G_k(T) - G_k(\overline{T})O\right\|_F^2$ (or $\left\|G_k(T_{\gamma^+,\gamma^-}) - G_k(\overline{T})O\right\|_F^2$ in the sparse setting). By an $(1+\xi)$-approximate algorithm, we mean an algorithm that is provably within an $(1+\xi)$ factor of the cost of the optimal solution achieved by k-means.

Lemma 35 (Lemma 5.3 of Lei and Rinaldo (2015), Approx. k-means error) *For any $\xi > 0$, and any two matrices $\overline{U}, U$, such that $\overline{U} = \overline{\Theta}\,\overline{X}$ with $(\overline{\Theta}, \overline{X}) \in \mathbb{M}_{n \times k} \times \mathbb{R}^{k \times k}$, let $(\widetilde{\Theta}, \widetilde{X}) \in \mathbb{M}_{n \times k} \times \mathbb{R}^{k \times k}$ be a $(1+\xi)$-approximate solution to the k-means problem $\min_{\Theta \in \mathbb{M}_{n \times k}, X \in \mathbb{R}^{k \times k}} \|\Theta X - U\|_F^2$ so that*

$$\left\|\widetilde{\Theta}\widetilde{X} - U\right\|_F^2 \leq (1+\xi) \min_{\Theta \in \mathbb{M}_{n \times k}, X \in \mathbb{R}^{k \times k}} \|\Theta X - U\|_F^2$$

and $\widetilde{U} = \widetilde{\Theta}\widetilde{X}$. For any $\delta_i \leq \min_{i' \neq i} \left\|\overline{X}_{i'*} - \overline{X}_{i*}\right\|$, define

$$S_i = \left\{ j \in C_i \; : \; \left\|\widetilde{U}_{j*} - \overline{U}_{j*}\right\| \geq \delta_i/2 \right\}, \qquad then,$$

$$\sum_{i=1}^{k} |S_i|\,\delta_i^2 \leq 4(4+2\xi)\left\|U - \overline{U}\right\|_F^2. \tag{60}$$

Moreover, if

$$(16+8\xi)\left\|U - \overline{U}\right\|_F^2 / \delta_i^2 < n_i \qquad \forall i \in [k], \tag{61}$$

then there exists a $k \times k$ permutation matrix π such that $\widetilde{\Theta}_G = \overline{\Theta}_G\pi$, where $G = \cup_{i=1}^{k}(C_i \backslash S_i)$.

Combining Lemma 35 with the perturbation results of Theorem 29 and Theorem 34, we readily arrive at mis-clustering error bounds for SPONGE$_{sym}$.

Theorem 36 (Mis-clustering error for SPONGE$_{sym}$) *Under the notation and assumptions of Theorem 29, let $(\widetilde{\Theta}, \widetilde{X}) \in \mathbb{M}_{n \times k} \times \mathbb{R}^{k \times k}$ be a $(1+\xi)$-approximate solution to the k-means problem $\min_{\Theta \in \mathbb{M}_{n \times k}, X \in \mathbb{R}^{k \times k}} \|\Theta X - G_k(T)\|_F^2$. Denoting*

$$S_i = \left\{ j \in C_i \; : \; \left\|(\widetilde{\Theta}\widetilde{X})_{j*} - (\Theta(C^-)^{-1/2}RO)_{j*}\right\| \geq \frac{1}{2\sqrt{n_i(\tau^+ + \frac{2}{1-l})}} \right\},$$

it holds with probability at least $1 - 2\varepsilon$ that

$$\sum_{i=1}^{k} \frac{|S_i|}{n_i} \leq \delta^2(64+32\xi)k\left(\tau^+ + \frac{2}{1-l}\right)\left(\frac{(\tau^+)^3 + 1}{(\tau^+)^4}\right).$$

In particular, if δ satisfies

$$\delta < \frac{(\tau^+)^2}{\sqrt{(64+32\xi)k(\tau^+ + \frac{2}{1-l})((\tau^+)^3 + 1)}},$$

then there exists a $k \times k$ permutation matrix π such that $\widetilde{\Theta}_G = \hat{\Theta}_G\pi$, where $G = \cup_{i=1}^{k}(C_i \backslash S_i)$.

*   In the sparse regime, the above statement holds under the notation and assumptions of Theorem 34 with $G_k(T)$ replaced with $G_k(T_{\gamma^+,\gamma^-})$, and with probability at least $1 - 2e^{-r}$.*

Proof Since $G_k(T) - G_k(\overline{T})O$ has rank at most $2k$, we obtain from Theorem 29 that

$$\left\| G_k(T) - G_k(\overline{T})O \right\|_F \leq \sqrt{2k} \left\| G_k(T) - G_k(\overline{T})O \right\| \leq \delta\sqrt{2k} \left(\frac{(\tau^+)^{3/2} + 1}{(\tau^+)^2} \right). \qquad (62)$$

We now use Lemma 35 with $U = G_k(T)$ and $\overline{U} = G_k(\overline{T})O$. It follows from (44) and Lemma 17 that $G_k(\overline{T}) = \Theta(C^-)^{-1/2}R = \Theta\Delta\,\Delta^{-1}(C^-)^{-1/2}R$ where $\Delta = \mathrm{diag}(\sqrt{n_1}, \ldots, \sqrt{n_k})$. Denoting $\overline{X} = \Delta^{-1}(C^-)^{-1/2}RO$, we can write $G_k(\overline{T})O = \hat\Theta\overline{X}$, where $\hat\Theta \in \mathbb{M}_{n\times k}$ is the ground truth membership matrix, and for each $i \neq i' \in [k]$, it holds true that

$$\left\| \overline{X}_{i*} - \overline{X}_{i'*} \right\| \geq \lambda_{\min}((C^-)^{-1/2})\sqrt{1/n_i + 1/n_{i'}} \geq \frac{1}{\sqrt{\lambda_{\max}(C^-)n_i}}\,.$$

From (18), one can verify using Weyl's inequality that

$$\lambda_{\max}(C^-) \leq 1 + \tau^+ + \max_i \frac{p}{d_i^-}(\eta_i + s_i n(1 - 2\eta)) \leq \tau^+ + \frac{2}{1 - l},$$

where the last inequality holds if $n \geq \frac{2\eta}{(1-l)(1-\eta)}$. The above considerations imply that $\delta_i = \frac{1}{\sqrt{n_i(\tau^+ + \frac{2}{1-l})}}$. Now with S_i as defined in the statement, we obtain from (60) and (62) that

$$\sum_{i=1}^k |S_i|\,\delta_i^2 = \frac{1}{\tau^+ + \frac{2}{1-l}} \sum_{i=1}^k \frac{|S_i|}{n_i} \leq \delta^2(32 + 16\xi)k\frac{((\tau^+)^{3/2} + 1)^2}{(\tau^+)^4} \leq \delta^2(64 + 32\xi)k\left(\frac{(\tau^+)^3 + 1}{(\tau^+)^4} \right),$$

where the last inequality uses $(a + b)^2 \leq 2(a^2 + b^2)$ for $a, b \geq 0$. This yields the first part of the Theorem.

For the second part, we need to ensure (61) holds. Using (62) and the expression for δ_i, it is easy to verify that (61) holds for the stated condition on δ.

Finally, the statement for the sparse regime readily follows in an analogous manner (replacing $G_k(T)$ with $G_k(T_{\gamma^+, \gamma^-})$), by following the same steps as above. $\blacksquare$

5. Concentration results for the symmetric Signed Laplacian

This section contains proofs of the main results for the symmetric Signed Laplacian, in both the dense regime $p \gtrsim \frac{\ln n}{n}$ and the sparse regime $p \gtrsim \frac{1}{n}$. Before proceeding with an overview of the main steps, for ease of reference, we summarize in the Table below the notation specific to this section.

Notation	Description
$\overline{L_{sym}}$	symmetric Signed Laplacian
$\mathcal{L}_{sym}$	population Signed Laplacian
L_γ	regularized Laplacian
$\mathcal{L}_\gamma$	population regularized Laplacian
$\gamma^+, \gamma^- > 0$	regularization parameters
$\gamma = \gamma^+ + \gamma^-$	
$\overline{\alpha} = 1 + \frac{p}{d}(1 - 2\eta)$	
$\overline{d} = p(n-1)$	expected signed degree
$\rho = \frac{n_{min}}{n_{max}} = \frac{s}{l}$	aspect ratio

The proof of Theorem 8 is built on the following steps. In Section 5.1, we compute the eigen-decomposition of the Signed Laplacian of the expected graph $\mathcal{L}_{sym}$. Then in Section 5.2, we show $\overline{L_{sym}}$ and $\mathcal{L}_{sym}$ are "close", and obtain an upper bound on the error $\left\|\overline{L_{sym}} - \mathcal{L}_{sym}\right\|$. Finally, in Section 5.3, we use the Davis-Kahan theorem (see Theorem 56) to bound the error between the subspaces $V_{k-1}(\overline{L_{sym}})$ and $V_{k-1}(\mathcal{L}_{sym})$. To prove Theorem 11, in Section 5.4, we first use a decomposition of the set of edges $[n] \times [n]$ and characterize the behaviour of the regularized Signed Laplacian on each subset. This leads in Section 5.5 to the error bounds of Theorem 11. Finally, the proof of Theorem 13, that bound the error on the eigenspace, relies on the same arguments as Theorem 8 and can be found in Section 5.6. Similarly to the approach for SPONGE$_{sym}$, the mis-clustering error is obtained using a $(1+\xi)$-approximate solution of the k-means problem applied to the rows of $V_{k-1}(\overline{L_{sym}})$ (resp. $V_{k-1}(L_\gamma)$). This solution contains, in particular, an estimated membership matrix $\widetilde{\Theta}$. The bound on the mis-clustering error of the algorithm given in Theorem 16 is derived using Lemma 35 (Lemma 5.3 of Lei and Rinaldo (2015)), in Section 5.7.

5.1 Analysis of the expected Signed Laplacian

In this section, we compute the eigen-decomposition of the matrix $\mathcal{L}_{sym}$. In particular, we aim at proving a lower bound on the eigengap between the $(k-1)^{th}$ and k^{th} smallest eigenvalues. For equal-size clusters, there is an explicit expression for this eigengap.

5.1.1 MATRIX DECOMPOSITION

Lemma 37 *Let $\Theta \in \mathbb{R}^{n \times k}$ denote the normalized membership matrix in the SSBM. Let $V^\perp \in \mathbb{R}^{n \times (n-k)}$ be a matrix whose columns are any orthonormal base of the subspace orthogonal to $\mathcal{R}(\Theta)$. The Signed Laplacian of the expected graph has the following decomposition*

$$\mathcal{L}_{sym} = [\Theta \; V^\perp] \begin{pmatrix} \overline{C} & 0 \\ 0 & \overline{\alpha}I_{n-k} \end{pmatrix} \begin{bmatrix} \Theta^T \\ (V^\perp)^T \end{bmatrix}, \tag{63}$$

with $\overline{C} = \overline{\alpha}I_k - \overline{B}$, $\overline{\alpha} = 1 + \frac{p}{d}(1 - 2\eta)$ and $\overline{B}$ is a $k \times k$ matrix such that

$$\overline{B}_{ii'} = \begin{cases} \frac{n_i p}{\overline{d}}(1 - 2\eta); & if \; i = i' \\ -\frac{\sqrt{n_i n_{i'}} p}{\overline{d}}(1 - 2\eta); & if \; i \neq i'. \end{cases} \tag{64}$$

Proof On one hand, we recall from Section 2.3 that the expected degree matrix is a scaled identity matrix $\mathbb{E}[\overline{D}] = \overline{d}I_n$, with $\overline{d} = p(n-1)$. Thus, any vector $v \in \mathbb{R}^n$ is an eigenvector

40

of $\mathbb{E}[\overline{D}]$ with corresponding eigenvalue $\overline{d}$, and it holds true that

$$\mathbb{E}[\overline{D}]^{-1/2} = \frac{1}{\sqrt{\overline{d}}}I_n = \frac{1}{\sqrt{\overline{d}}}[\Theta\,(V^{\perp})]\,I_n\begin{bmatrix}\Theta^T\\(V^{\perp})^T\end{bmatrix}. \tag{65}$$

On the other hand, the signed adjacency matrix can be written in the form

$$\mathbb{E}[A] = \mathbb{E}[A^+] - \mathbb{E}[A^-] = M - p(1-2\eta)I_n, \tag{66}$$

where

$$M = \begin{bmatrix} p(1-2\eta)J_{n_1} & -p(1-2\eta)J_{n_1\times n_2} & \cdots & -p(1-2\eta)J_{n_1\times n_k} \\ -p(1-2\eta)J_{n_2\times n_1} & p(1-2\eta)J_{n_2} & \cdots & -p(1-2\eta)J_{n_2\times n_k} \\ \vdots & \vdots & \ddots & \vdots \\ -p(1-2\eta)J_{n_k\times n_1} & \cdots & \cdots & p(1-2\eta)J_{n_k} \end{bmatrix}.$$

The matrix M has the following decomposition

$$M = \overline{d}\Theta\overline{B}\Theta^T = \overline{d}[\Theta\,V^{\perp}]\begin{pmatrix}\overline{B} & 0\\0 & 0\end{pmatrix}\begin{bmatrix}\Theta^T\\(V^{\perp})^T\end{bmatrix},$$

with $\overline{B}$ defined in (64). Thus, combining (65) and (66), we arrive at

$$\mathbb{E}[\overline{D}]^{-1/2}\mathbb{E}[A]\mathbb{E}[\overline{D}]^{-1/2} = \frac{1}{\overline{d}}M - p(1-2\eta)\frac{1}{\overline{d}}I_n = [\Theta\,V^{\perp}]\begin{pmatrix}\overline{B} & 0\\0 & 0\end{pmatrix}\begin{bmatrix}\Theta^T\\(V^{\perp})^T\end{bmatrix} - (1-2\eta)\frac{p}{\overline{d}}I_n.$$

This finally leads to the decomposition of $\mathcal{L}_{sym}$

$$\mathcal{L}_{sym} = I - \mathbb{E}[\overline{D}]^{-1/2}\mathbb{E}[A]\mathbb{E}[\overline{D}]^{-1/2} = [\Theta\,V^{\perp}]\begin{pmatrix}\overline{C} & 0\\0 & \overline{\alpha}I_{n-k}\end{pmatrix}\begin{bmatrix}\Theta^T\\(V^{\perp})^T\end{bmatrix},$$

with $\overline{C} = \overline{\alpha}I_k - \overline{B}$ and $\overline{\alpha} = 1 + p(1-2\eta)$. ∎

We can infer from Lemma 37 that the spectrum of $\mathcal{L}_{sym}$ is the union of the spectrum of the matrix $\overline{C} \in \mathbb{R}^{k\times k}$ and $\{\overline{\alpha}\}$. Moreover, denoting $u = \frac{1}{\sqrt{\overline{d}}}(\sqrt{n_1},\ldots,\sqrt{n_k})^T$, we have $\overline{C} = p(1-2\eta)uu^T + \mathrm{diag}\left(1 + \frac{p}{\overline{d}}(1-2\eta)(1-2n_i)\right)$. For a SSBM with equal-size clusters, we are able to find explicit expressions for the eigenvalues of $\overline{C}$.

5.1.2 Spectrum of the Signed Laplacian: equal-size clusters

In this section, we assume that the clusters in the SSBM have equal sizes $n_1 = n_2 = \cdots = n_k = \frac{n}{k}$. In this case,

$$\frac{1}{\sqrt{\overline{d}}}(\sqrt{n_1},\ldots,\sqrt{n_k})^T = \sqrt{\frac{n}{\overline{d}}}\chi_1,$$

and denoting by $\overline{C}_e$ the matrix $\overline{C}$ in this setting of equal clusters, we may write

$$\overline{C}_e = \frac{np}{\overline{d}}(1-2\eta)\chi_1\chi_1^T + \left(1 + \frac{p}{\overline{d}}(1-2\eta)\left(1 - 2\frac{n}{k}\right)\right)I_k. \tag{67}$$

41

Hence, the spectrum of $\overline{C}_e$ contains only two different values. The largest one has multiplicity 1, and χ_1 is the corresponding largest eigenvector. The $k-1$ remaining eigenvalues are all equal. In fact, we have

$$\lambda_i(\overline{C}_e) = \begin{cases} 1 + \frac{p}{\bar{d}}(1 - 2\eta)(n + 1 - 2\frac{n}{k}); & \text{if } i = 1 \\ 1 + \frac{p}{\bar{d}}(1 - 2\eta)\left(1 - 2\frac{n}{k}\right); & \text{if } 2 \le i \le k. \end{cases}$$

One can easily check that these eigenvalues are positive, and that the following inequality holds true

$$\lambda_1(\overline{C}_e) = \overline{\alpha} + \frac{p}{\bar{d}}(1 - 2\eta)(n - 2\frac{n}{k}) \ge \overline{\alpha} > \overline{\alpha} - 2\frac{n}{k}(1 - 2\eta) = \lambda_2(\overline{C}_e).$$

We finally have

$$\lambda_j(\mathcal{L}_{sym}) = \begin{cases} 1 + \frac{p}{\bar{d}}(1 - 2\eta)(n + 1 - 2\frac{n}{k}); & \text{if } j = 1 \\ \overline{\alpha}; & \text{if } 2 \le j \le n - k + 1 \\ \lambda_2(\overline{C}_e); & \text{if } n - k + 2 \le j \le n. \end{cases}$$

Note that for $k = 2$, $\lambda_1(\overline{C}_e) = \overline{\alpha}$ and the spectrum of $\mathcal{L}_{sym}$ contains only two values $\{\overline{\alpha}, \lambda_2(\overline{C}_e)\}$. For $k > 2$, $\lambda_1(\mathcal{L}_{sym}) > \overline{\alpha} > \lambda_2(\overline{C}_e)$. Writing the spectral decomposition

$$\overline{C}_e = R\,\Lambda\,R^T = [R_{k-1}\ \gamma_1]\,\Lambda\,\begin{bmatrix} R_{k-1}^T \\ \gamma_1^T \end{bmatrix},$$

with $\gamma_1 = \chi_1$ and $R_{k-1} \in \mathbb{R}^{k \times (k-1)}$ being the matrix of eigenvectors associated to $\lambda_2(\overline{C}_e)$, we conclude that $V_{k-1}(\mathcal{L}_{sym}) = \Theta R_{k-1}$. In fact, since Θ has k distinct rows and R is a unitary matrix, ΘR also has k distinct rows. As χ_1 is the all one's vector , ΘR_{k-1} has k distinct rows as well. These observations are summarized in the following lemma and lead to the expression of the eigengap.

