Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Bias-Adjusted Spectral Clustering in Multi-Layer Stochastic Block Models

Jing Lei & Kevin Z. Lin

To cite this article: Jing Lei & Kevin Z. Lin (2022): Bias-Adjusted Spectral Clustering in Multi-Layer Stochastic Block Models, Journal of the American Statistical Association, DOI: 10.1080/01621459.2022.2054817

To link to this article: https://doi.org/10.1080/01621459.2022.2054817

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–13: Theory and Methods
https://doi.org/10.1080/01621459.2022.2054817

Bias-Adjusted Spectral Clustering in Multi-Layer Stochastic Block Models

Jing Lei[a] and Kevin Z. Lin[b]

[a]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA; [b]Department of Statistics, Wharton School of Business, University of Pennsylvania, Philadelphia, PA

ABSTRACT

We consider the problem of estimating common community structures in multi-layer stochastic block models, where each single layer may not have sufficient signal strength to recover the full community structure. In order to efficiently aggregate signal across different layers, we argue that the sum-of-squared adjacency matrices contain sufficient signal even when individual layers are very sparse. Our method uses a bias-removal step that is necessary when the squared noise matrices may overwhelm the signal in the very sparse regime. The analysis of our method relies on several novel tail probability bounds for matrix linear combinations with matrix-valued coefficients and matrix-valued quadratic forms, which may be of independent interest. The performance of our method and the necessity of bias removal is demonstrated in synthetic data and in microarray analysis about gene co-expression networks. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received March 2021
Accepted March 2022

KEYWORDS
Community detection; Gene co-expression network; Matrix concentration inequalities; Network data; Spectral clustering; Stochastic block models

1. Introduction

A network records the interactions among a collection of individuals, such as gene coexpression, functional connectivity among brain regions, and friends on social media platforms. In the simplest form, a network can be represented by a binary symmetric matrix $A \in \{0,1\}^{n\times n}$ where each row/column represents an individual and the (i,j)entry of A represents the presence/absence of interaction between the two individuals. In the more general case, A_{ij} may take values in $\mathbb{R}^1$ to represent different magnitudes or counts of the interaction. We refer to Kolaczyk (2009), Newman (2009), and Goldenberg et al. (2010) for general introduction of statistical analysis of network data.

In many applications, the interaction between individuals are recorded multiple times, resulting in multi-layer network data. For example, in this article, we study the temporal gene coexpression networks in the medial prefrontal cortex of rhesus monkeys at 10 different developmental stages (Bakken et al. 2016). The medial prefrontal cortex is believed to be related to developmental brain disorders, and many of the genes we study are suspected to be associated with autism spectrum disorder at different stages of development. Other examples of multi-layer network data are brain imaging, where we may infer one set of interactions among different brain regions from electroencephalography (EEG), and another set of interactions using resting-state functional magnetic resonance imaging (fMRI) measures. Similarly, one may expect the brain regions to form groups in terms of connectivity. The wide applicability and rich structures of multi-layer networks make it an active research area in the statistics, machine learning, and signal processing community. See Tang, Lu, and Dhillon (2009), Dong et al. (2012), Kivelä et al. (2014), Xu and Hero (2014), Han, Xu, and Airoldi (2015), Zhang and Cao (2017), Matias and Miele (2017) and references within.

In this article, we study multi-layer network data through the lens of multi-layer stochastic block models, where we observe many simple networks on a common set of nodes. The stochastic block model (SBM) and its variants (Holland, Laskey, and Leinhardt 1983; Bickel and Chen 2009; Karrer and Newman 2011; Airoldi et al. 2008) are an important prototypical class of network models that allow us to mathematically describe the community structure and understand the performance of popular algorithms such as spectral clustering (McSherry 2001; Rohe, Chatterjee, and Yu 2011; Jin 2015; Lei and Rinaldo 2015) and other methods (Latouche, Birmele, and Ambroise 2012; Peixoto 2013; Abbe and Sandon 2015). Roughly speaking, in an SBM, the nodes in a network are partitioned into disjoint communities (i.e., clusters), and nodes in the same community have similar connectivity patterns with other nodes. A key inference problem in the study of SBM is estimating the community memberships given an observed network.

Compared to an individual layer, a multi-layer network contains more data and hopefully enables us to extract salient structures, such as communities, more easily. On the other hand, new methods must be developed in order to efficiently combine the signal from individual layers. To demonstrate the necessity for these methods, we plot the observed gene coexpression networks collected from Bakken et al. (2016) in Figure 1. The three networks correspond to gene co-expression patterns within the medial prefrontal cortex tissue of rhesus monkeys collected at

CONTACT Jing Lei *jinglei@andrew.cmu.edu* Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