Lemma 38 (Eigengap for equal-size clusters) *For the SSBM with $k \ge 2$ clusters of equal-size $\frac{n}{k}$, we have that $V_{k-1}(\mathcal{L}_{sym}) = \Theta R_{k-1} \in \mathbb{R}^{n \times (k-1)}$, where R_{k-1} corresponds to the $(k-1)$ smallest eigenvectors of $\overline{C}_e$. Moreover, with the eigengap defined as*

$$\lambda_{gap} := \lambda_{n-k+1}(\mathcal{L}_{sym}) - \lambda_{n-k+2}(\mathcal{L}_{sym}),$$

it holds true that

$$\lambda_{gap} = \overline{\alpha} - \lambda_2(\overline{C}_e) = \frac{2np}{k\bar{d}}(1 - 2\eta) \ge \frac{2}{k}(1 - 2\eta). \tag{68}$$

5.1.3 Non-equal-size clusters

In the general setting of non-equal-size clusters, it is difficult to obtain an explicit expression of the spectrum of $\mathcal{L}_{sym}$. Thus, using a perturbation method, we establish a lower bound on the eigengap, provided that the aspect ratio ρ is close to 1. Recall that

$$\overline{C} = p(1 - 2\eta)uu^T + \text{diag}\left(1 + \frac{p}{\bar{d}}(1 - 2\eta)(1 - 2n_i)\right)$$

$$= p(1 - 2\eta)uu^T - 2p(1 - 2\eta)\text{diag}(u_i^2)_{i=1}^n + \text{diag}\left(1 + \frac{p}{\bar{d}}(1 - 2\eta)\right). \tag{69}$$

We note that this matrix is of the form $\Lambda + vv^T$, with Λ being a diagonal matrix and $v \in \mathbb{R}^k$ a vector. Using again the spectral decomposition

$$\overline{C} = R \, \Lambda \, R^T = [R_{k-1} \; \gamma_1] \, \Lambda \, \begin{bmatrix} R_{k-1}^T \\ \gamma_1^T \end{bmatrix}, \tag{70}$$

where γ_1 is the largest eigenvector and $R_{k-1} \in \mathbb{R}^{k \times (k-1)}$ contains the smallest $(k-1)$ eigenvectors of $\overline{C}$, we would like to ensure that the smallest $(k-1)$ eigenvectors of $\mathcal{L}_{sym}$ are related to the $(k-1)$ eigenvectors of $\overline{C}$ in the following way $V_{k-1}(\mathcal{L}_{sym}) = \Theta R_{k-1}$. Note that γ_1 is not necessarily the all one's vector, and ΘR_{k-1} has at least $k-1$ distinct rows. To this end, we will like to ensure that

$$\{\lambda_2(\overline{C}), \ldots, \lambda_{k-1}(\overline{C}), \lambda_k(\overline{C})\} = \{\lambda_{n-k+2}(\mathcal{L}_{sym}), \ldots, \lambda_{n-1}(\mathcal{L}_{sym}), \lambda_n(\mathcal{L}_{sym})\}. \tag{71}$$

From Weyl's inequality (see Theorem 55), we know that

$$|\lambda_i(\overline{C}_e) - \lambda_i(\overline{C})| \leq \|\overline{C} - \overline{C}_e\| \quad \forall i = 1, \ldots k,$$

which in particular implies

$$\lambda_2(\overline{C}) \leq \lambda_2(\overline{C}_e) + \|\overline{C} - \overline{C}_e\|, \qquad \lambda_1(\overline{C}) \geq \lambda_1(\overline{C}_e) - \|\overline{C} - \overline{C}_e\|.$$

Moreover, $\lambda_1(\overline{C}) = \overline{\alpha}$ when $k = 2$, and $\lambda_1(\overline{C}) > \overline{\alpha}$ when $k > 2$. Thus, for Condition 71 to be true, it suffices to ensure

$$\lambda_2(\overline{C}_e) + \|\overline{C} - \overline{C}_e\| < \overline{\alpha} + \|\overline{C} - \overline{C}_e\| \iff \|\overline{C} - \overline{C}_e\| < \frac{\overline{\alpha} - \lambda_2(\overline{C}_e)}{2}$$

$$\iff \|\overline{C} - \overline{C}_e\| < \frac{np}{k\overline{d}}(1 - 2\eta),$$

using (68). In this case, we indeed have that $V_{k-1}(\mathcal{L}_{sym}) = \Theta R_{k-1}$. As it will be convenient later, we will ensure a slightly stronger condition, i.e.

$$\|\overline{C} - \overline{C}_e\| < \frac{\overline{\alpha} - \lambda_2(\overline{C}_e)}{4} = \frac{np}{2k\overline{d}}(1 - 2\eta). \tag{72}$$

Now we compute the error $\|\overline{C} - \overline{C}_e\|$. We recall that $\|u\| = \sqrt{\frac{n}{d}}$ and denote $D_u =: \frac{1}{\|u\|^2} \text{diag}(u_i^2)_{i=1}^n$, then (69) becomes

$$\overline{C} = \overline{\alpha} I_k + \frac{np}{\overline{d}}(1 - 2\eta) \left(\frac{u}{\|u\|}\right) \left(\frac{u}{\|u\|}\right)^T - 2\frac{np}{\overline{d}}(1 - 2\eta) D_u.$$

Using (67), we obtain

$$\|\overline{C} - \overline{C}_e\| = \left\| \frac{np}{\overline{d}}(1 - 2\eta) \left(\left(\frac{u}{\|u\|}\right) \left(\frac{u}{\|u\|}\right)^T - \chi_1 \chi_1^T \right) - 2\frac{np}{\overline{d}}(1 - 2\eta) \left(D_u - \frac{1}{k} I_n \right) \right\|$$

$$\leq \frac{np}{\overline{d}}(1 - 2\eta) \left\| \left(\frac{u}{\|u\|}\right) \left(\frac{u}{\|u\|}\right)^T - \chi_1 \chi_1^T \right\| + 2\frac{np}{\overline{d}}(1 - 2\eta) \left\| D_u - \frac{1}{k} I_n \right\|. \tag{73}$$

43

For the first term on the RHS, we have

$$\left\| \left(\frac{u}{\|u\|} \right) \left(\frac{u}{\|u\|} \right)^T - \chi_1 \chi_1^T \right\| \le 2 \left\| \frac{u}{\|u\|} - \chi_1 \right\| \le 2\sqrt{k} \max_i \left| \sqrt{\frac{n_i}{n}} - \sqrt{\frac{1}{k}} \right|$$

$$\le 2\sqrt{k}(\sqrt{l} - \sqrt{s}) \le 2\sqrt{k}(1 - \sqrt{\rho}), \tag{74}$$

while for the second term on the RHS, we have

$$\left\| D_u - \frac{1}{k} I_n \right\| = \max_i \left| \sqrt{\frac{n_i}{n}} - \sqrt{\frac{1}{k}} \right| \le 1 - \sqrt{\rho}. \tag{75}$$

By combining (74) and (75) into (73), we arrive at

$$\|\overline{C} - \overline{C}_e\| \le \frac{np}{\overline{d}}(1 - 2\eta)\sqrt{k}(1 - \sqrt{\rho}) + \frac{2np}{\overline{d}}(1 - 2\eta)(1 - \sqrt{\rho})$$

$$\le \frac{np}{\overline{d}}(1 - 2\eta)(1 - \sqrt{\rho})\left(\sqrt{k} + 2\right)$$

$$\le 2(2 + \sqrt{k})(1 - 2\eta)(1 - \sqrt{\rho}),$$

using that $\frac{np}{\overline{d}} = \frac{n}{n-1} \le 2$. Now since $\frac{np}{2k\overline{d}} \ge \frac{1-2\eta}{2k}$ and from Condition 72, it suffices that ρ satisfies

$$2(2 + \sqrt{k})(1 - 2\eta)(1 - \sqrt{\rho}) \le \frac{1 - 2\eta}{2k} \iff 1 - \sqrt{\rho} \le \frac{1}{4k(2 + \sqrt{k})}.$$

Finally, we can compute

$$\lambda_{gap} := \lambda_{n-k+1}(\mathcal{L}_{sym}) - \lambda_{n-k+2}(\mathcal{L}_{sym})$$

$$\ge \overline{\alpha} - \|\overline{C} - \overline{C}_e\| - (\lambda_2(\overline{C}_e) + \|\overline{C} - \overline{C}_e\|)$$

$$\ge \overline{\alpha} - \lambda_2(\overline{C}_e) - 2\|\overline{C} - \overline{C}_e\|$$

$$\ge \frac{\overline{\alpha} - \lambda_2(\overline{C}_e)}{2} = \frac{np}{k\overline{d}}(1 - 2\eta) \ge \frac{1 - 2\eta}{k}.$$

Hence we arrive at the following lemma.

Lemma 39 (General lower-bound on the eigengap) *For a SSBM with $k \ge 2$ clusters of general sizes $(n_1, \dots, n_k)$ and aspect ratio ρ satisfying*

$$\sqrt{\rho} > 1 - \frac{1}{4k(2 + \sqrt{k})},$$

it holds true that $V_{k-1}(\mathcal{L}_{sym}) = \Theta R_{k-1}$, where $R_{k-1} \in \mathbb{R}^{k \times k-1}$ corresponds to the $(k-1)$ smallest eigenvectors of $\overline{C}$. Furthermore, we can lower-bound the spectral gap λ_{gap} as

$$\lambda_{gap} := \lambda_{n-k+1}(\mathcal{L}_{sym}) - \lambda_{n-k+2}(\mathcal{L}_{sym}) \ge \frac{1 - 2\eta}{k}.$$

We will now show that $\overline{L_{sym}}$ concentrates around the population Laplacian $\mathcal{L}_{sym}$, provided the graph is dense enough.

5.2 Concentration of the Signed Laplacian in the dense regime

In the moderately dense regime where $p \gtrsim \frac{\ln n}{n}$, the adjacency and the degree matrices concentrate towards their expected counterparts, as n increases. This can be established using standard concentration tools from the literature.

Lemma 40 *We have the following concentration inequalities for A and $\overline{D}$*

1. $\forall 0 < \varepsilon \leq \frac{1}{2}, \exists c_\varepsilon > 0,$

$$\mathbb{P}\left(\|A - \mathbb{E}[A]\| \leq ((1+\varepsilon)4\sqrt{2} + 2)\sqrt{np} \right) \geq 1 - n \exp\left(-\frac{np}{c_\varepsilon} \right).$$

   In particular, there exists a universal constant $c > 0$ such that

$$\mathbb{P}\left(\|A - \mathbb{E}[A]\| \leq 12\sqrt{np} \right) \geq 1 - n \exp\left(-\frac{np}{c} \right).$$

2. *If $p > 12\frac{\ln n}{n}$,*

$$\mathbb{P}\left(\|\overline{D} - \mathbb{E}[\overline{D}]\| \leq \sqrt{3np \ln n} \right) \geq 1 - \frac{2}{n}.$$

Proof For the first statement, we recall that A is a symmetric matrix, with $A_{jj'} = 0$ and with independent entries above the diagonal $(A_{jj'})_{j<j'}$. We denote $Z_{jj'} = A_{jj'} - \mathbb{E}[A_{jj'}]$. If j, j' lie in the same cluster,

$$Z_{jj'} = \begin{cases} 1 - p(1-2\eta) & ; \quad \text{w. p. } p(1-\eta) \\ -1 - p(1-2\eta) & ; \quad \text{w. p. } p\eta \\ -p(1-2\eta) & ; \quad \text{w. p. } 1-p \end{cases}.$$

If j, j' lie in different clusters,

$$Z_{jj'} = \begin{cases} 1 + p(1-2\eta) & ; \quad \text{w. p. } p\eta \\ -1 + p(1-2\eta) & ; \quad \text{w. p. } p(1-\eta) \\ p(1-2\eta) & ; \quad \text{w. p. } 1-p \end{cases}.$$

One can easily check that in both cases, it holds true that

$$\mathbb{E}[(Z_{jj'})^2] = p\left[(1-\eta)(1-p(1-2\eta))^2 + \eta(1+p(1-2\eta))^2 + p(1-2\eta)^2)(1-p) \right]$$
$$\leq p(1 + \eta(1+p)^2 + p) \leq 4p.$$

Thus we can conclude that for each $j \in [n]$, the following holds

$$\sqrt{\sum_{j'=1}^{n} \mathbb{E}[(Z_{jj'})^2]} \leq \sqrt{4np} = 2\sqrt{np}.$$

Hence, $\widetilde{\sigma} := \max_j \sqrt{\sum_{j'=1}^{n} \mathbb{E}[(Z_{jj'})^2]} \leq 2\sqrt{np}$. Moreover, $\widetilde{\sigma}_* := \max_{j,j'} \left\| Z_{jj'}^+ \right\|_\infty = 1 + p(1-2\eta) \leq 2$. Therefore, we can apply the concentration bound for the norm of symmetric

45

matrices by Bandeira and van Handel (2016, Corollary 3.12, Remark 3.13) (recalled in Appendix 53) with $t = 2\sqrt{np}$, in order to bound $\|Z\| = \|A - \mathbb{E}[A]\|$. For any given $0 < \varepsilon \leq 1/2$, we have that

$$\|A - \mathbb{E}[A]\| \leq \left((1+\varepsilon)4\sqrt{2} + 2\right)\sqrt{np},$$

with probability at least $1 - n\exp\left(\frac{-pn}{c_\varepsilon}\right)$, where c_ε only depends on ε.

For the second statement, we apply Chernoff's bound (see Appendix A.1) to the random variables $\overline{D}_{jj} = \sum_{j'=1}^n \left(A_{jj'}^+ + A_{jj'}^-\right)$, where we note that $(A_{jj'}^+ + A_{jj'}^-)_{j'=1}^n$ are independent Bernoulli random variables with mean p. Hence, $\mathbb{E}[D_{jj}] = \overline{d} = p(n-1)$. Let $\delta = \sqrt{\frac{6\ln n}{\overline{d}}}$ and assuming that $p > 12\frac{\ln n}{n}$ (so that $\delta < 1$), we obtain

$$\mathbb{P}\left[\left|\overline{D}_{jj} - \overline{d}\right| \geq \sqrt{6\overline{d}\ln n}\right] \leq \mathbb{P}\left[\left|\overline{D}_{jj} - \overline{d}\right| \geq \sqrt{3np\ln n}\right] \leq 2\exp\left(-2\ln n\right) = \frac{2}{n^2},$$

using that $n - 1 \geq \frac{n}{2}$. Applying the union bound, we finally obtain that

$$\mathbb{P}\left(\|\overline{D} - \mathbb{E}[\overline{D}]\| \geq \sqrt{3np\ln n}\right) \leq \frac{2}{n}.$$

∎

Lemma 41 *If $\|A - \mathbb{E}[A]\| \leq \Delta_A$, $\|\overline{D} - \mathbb{E}[\overline{D}]\| \leq \Delta_D$ and $p > 12\frac{\ln n}{n}$, then with probability at least $1 - \frac{2}{n}$, it follows that*

$$\|\overline{L_{sym}} - \mathcal{L}_{sym}\| \leq \frac{\Delta_A}{\overline{d}} + 2\frac{\Delta_D}{\overline{d}} + \frac{\Delta_D^2}{\overline{d}^2}.$$

Proof We first note that using the proof of Lemma 40, with probability at least $1 - \frac{2}{n}$, we have that $\left|\overline{D}_{jj} - \overline{d}\right| \leq \delta\overline{d}, \forall j \in [n]$, with $\delta < 1$. Consequently,

$$\|(\mathbb{E}[\overline{D}])^{-1/2}\overline{D}^{1/2} - I\| = \max_j \left|\sqrt{\frac{\overline{D}_{jj}}{\overline{d}}} - 1\right| \leq \max_j \frac{|\overline{D}_{jj} - \overline{d}|}{\overline{d}} = \frac{\Delta_D}{\overline{d}},$$

since $|\sqrt{x} - 1| \leq |x - 1|$ for $0 < x < 1$. We now apply the first inequality of Proposition 59 with $A^- = \overline{D}, A^+ = A, B^- = \mathbb{E}\left[\overline{D}\right], B^+ = \mathbb{E}[A]$. We obtain

$$\|\overline{L_{sym}} - \mathcal{L}_{sym}\| \leq \frac{\Delta_A}{\overline{d}} + \left\|\overline{D}^{-1}\right\| \|A\| \left(\frac{\Delta_D^2}{\overline{d}^2} + 2\frac{\Delta_D}{\overline{d}}\right).$$

It remains to prove that $\left\|\overline{D}^{-1}\right\| \|A\| \leq 1$. It holds since $\overline{D}$ is a diagonal matrix, thus $\left\|\overline{D}^{-1}\right\| \|A\| = \left\|\overline{D}^{-1}A\right\|$ and similarly to Lemma 61, it is straightforward to prove that $I - \left\|\overline{D}^{-1}A\right\| \leq 2$, therefore $\left\|\overline{D}^{-1}A\right\| \leq 1$.

∎

Combining the results from Lemma 40 and Lemma 41, we arrive at the concentration bound for $\|\overline{L_{sym}} - \mathcal{L}_{sym}\|$.

Lemma 42 *Under the assumptions of Theorem 8, if $n \geq 10$, then with probability at least $1 - n\exp(-\frac{np}{c_\epsilon}) - \frac{2}{n}$ there exists a universal constant $0 < C < 43$ such that*

$$\|\overline{L_{sym}} - \mathcal{L}_{sym}\| \leq C\sqrt{\frac{\ln n}{np}}.$$

Proof If $p \geq \frac{12\ln n}{n}$, the bounds in Lemma 40 hold simultaneously with probability at least $1 - n\exp(-\frac{np}{c}) - \frac{2}{n}$ and we have, with the notations of Lemma 41, $\Delta_A \leq 12\sqrt{np}$ and $\Delta_D \leq \sqrt{3np\ln n}$. Applying Lemma 41, we then obtain

$$\|\overline{L_{sym}} - \mathcal{L}_{sym}\| \leq \frac{12\sqrt{np}}{\overline{d}} + 2\frac{\sqrt{3np\ln n}}{\overline{d}} + \frac{3np\ln n}{\overline{d}^2} \leq \frac{24}{\sqrt{np}} + 4\sqrt{3}\sqrt{\frac{\ln n}{np}} + \frac{12\ln n}{np}.$$

If $n \geq 10$, $\ln n \geq 1$ and $\sqrt{\frac{\ln n}{np}} \geq \frac{1}{\sqrt{np}}$. Moreover, since $p \geq 12\frac{\ln n}{n}$, then $\frac{\ln n}{np} \leq \frac{1}{12} < 1$ and $\sqrt{\frac{\ln n}{np}} \geq \frac{\ln n}{np}$. We finally obtain

$$\|\overline{L_{sym}} - \mathcal{L}_{sym}\| \leq (24 + 4\sqrt{3} + 12)\sqrt{\frac{\ln n}{np}} = C\sqrt{\frac{\ln n}{np}},$$

with $C = 24 + 4\sqrt{3} + 12 \leq 43$. ∎

5.3 Proof of Theorem 8

The proof of this theorem relies on the Davis-Kahan theorem. Using Weyl's inequality (see Theorem 55) and Lemma 42, we obtain for all $1 \leq j \leq n$,

$$|\lambda_j(\overline{L_{sym}}) - \lambda_j(\mathcal{L}_{sym})| \leq C\left(\frac{\ln n}{np}\right)^{1/2}.$$

In particular, for the k-th smallest eigenvalue,

$$\lambda_{n-k+1}(\overline{L_{sym}}) \geq \lambda_{n-k+1}(\mathcal{L}_{sym}) - C\left(\frac{\ln n}{np}\right)^{1/2},$$

$$\lambda_{n-k+1}(\overline{L_{sym}}) - \lambda_{n-k+2}(\mathcal{L}_{sym}) \geq \lambda_{n-k+1}(\mathcal{L}_{sym}) - \lambda_{n-k+2}(\mathcal{L}_{sym}) - C\left(\frac{\ln n}{np}\right)^{1/2}$$

$$= \lambda_{gap} - C\left(\frac{\ln n}{np}\right)^{1/2}.$$

For $\delta \in (0,1)$, we will like to ensure that

$$\lambda_{gap} - C\left(\frac{\ln n}{np}\right)^{1/2} > \lambda_{gap}\left(1 - \frac{\delta}{2}\right). \tag{76}$$

From Lemma 39, if $\sqrt{\rho} > 1 - \frac{1}{4k(2+\sqrt{k})}$, then $\lambda_{gap} \geq \frac{1}{k}(1 - 2\eta)$. Then for the previous condition (76) to hold, it is sufficient that

$$C\left(\frac{\ln n}{np}\right)^{1/2} < \frac{\delta}{2k}(1 - 2\eta) \iff p > \left(\frac{2Ck}{\delta(1-2\eta)}\right)^2 \frac{\ln n}{n} = C(k,\eta,\delta)\frac{\ln n}{n}, \quad (77)$$

with $C(k,\eta,\delta) = \left(\frac{2Ck}{\delta(1-2\eta)}\right)^2$. We note that since $C(k,\eta,\delta) \geq C \geq 12$, hence (77) implies that $p > 12\frac{\ln n}{n}$.

With this condition, we now apply the Davis-Kahan theorem (Theorem 56)

$$\|(I - V_{k-1}(\overline{L_{sym}})V_{k-1}(\overline{L_{sym}})^T)V_{k-1}(\mathcal{L}_{sym})\| \leq \frac{\left\|\overline{L_{sym}} - \mathcal{L}_{sym}\right\|}{\lambda_{gap} - C\left(\frac{\ln n}{np}\right)^{1/2}}$$

$$\leq \frac{\delta\lambda_{gap}/2}{\lambda_{gap}(1 - \delta/2)} = \frac{\delta/2}{1 - \delta/2} \leq \delta.$$

Using Proposition 57, there then exists an orthogonal matrix $O \in \mathbb{R}^{(k-1)\times(k-1)}$ so that

$$\|V_{k-1}(\overline{L_{sym}}) - \Theta R_{k-1}O\| \leq 2\delta.$$

5.4 Properties of the regularized Laplacian in the sparse regime

The analysis of the signed regularized Laplacian differs from the one of unsigned regularized Laplacian. In particular, Lemma 30 cannot be directly applied, since the trimming approach of the adjacency matrix for unsigned graphs is not available in this case. However, we will also use arguments of Le et al. (2015) and Le et al. (2017) for unsigned directed adjacency matrices in the inhomogeneous Erdős-Rényi model $G(n, (p_{jj'})_{j,j'})$. More precisely, in Section 5.4.1, we will prove that the adjacency matrix concentrates on a large subset of edges called the *core*. On this subset, the unregularized (resp. regularized) Laplacian also concentrates towards the expected matrix $\mathcal{L}_{sym}$ (resp. $\mathcal{L}_\gamma$). In Section 5.4.2, we will show that on the remaining subset of nodes, the norm of the regularized Laplacian is relatively small.

5.4.1 PROPERTIES OF THE SIGNED ADJACENCY AND DEGREE MATRICES

In this section, we adapt the results by Le et al. (2017) for the signed adjacency matrix and the degree matrix in our SSBM. Similarly to Le et al. (2017, Theorem 2.6) (see Theorem 54), the following lemma shows that the set of edges can be decomposed into a large block, and two blocks with respectively few columns and few rows.

Lemma 43 *(Decomposition of the set of edges for the SSBM) Let A be the signed adjacency matrix of a graph sampled from the SSBM. For any $r \geq 1$, with probability at least $1 - 6n^{-r}$, the set of edges $[n] \times [n]$ can be partitioned into three classes $\mathcal{N}, \mathcal{R}$ and $\mathcal{C}$ such that*

1. *the signed adjacency matrix concentrates on $\mathcal{N}$*

$$\|(A - \mathbb{E}A)_{\mathcal{N}}\| \leq Cr^{3/2}\sqrt{\overline{d}(1 - \eta)},$$

   with $C > 1$ a constant;

2. $\mathcal{R}$ (resp. $\mathcal{C}$) intersects at most $4n/\overline{d}$ columns (resp. rows) of $[n] \times [n]$;

3. each row (resp. column) of $A_{\mathcal{R}}$ (resp. $A_{\mathcal{C}}$) has at most $128r$ non-zero entries.

Remark 44 *We underline that this lemma is valid because the unsigned adjacency matrices A^+ and A^- have disjoint support. We do not know if similar results could be obtained for the Signed Stochastic Block Model defined by Mercado et al. (2016).*

Proof We denote $A^{\pm}_{sup}$ (resp. $A^{\pm}_{inf}$) the upper (resp. lower) triangular part of the unsigned adjacency matrices. Using this decomposition, we have

$$A = A^+_{inf} + A^+_{sup} - A^-_{inf} - A^-_{sup}.$$

We note that $A^+_{inf}, A^+_{sup}, A^-_{inf}, A^-_{sup}$ have disjoint supports, and each of them has independent entries. We can hence apply Theorem 54 to each of these matrices, where we note that for each matrix

$$d := n \max_{j,j'} \mathbb{E}[A_{jj'}] = np(1 - \eta) \leq 2\overline{d}(1 - \eta).$$

With probability at least $1 - 2 \times 3n^{-r}$, there exists $\mathcal{N}^{\pm}_{inf}, \mathcal{R}^{\pm}_{inf}, \mathcal{C}^{\pm}_{inf}, \mathcal{N}^{\pm}_{sup}, \mathcal{R}^{\pm}_{sup}, \mathcal{C}^{\pm}_{sup}$ four partitions of $[n] \times [n]$ that have the subsequent properties. For e.g., for A^+_{inf},

- $\|(A^+_{inf} - \mathbb{E}A^+_{inf})_{\mathcal{N}}\| \leq Cr^{3/2}\sqrt{d} \leq Cr^{3/2}\sqrt{2\overline{d}(1 - \eta)}$;

- $\mathcal{R}^+_{inf}$ (resp. $\mathcal{C}^+_{inf}$) intersects at most $n/d \leq n/\overline{d}$ columns (resp. rows) of $[n] \times [n]$;

- each row (resp. column) of $(A^+_{inf})_{\mathcal{R}}$ (resp. $(A^+_{inf})_{\mathcal{C}}$) have at most $32r$ ones.

We note that this decomposition holds simultaneously for $A^{\pm}_{inf}$ and $A^{\pm}_{sup}$. Taking the unions of these subsets,

$$\mathcal{N} = \mathcal{N}^+_{inf} \cup \mathcal{N}^+_{sup} \cup \mathcal{N}^-_{inf} \cup \mathcal{N}^-_{sup},$$

and similarly for $\mathcal{R}$ and $\mathcal{C}$, we have, with the triangle inequality

$$\begin{aligned}
&\|(A - \mathbb{E}A)_{\mathcal{N}}\| \\
&= \|(A^+_{inf} - \mathbb{E}A^+_{inf})_{\mathcal{N}^+_{inf}} + (A^+_{sup} - \mathbb{E}A^+_{sup})_{\mathcal{N}^+_{sup}} - (A^-_{inf} - \mathbb{E}A^-_{inf})_{\mathcal{N}^-_{inf}} \\
&\qquad - (A^-_{sup} - \mathbb{E}A^-_{sup})_{\mathcal{N}^-_{sup}}\| \\
&\leq \|(A^+_{inf} - \mathbb{E}A^+_{inf})_{\mathcal{N}^+_{inf}}\| + \|(A^+_{sup} - \mathbb{E}A^+_{sup})_{\mathcal{N}^+_{sup}}\| + \|(A^-_{inf} - \mathbb{E}A^-_{inf})_{\mathcal{N}^-_{inf}}\| \\
&\qquad + \|(A^-_{sup} - \mathbb{E}A^-_{sup})_{\mathcal{N}^-_{sup}}\| \\
&\leq 4Cr^{3/2}\sqrt{d} \leq C_1 r^{3/2}\sqrt{\overline{d}(1 - \eta)},
\end{aligned}$$

with $C_1 = 4C\sqrt{2}$. Moreover, each row of $\mathcal{R}$ (resp. each column of $\mathcal{C}$) has at most $2 \times 32r$ entries equal to 1 and $2 \times 32r$ entries equal to -1, which means at most $128r$ non-zero entries. Finally $\mathcal{R}$ (resp. $\mathcal{C}$) intersects at most $4n/\overline{d}$ rows (resp. columns) of $[n] \times [n]$. ∎

For the degree matrix $\overline{D}$, we use inequality (4.3) from (Le et al., 2017). Recall that the degree of node j is $\overline{D}_{jj} = \sum_{j'=1}^{n}(A_{jj'}^{+} + A_{jj'}^{-})$ which is a sum of n independent Bernoulli variables with bounded variance d/n. We can thus find an upper bound on the error $\|\overline{D} - \mathbb{E}[\overline{D}]\|_F$. This bound is weaker than the one obtained in Lemma 40 with the assumption $p \gtrsim \frac{\ln n}{n}$.

Lemma 45 *There exists a constant $C' > 0$ such that for any $r \geq 1$, with probability at least $1 - e^{-2r}$, it holds true*

$$\sum_{j=1}^{n}(\overline{D}_{jj'} - \overline{d})^2 \leq C'r^2 nd \leq 2C'r^2 n\overline{d}(1 - \eta).$$

5.4.2 PROPERTIES OF THE REGULARIZED LAPLACIAN OUTSIDE THE CORE

In this section, we will bound the norm of the Signed Laplacian restricted to the subsets of edges $\mathcal{N}$ and $\mathcal{C}$. The following "restriction lemma" is an extension of Lemma 8.1 in Le et al. (2015) for Signed Laplacian matrices.

Lemma 46 *(Restriction of Signed Laplacian) Let B be a $n \times n$ symmetric matrix, B_γ its regularized form as described in Section 2.2, and $\mathcal{C} \subset [n] \times [n]$. We denote $\overline{D}_\gamma$ the regularized degree matrix , and $\overline{L}_\gamma = \overline{D}_\gamma^{-1/2} B_\gamma \overline{D}_\gamma^{-1/2}$ the modified "Laplacian" and $B_\mathcal{C}$ the $n \times n$ matrix such that the entries outside of $\mathcal{C}$ are set to 0. Let $0 < \varepsilon < 1$ such that the degree of each node in $(B_\gamma)_\mathcal{C}$ is less that ε times the the corresponding degree in B_γ. Then we have*

$$\|(\overline{L}_\gamma)_\mathcal{C}\| \leq \sqrt{\varepsilon}.$$

Proof We denote $\overline{D}_r$ (resp. $\overline{D}_c$) the degree matrix of $(B_\gamma)_\mathcal{C}$ (resp. $(B_\gamma)_\mathcal{C}^T$) and $\widetilde{L}$ its regularized "Laplacian" (it is not necessarily a symmetric matrix) where

$$\widetilde{L} = (\overline{D}_r^{1/2})^\dagger (B_\gamma)_\mathcal{C} (\overline{D}_c^{1/2})^\dagger.$$

By definition of $\overline{L}_\gamma$, $(\overline{L}_\gamma)_\mathcal{C} = \overline{D}_\gamma^{-1/2}(B_\gamma)_\mathcal{C}\overline{D}_\gamma^{-1/2}$. Since in $(B_\gamma)_\mathcal{C}$, some entries in B are set to 0, we have that for all $1 \leq j \leq n$,

$$(\overline{D}_c)_{jj} \leq [\overline{D}_\gamma]_{jj}.$$

Moreover, by assumption, $(\overline{D}_r)_{jj} \leq \varepsilon[\overline{D}_\gamma]_{jj}$. We denote $X = (\overline{D}_r^{1/2})^\dagger$, $Y = (\overline{D}_c^{1/2})^\dagger$ and $Z = \overline{D}_\gamma^{-1/2}$, and now we have

$$\overline{L}_\mathcal{C} = ZB_\mathcal{C}Z = ZX^\dagger XB_\mathcal{C}YY^\dagger Z = ZX^\dagger\widetilde{L}Y^\dagger Z.$$

Because $\|ZX^\dagger\| \leq \sqrt{\varepsilon}$ and $\|Y^\dagger Z\| \leq 1$, by sub-multiplicativity of the norm, we thus obtain

$$\|\overline{L}_\mathcal{C}\| \leq \|ZX^\dagger\| \cdot \|\widetilde{L}\| \cdot \|Y^\dagger Z\| \leq \sqrt{\varepsilon}\|\widetilde{L}\|.$$

In addition, by considering the $2n \times 2n$ symmetric matrix $\widetilde{L}'$

$$\widetilde{L}' = \begin{pmatrix} 0_n & \widetilde{L} \\ \widetilde{L} & 0_n \end{pmatrix},$$

50

we have $\|\widetilde{L}'\| = \|\widetilde{L}\| \leq 1$. In fact, $\widetilde{L}'$ is equal to the identity matrix minus the regularized Laplacian of

$$\begin{pmatrix} 0_n & (B_\gamma)_\mathcal{C} \\ (B_\gamma)_\mathcal{C}^T & 0_n \end{pmatrix}.$$

Using Appendix E, we can conclude that the eigenvalues of $\widetilde{L}'$ are between -1 and 1, leading to $\|\widetilde{L}'\| \leq 1$. Hence, we finally arrive at $\|(\overline{L}_\gamma)_\mathcal{C}\| \leq \sqrt{\varepsilon}$. ∎

Remark 47 *We note that this lemma is not specific to the rows of the matrix B, and one could also derive the same lemma with the assumptions on the columns of the matrix.*

5.5 Error bounds w.r.t the expected regularized Laplacian and expected Signed Laplacian

In this section, we prove an upper bound on the errors $\|L_\gamma - \mathcal{L}_\gamma\|$ and $\|L_\gamma - \mathcal{L}_{sym}\|$ from Theorem 11. We will use the decomposition of the set of edges $(\mathcal{N}, \mathcal{R}, \mathcal{C})$ from Lemma 43, and sum the errors on each of these subsets of edges. We recall that on the subset $\mathcal{N}$, we have an upper bound on $\|(A - \mathbb{E}A)_\mathcal{N}\|$. We will also use the fact that the regularized degrees $[\overline{D}_\gamma]_{jj}$ are lower-bounded by the regularization parameter γ. On the subsets $\mathcal{R}$ and $\mathcal{C}$, we will use Lemma 46 to upper bound the norm of the regularized Laplacian.