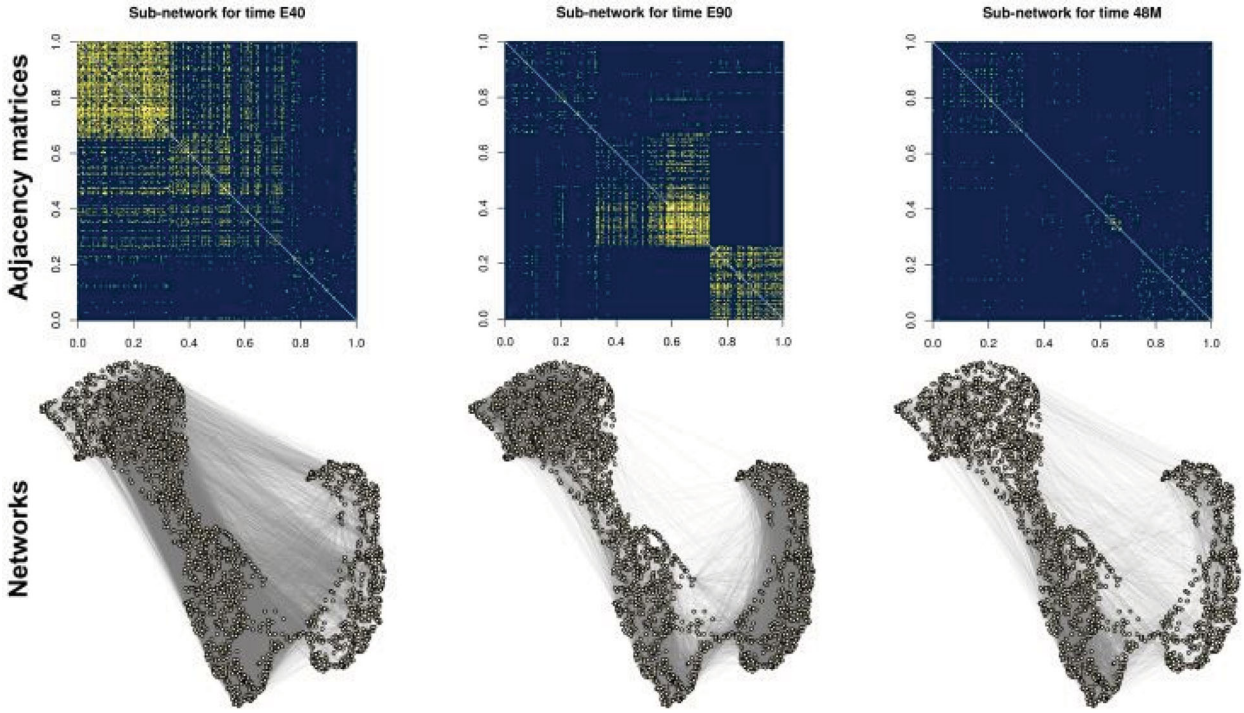

Figure 1. The adjacency matrix (top row, yellow denoting the presence of an edge and blue denoting the lack of) and the corresponding network (bottom row) for three different developmental times of the rhesus monkey's gene coexpression in the medial prefrontal cortex based on selected set of genes to visually demonstrate the varying network structures. Likewise, the ordering of the genes in the adjacency matrices is chosen to visually demonstrate the clustering structure, and persist throughout all three adjacency matrices. The three developmental times are E40, E90 (for 40 or 90 days in the embryo), and 48M (for 48 months after birth), corresponding to the pair of plots on the left to the pair of plots on the right. The full dataset is analyzed in Section 6.

different stages of development. We plot only the sub-network formed by a small collection genes for simplicity. A quick visual inspection across the three networks suggests that the genes can be approximately divided into four common communities (i.e., clusters that persist throughout all three networks), where genes in the same community exhibit similar connectivity patterns. However, different gene communities are more visually apparent in different layers. For example, in the layer labeled as "E40" (for tissue collected 40 days of development in the embryo), the last three communities are indistinguishable. In contrast, in the layer labeled as "E90," the first community is less distinguishable, and in the layer labeled "48M" (for the tissue collected 48 months after birth), nearly all of the communities are indistinguishable. These qualitative observations are of scientific interest since these time-dependent densely-connected communities are evidence of "gene coordination," a biological concept that describes when a community of genes is synchronized in ramping up or down in gene expression at certain stages of development (Paul et al. 2012; Werling et al. 2020). Hence, we can infer two potential advantages of analyzing such multi-layer network data in an aggregated manner. First, an aggregated analysis is able to reveal global structures that are not exhibited by any individual layer. Second, the common structure across different layers can help us to better filter out the noise, which allows us to obtain more accurate inference results. We describe the analysis in more detail and return to analyze the full dataset in Section 6.

The theoretical understanding of estimating common communities in multi-layer SBMs is relatively limited compared to those in single-layer SBMs. Bhattacharyya and Chatterjee (2018) and Paul and Chen (2020) studied variants of spectral clustering for multi-layer SBMs, but the strong theoretical guarantee requires a so-called layer-wise positivity assumption, meaning each matrix encoding the probability of an edge among the communities must have only positive eigenvalues bounded away from zero. In contrast, Pensky and Zhang (2019) studied a different variant of spectral clustering, but established estimation consistency under conditions similar to those for single-layer SBMs. These results only partially describe the benefits of multi-layer network aggregation. Alternatively, Lei, Chen, and Lynch (2019) considered a least-squares estimator, and proved consistency of the global optima for general block structures without imposing the positivity assumption for individual layers, but that method is computationally intractable in the worst-case.

The first main contribution of this article is a simple, novel, and computationally-efficient aggregated spectral clustering method for multi-layer SBMs, described in Section 2. The estimator applies spectral clustering to the sum of squared adjacency matrices after removing the bias by setting the diagonal entries to 0. In addition to its simplicity, this estimator has two appealing features. First, summing over the squared adjacency matrices enables us to prove its consistency without requiring a layer-wise positivity assumption. Second, compared with single-layer SBMs, the consistency result reflects a boost of signal strength by a factor of $L^{1/2}$, where L is the number of layers. Such a $L^{1/2}$ signal boost is comparable to that obtained in Lei, Chen, and Lynch (2019), but is now achieved by a simple and computationally tractable algorithm. The removal of the diagonal bias in the squared matrices is shown to be crucial in both theory (Section 3) and simulations (Section 5), especially in the most interesting regime where the network density is too

low for any single layer to carry sufficient signal for community estimation. Interestingly, similar diagonal-removal techniques have also been discovered and studied in other contexts, such as Gaussian mixture model clustering (Ndaoud 2018), principal components analysis (Zhang, Cai, and Wu 2022), and centered distance matrices (Székely and Rizzo 2014).