Lemma 48 *Under the assumptions of Theorem 11, for any $r \geq 1$, with probability at least $1 - 7e^{-2r}$, we have*

$$\|L_\gamma - \mathcal{L}_\gamma\| \leq \frac{Cr^2}{\sqrt{\gamma}} \left(1 + \frac{\overline{d}}{\gamma}\right)^{5/2} + \frac{32\sqrt{2r}}{\sqrt{\gamma}} + \frac{8}{\sqrt{\overline{d}}}. \tag{78}$$

Proof Let $L_\gamma - \mathcal{L}_\gamma = S + T$ with

$$S = (\overline{D}_\gamma)^{-1/2} A_\gamma (\overline{D}_\gamma)^{-1/2} - (\overline{D}_\gamma)^{-1/2} \mathbb{E}A_\gamma (\overline{D}_\gamma)^{-1/2} = (\overline{D}_\gamma)^{-1/2}(A_\gamma - \mathbb{E}A_\gamma)(\overline{D}_\gamma)^{-1/2},$$
$$T = (\overline{D}_\gamma)^{-1/2} \mathbb{E}A_\gamma (\overline{D}_\gamma)^{-1/2} - (\mathbb{E}\overline{D}_\gamma)^{-1/2} \mathbb{E}A_\gamma (\mathbb{E}\overline{D}_\gamma)^{-1/2}.$$

We will bound the norm of $S + T$ on $\mathcal{N}$, and the norms of L_γ and $\mathcal{L}_\gamma$ on the residuals $\mathcal{R}, \mathcal{C}$. We first use the triangle inequality to obtain

$$\begin{aligned}
&\|L_\gamma - \mathcal{L}_\gamma\| \\
&\leq \|(L_\gamma - \mathcal{L}_\gamma)_\mathcal{N}\| + \|((L_\gamma - I) - (\mathcal{L}_\gamma - I))_\mathcal{R}\| + \|(L_\gamma - \mathcal{L}_\gamma)_\mathcal{C}\| \\
&\leq \|(L_\gamma - \mathcal{L}_\gamma)_\mathcal{N}\| + \|(I - L_\gamma)_\mathcal{R}\| + \|(I - \mathcal{L}_\gamma)_\mathcal{R}\| + \|(I - L_\gamma)_\mathcal{C}\| + \|(I - \mathcal{L}_\gamma)_\mathcal{C}\| \\
&= \|(S + T)_\mathcal{N}\| + \|(I - L_\gamma)_\mathcal{R}\| + \|(I - \mathcal{L}_\gamma)_\mathcal{R}\| + \|(I - L_\gamma)_\mathcal{C}\| + \|(I - \mathcal{L}_\gamma)_\mathcal{C}\| \\
&\leq \|S_\mathcal{N}\| + \|T_\mathcal{N}\| + \|(I - L_\gamma)_\mathcal{R}\| + \|(I - \mathcal{L}_\gamma)_\mathcal{R}\| + \|(I - L_\gamma)_\mathcal{C}\| + \|(I - \mathcal{L}_\gamma)_\mathcal{C}\|.
\end{aligned}$$

1. Bounding the norm $\|T_\mathcal{N}\|$. Denoting $\gamma = \gamma^+ + \gamma^-$, we have that

$$
\begin{aligned}
\|T_\mathcal{N}\|^2 &\leq \|T_\mathcal{N}\|_F^2 \\
&= \sum_{j,j'=1}^n T_{jj'}^2 \\
&= \sum_{j,j'=1}^n \left(\mathbb{E}A_{jj'} + (\gamma^+ - \gamma^-)/n\right)^2 \left[\frac{1}{\sqrt{(\overline{D}_{jj} + \gamma)(\overline{D}_{j'j'} + \gamma)}} - \frac{1}{\overline{d} + \gamma}\right]^2 \quad (79) \\
&\leq \frac{(\overline{d} + \gamma)^2}{2n^2\gamma^6} \left[\sum_{j=1}^n (\overline{D}_{jj} + \gamma)^2 \sum_{j'=1}^n (\overline{D}_{j'j'} - \overline{d})^2 + n(\overline{d} + \gamma)^2 \sum_{i=1}^n (\overline{D}_{jj} - \overline{d})^2\right]. \quad (80)
\end{aligned}
$$

To upper bound (79) by (80), we have used the simplification trick in the proof of (Le et al., 2017, Theorem 4.1) which we now recall. Firstly, the second factor of (79) can be upper bounded in the following way. For $1 \leq j, j' \leq n$,

$$
\begin{aligned}
&\left|\frac{1}{\sqrt{(\overline{D}_{jj} + \gamma)(\overline{D}_{j'j'} + \gamma)}} - \frac{1}{\overline{d} + \gamma}\right| \\
&= \frac{|(\overline{D}_{jj} + \gamma)(\overline{D}_{j'j'} + \gamma) - (\overline{d} + \gamma)^2|}{(\overline{D}_{jj} + \gamma)(\overline{D}_{j'j'} + \gamma)(\overline{d} + \gamma) + \sqrt{(\overline{D}_{jj} + \gamma)(\overline{D}_{j'j'} + \gamma)}(\overline{d} + \gamma)^2} \\
&\leq \frac{|(\overline{D}_{jj} + \gamma)(\overline{D}_{j'j'} + \gamma) - (\overline{d} + \gamma)^2|}{2\gamma^3} \\
&= \frac{|(\overline{D}_{jj} + \gamma)(\overline{D}_{j'j'} + \gamma) - (\overline{d} + \gamma)(\overline{D}_{jj} + \gamma) + (\overline{d} + \gamma)(\overline{D}_{jj} + \gamma) - (\overline{d} + \gamma)^2|}{2\gamma^3} \\
&= \frac{|(\overline{D}_{jj} - \overline{d})(\overline{D}_{j'j'} + \gamma) + (\overline{d} + \gamma)(\overline{D}_{jj} - \overline{d})|}{2\gamma^3}, \quad (81)
\end{aligned}
$$

where the inequality comes from the fact that $\overline{D}_{jj} + \gamma \geq \gamma$. Secondly, we use the inequality $(a + b)^2 \leq 2(a^2 + b^2)$ and we recall that by definition, we can bound the first factor of (79) by $|\mathbb{E}(A_\gamma)_{jj'}| \leq \frac{\overline{d}+\gamma}{n}$. This finally leads to (80).

Now we will bound each term of (80). Using Lemma 45, we have, for any $r \geq 1$, with probability at least $1 - e^{-2r}$,

$$
\sum_{j=1}^n (\overline{D}_{jj} - \overline{d})^2 \leq 2C'r^2 n\overline{d}(1 - \eta) \leq 2C'r^2 n\overline{d}.
$$

If this holds, then the first term of (80) is upper bounded by

$$\sum_{i=1}^{n}(\overline{D}_{jj}+\gamma)^2\sum_{j=1}^{n}(\overline{D}_{j'j'}-\overline{d})^2 \leq \left(2\sum_{j=1}^{n}(\overline{D}_{jj}-\overline{d})^2 + 2n(\overline{d}+\gamma)^2\right)\sum_{j'=1}^{n}(\overline{D}_{j'j'}-\overline{d})^2$$

$$\leq 2C'r^2n\overline{d}\left(4C'r^2n\overline{d}+2n(\overline{d}+\gamma)^2\right)$$

$$\leq 2C'r^2n(\overline{d}+\gamma)(1-\eta)\left(4C'r^2nd+2n(\overline{d}+\gamma)^2\right)$$

$$\leq 2C'r^2n(\overline{d}+\gamma)\left(2(2C'+1)r^2n(d+\gamma)^2\right)$$

$$\leq C_1r^4n^2(\overline{d}+\gamma)^3,$$

with $C_1 = 4C'(2C'+1)$. Similarly, we can bound the second term of (80)

$$n(\overline{d}+\gamma)^2\sum_{j=1}^{n}(\overline{D}_{jj}-\overline{d})^2 \leq 2C'(\overline{d}+\gamma)^2r^2n^2\overline{d} \leq 2C'(\overline{d}+\gamma)^3r^2n^2.$$

Hence, we obtain the following upper bound of (80)

$$\|T_{\mathcal{N}}\|^2 \leq \frac{(C_1+2C')r^4}{2\gamma^6}(\overline{d}+\gamma)^5 = \frac{C_2r^4}{\gamma}\left(1+\frac{\overline{d}}{\gamma}\right)^5, \tag{82}$$

with $C_2 = (C_1+2C')/2$.

2. Bounding the norm $\|S_{\mathcal{N}}\|$. We first note that

$$S = (\overline{D}_\gamma)^{-1/2}(A_\gamma - \mathbb{E}A_\gamma)(\overline{D}_\gamma)^{-1/2} = (\overline{D}_\gamma)^{-1/2}(A - \mathbb{E}A)(\overline{D}_\gamma)^{-1/2}.$$

We also recall that $\|\overline{D}_\gamma\| \geq \gamma$. Hence, using Lemma 43, with probability at least $1 - 6n^{-r}$, we have

$$\|S_{\mathcal{N}}\| \leq \|\overline{D}_\gamma^{-1/2}\|\,\|(A-\mathbb{E}A)_{\mathcal{N}}\|\,\|\overline{D}_\gamma^{-1/2}\| \leq \|(A-\mathbb{E}A)_{\mathcal{N}}\|/\gamma \leq \frac{Cr^{3/2}}{\gamma}\sqrt{\overline{d}(1-\eta)}$$

$$\leq \frac{Cr^{3/2}}{\gamma}\sqrt{\overline{d}}. \tag{83}$$

Summing the bounds in (82) and (83), we have the intermediate result

$$\|(L_\gamma - \mathcal{L}_\gamma)_{\mathcal{N}}\| \leq \frac{Cr^{3/2}}{\gamma}\sqrt{\overline{d}} + \frac{\sqrt{C_2}r^2}{\sqrt{\gamma}}\left(1+\frac{\overline{d}}{\gamma}\right)^{5/2} \tag{84}$$

$$\leq \frac{r^2}{\sqrt{\gamma}}\left(C\sqrt{\frac{\overline{d}}{\gamma}} + \sqrt{C_2}\left(1+\frac{\overline{d}}{\gamma}\right)^{5/2}\right) \tag{85}$$

$$\leq \frac{r^2}{\sqrt{\gamma}}(C+\sqrt{C_2})\left(1+\frac{\overline{d}}{\gamma}\right)^{5/2} = \frac{C_3r^2}{\sqrt{\gamma}}\left(1+\frac{\overline{d}}{\gamma}\right)^{5/2}, \tag{86}$$

with $C_3 = C + \sqrt{C_2}$.

3. Bounding $\|(L_\gamma)_{\mathcal{R}}\|, \|(L_\gamma)_{\mathcal{C}}\|, \|(\mathcal{L}_\gamma)_{\mathcal{R}}\|, \|(\mathcal{L}_\gamma)_{\mathcal{C}}\|$. Using the proof of Lemma 43, each row of $A_{\mathcal{R}}$ has at most $128r$ non-zeros entries and intersects at most $4n/\bar{d}$ columns. Thus, for all $1 \leq j \leq n$

$$\sum_{j'=1}^{n} \left[(A_\gamma^+ + A_\gamma^-)_{\mathcal{R}}\right]_{jj'} \leq 128r + \frac{4\gamma}{\bar{d}} = \gamma\left(\frac{128r}{\gamma} + \frac{4}{\bar{d}}\right) \leq \sum_{j'} \left[A_\gamma^+ + A_\gamma^-\right]_{jj'}\left(\frac{128r}{\gamma} + \frac{4}{\bar{d}}\right),$$

as $\sum_{j'}[A_\gamma^+ + A_\gamma^-]_{jj'} \geq n \times \left(\frac{\gamma^+}{n} + \frac{\gamma^-}{n}\right) = \gamma$. We can thus apply Lemma 46 with $\varepsilon = \frac{128r}{\gamma} + \frac{4}{\bar{d}}$, and we arrive at

$$\|(L_\gamma)_{\mathcal{R}}\| \leq \sqrt{\frac{128r}{\gamma} + \frac{4}{\bar{d}}}.$$

We also obtain the same bound for $\|(L_\gamma)_{\mathcal{C}}\|$. Similarly, we have $\sum_{j'} \left[\mathbb{E}[A_\gamma^+] + \mathbb{E}[A_\gamma^-]\right]_{jj'} = (n-1)p + \gamma = \bar{d} + \gamma \geq \gamma$ and

$$\sum_{j'=1}^{n} \left[(\mathbb{E}[A_\gamma^+] + \mathbb{E}[A_\gamma^-])_{\mathcal{R}}\right]_{jj'} \leq 4\frac{np}{\bar{d}} + \frac{4\gamma}{\bar{d}} \leq 8 + \frac{4\gamma}{\bar{d}} = \gamma\left(\frac{8}{\gamma} + \frac{4}{\bar{d}}\right)$$

$$\leq \sum_{j'} \left[\mathbb{E}[A^+]_\gamma + \mathbb{E}[A^-]_\gamma\right]_{jj'}\left(\frac{8}{\gamma} + \frac{4}{\bar{d}}\right).$$

We arrive at $\|(\mathcal{L}_\gamma)_{\mathcal{R}}\| \leq \sqrt{\frac{8}{\gamma} + \frac{4}{\bar{d}}}$, and finally, we also have $\|(\mathcal{L}_\gamma)_{\mathcal{C}}\| \leq \sqrt{\frac{8}{\gamma} + \frac{4}{\bar{d}}}$.

4. Bounding $\|L_\gamma - \mathcal{L}_\gamma\|$. Summing up the bounds obtained in the first three steps, with probability at least $1 - e^{-2r} - 6n^{-r} \geq 1 - 7e^{-2r}$, we finally arrive at the bound

$$\|L_\gamma - \mathcal{L}_\gamma\| \leq \frac{C_3 r^2}{\sqrt{\gamma}}\left(1 + \frac{\bar{d}}{\gamma}\right)^{5/2} + 2\sqrt{\frac{128r}{\gamma} + \frac{4}{\bar{d}}} + 2\sqrt{\frac{8}{\gamma} + \frac{4}{\bar{d}}}$$

$$\leq \frac{C_3 r^2}{\sqrt{\gamma}}\left(1 + \frac{\bar{d}}{\gamma}\right)^{5/2} + 4\sqrt{\frac{128r}{\gamma} + \frac{4}{\bar{d}}}$$

$$\leq \frac{C_3 r^2}{\sqrt{\gamma}}\left(1 + \frac{\bar{d}}{\gamma}\right)^{5/2} + \frac{32\sqrt{2r}}{\sqrt{\gamma}} + \frac{8}{\sqrt{\bar{d}}}.$$

$\blacksquare$

This bound also provides easily a bound on the norm of $L_\gamma - \mathcal{L}_{sym}$.

Corollary 49 *(Error bound of the regularized Laplacian) With the notations of Theorem 8 and Theorem 11, and $\gamma = \gamma^+ + \gamma^-$, we have*

$$\|L_\gamma - \mathcal{L}_{sym}\| \leq \frac{Cr^2}{\sqrt{\gamma}}\left(1 + \frac{\bar{d}}{\gamma}\right)^{5/2} + \frac{32\sqrt{2r}}{\sqrt{\gamma}} + \frac{8}{\sqrt{\bar{d}}} + \frac{\gamma}{\bar{d} + \gamma} =: \Delta_L(\gamma, \bar{d}). \qquad (87)$$

In particular, for the choice $\gamma = \bar{d}^{7/8}$, if $p \geq 2/n$, we obtain

$$\|L_\gamma - \mathcal{L}_{sym}\| \leq \left(128Cr^2 + 1\right)\bar{d}^{-1/8}.$$

Proof By triangular inequality,

$$\|L_\gamma - \mathcal{L}_{sym}\| \le \|L_\gamma - \mathcal{L}_\gamma\| + \|\mathcal{L}_\gamma - \mathcal{L}_{sym}\|.$$

For the second term on the RHS, we have

$$\|\mathcal{L}_\gamma - \mathcal{L}_{sym}\| = \left\| \frac{1}{\overline{d} + \gamma} \mathbb{E}A - \frac{1}{\overline{d}} \mathbb{E}A \right\| = \frac{\gamma}{\overline{d}(\overline{d} + \gamma)} \|\mathbb{E}A\| \le \frac{\gamma}{\overline{d} + \gamma}. \tag{88}$$

The last inequality comes from the fact that $\|\mathbb{E}A\| \le (n-1)p(1-\eta) \le \overline{d}$. Thus, by summing the bound obtained in Lemma 48 and (88), we arrive at the expected result in (87). Moreover, if $\gamma \le \overline{d}$, since $C > 1$, one can readily verify that

$$\|\mathcal{L}_\gamma - \mathcal{L}_{sym}\| \le 128 C r^2 \frac{\overline{d}^{\frac{5}{2}}}{\gamma^3} + \frac{\gamma}{\overline{d}}. \tag{89}$$

If $\gamma = \overline{d}^{7/8}$, then $\gamma \le \overline{d}$ holds provided $\overline{d} \ge 1$ or equivalently, $p \ge \frac{1}{n-1}$. The latter is ensured if $p \ge 2/n$ (since $n \ge 2$). Plugging this in (89), we then obtain the bound

$$\|\mathcal{L}_\gamma - \mathcal{L}_{sym}\| \le \left(128 C r^2 + 1 \right) \overline{d}^{-1/8}.$$

This concludes the proof of Corollary 49 and Theorem 11. ∎

5.6 Error bound on the eigenspaces and mis-clutering rate in the sparse regime

This section provides a bound on the misalignment error of the eigenspaces of L_γ and $\mathcal{L}_{sym}$, which then leads to a bounds on the mis-clustering rate of the k-means clustering step.