Another contribution of this article is a collection of concentration inequalities for matrix-valued linear combinations and quadratic forms. These are described in Section 4, which are an important ingredient for the aforementioned theoretical results. Specifically, an important step in analyzing our matrix-valued data is to understand the behavior of the matrix-valued measurement errors. Toward this end, many powerful concentration inequalities have been obtained for matrix operator norms under various settings, such as random matrix theory (Bai and Silverstein 2010), eigenvalue perturbation and concentration theory (Feige and Ofek 2005; O'Rourke, Vu, and Wang 2018; Lei and Rinaldo 2015; Le, Levina, and Vershynin 2017; Cape, Tang, and Priebe 2017), and matrix deviation inequalities (Bandeira and Van Handel 2016; Vershynin 2011). The matrix Bernstein inequality and related results (Tropp 2012) are also applicable to linear combinations of noise matrices with scalar coefficients. In order to provide technical tools for our multi-layer network analysis, we extend these matrix-valued concentration inequalities in two directions. First, we provide upper bounds for linear combinations of noise matrices with matrix-valued coefficients. This can be viewed as an extension of the matrix Bernstein inequality to allow for matrix-valued coefficients. Second, we provide concentration inequalities for sums of matrix-valued quadratic forms, extending the scalar case known as the Hanson–Wright inequality (Hanson and Wright 1971; Rudelson and Vershynin 2013) in several directions. A key intermediate step in relating linear cases to quadratic cases is deriving a deviation bound for matrix-valued U-statistics of order two.

2. Community Estimation in Multi-Layer SBM

Throughout this section, we describe the model, theoretical motivation, and our estimator for clustering nodes in a multi-layer SBM. Motivated by such multi-layer network data with a common community structure as demonstrated in Figure 1, we consider the L-layer SBM containing n nodes assigned to K different communities,

$$A_{\ell,ij} \sim \text{Bernoulli}(\rho B_{\ell,\theta_i\theta_j}) \quad \text{for} \quad 1 \le i < j \le n, \quad 1 \le \ell \le L, \tag{1}$$

where ℓ is the layer index, $\theta_i \in \{1,\dots,K\}$ is the membership index of node i for $i \in \{1,\dots,n\}$, $\rho \in (0,1]$ is an overall edge density parameter, and $B_\ell \in [0,1]^{K\times K}$ is a symmetric matrix of community-wise edge probabilities in layer ℓ. We assume A_ℓ is symmetric and $A_{\ell,ii} = 0$ for all $\ell \in \{1,\dots,L\}$ and $i \in \{1,\dots,n\}$.

Our statistical problem is to estimate the membership vector $\theta = (\theta_1,\dots,\theta_n) \in \{1,\dots,K\}^n$ given the observed adjacency matrices $A_1,\dots,A_L$. Let $\widehat{\theta} \in \{1,\dots,K\}^n$ be an estimated membership vector, and the estimation error is the number of mis-clustered nodes based on the Hamming distance,

$$d(\widehat{\theta},\theta) = \min_{\pi} \sum_{i=1}^{n} \mathbb{1}(\theta_i \ne \pi(\widehat{\theta}_i)), \tag{2}$$

for the indicator function $\mathbb{1}(\cdot)$, where the minimum is taken over all label permutations $\pi : \{1,\dots,K\} \mapsto \{1,\dots,K\}$. An estimator $\widehat{\theta}$ is consistent if $n^{-1}d(\widehat{\theta},\theta) = o_P(1)$.

The assumption of a fixed common membership vector θ can be relaxed to each layer having its own membership vector but close to a common one. The theoretical consequence of this relaxation is discussed in Remark 1, after the main theorem in Section 3. We assume that K is known. The problem of selecting K from the data is an important problem and will not be pursued in this article. Further discussion will be given in Section 7.

When $L = 1$, the community estimation problem for single-layer SBM is well-understood (Bickel and Chen 2009; Lei and Rinaldo 2015; Abbe 2017). If K is fixed as a constant while $n \to \infty$, $\rho \to 0$ with balanced community sizes lower bounded by a constant fraction of n, and B is a constant matrix with distinct rows, then the community memberships can be estimated with vanishing error when $n\rho \to \infty$. Practical estimators include variants of spectral clustering, message passing, and likelihood-based estimators.

As mentioned in Section 1, in the multi-layer case, consistent community estimation has been studied in some recent works. The theoretical focus is to understand how the number of layers L affects the estimation problem. Paul and Chen (2020) and Bhattacharyya and Chatterjee (2018) show that consistency can be achieved if $Ln\rho$ diverges, but under the aforementioned positivity assumption, meaning that each B_ℓ is positive definite with a minimum eigenvalue bounded away from zero. Such assumptions are plausible in networks with strong associativity patterns where nodes in the same communities are much more likely to connect to one another than nodes in different communities. But there are networks observed in practice that do not satisfy this assumption, such as those in Newman (2002) and Litvak and Van Der Hofstad (2013). See Lei (2018) and the references within for additional discussion on such positivity assumptions in a more general context. To remove the positivity assumption, Lei, Chen, and Lynch (2019) considered a least-squares estimator, and proved consistency when $L^{1/2}n\rho$ diverges (up to a small poly-logarithmic factor) and the smallest eigenvalue of $\sum_\ell B_\ell^2$ grows linearly in L. A caveat is that the least-squares estimator is computationally challenging, and in practice, one may only be able to find a local minimum using greedy algorithms.