5.6.1 EIGENSPACE ALIGNMENT

Using the bound from Corollary 49, we can perform the same analysis of the eigenspaces of L_γ and $\mathcal{L}_{sym}$, as in Theorem 8, which will prove Theorem 13. We apply, once again, Weyl's inequality and the Davis-Kahan theorem to bound the distance between the two subspaces $\mathcal{R}(V_{k-1}(L_\gamma))$ and $\mathcal{R}(V_{k-1}(\mathcal{L}_{sym}))$. We have that

$$\lambda_{n-k+1}(L_\gamma) - \lambda_{n-k+2}(\mathcal{L}_{sym}) \ge \lambda_{gap} - \|L_\gamma - \mathcal{L}_{sym}\| \ge \lambda_{gap} - \Delta_L(\gamma, \overline{d}),$$

using Corollary 49. If $\gamma = \gamma_0 \overline{d}^{7/8}$, then

$$\Delta_L(\gamma, \overline{d}) \le \left(128 C r^2 + 1 \right) (\overline{d})^{-1/8} := \frac{C_4}{\overline{d}^{1/8}},$$

with $C_4 = 128 C r^2 + 1$. For $0 < \delta < 1/2$, we would like to ensure that

$$\lambda_{gap} - \Delta_L(\gamma, \overline{d}) \ge \lambda_{gap} \left(1 - \frac{\delta}{2} \right).$$

Hence, using the lower bound on the eigengap from Lemma 39, it suffices that

$$\lambda_{gap} - \frac{C_4}{\overline{d}^{1/8}} \geq \lambda_{gap}\left(1 - \frac{\delta}{2}\right) \iff \overline{d}^{1/8} \geq \frac{2kC_4}{\delta(1-2\eta)} \iff p \geq \left(\frac{2kC_4}{\delta(1-2\eta)}\right)^8 \frac{1}{n-1}.$$

Thus, the condition $p \geq \left(\frac{2kC_4}{\delta(1-2\eta)}\right)^8 \frac{2}{n}$ is sufficient. Applying the Davis-Kahan theorem, we arrive at

$$\|(I - V_{k-1}(L_\gamma)V_{k-1}(L_\gamma)^T)V_{k-1}(\mathcal{L}_{sym})\| \leq \frac{\delta\lambda_{gap}/2}{\lambda_{gap}(1-\delta/2)} \leq \frac{\delta/2}{1-\delta/2} \leq \delta,$$

and using once again Proposition 57, there exists an orthogonal matrix $O \in \mathbb{R}^{(k-1)\times(k-1)}$ such that

$$\|V_{k-1}(L_\gamma) - \Theta R_{k-1}O\| \leq 2\delta.$$

5.7 Proof of Theorem 16

In this section, we finally prove our result on the clustering performance of the Signed Laplacian and regularized Laplacian algorithms. The proof essentially relies on the following lemma, which provides a lower bound on the distance between two rows of $\Delta^{-1}R_{k-1}$, with $\Delta = \mathrm{diag}(\sqrt{n_i})$.

Lemma 50 *For all $1 \leq i \neq i' \leq k$, we have $\|(R_{k-1})_{i*} - (R_{k-1})_{i'*}\| \geq 1$. Moreover, for $i \in [k]$, it holds that*

$$\min_{\substack{i,i'\in[k],i\neq i' \\ j\in C_i, j'\in C_{i'}}} \left\|(\Delta^{-1}R_{k-1})_{j*} - (\Delta^{-1}R_{k-1})_{j'*}\right\|^2 \geq \frac{2}{3n_i}.$$

Proof Recall from (69) that $\overline{C} = p(1-2\eta)uu^T + \mathrm{diag}(d_i)$, with $d_i = u_i^2 + \left(1 + \frac{p}{d}(1-2\eta)\right)$ and $u_i = \sqrt{\frac{n_i}{d}}, 1 \leq i \leq k$. Moreover, from (70), $\overline{C} = R\Lambda R$ with $R = [R_{k-1} \ \gamma_1]$ and γ_1 the largest eigenvector of $\overline{C}$. We first show that the entries of γ_1 are necessarily of the same sign, i.e. $(\gamma_1)_i \geq 0, \forall i$ or $(\gamma_1)_i \leq 0, \forall i$. In fact, by definition, γ_1 is the solution of

$$\max_{\|v\|=1} v^T\overline{C}v = \max_{\|v\|=1} p(1-2\eta)(v^u)^2 + \sum_{i=1}^{k} d_i v_i^2. \tag{90}$$

Since all the entries of u are positive, it is easy to see that any solution γ_1 of (90) necessarily has entries of the same sign (otherwise you could replace some $(\gamma_1)_i$ by $-(\gamma_1)_i$ and increase the objective function).

Let $i \neq i' \in [k]$. As R has orthonormal rows,

$$\langle R_{i*}, R_{i'*}\rangle = 0 \iff \langle (R_{k-1})_{i*}, (R_{k-1})_{i'*}\rangle + \underbrace{(\gamma_1)_i(\gamma_1)_{i'}}_{\geq 0} = 0$$

$$\implies \langle (R_{k-1})_{i*}, (R_{k-1})_{i'*}\rangle \leq 0.$$

Hence,

$$
\begin{aligned}
\left\|(R_{k-1})_{i*} - (R_{k-1})_{i'*}\right\|^2 &= \left\|(R_{k-1})i*\right\|^2 + \left\|(R_{k-1})i'*\right\|^2 - 2\underbrace{\langle (R_{k-1})_{i*}, (R_{k-1})_{i'*}\rangle}_{\leq 0}\\
&\geq \left\|(R_{k-1})_{i*}\right\|^2 + \left\|(R_{k-1})_{i'*}\right\|^2\\
&= 2 - \underbrace{[(\gamma_1)_i^2 + (\gamma_1)_{i'}^2]}_{\leq 1} \geq 1.
\end{aligned}
$$

In particular, this implies that R_{k-1} has k distinct rows. Now let $j, j' \in [n]$ such that $j \in C_i$ and $j' \in C_{i'}$. Recalling that with $\Delta = \mathrm{diag}(\sqrt{n_i})$, $V_{k-1}(\mathcal{L}_{sym}) = \Theta R_{k-1} = \Theta \Delta \Delta^{-1} R_{k-1} = \hat{\Theta} \Delta^{-1} R_{k-1}$, we have

$$
\begin{cases}
(\Delta^{-1} R_{k-1})_{j*} = \frac{1}{\sqrt{n_i}}(R_{k-1})_{i*},\\
(\Delta^{-1} R_{k-1})_{j'*} = \frac{1}{\sqrt{n_{i'}}}(R_{k-1})_{i'*}.
\end{cases}
$$

Hence,

$$
\begin{aligned}
&\left\|(\Delta^{-1} R_{k-1})_{j*} - (\Delta^{-1} R_{k-1})_{j'*}\right\|^2\\
&= \frac{1}{n_i}\left\|(R_{k-1})i*\right\|^2 + \frac{1}{n_{i'}}\left\|(R_{k-1})_{i'*}\right\|^2 - 2\frac{1}{\sqrt{n_i n_{i'}}}\underbrace{\langle (R_{k-1})_{i*}, (R_{k-1})_{i'*}\rangle}_{\leq 0}\\
&\geq \frac{1}{n_i}\left\|(R_{k-1})_{i*}\right\|^2 + \frac{1}{n_{i'}}\left\|(R_{k-1})_{i'*}\right\|^2\\
&\geq \frac{1}{n_i} + \frac{1}{n_{i'}} - \frac{(\gamma_1)_i^2}{n_i} - \frac{(\gamma_1)_{i'}^2}{n_i'}\\
&\geq \frac{1}{n_i} + \frac{1}{n_{i'}} - \frac{(\gamma_1)_i^2 + (\gamma_1)_{i'}^2}{ns} \geq \frac{1}{n_i} + \frac{1}{n_{i'}} - \frac{1}{ns} \geq \frac{1}{n_i} + \frac{1}{nl} - \frac{1}{ns}.
\end{aligned}
$$

Besides, we know that $\frac{1}{nl} \geq \frac{\rho}{n_i}$ and $\frac{1}{ns} \leq \frac{1}{\rho n_i}$. Therefore, we obtain the bound

$$
\left\|(\Delta^{-1} R_{k-1})_{j*} - (\Delta^{-1} R_{k-1})_{j'*}\right\|^2 \geq \frac{1}{n_i}\left(1 + \rho - \frac{1}{\rho}\right).
$$

We will now prove that with the condition $\sqrt{\rho} > 1 - \frac{1}{4k(2+\sqrt{k})}$, we have $1 + \rho - \frac{1}{\rho} \geq \frac{2}{3}$ and this will lead to the final result. First, we note that $\rho > 1 - \frac{1}{2k(2+\sqrt{k})}$ and $2k(2+\sqrt{k}) \geq 12$, and $\frac{2k(2+\sqrt{k})}{2k(2+\sqrt{k})-1} \leq \frac{5}{4}$ for $k \geq 2$. Thus,

$$
1 + \rho - \frac{1}{\rho} \geq 2 - \frac{1}{2k(2+\sqrt{k})} - \frac{2k(2+\sqrt{k})}{2k(2+\sqrt{k}) - 1} \geq 2 - \frac{1}{12} - \frac{5}{4} = \frac{2}{3}.
$$

<div align="right">■</div>

Remark 51 *In the equal-size case $n_i = \frac{n}{k}, \forall 1 \le i \le k$, since $\gamma_1 = \chi_1$, R_{k-1} has orthogonal rows and*

$$\left\| (R_{k-1})_{i*} - (R_{k-1})_{i'*} \right\|^2 = \left\| R_{i*} - R_{i'*} \right\|^2 = 2.$$

This implies that

$$\left\| (\Delta^{-1} R_{k-1})_{j*} - (\Delta^{-1} R_{k-1})_{j'*} \right\|^2 = \frac{2k}{n}.$$

From Lemma 50, we have that $\forall 1 \le i \le k,$ $\min\limits_{\substack{i,i' \in [k], i \ne i' \\ j \in C_i, j' \in C_{i'}}} \left\| (\Delta^{-1} R_{k-1})_{j*} - (\Delta^{-1} R_{k-1})_{j'*} \right\|^2 \ge$

$\frac{2}{3n_i}$. Hence with $\delta_i^2 := \frac{2}{3n_i}$ and using Lemma 35, we obtain

$$\sum_{i=1}^{k} \delta_i^2 |S_i| = \sum_{i=1}^{k} \frac{2|S_i|}{3n_i} \le 4(4 + 2\xi) \left\| V_{k-1}(\overline{L_{sym}}) - V_{k-1}(\mathcal{L}_{sym}) \right\|_F^2$$

$$\le 4(16 + 8\xi)(k-1) \left\| V_{k-1}(\overline{L_{sym}}) - V_{k-1}(\mathcal{L}_{sym})O \right\|^2$$

$$\le 8(16 + 8\xi)(k-1)\delta^2,$$

using Theorem 8 . Moreover, we have

$$\left\| V_{k-1}(\overline{L_{sym}}) - V_{k-1}(\mathcal{L}_{sym}) \right\|_F^2 \le 2(k-1) \left\| V_{k-1}(\overline{L_{sym}}) - V_{k-1}(\mathcal{L}_{sym})O \right\|^2$$

$$\le 8(k-1)\delta^2$$

$$< 8(k-1) \cdot \frac{1}{12(16 + 8\xi)(k-1)}$$

$$= \frac{n_i \delta_i^2}{16 + 8\xi}, \qquad\qquad \forall 1 \le i \le k.$$

Therefore, we can use the second part of Lemma 35 and finally conclude that

$$\sum_{i=1}^{k} \frac{|S_i|}{n_i} \le 96(2 + \xi)\delta^2.$$

For the regularized Laplacian algorithm, the same computations are valid using the result from Theorem 13.

6. Numerical experiments

In this section, we report on the outcomes of several numerical experiments that compare our two proposed algorithms with a suite of state-of-the-art methods from the signed clustering literature. We test the performances of the different algorithms on signed graphs drawn from our Signed Stochastic Block Model, as well as on three real-world data sets that are standard benchmarks in the signed networks literature. We rely on a previous Python implementation of SPONGE and Signed Laplacian (along with their respective normalized

versions), and of other methods from the literature[3], made available in the context of previous work of a subset of the authors of the present paper (Cucuringu et al., 2019). More specifically, we consider algorithms based on the adjacency matrix A, the Signed Laplacian matrix $\overline{L}$, its symmetrically normalized version $\overline{L}_{sym}$ (Kunegis et al., 2010), SPONGE and its normalized version SPONGE_{sym}, and the two algorithms introduced in Chiang et al. (2012) that optimize the Balanced Ratio Cut and the Balanced Normalized Cut objectives.

We note that once the low-dimensional embedding has been computed by any of the considered algorithms, the final partition is obtained after running k-means++ (Arthur and Vassilvitskii, 2007), which improves over the popular k-means algorithm by employing a careful seeding initialization procedure and is the typical choice in practice.

6.1 Grid search for choosing the parameters τ^+, τ^-

In the following experiments, the Signed Stochastic Block Model will be sampled with the following set of parameters

- the number of nodes $n = 5000$,

- the number of communities $k \in \{3, 5, 10, 20\}$,

- the relative size of communities $\rho = 1$ (equal-size clusters) and $\rho = 1/k$ (non-equal size clusters).

For the edge density parameter p, we choose two sparsity regimes, *Regime I* and *Regime II*, where *Regime II* is strictly harder than *Regime I*, in the sense than for the same value of k, the edge density in *Regime I* is significantly larger compared to *Regime II*. The noise level η is chosen such that the recovery of the clusters is unsatisfactory for a subset of pairs of parameters (τ^+, τ^-). For each set of parameters, we sample 20 graphs from the SSBM and average the resulting ARI.

Our experimental setup is summarized in the following steps.

1. Select a set of parameters (k, ρ, p, η) from the regime of interest;

2. Sample a graph from the $\text{SSBM}(n, k, \rho, p, \eta)$;

3. Extract the largest connected component of the measurement graph (regardless of the sign of the edges);

4. If the size of the latter is too small $(< n/2)$, resample a graph until successful;

5. For each pair of parameters (τ^+, τ^-), compute the k-dimensional embeddings using the SPONGE_{sym} algorithm (with the implementation in the *signet* package (Cucuringu et al., 2019));

6. Obtain a partition of the graph into k clusters, and compute the *ARI* between this estimated partition and the ground-truth clusters using the implementation in scikit-learn of the k-means++ algorithm;

---

[3]Python implementations of a suite of algorithms for signed clustering are available at `https://github.com/alan-turing-institute/signet`

7. Repeat 20 times the steps $2 - 7$ mentioned above, and record the average performance over the 20 runs.

The results in the dense regimes are reported in Figure 1, while those for the sparse regimes in Figure 2. This set of results indicate that the gradient of the ARI in the space of parameters (τ^+, τ^-) is larger when the cluster sizes are very unbalanced and the edge density is low. We attribute this to the fact that, for suitably chosen values, the parameters (τ^+, τ^-) are performing a form of regularization of the graph that can significantly improve the clustering performance.

6.2 Comparison of a suite of spectral methods

This section performs a comparison of the performance of the following spectral clustering algorithms. We rely on the same notation used in Cucuringu et al. (2019), when mentioning the names of the SPONGE algorithms, namely SPONGE and SPONGE$_{sym}$. The complete list of algorithms compared is as follows.

- the combinatorial (un-normalized) Signed Laplacian $\overline{L} = \overline{D} - A$,

- the symmetric Signed Laplacian $\overline{L}_{sym} = I - \overline{D}^{-1/2} A \overline{D}^{-1/2}$,

- SPONGE and SPONGE$_{sym}$ with a suitably chosen pair of parameters (τ^+, τ^-)

- the Balanced Ratio Cut $L_{BRC} = D^+ - A$

- the Balanced Normalized Cut $L_{BNC} = D^{-1/2}(D^+ - A)D^{-1/2}$.

For the combinatorial and symmetric Signed Laplacians $\overline{L}$ and $\overline{L_{sym}}$, we compute $(k-1)$-dimensional embeddings before applying the k-means++ algorithm. For all other methods, we use the k smallest eigenvectors.