In the following sections, we will motivate a spectral clustering method from the least-squares perspective, investigate its bias, and derive our estimator with a data-driven bias adjustment.

2.1. From Least Squares to Spectral Clustering

In this section, we motivate how least-squares estimators is well-approximated by spectral clustering, which lays down the intuition of our estimator in Section 2.3. Let $\psi \in \{1,\dots,K\}^n$ be a membership vector and $\Psi = [\Psi_1,\dots,\Psi_K]$ be the corresponding $n \times K$ membership matrix where each $\Psi_k = (\Psi_{1,k},\dots,\Psi_{n,k})^T$ is an $n \times 1$ vector with $\Psi_{i,k} = \mathbb{1}(\psi_i = k)$. Let $I_k(\psi) = \{i \in \{1,\dots,n\} : \psi_i = k\}$ and $n_k(\psi) = |I_k(\psi)|$, the size of the set $I_k(\psi)$.

The least-squares estimator of Lei, Chen, and Lynch (2019) seeks to minimize the residual sum of squares,

$$\widehat{\theta} = \operatorname{argmin}_{\psi\in\{1,\dots,K\}^n} \sum_{\ell=1}^{L} \sum_{1\le i<j\le n} (A_{\ell,ij} - \widehat{B}_{\ell,\psi_i\psi_j}(\psi))^2 \quad (3)$$

where

$$\widehat{B}_{\ell,kl}(\psi) = \begin{cases} \frac{\sum_{i,j\in I_k(\psi)} A_{\ell,ij}}{n_k(\psi)(n_k(\psi)-1)} & \text{when} \quad k=l, \\ \frac{\sum_{i\in I_k(\psi), j\in I_l(\psi)} A_{\ell,ij}}{n_k(\psi)n_l(\psi)} & \text{when} \quad k\neq l, \end{cases}$$

is the sample mean estimate of B_ℓ under a given membership vector ψ. Recall that the total-variance decomposition implies the equivalence between minimizing within-block sum of squares and maximizing between-block sum of squares. Hence, if we accept the approximation $n_k(\psi)(n_k(\psi)-1) \approx n_k^2(\psi)$, then after multiplying the least-squares objective function (3) by 2 and using the total-variance decomposition, the objective function becomes

$$\max_{\psi\in\{1,\dots,K\}^n} \sum_{\ell=1}^{L} \sum_{1\le k,l\le K} \frac{(\Psi_k^T A_\ell \Psi_l)^2}{n_k(\psi)n_l(\psi)},$$

which is equivalent to

$$\max_{\psi\in\{1,\dots,K\}^n} \sum_{\ell=1}^{L} \sum_{1\le k,l\le K} \left(\widetilde{\Psi}_k^T A_\ell \widetilde{\Psi}_l\right)^2 = \max_{\psi} \sum_{\ell=1}^{L} \left\|\widetilde{\Psi}^T A_\ell \widetilde{\Psi}\right\|_F^2,$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm, and $\widetilde{\Psi} = [\widetilde{\Psi}_1,\dots,\widetilde{\Psi}_K]$ with $\widetilde{\Psi}_k = \Psi_k/\sqrt{n_k(\psi)}$ is the column-normalized version of Ψ where each column of $\widetilde{\Psi}$ has norm 1. This means $\widetilde{\Psi}$ is orthonormal, that is, $\widetilde{\Psi}^T\widetilde{\Psi} = I_K$. The benefit of considering orthonormal matrices is that for any orthonormal matrix $U \in \mathbb{R}^{n\times K}$ and symmetric matrix $A \in \mathbb{R}^{n\times n}$,

$$\|U^T AU\|_F^2 = \operatorname{tr}(U^T AUU^T AU) \le \operatorname{tr}(U^T A^2 U).$$

The right-hand side of the above inequality is maximized by the leading K eigenvectors of A, where the eigenvalues ordered by absolute value. For this U, the inequality becomes equality. Additionally, under the multi-layer SBM, the expected values of adjacency matrices $\{P_1,\dots,P_L\}$ (where $P_\ell = \mathbb{E}A_\ell$ for $\ell \in \{1,\dots,L\}$) share roughly the same leading principal subspace as determined by the common community structure. Putting all these facts together, we intuitively expect $U = \widetilde{\Theta}$ to correspond to an approximate solution of the original least-squares problem, where $\widetilde{\Theta}$ is the column-normalized version of the true membership matrix Θ.

Therefore, a relaxation of the approximate version of the original problem (3) is

$$\max_{U\in\mathbb{R}^{n\times K}: U^TU=I_K} \operatorname{tr}\left[U^T\left(\sum_{\ell=1}^{L} A_\ell^2\right)U\right], \quad (4)$$

which is a standard spectral problem. For this reason, we often call U the "spectral embedding." The community estimation is then obtained by applying a clustering algorithm to the rows of $\widehat{U}$, a solution to (4).

2.2. The Necessity of Bias Adjustment

Let $P_\ell = \mathbb{E}A_\ell$ denote the expected adjacency matrix, meaning that P_ℓ is the matrix obtained by zeroing out the diagonal entries of $\widetilde{P}_\ell = \rho\Theta B_\ell \Theta^T$. We now show that $\sum_\ell A_\ell^2$ is a biased estimate of $\sum_\ell P_\ell^2$, and that we can correct for this bias by simply removing its diagonal entries. Let $X_\ell = A_\ell - P_\ell$ be the noise matrix. Then

$$\sum_{\ell=1}^{L} A_\ell^2 = \left(\sum_{\ell=1}^{L} P_\ell^2\right) + \left(\sum_{\ell=1}^{L}(X_\ell P_\ell + P_\ell X_\ell)\right) + S, \quad (5)$$

where $S = \sum_\ell X_\ell^2$. The first term is the signal term, with each summand close to $\widetilde{P}_\ell^2 = \rho^2\Theta B_\ell^2\Theta^T$, and will add up over the layers, because each matrix B_ℓ^2 is positive semidefinite. The second term is a mean-0 noise matrix, which can be controlled using matrix concentration inequalities developed in Section 4. The third term $S = \sum_\ell X_\ell^2$ is a squared error matrix and will also add up over the layers, which may introduce bias if the overall edge density parameter ρ is too small.

We use a simple simulation study to illustrate the necessity of bias adjustment in spectral clustering applied to the sum of squared adjacency matrices. We set $K = 2$ and consider two edge-probability matrices,

$$B^{(1)} = \begin{bmatrix} 3/4 & \sqrt{3}/8 \\ \sqrt{3}/8 & 1/2 \end{bmatrix}, \quad \text{and} \quad B^{(2)} = \begin{bmatrix} 7/8 & 3\sqrt{3}/8 \\ 3\sqrt{3}/8 & 1/8 \end{bmatrix}.$$

These two matrices are chosen such that spectral clustering applied to the sum of the adjacency matrices and the sum of squared adjacency matrices would be either sub-optimal or inconsistent in the very sparse regime. We set $n = 200$ nodes with 100 nodes in each community, the number of layers to be $L = 30$, and for each layer ℓ, B_ℓ is randomly and independently chosen from $B^{(1)}$ and $B^{(2)}$ with equal probability. We use five different values of the overall edge density parameter ρ between 0.02 and 0.06. For each value of ρ, we generate a multi-layer SBM according to (1) and apply spectral clustering to three matrices: (a) the sum of adjacency matrices without squaring (i.e., "Sum"), (b) the sum of squared adjacency matrices (i.e., "SoS"), and (c) a bias-adjusted sum of squared adjacency matrices (i.e., "Bias-adjusted SoS"), which will be introduced in the next section. The results across 100 trials are reported in Figure 2. By construction, the "Sum" method performs poorly since the sum of adjacency matrices has only one significant eigen-component, meaning the result is sensitive to noise when $K = 2$ eigenvectors are used for spectral clustering. In fact, as described in Example 1, it is also easy to generate cases in which the sum of adjacency matrices carries no signal at all. The "SoS" method also performs poorly. This is because although the sum of squared adjacency matrices contains signal for clustering, the aforementioned bias is large when ρ is small. In contrast, our method "Bias-adjusted SoS" performs the best. A more detailed simulation study is presented in Section 5.

2.3. Bias-Adjusted Sum-of-Squared Spectral Clustering

We are now ready to quantify the amount of bias, and to describe our aforementioned bias-adjusted sum-of-squared method to cluster nodes in a multi-layer SBM. From (5), we

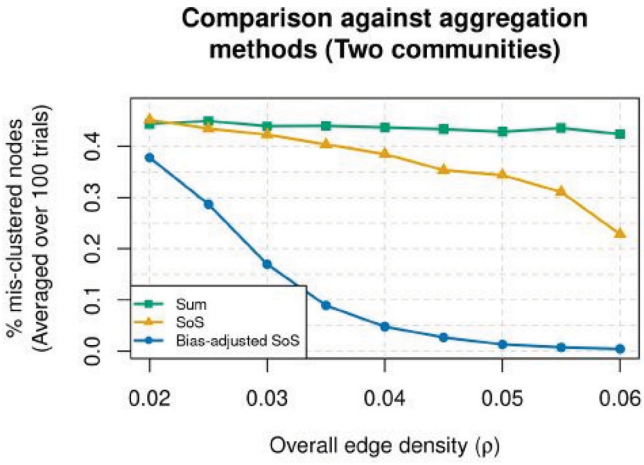

Figure 2. The average proportion of misclustered nodes for three methods (measured via Hamming distance $n^{-1}d(\widehat{\theta},\theta)$ shown in (2), averaged over 100 trials), with $n = 200$ and two equal-sized communities among overall edge densites ranging from $\rho \in [0.02, 0.06]$ and $L = 30$ layers. Three methods' performance are shown: Sum (green squares), SoS (orange triangles), and Bias-adjusted SoS (blue circles).

see that the diagonal entries of the squared error term S have positive expected value and hence may cause systematic bias in the principal subspace of $\sum_\ell A_\ell^2$. Now consider a further decomposition $S = S_1 + S_2$ where S_1 and S_2 correspond to the off-diagonal and diagonal parts of S, respectively. Observe that only the diagonal entries of S have positive expected value, so our effort will focus on removing the bias caused by S_2. Toward this end, observe that by construction, we have

$$\begin{aligned}(S_2)_{ii} = S_{ii} &= \sum_{\ell=1}^{L}\sum_{j=1}^{n} X_{\ell,ij}^2 \\ &= \sum_{\ell=1}^{L}\sum_{j=1}^{n} P_{\ell,ij}^2 \mathbb{1}(A_{\ell,ij}=0) + (1-P_{\ell,ij})^2\mathbb{1}(A_{\ell,ij}=1) \\ &\le Ln \max_{\ell,ij} P_{\ell,ij}^2 + \sum_{\ell=1}^{L} d_{\ell,i} \end{aligned} \tag{6}$$

where $d_{\ell,i} = \sum_j A_{\ell,ij}$ is the degree of node i in layer ℓ. The expected value of $\sum_\ell d_{\ell,i}$ is $\sum_{\ell,j} P_{\ell,ij} \asymp Ln \max_{\ell,ij} P_{\ell,ij}$. In the very sparse regime, $\max_{\ell,ij} P_{\ell,ij}$ is very small so $\sum_\ell d_{\ell,i}$ is the leading term in $(S_2)_{ii}$.