In this experiment, we fix the parameters $n = 5000$, $k \in \{3, 5, 10, 20\}$ and p, η in a certain set, and for each plot, we vary the aspect ratio $\rho \in [0, 1]$. The relative proportions of the classes $s_i = \frac{n_i}{n}$ are chosen according to the following procedure

1. Fix $s_1' = 1/k$, pick a value for ρ and compute $s_k' = s_1'/\rho$.

2. For $i \in [2, k-1]$, sample s_i' from the uniform distribution in the interval $[s_1', s_k']$.

3. Compute the proportions $s_i = \frac{s_i'}{\sum_{i=1}^{k} s_i'}$, and then sample the graph from the resulting SSBM.

4. Repeat 20 times the steps 1-3 mentioned above, and record the average performance over the 20 runs.

The results are reported in Figure 3. We note that in almost all settings, the SPONGE$_{sym}$ algorithm outperforms the other clustering methods, in particular for low values of the aspect ratio ρ. With the exception of the symmetric Signed Laplacian, most methods seem to perform worse when the aspect ratio is higher, meaning that the clusters are more unbalanced, which is a more challenging regime.
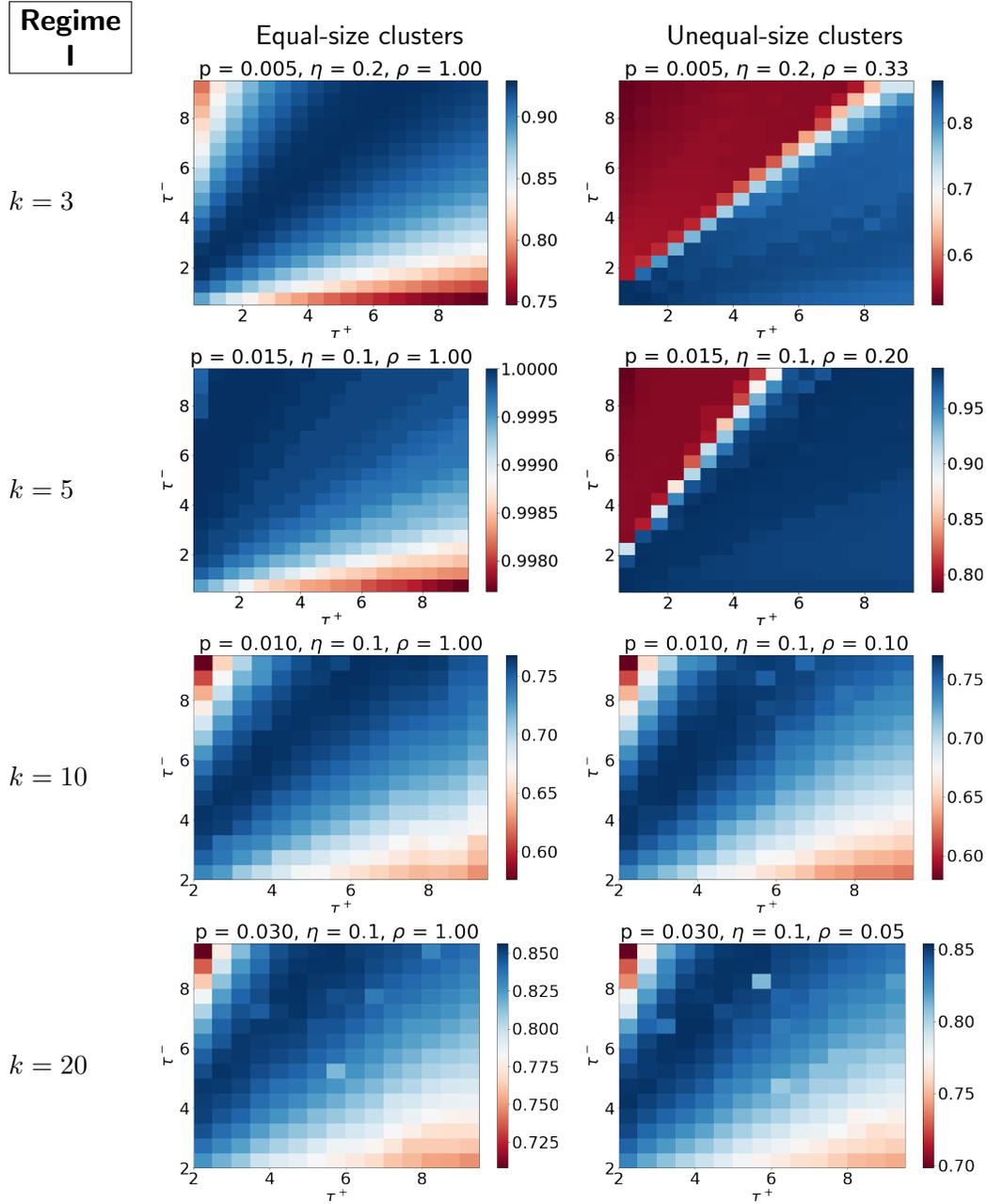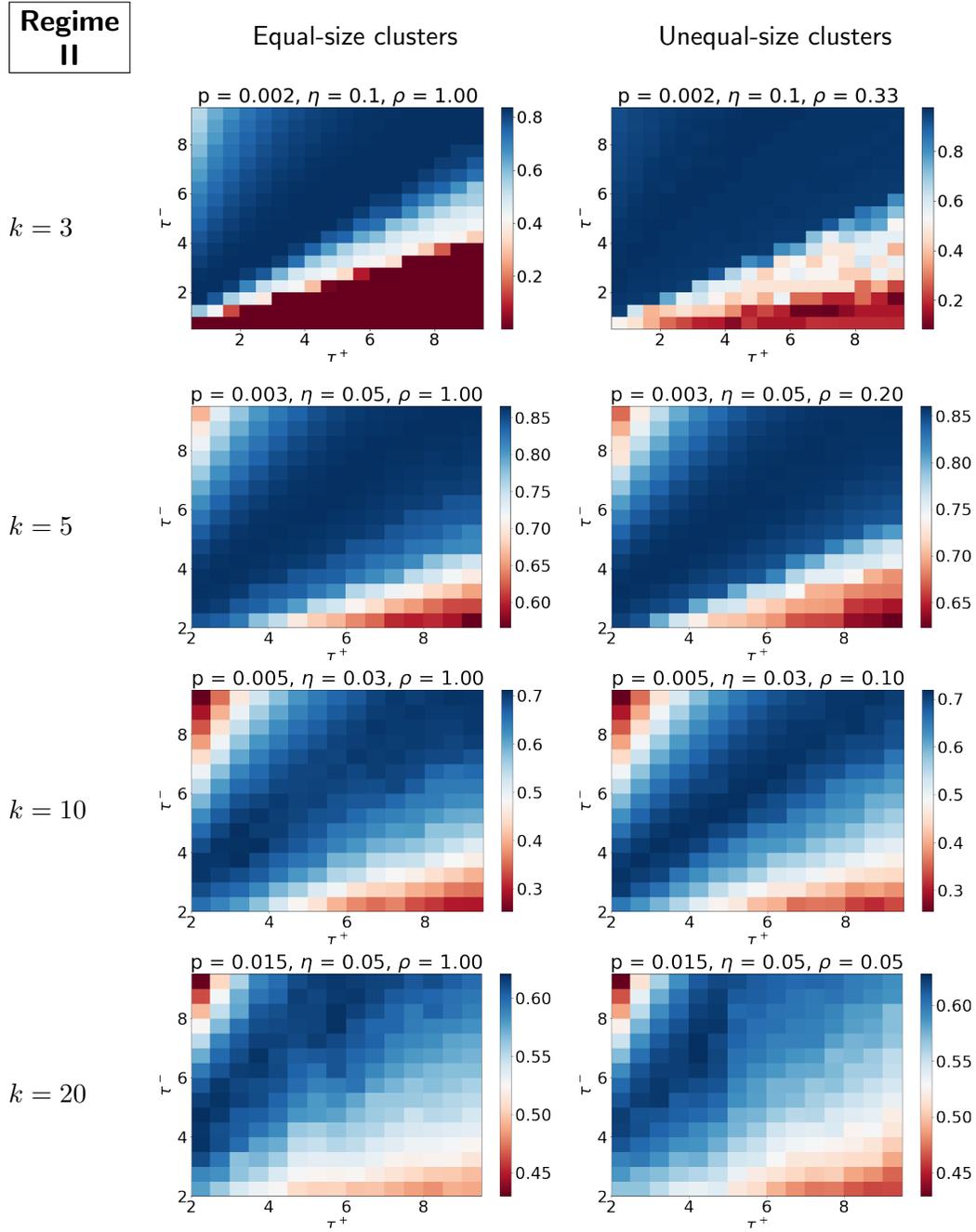
Figure 1: Heatmaps of the Adjusted Rand Index between the ground truth and the partition obtained using the SPONGE$_{sym}$ algorithm with varying regularization parameters (τ^+, τ^-), for a SSBM in *Regime I*, with $n = 5000$ and $k = \{3, 5, 10, 20\}$ clusters of equal sizes (left column) and unequal sizes (right column).

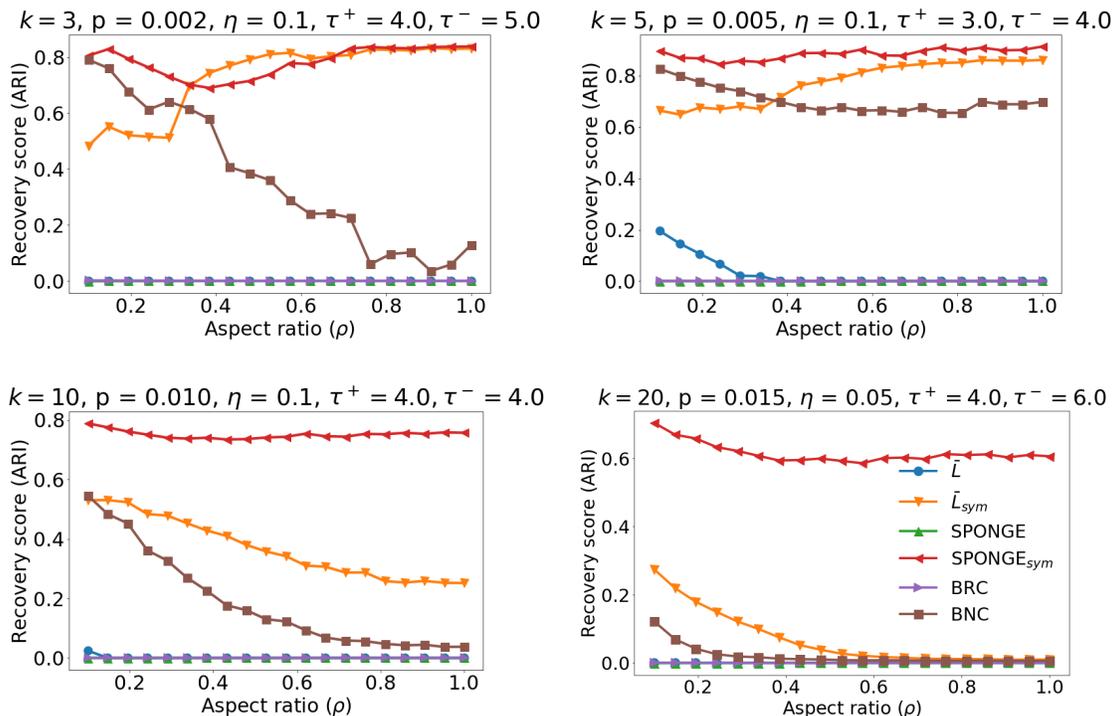Figure 2: Heatmaps of the Adjusted Rand Index between the ground truth and the partition obtained using the SPONGE$_{sym}$ algorithm with varying regularization parameters (τ^+, τ^-), for a SSBM in *Regime II* with $n = 5000$ and $k = \{3, 5, 10, 20\}$ clusters of equal sizes (left column) and unequal sizes (right column).

Figure 3: Performance of the various clustering algorithms, as measured by the Adjusted Rand Index, versus the aspect ratio ρ for a SSBM with $k = \{3, 5, 10, 20\}$ for $n = 5000$. For larger number of clusters, $k = 10$ and especially $k = 20$, SPONGE$_{sym}$ is essentially the only algorithm able to produce meaningful results, and clearly outperforms all the other methods. Note that no regularization has been used throughout this set of experiments.

6.3 Performance of the regularized algorithms in the sparse regime

In this batch of experiments, we study how the *regularized* Signed Laplacian and the SPONGE$_{sym}$ sparse algorithms perform. We consider sparse settings of the SSBM ($p \leq 0.003$) with $n = 5000$ nodes. For the SPONGE$_{sym}$ algorithm, we fix the parameters (τ^+, τ^-) in each setting. Our parameter selection procedure is to chose a pair of parameters that leads to a "good" recovery of the clusters for the unregularized algorithm (see Figure 2). We perform a grid search on the parameters (γ^+, γ^-) for each of the two regularized algorithms (see Figure 4 and Figure 5). For the regularized Signed Laplacian algorithm, we observe distinct regions of performance on the space of parameters (γ^+, γ^-). This is not predictable from our theoretical results, where the positive and negative regularization parameters play symmetric roles. We conjecture this to be due to the difference of density of the positive and negative subgraphs in our signed random graph model. For the SPONGE$_{sym}$ sparse algorithm, we note that the gradient of performances in the heatmaps (Figure 4, Figure 5) is similar to what was reported in Figure 2, which could be due to the fact that the parameters (τ^+, τ^-) already have a regularization effect.
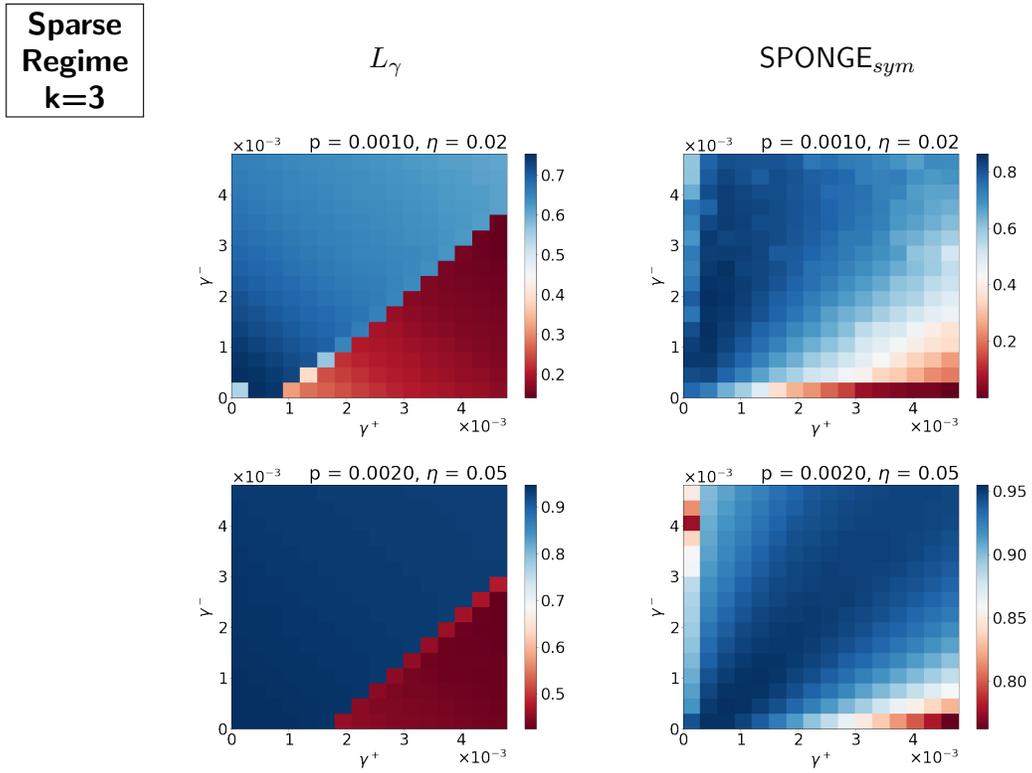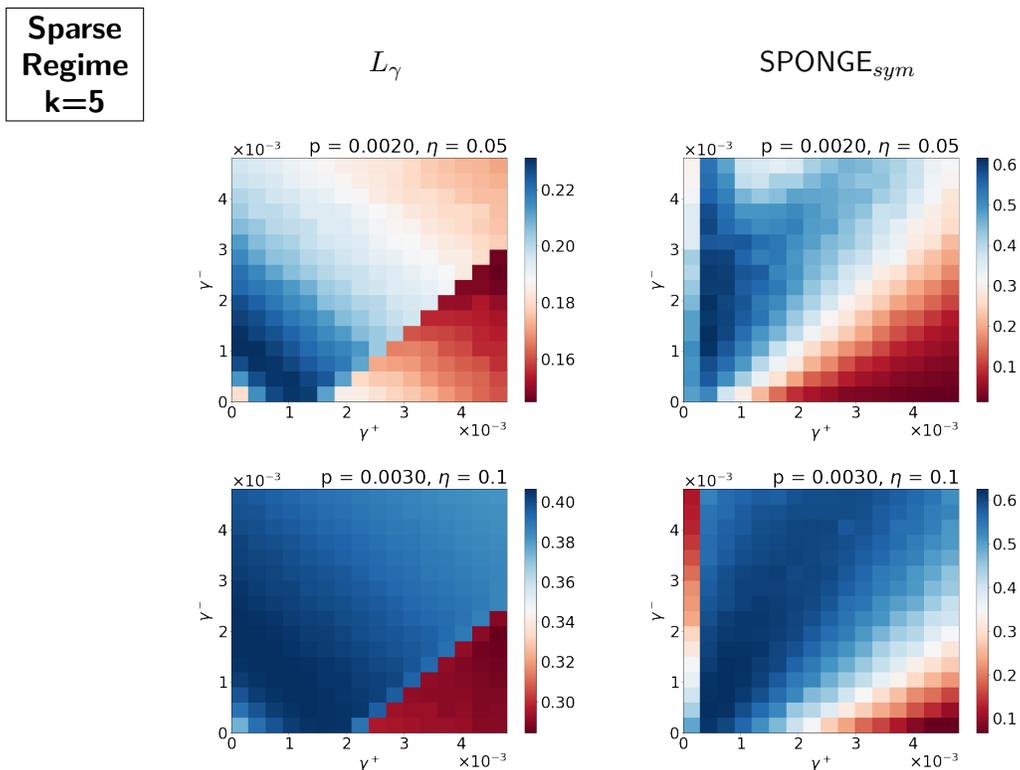
Figure 4: Heatmaps of the Adjusted Rand Index between the ground truth and the partition obtained using the L_γ and SPONGE$_{sym}$ algorithm with fixed parameters (τ^+, τ^-) and varying **regularization** parameters (γ^+, γ^-), for a SSBM in two **sparse** regimes, with $n = 5000$ and $k = 3$ clusters.