Combining this calculation with a key observation that $\sum_\ell d_{\ell,i}$ can be computed from the data, we arrive at the following bias-adjusted sum-of-squared spectral clustering algorithm. Let D_ℓ be the diagonal matrix consisting of the degrees of A_ℓ where $(D_\ell)_{ii} = d_{\ell,i}$. The bias-adjusted sum of squared adjacency matrices is

$$S_0 = \sum_{\ell=1}^{L} (A_\ell^2 - D_\ell)\,. \tag{7}$$

The community membership is estimated by applying a clustering algorithm to the rows of the matrix whose columns are the leading K eigenvectors of S_0 given in (7).

3. Consistency of Bias-Adjusted Sum-of-Squared Spectral Clustering

We now describe our theoretical result characterizing how multi-layer networks benefit community estimation. The hardness of community estimation is determined by many aspects of the problem, including number of communities, community sizes, number of nodes, separation of communities, and overall edge density. Here, we need to consider all of these aspects jointly across the L layers. To simplify the discussion, we primarily focus on the following setting but discuss additional settings in later remarks.

Assumption 1. (a) The number of communities K is fixed and community sizes are balanced. That is, there exists a constant c such that each community size is in $[c^{-1}n/K, cn/K]$.
(b) The relative community separation is constant. That is, $B_\ell = \rho B_{\ell,0}$ where $B_{\ell,0}$ is a $K \times K$ symmetric matrix with constant entries in $[0, 1]$. Furthermore, the minimum eigenvalue of $\sum_\ell B_{\ell,0}^2$ is at least cL for some constant $c > 0$.

Part (a) simplifies the effect of the community sizes and the number of communities. This setting has been well-studied in the SBM literature for $L = 1$ (Lei and Rinaldo 2015). Part (b) puts the focus on the effect of the overall edge density parameter ρ, and requires a linear growth of the aggregated squared edge-probability matrices in terms of the minimum eigenvalue. This is much less restrictive than the layer-wise positivity assumption used in other work mentioned in Section 2 which require each $B_{\ell,0}$ to be positive definite. We give two examples in which Assumption 1(b) is satisfied but the layer-wise positivity is not.

Example 1 (Identically distributed random layers). Consider a theoretical scenario in which the $B_{\ell,0}$'s have iid Uniform$(0, 1)$ entries subject to symmetry. It is easy to verify that the expected sum matrix $\mathbb{E}\sum_\ell B_\ell$ is a constant matrix with each entry being $L\rho/2$. Therefore, it is impossible to reconstruct the block structure from the sum of adjacency matrices $\sum_\ell A_\ell$ when ρ is small.

Example 2 (Community merge and split). Consider a more realistic scenario in which for $\{B_\ell : 1 \le \ell \le L\}$, some layers ℓ and community indices k, k' have $B_{\ell,kj} = B_{\ell,k'j}$ for all j. This can be interpreted as the merge of communities k and k' at layer ℓ. In such cases, each layer may not contain full community information, and we must aggregate the layers to recover the full community structure. In our real data example, we actually observe that in most layers, all but one or two communities merge with a large, null community, and each nonnull community is active in one or two layers.

Based on these assumptions, in the asymptotic regime $n \to \infty$ and $\rho \to 0$, it is well-known that consistent community estimation is possible for $L = 1$ when $n\rho \to \infty$. Hence, in the multi-layer setting when $L \to \infty$, one should expect a lower requirement on overall density as we aggregate information across layers. This is shown in our following result.

Theorem 1. Under Assumption 1, if $L^{1/2}n\rho \ge C_1 \log^{1/2}(L+n)$ and $n\rho \le C_2$ for a large enough positive constant C_1 and a positive constant C_2, then spectral clustering with a constant

factor approximate K-means clustering algorithm applied to S_0, the bias-adjusted sum of squared adjacency matrices in (7), correctly estimates the membership of all but a

$$C\left(\frac{1}{n^2}+\frac{\log(L+n)}{Ln^2\rho^2}\right)$$

proportion of nodes for some constant C with probability at least $1-O((L+n)^{-1})$.

An immediate consequence of Theorem 1 is the Hamming distance consistency of the bias-adjusted sum-of-squared spectral clustering, provided that $L^{1/2}n\rho/\log^{1/2}(L+n)\to\infty$. This demonstrates the boost of signal strength by a factor of $L^{1/2}$ made possibly due to aggregating layers (up to a poly-logarithmic factor) that we alluded to in Section 1.

The proof of Theorem 1 is given in Appendix D, supplementary materials where the main effort is to establish sharp operator norm bounds for the linear noise term $\sum_\ell X_\ell P_\ell$ and the quadratic noise term $\sum_\ell(X_\ell^2-D_\ell)$. A refined operator norm bound for the off-diagonal part of $\sum_\ell(X_\ell^2-D_\ell)$ plays an important role (Theorem 5). Once the operator norm bound is established, the clustering consistency follows from a standard analysis of the K-means algorithm (Lemma D.1, supplementary materials). These concentration inequalities indeed hold for more general classes of matrices, and we provide a systematic development in the next section.

Theorem 1 is stated in a simple form for brevity. It can be generalized in several directions to better suit practical scenarios with more careful bookkeeping in the proof. We describe some important extensions in the remarks below, where $\|\cdot\|$ denotes the operator norm (i.e., largest singular value).

Remark 1 (*Varying membership across layers*). Theorem 1 can be extended to accommodate varying membership across the layers. In particular, assume that the ℓth layer has membership matrix $\Psi_\ell\in\{0,1\}^{n\times K}$, such that each Ψ_ℓ is close to a common membership matrix $\Psi\in\{0,1\}^{n\times K}$,

$$\|\Psi_\ell-\Psi\|\le\epsilon_\ell\sqrt{n}, \tag{8}$$

for some positive constant ϵ_ℓ. Then we have the following generalization of Theorem 1.

Corollary 2 (*Consistency under varying membership*). Assume the multilayer adjacency matrices $A_1,\dots,A_L$ are generated from individual membership matrices $\Psi_1,\dots,\Psi_L$ satisfying (8) for some sequence $\epsilon_1,\dots,\epsilon_L$ and common membership matrix Ψ. Under the same condition as in Theorem 1, if in addition $\bar\epsilon:=L^{-1}\sum_\ell\epsilon_\ell\le C_3$ for some positive constant C_3, then the error bound of the bias-adjusted sum of squared spectral clustering is no more than

$$C\left(\frac{1}{n^2}+\bar\epsilon^2+\frac{\log(L+n)}{Ln^2\rho^2}\right)$$

with high probability.

Remark 2 (*Other regimes of network density*). The condition $L^{1/2}n\rho\ge C_1\log^{1/2}(L+n)$ is required in order for the error bound in Theorem 1 to imply consistency, and is suitable for the linear squared signal accumulation assumed in Part (b) of Assumption 1. If we assume a different growth speed of the minimum eigenvalue of $\sum_\ell B_{\ell,0}^2$, this requirement needs to be changed accordingly. Second, the condition $n\rho\lesssim 1$ is used for notational simplicity. The regime $n\rho\gg 1$ would allow for consistent community recovery even when $L=1$. For multilayer models, if $n\rho\ge C_2$ for some constant C_2, the error bound in Theorem 1 becomes

$$C\left(\frac{1}{n^2}+\frac{\log(L+n)}{Ln\rho}\right).$$

for some constant C with high probability. Detailed explanations of this claim are given in Appendix D, supplementary materials.

Remark 3 (*More general conditions on community sizes*). Let $n_{\min}=\min_{1\le k\le K}\|\Psi_{\cdot k}\|_1$ be the size of the smallest community, and denote $\alpha=n_{\min}/n$. Our analysis can also allow the number of communities, K, and α to change with other model parameters (n,L,ρ). In particular, the lower bound of the signal term in (5) will be multiplied by α since the operator norm of Ψ is proportional to α. All the matrix concentration results, such as Theorem 5 and Lemma C.1, supplementary materials still hold as they do not rely on any block structures. Therefore, under the same setting as Theorem 1, if we allow K and α to vary with (n,L,ρ), but have $\alpha L^{1/2}n\rho\ge C_1\log^{1/2}(L+n)$ for some constant C_1, then with high probability,Theorem 1 holds with error bound

$$CK\alpha^{-2}\left(\frac{1}{n^2}+\frac{\log(L+n)}{Ln^2\rho^2}\right).$$

4. Matrix Concentration Inequalities

We generically consider a sequence of independent matrices $X_1,\dots,X_L\in\mathbb{R}^{n\times r}$ with independent mean-0 entries. The goal is to provide upper bounds for operator norms of linear combinations of the form $\sum_\ell X_\ell H_\ell$ with $H_\ell\in\mathbb{R}^{r\times m}$ for $\ell\in\{1,\dots,L\}$, and quadratic forms $\sum_\ell X_\ell G_\ell X_\ell^T$ with $G_\ell\in\mathbb{R}^{r\times r}$ for $\ell\in\{1,\dots,L\}$. Here, H_ℓ and G_ℓ are nonrandom. To connect with the notations in previous sections, let $H_\ell=P_\ell$, then an operator norm bound of $\sum_\ell X_\ell P_\ell$ will help control the second term in (5). Let $G_\ell=I_r$ be the $r\times r$ identity matrix, then $\sum_\ell X_\ell G_\ell X_\ell^T$ corresponds to the third term in (5). Our general results cover both the symmetric and asymmetric cases, as well as more general entries of X_ℓ beyond the Bernoulli case.

Concentration inequalities usually require tail conditions on the entries of X_ℓ. A standard tail condition for scalar random variables is the Bernstein tail condition.

Definition 1. We say a random variable Y satisfies a (v,R)-Bernstein tail condition (or is (v,R)-Bernstein), if $\mathbb{E}[|Y|^k]\le\frac{v}{2}k!R^{k-2}$ for all integers $k\ge 2$.

The Bernstein tail condition leads to concentration inequalities for sums of independent random variables (van der Vaart and Wellner 1996, chap. 2). Since we are interested not only in linear combinations of X_ℓ's, but also the quadratic forms involving $X_\ell G_\ell X_\ell^T$, we need the Bernstein condition to hold for the squared entries of $X_1,\dots,X_L$. Specifically we consider the following three assumptions.