Figure 5: Heatmaps of the Adjusted Rand Index between the ground truth and the partition obtained using the L_γ and SPONGE_{sym} algorithm with fixed parameters (τ^+, τ^-) and varying **regularization** parameters (γ^+, γ^-), for a SSBM in two **sparse** regimes, with $n = 5000$ and $k = 5$ clusters.

Dataset	Number of nodes	Edge density
Wikipedia	$11,259$	2.2×10^{-3}
Slashdot	$82,140$	1.3×10^{-4}
Bitcoin	$5,875$	3.6×10^{-3}

Table 1: Characteristics of the three benchmark data sets.

6.4 Performances on real-world data sets

Finally, we measure the performances of our unregularized and regularized algorithms on three benchmark data sets in the signed clustering problem: the Wikipedia Requests for Adminship, the Slashbot Zoo and Bitcoin data sets from Leskovec and Krevl (2014). These networks are large and sparse (see Table 1 for a summary of the number of nodes and edge densities). Since no ground-truth clusters are available for these networks, we measure the quality of the clustering using an objective score, namely the *normalized adjacency score*. This metric is defined as the sum of ratios of the number of positive edges over the number of negative edges within each cluster. We assume that a higher value of this score indicates a better partition of the node set. Our results are reported in Figure 6. We observe that the regularized versions of our algorithms, namely SPONGE$_{sym}$ and the Symmetric Signed Laplacian, perform much better that their respective unregularized versions, confirming the fact that regularization improves the performance of spectral algorithms in the sparse regime. We also note that the standard Spectral Clustering algorithm based on the signed adjacency matrix - denoted by A in the figure legend - also performs well on these real-world data sets.

7. Concluding remarks and future research directions

In this work, we provided a thorough theoretical analysis of the robustness of the SPONGE$_{sym}$ and symmetric Signed Laplacian algorithms, for graphs generated from a Signed Stochastic Block Model. Under this model, the sign of the edges (rather than the usual discrepancy of the edge densities across clusters versus within clusters) is an essential attribute which induces the underlying cluster structure of the graph. We proved that our signed clustering algorithms, based on suitably defined matrix operators, are able to recover the clusters under certain favorable noise regimes, and under two regimes of edge sparsity. Although the sparse setting is particularly challenging, our algorithms based on regularized graphs perform well, provided that the regularization parameters are suitably chosen. We also expect that the same type of analysis could be adapted to other probabilistic generative models for signed networks. For instance, extensions of the unsigned Stochastic Block Models, such as the Degree-Corrected Stochastic Block Model, that includes degree-heterogeneity could be considered, as well extensions to the setting of polarized communities, in the spirit of those proposed by Bonchi et al. (2019) and Xiao et al. (2020).

One theoretical question that has been not been answered yet relates to the choice of the positive and negative regularization parameters γ_+, γ_-. Having a data-driven approach to tune the regularization parameters would be of great use in many practical applications involving very sparse graphs. An interesting future line of work would be to study the latest
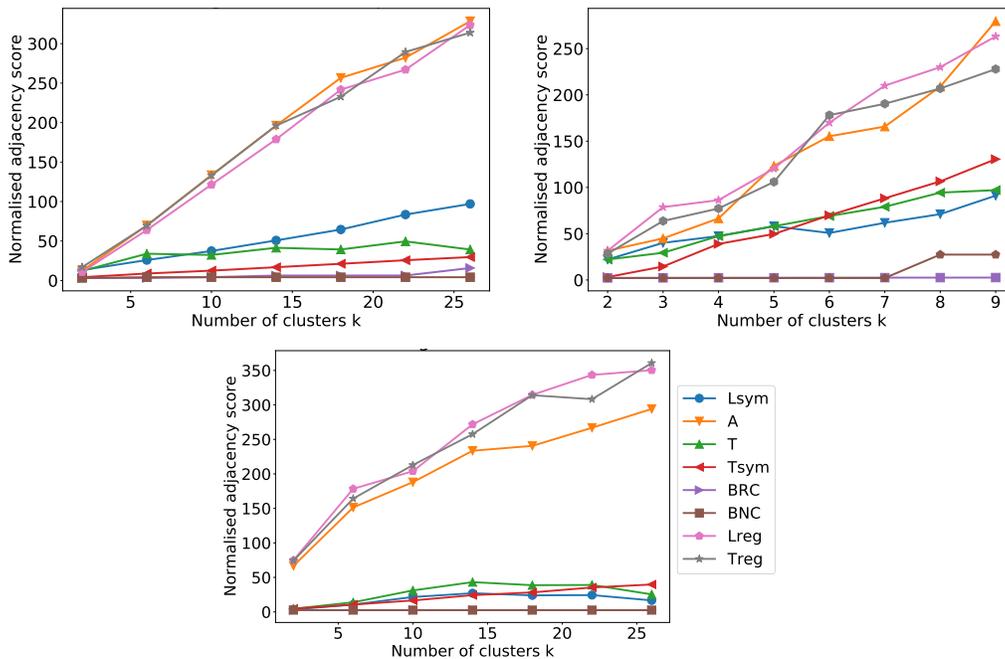
Figure 6: Objective clustering scores attained by the different spectral clustering algorithms, as we vary the number of clusters k on the Wikipedia (top left panel), the Bitcoin (top right panel), and Slashdot (bottom panel) data sets.

regularizing techniques based on powers of adjacency matrices or certain graph distance matrices, in the context of sparse signed graphs.

Yet another approach is to consider a pre-processing stage that performs low-rank matrix completion on the adjacency matrix, whose output could subsequently be used as input for our proposed algorithms. An extension of the Cheeger inequality to the setting of signed graphs, analogue to the generalized Cheeger inequality previously explored in Cucuringu et al. (2016), is another interesting research question. Extensions to the time-dependent setting and online clustering (Liberty et al., 2016; Mansfield et al., 2018), or when covariate information is available (Yan and Sarkar, 2020), are further research directions worth exploring, well motivated by real world applications involving signed networks.

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

Emmanuel Abbe, Enric Boix-Adserà, Peter Ralli, and Colin Sandon. Graph powering and spectral robustness. *SIAM Journal on Mathematics of Data Science*, 2(1):132–157, 2020.

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

Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

Arash A. Amini, Aiyou Chen, Peter J. Bickel, and Elizaveta Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41 (4):2097–2122, 2013.

David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics.

Afonso S. Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab.*, 44(4):2479–2506, 07 2016.

Afonso S Bandeira, Amit Singer, and Daniel A Spielman. A Cheeger inequality for the graph Connection Laplacian. *SIAM Journal on Matrix Analysis and Applications*, 34(4): 1611–1630, 2013.

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.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Mach. Learn.*, 56 (1-3):89–113, June 2004.

R. Bhatia. *Matrix Analysis*. Springer New York, 1996.

Francesco Bonchi, Edoardo Galimberti, Aristides Gionis, Bruno Ordozgoiti, and Giancarlo Ruffo. Discovering polarized communities in signed networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 961–970, 2019.

Kamalika Chaudhuri, Fan Chung, and Alexander Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. In *25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 35.1–35.23, Edinburgh, Scotland, 2012. JMLR Workshop and Conference Proceedings.

Kai-Yang Chiang, Joyce Jiyoung Whang, and Inderjit S. Dhillon. Scalable clustering of signed networks using balance normalized cut. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, page 615–624, New York, NY, USA, 2012. Association for Computing Machinery.

Kai-Yang Chiang, Cho-Jui Hsieh, Nagarajan Natarajan, Inderjit S. Dhillon, and Ambuj Tewari. Prediction and clustering in signed networks: A local to global perspective. *Journal of Machine Learning Research*, 15:1177–1213, 2014.

Lingyang Chu, Zhefeng Wang, Jian Pei, Jiannan Wang, Zijin Zhao, and Enhong Chen. Finding gangs in war from signed networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1505–1514, 2016.

Fan Chung. Laplacians and the Cheeger inequality for directed graphs. *Annals of Combinatorics*, 9(1):1–19, 2005.

Fan Chung and Mary Radcliffe. On the spectra of general random graphs. *Electronic Journal of Combinatorics*, 18(1):Paper 215, 14, 2011.

Fan RK Chung. Laplacians of graphs and Cheeger's inequalities. *Combinatorics, Paul Erdos is Eighty*, 2(157-172):13–2, 1996.

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

Mihai Cucuringu, Ioannis Koutis, Sanjay Chawla, Gary Miller, and Richard Peng. Simple and scalable constrained clustering: a generalized spectral method. *Artificial Intelligence and Statistics Conference (AISTATS) 2016*, 51:445–454, 2016.

Mihai Cucuringu, Peter Davies, Aldo Glielmo, and Hemant Tyagi. SPONGE: A generalized eigenproblem for clustering signed networks. In *Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1088–1098. PMLR, 2019.

Mihai Cucuringu, Andrea Pizzoferrato, and Yves van Gennip. An MBO scheme for clustering and semi-supervised clustering of signed networks. *Communications in Mathematical Sciences*, 19(1):73–109, 2021.

Sanjoy Dasgupta. The hardness of k-means clustering. Technical Report CS2007-0890, University of California, San Diego, 2008.

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

Erik D. Demaine, Dotan Emanuel, Amos Fiat, and Nicole Immorlica. Correlation clustering in general weighted graphs. *Theor. Comput. Sci.*, 361(2):172–187, September 2006.

Tyler Derr, Yao Ma, and Jiliang Tang. Signed graph convolutional networks. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 929–934. IEEE, 2018.

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

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.

Jean Gallier. Spectral theory of unsigned and signed graphs. applications to graph clustering: a survey. *CoRR*, abs / 1601.04692:1–122, 2016.

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

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, 2015.

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.

Roberto Imbuzeiro Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. *arXiv e-prints*, art. arXiv:0911.0600, November 2009.

Jiashun Jin. Fast community detection by SCORE. *The Annals of Statistics*, 43(1):57 – 89, 2015.

Antony Joseph and Bin Yu. Impact of regularization on spectral clustering. *Annals of Statistics*, 44(4):1765–1791, 2016.

Jinhong Jung, Woojeong Jin, Lee Sael, and U Kang. Personalized ranking in signed networks using signed random walk with restart. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 973–978. IEEE, 2016.

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

A. Kumar, Y. Sabharwal, and S. Sen. A simple linear time $(1 + \varepsilon)$-approximation algorithm for k-means clustering in any dimensions. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 454–462, 2004.

Srijan Kumar, Francesca Spezzano, and VS Subrahmanian. Accurately detecting trolls in Slashdot Zoo via decluttering. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 188–195. IEEE, 2014.

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

Jérôme Kunegis, Stephan Schmidt, Andreas Lommatzsch, Jürgen Lerner, Ernesto W. De Luca, and Sahin Albayrak. *Spectral Analysis of Signed Graphs for Clustering, Prediction and Visualization*, pages 559–570. SIAM, 2010.

Can M. Le, Elizaveta Levina, and Roman Vershynin. Sparse random graphs: regularization and concentration of the laplacian, 2015.

Can M. Le, Elizaveta Levina, and Roman Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 51(3):538–561, 2017.

James R Lee, Shayan Oveis Gharan, and Luca Trevisan. Multiway spectral partitioning and higher-order Cheeger inequalities. *Journal of the ACM (JACM)*, 61(6):1–30, 2014.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237, 02 2015.

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

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

Ren-Cang Li. On perturbations of matrix pencils with real spectra. *Mathematics of Computation*, 62(205):231–265, 1994.

Xiaoming Li, Hui Fang, and Jie Zhang. Supervised user ranking in signed social networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 184–191, 2019.

Yu Li, Yuan Tian, Zhang Jiawei, and Yi Chang. Learning signed network embedding via graph attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:4772–4779, 04 2020.

Edo Liberty, Ram Sriharsha, and Maxim Sviridenko. An algorithm for online k-means clustering. In *2016 Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 81–89. SIAM, 2016.

Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is NP-hard. *Theoretical Computer Science*, 442:13 – 21, 2012.

Philip Andrew Mansfield, Quan Wang, Carlton Downey, Li Wan, and Ignacio Lopez Moreno. Links: A high-dimensional online clustering method, 2018.

Pedro Mercado, Francesco Tudisco, and Matthias Hein. Clustering Signed Networks with the Geometric Mean of Laplacians. In *Advances in Neural Information Processing Systems 29*, pages 4421–4429. Curran Associates, Inc., 2016.

Pedro Mercado, Francesco Tudisco, and Matthias Hein. Spectral clustering of signed graphs via matrix power means. In *36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4526–4536, Long Beach, California, USA, 2019. PMLR.

Ekaterina Merkurjev, Justin Sunu, and Andrea L. Bertozzi. Graph MBO method for multi-class segmentation of hyperspectral stand-off detection video. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 689–693. IEEE, 2014.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA, 2005.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 849–856. MIT Press, 2001.

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.

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.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878 – 1915, 2011.

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. Numerical Analysis 2000. Vol. III: Linear Algebra.

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

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.

Ludovic Stephan and Laurent Massoulié. Robustness of spectral methods for community detection. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2831–2860, Phoenix, USA, 2019. PMLR.

G.W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

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

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

Han Xiao, Bruno Ordozgoiti, and Aristides Gionis. Searching for polarization in signed graphs: a local spectral approach. *Proceedings of The Web Conference 2020*, pages 362–372, 2020. doi: 10.1145/3366423.3380121.

Bowei Yan and Purnamrita Sarkar. Covariate regularized community detection in sparse graphs. *Journal of the American Statistical Association*, 0(0):1–12, 2020.

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

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Yilin Zhang and Karl Rohe. Understanding regularized spectral clustering via graph conductance. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Zhixin Zhou and Arash A. Amini. Analysis of spectral clustering algorithms for community detection: the general bipartite setting, 2018.

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.

Appendix A. Useful concentration inequalities

A.1 Chernoff bounds

Recall the following Chernoff bound for sums of independent Bernoulli random variables.

Theorem 52 ((Mitzenmacher and Upfal, 2005, Corollary 4.6)) *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with $\mathbb{P}[X_i = 1] = p_i$. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}[X]$. For $\delta \in (0,1)$, it holds true that*

$$\mathbb{P}[|X - \mu| \geq \delta\mu] \leq 2\exp(-\mu\delta^2/3).$$

A.2 Spectral norm of random matrices

We will make use of the following result for bounding the spectral norm of symmetric matrices with independent, centered and bounded random variables.

Theorem 53 ((Bandeira and van Handel, 2016, Corollary 3.12, Remark 3.13)) *Let X be an $n \times n$ symmetric matrix whose entries X_{ij} $(i \leq j)$ are independent, centered random variables. There there exists for any $0 < \varepsilon \leq 1/2$ a universal constant c_ε such that for every $t \geq 0$,*

$$\mathbb{P}\left[\|X\| \geq (1+\varepsilon)2\sqrt{2}\widetilde{\sigma} + t\right] \leq n\exp\left(-\frac{t^2}{c_\varepsilon\widetilde{\sigma}_*^2}\right) \tag{91}$$

where

$$\widetilde{\sigma} := \max_i \sqrt{\sum_j \mathbb{E}[X_{ij}^2]}, \quad \widetilde{\sigma}_* := \max_{i,j} \|X_{ij}\|_\infty.$$

Note that it suffices to employ upper bound estimates on $\widetilde{\sigma}, \widetilde{\sigma}_*$ in (91). Indeed, if $\widetilde{\sigma} \leq \widetilde{\sigma}^{(u)}$ and $\widetilde{\sigma}_* \leq \widetilde{\sigma}_*^{(u)}$, then

$$\mathbb{P}\left[\|X\| \geq (1+\varepsilon)2\sqrt{2}\widetilde{\sigma}^{(u)} + t\right] \leq \mathbb{P}\left[\|X\| \geq (1+\varepsilon)2\sqrt{2}\widetilde{\sigma} + t\right] \leq n\exp\left(-\frac{t^2}{c_\varepsilon\widetilde{\sigma}_*^2}\right)$$

$$\leq n\exp\left(-\frac{t^2}{c_\varepsilon(\widetilde{\sigma}_*^{(u)})^2}\right).$$

A.3 A graph decomposition result

The following graph decomposition result for inhomogeneous Erdős-Rényi graphs was established in (Le et al., 2017, Theorem 2.6).

Theorem 54 *(Le et al., 2017, Theorem 2.6) Let A be a directed adjacency matrix sampled from an inhomogeneous Erdős-Rényi $G(n, (p_{jj'})_{j,j'})$ model and let $d = n\max_{j,j'} p_{jj'}$. For any $r \geq 1$, with probability at least $1 - 3n^{-r}$, the set of edges $[n] \times [n]$ can be partitioned into three classes $\mathcal{N}, \mathcal{R}$ and $\mathcal{C}$, such that*

1. *the signed adjacency matrix concentrates on $\mathcal{N}$*

$$\|(A - \mathbb{E}A)_{\mathcal{N}}\| \leq Cr^{3/2}\sqrt{d},$$

2. *$\mathcal{R}$ (resp. $\mathcal{C}$) intersects at most n/d columns (resp. rows) of $[n] \times [n]$,*

3. *each row (resp. column) of $A_{\mathcal{R}}$ (resp. $A_{\mathcal{C}}$) have at most $32r$ non-zero entries.*

Appendix B. Matrix perturbation analysis

In this section, we recall several standard tools from matrix perturbation theory for studying the perturbation of the spectra of Hermitian matrices. The reader is referred to Stewart and Sun (1990) for a more comprehensive overview of this topic.

Let $A \in \mathbb{C}^{n \times n}$ be Hermitian with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and corresponding eigenvectors $v_1, v_2, \ldots, v_n \in \mathbb{C}^n$. Let $\widetilde{A} = A + W$ be a perturbed version of A, with the perturbation matrix $W \in \mathbb{C}^{n \times n}$ being Hermitian. Let us denote the eigenvalues of $\widetilde{A}$ and W by $\widetilde{\lambda}_1 \geq \cdots \geq \widetilde{\lambda}_n$, and $\epsilon_1 \geq \epsilon_2 \geq \cdots \geq \epsilon_n$, respectively.

To begin with, one can quantify the perturbation of the eigenvalues of $\widetilde{A}$ with respect to the eigenvalues of A. Weyl's inequality (Weyl, 1912) is a very useful result in this regard.

Theorem 55 (Weyl's Inequality (Weyl, 1912)) *For each $i = 1, \ldots, n$, it holds that*

$$\lambda_i + \epsilon_n \leq \widetilde{\lambda}_i \leq \lambda_i + \epsilon_1. \tag{92}$$

In particular, this implies that $\widetilde{\lambda}_i \in [\lambda_i - \|W\|, \lambda_i + \|W\|]$.

One can also quantify the perturbation of the subspace spanned by eigenvectors of A, which was established by Davis and Kahan (1970). Before introducing the theorem, we need some definitions. Let $U, \widetilde{U} \in \mathbb{C}^{n \times k}$ (for $k \leq n$) have orthonormal columns respectively, and let $\sigma_1 \geq \cdots \geq \sigma_k$ denote the singular values of $U^* \widetilde{U}$. Also, let us denote $\mathcal{R}(U)$ to be the range space of the columns of U, and similarly for $\mathcal{R}(\widetilde{U})$. Then the k principal angles between $\mathcal{R}(U), \mathcal{R}(\widetilde{U})$ are defined as $\theta_i := \cos^{-1}(\sigma_i)$ for $1 \leq i \leq k$, with each $\theta_i \in [0, \pi/2]$. It is usual to define $k \times k$ diagonal matrices $\Theta(\mathcal{R}(U), \mathcal{R}(\widetilde{U})) := \mathrm{diag}(\theta_1, \ldots, \theta_k)$ and $\sin \Theta(\mathcal{R}(U), \mathcal{R}(\widetilde{U})) := \mathrm{diag}(\sin \theta_1, \ldots, \sin \theta_k)$. Denoting $||| \cdot |||$ to be any unitarily invariant norm (Frobenius, spectral, etc.), the following relation holds (see for eg., (Li, 1994, Lemma 2.1), (Stewart and Sun, 1990, Corollary I.5.4)).

$$||| \sin \Theta(\mathcal{R}(U), \mathcal{R}(\widetilde{U})) ||| = |||(I - \widetilde{U}\widetilde{U}^*)U|||.$$

With the above notation in mind, we now introduce a version of the Davis-Kahan theorem taken from (Yu et al., 2015, Theorem 1) (see also (Stewart and Sun, 1990, Theorem V.3.6)).

Theorem 56 (Davis-Kahan) *Fix $1 \leq r \leq s \leq n$, let $d = s - r + 1$, and let $U = (u_r, u_{r+1}, \ldots, u_s) \in \mathbb{C}^{n \times d}$ and $\widetilde{U} = (\widetilde{u}_r, \widetilde{u}_{r+1}, \ldots, \widetilde{u}_s) \in \mathbb{C}^{n \times d}$. Write*

$$\delta = \inf \left\{ \left| \hat{\lambda} - \lambda \right| : \lambda \in [\lambda_s, \lambda_r], \hat{\lambda} \in (-\infty, \widetilde{\lambda}_{s+1}] \cup [\widetilde{\lambda}_{r-1}, \infty) \right\}$$

where we define $\widetilde{\lambda}_0 = \infty$ and $\widetilde{\lambda}_{n+1} = -\infty$ and assume that $\delta > 0$. Then

$$||| \sin \Theta(\mathcal{R}(U), \mathcal{R}(\widetilde{U})) ||| = |||(I - \widetilde{U}\widetilde{U}^*)U||| \leq \frac{|||W|||}{\delta}.$$

For instance, if $r = s = j$, then by using the spectral norm $\|\cdot\|$, we obtain

$$\sin \Theta(\mathcal{R}(\widetilde{v}_j), \mathcal{R}(v_j)) = \left\|(I - v_j v_j^*)\widetilde{v}_j\right\| \leq \frac{\|W\|}{\min \left\{ \left| \widetilde{\lambda}_{j-1} - \lambda_j \right|, \left| \widetilde{\lambda}_{j+1} - \lambda_j \right| \right\}}. \tag{93}$$

Finally, we recall the following standard result which states that given any pair of k-dimensional subspaces with orthonormal basis matrices $U, \widetilde{U} \in \mathbb{R}^{n \times k}$, there exists an alignment of $U, \widetilde{U}$ with the error after alignment bounded by the distance between the subspaces. We provide the proof for completeness.

Proposition 57 *Let $U, \widetilde{U} \in \mathbb{R}^{n \times k}$ respectively consist of orthonormal vectors. Then there exists a $k \times k$ rotation matrix O such that*

$$\left\| \widetilde{U} - UO \right\| \leq 2 \left\| (I - UU^T)\widetilde{U} \right\|.$$

Proof Write the SVD as $U^T \widetilde{U} = V\Sigma(V')^T$, where we recall that the ith largest singular value $\sigma_i = \cos\theta_i$ with $\theta_i \in [0, \pi/2]$ denoting the principal angles between $\mathcal{R}(U)$ and $\mathcal{R}(\widetilde{U})$. Choosing $O = V(V')^T$, we then obtain

$$\begin{aligned}
\left\| \widetilde{U} - UV(V')^T \right\| &\leq \left\| \widetilde{U} - UU^T\widetilde{U} \right\| + \left\| UU^T\widetilde{U} - UV(V')^T \right\| \\
&= \left\| (I - UU^T)\widetilde{U} \right\| + \left\| U^T\widetilde{U} - V(V')^T \right\| \\
&= \left\| (I - UU^T)\widetilde{U} \right\| + \| I - \Sigma \| \\
&\leq 2 \left\| (I - UU^T)\widetilde{U} \right\|,
\end{aligned}$$

where the last inequality follows from the fact $\| I - \Sigma \| = 1 - \cos\theta_k \leq \sin\theta_k$. ∎

Appendix C. Summary of main technical tools

This section collects certain technical results that were used in the course of proving our main results.

Proposition 58 ((Bhatia, 1996, Theorem X.1.1)) *For matrices $A, B \succ 0$,*

$$\left\| A^{1/2} - B^{1/2} \right\| \leq \| A - B \|^{1/2}$$

holds as $(\cdot)^{1/2}$ is operator monotone.

Proposition 59 *For symmetric matrices A^+, A^-, B^+ and B^- where $A^-, B^- \succ 0$, the following holds.*

$$\begin{aligned}
&\left\| (A^-)^{-1/2} A^+ (A^-)^{-1/2} - (B^-)^{-1/2} B^+ (B^-)^{-1/2} \right\| \\
&\leq \| (A^-)^{-1} \| \, \| A^+ \| \left(\left\| I - (B^-)^{-1/2}(A^-)^{1/2} \right\|^2 + 2 \left\| I - (B^-)^{-1/2}(A^-)^{1/2} \right\| \right) \\
&\qquad + \| (B^-)^{-1} \| \, \| A^+ - B^+ \| \\
&\leq \| (A^-)^{-1} \| \, \| A^+ \| \left(\| (B^-)^{-1} \| \, \| (B^-) - (A^-) \| + 2 \left\| (B^-)^{-1/2} \right\| \, \| (B^-) - (A^-) \|^{1/2} \right) \\
&\qquad + \| (B^-)^{-1} \| \, \| A^+ - B^+ \| .
\end{aligned}$$

Proof

$$\left\| (A^-)^{-1/2} A^+ (A^-)^{-1/2} - (B^-)^{-1/2} B^+ (B^-)^{-1/2} \right\|$$

$$= \left\| (A^-)^{-1/2} A^+ (A^-)^{-1/2} - (B^-)^{-1/2} A^+ (B^-)^{-1/2} \right.$$

$$\left. + (B^-)^{-1/2} A^+ (B^-)^{-1/2} - (B^-)^{-1/2} B^+ (B^-)^{-1/2} \right\|$$

$$\leq \left\| (B^-)^{-1/2} (A^+ - B^+)(B^-)^{-1/2} \right\| + \left\| (A^-)^{-1/2} A^+ (A^-)^{-1/2} - (B^-)^{-1/2} A^+ (B^-)^{-1/2} \right\| .$$

Now, we bound the two terms separately. The first term is easy to bound.

$$\left\| (B^-)^{-1/2} (A^+ - B^+)(B^-)^{-1/2} \right\| \leq \left\| (B^-)^{-1/2} \right\| \left\| A^+ - B^+ \right\| \left\| (B^-)^{-1/2} \right\|$$

$$= \left\| (B^-)^{-1} \right\| \left\| A^+ - B^+ \right\| . \tag{94}$$

To bound the second term, we do the following manipulations,

$$\left\| (A^-)^{-1/2} A^+ (A^-)^{-1/2} - (B^-)^{-1/2} A^+ (B^-)^{-1/2} \right\|$$

$$= \left\| (A^-)^{-1/2} A^+ (A^-)^{-1/2} - (A^-)^{-1/2} (A^-)^{1/2} (B^-)^{-1/2} A^+ (B^-)^{-1/2} (A^-)^{1/2} (A^-)^{-1/2} \right\|$$

$$= \left\| (A^-)^{-1/2} \left(A^+ - (A^-)^{1/2} (B^-)^{-1/2} A^+ (B^-)^{-1/2} (A^-)^{1/2} \right) (A^-)^{-1/2} \right\|$$

$$= \left\| (A^-)^{-1/2} \left(A^+ - \left((A^-)^{1/2} (B^-)^{-1/2} - I + I \right) A^+ \left((B^-)^{-1/2} (A^-)^{1/2} - I + I \right) \right) (A^-)^{-1/2} \right\|$$

$$= \left\| (A^-)^{\frac{-1}{2}} \left(\left((A^-)^{\frac{1}{2}} (B^-)^{-\frac{1}{2}} - I \right) A^+ \left((B^-)^{-\frac{1}{2}} (A^-)^{\frac{1}{2}} - I \right) + A^+ \left((B^-)^{-\frac{1}{2}} (A^-)^{\frac{1}{2}} - I \right) \right. \right.$$

$$\left. \left. + \left((A^-)^{\frac{1}{2}} (B^-)^{-\frac{1}{2}} - I \right) A^+ \right) (A^-)^{-\frac{1}{2}} \right\|$$

$$\leq \left\| (A^-)^{-1} \right\| \left\| A^+ \right\| \left(\left\| I - (B^-)^{-1/2} (A^-)^{1/2} \right\|^2 + 2 \left\| I - (B^-)^{-1/2} (A^-)^{1/2} \right\| \right) . \tag{95}$$

The first inequality of the lemma follows by adding (95) and (94). To see the second inequality of the lemma, observe that,

$$\left\| I - (B^-)^{-1/2} (A^-)^{1/2} \right\| = \left\| (B^-)^{-1/2} \left((B^-)^{1/2} - (A^-)^{1/2} \right) \right\|$$

$$\leq \left\| (B^-)^{-1/2} \right\| \left\| (B^-)^{1/2} - (A^-)^{1/2} \right\|$$

$$\leq \left\| (B^-)^{-1/2} \right\| \left\| B^- - A^- \right\|^{1/2} \quad \text{(using Proposition 58)} . \tag{96}$$

The second inequality of the lemma follows by substituting (96) in the first inequality of the lemma.

∎

Appendix D. Proofs from Section 4

Lemma 60 (Expression for C_e^+ & C_e^-)

$$C_e^+ = -p\eta \frac{n}{d^+} \chi_1 \chi_1^\top + \left(1 + \tau^- + \frac{p}{d^+} \left(1 - \eta - \frac{n}{k}(1 - 2\eta) \right) \right) I ,$$

$$C_e^- = -p(1-\eta)\frac{n}{d^-}\chi_1\chi_1^\top + \left(1+\tau^+ + \frac{p}{d^-}\left(\eta + \frac{n}{k}(1-2\eta)\right)\right)I\,.$$

It follows that can be written as $C_e^+ = R\Sigma^+ R^\top$ and $C_e^- = R\Sigma^- R^\top$, where R is a rotation matrix, and

$$\Sigma^+ = \begin{bmatrix} \left(1+\tau^- + \frac{p}{d^+}\left(1-\eta-n\left(\eta+\frac{1-2\eta}{k}\right)\right)\right) & \\ & \left(1+\tau^- + \frac{p}{d^+}\left(1-\eta-n\left(\frac{1-2\eta}{k}\right)\right)\right)I_{k-1} \end{bmatrix},$$

$$\Sigma^- = \begin{bmatrix} \left(1+\tau^+ + \frac{p}{d^-}\left(\eta-n\left(1-\eta-\frac{1-2\eta}{k}\right)\right)\right) & \\ & \left(1+\tau^+ + \frac{p}{d^-}\left(\eta+n\left(\frac{1-2\eta}{k}\right)\right)\right)I_{k-1} \end{bmatrix}.$$

The above lemma shows that we know the spectrum of $(C^-)^{-1/2}C^+(C^-)^{-1/2}$ exactly, in the case of equal-sized clusters.

Proof [Proof of Lemma 22] From (17) it follows that,

$$\lambda_{\max}(C^+) \le \max_{i\in[k]}\left(1+\tau^- + \frac{p}{d_i^+}(1-\eta-n_i(1-2\eta))\right)\,.$$

The maximum is achieved for the smallest sized cluster. This shows the proof for (33).

The proof of (34) follows from the fact that in (24) we had decomposed the matrix $\overline{L_{sym}^-} + \tau^+ I$ as a block-diagonal matrix, with block of $C^-, \alpha_1^- I_{n_1-1}, \ldots, \alpha_k^- I_{n_k-1}$. Since $\overline{L_{sym}^-}$ is a symmetric Laplacian, we know that $\lambda_{\min}(\overline{L_{sym}^-} + \tau^+ I) = \tau^+$. Also, $\alpha_i^- > \tau^+$ for $i \in [k]$. Thus the equation follows. ∎

Appendix E. Spectrum of Signed Laplacians

This section extends some classical results for the unsigned Laplacian to the symmetric Signed Laplacian and the regularized Laplacian.

Lemma 61 *For all $x \in \mathbb{R}^n$,*

$$x^T\overline{L_{sym}}x = \frac{1}{2}\sum_{j,j'}|A_{jj'}|\left(\frac{x_j}{\sqrt{d_j}} - sgn(A_{jj'})\frac{x_{j'}}{\sqrt{d_{j'}}}\right)^2 \tag{97}$$

Moreover, the eigenvalues of $\overline{L_{sym}}$ and L_γ are in the interval $[0,2]$.

Proof Equation (97) is adapted from Proposition 5.2 from Gallier (2016) and is obtained by replacing x by $\overline{D}^{-1/2}x$. The second part of the lemma comes from the fact that $(a\pm b)^2 \le 2(a^2+b^2)$. In fact, for $x \in \mathbb{R}^n$ such that $\|x\| = 1$, we have

$$x^T\overline{L_{sym}}x \le \sum_{j,j'}|A_{jj'}|\left(\frac{x_j^2}{d_j} + \frac{x_{j'}^2}{d_{j'}}\right)$$

$$= 2\sum_{j,j'}|A_{jj'}|\frac{x_j^2}{d_j} = 2\sum_j x_j^2 = 2.$$

Similarly, we have

$$x^T L_\gamma x \leq \sum_{j,j'} |(A_\gamma)jj'| \left(\frac{x_j^2}{\overline{D}_{jj} + \gamma} + \frac{x_{j'}^2}{\overline{D}_{j'j'} + \gamma} \right)$$

$$\leq 2 \sum_{j,j'} (|A_{jj'}| + \frac{\gamma}{n}) \frac{x_j^2}{\overline{D}_{jj} + \gamma}$$

$$= 2 \sum_j \frac{(\overline{D}_{jj} + \gamma) x_j^2}{\overline{D}_{jj} + \gamma} = 2.$$

Moreover $\overline{L_{sym}}$ and L_γ are positive semi-definite, thus we can conclude that their eigenvalues are between 0 and 2. ∎