Assumption 2. Each entry $X_{\ell,ij}$ is (ν_1, R_1)-Bernstein, for all $\ell \in \{1,\dots,L\}$ and $i,j \in \{1,\dots,n\}$.

Assumption 3. Each squared entry $X^2_{\ell,ij}$ is (ν_2, R_2)-Bernstein, for all $\ell \in \{1,\dots,L\}$ and $i,j \in \{1,\dots,n\}$.

Assumption 3'. The product $X_{\ell,ij}\widetilde{X}_{\ell,ij}$ is (ν_2', R_2')-Bernstein, for all $\ell \in \{1,\dots,L\}$ and $i,j \in \{1,\dots,n\}$, where $\widetilde{X}_\ell$ is an independent copy of X_ℓ.

There are two typical scenarios in which such a squared Bernstein condition in Assumption 3 holds. The first is the sub-Gaussian case: If a random variable Y satisfies the sub-Gaussian condition $\mathbb{E}e^{Y^2/\sigma^2} \le 2$ for some $\sigma > 0$, then we have $\mathbb{E}Y^{2k} \le 2\sigma^4(\sigma^2)^{k-2}k!$, and hence Y^2 is $(4\sigma^4, \sigma^2)$-Bernstein. The second scenario is centered Bernoulli: If a random variable Y satisfies $\mathbb{P}(Y = 1-p) = 1-\mathbb{P}(Y=-p) = p$ for some $p \in [0,1/2]$, then we have $\mathbb{E}Y^{2k} = p(1-p)^{2k} + (1-p)p^{2k} \le p$, and hence Y^2 is $(2p, 1)$-Bernstein. Our proof will also use the fact that if Y^2 is (ν_2, R_2)-Bernstein, then the centered version $Y^2 - \mathbb{E}(Y^2)$ is also (ν_2, R_2)-Bernstein (Wang, Berthet, and Plan 2016, Lemma 3).

We require Assumption 3' in order to use a decoupling technique in establishing concentration of quadratic forms. One can show that if Assumption 3 holds then Assumption 3' holds with $(\nu_2', R_2') = (\nu_2, R_2)$. However, when $X_{\ell,ij}$'s are centered Bernoulli random variables with parameters bounded by $p \le 1/2$, then Assumption 3' holds with $\nu_2' = 2p^2$ and $R_2' = 1$, while Assumption 3 holds with $\nu_2 = 2p$ and $R_2 = 1$, so that ν_2' can potentially be much smaller than ν_2. We will explicitly keep track of the Bernstein parameters in our results for the sake of generality.

4.1. Linear Combinations with Matrix Coefficients

Theorem 3. Let $X_1,\dots,X_L$ be a sequence of independent $n \times r$ matrices with mean-0 independent entries satisfying Assumption 2, and H_ℓ be any sequence of $r \times m$ nonrandom matrices. Then for all $t > 0$,

$$\mathbb{P}\left[\left\|\sum_{\ell=1}^L X_\ell H_\ell\right\| \ge t\right] \le 2(m+n) \times \exp\left(-\frac{t^2/2}{\nu_1\left(n\left\|\sum_\ell H_\ell^T H_\ell\right\| \vee \sum_\ell \|H_\ell\|_F^2\right) + R_1 \max_\ell \|H_\ell\|_{2,\infty} t}\right). \tag{9}$$

A similar result holds, with $t^2/2$ replaced by $t^2/8$ and $2(m+n)$ replaced by $4(m+n)$ in (9), for symmetric X_ℓ's of size $n \times n$ with independent (ν_1, R_1)-Bernstein diagonal and upper-diagonal entries and H_ℓ of size $n \times m$.

The proof of Theorem 3, given in Appendix B, supplementary materials, combines the matrix Bernstein inequality (Tropp 2012) for symmetric matrices and a rank-one symmetric dilation trick (Lemma B.1, supplementary materials) to take care of the asymmetry in $X_\ell H_\ell$.

Remark 4. If $n = m = r = 1$, then Theorem 3 recovers the well-known Bernstein's inequality as a special case with a different prefactor.

If $n \ge \min\{m, Lr\}$, then $n\|\sum_\ell H_\ell^T H_\ell\| \ge \sum_\ell \|H_\ell\|_F^2$ and the probability upper bound in Theorem 3 reduces to

$$\mathbb{P}\left[\left\|\sum_{\ell=1}^L X_\ell H_\ell\right\| \ge t\right] \le 2(m+n) \times \exp\left(-\frac{t^2/2}{\nu_1 n\left\|\sum_\ell H_\ell^T H_\ell\right\| + R_1 \max_\ell \|H_\ell\|_{2,\infty} t}\right). \tag{10}$$

If $n = 1$, then $n\|\sum_\ell H_\ell^T H_\ell\| \le \sum_\ell \|H_\ell\|_F^2$ and the probability bound reduces to

$$\mathbb{P}\left[\left\|\sum_{\ell=1}^L X_\ell H_\ell\right\| \ge t\right] \le 2(m+n) \times \exp\left(-\frac{t^2/2}{\nu_1 \sum_\ell \|H_\ell\|_F^2 + R_1 \max_\ell \|H_\ell\|_{2,\infty} t}\right). \tag{11}$$

Remark 5. When $L = 1$, the setting is similar to that considered in Vershynin (2011). In the constant variance case (e.g., sub-Gaussian), $\nu_1^{1/2} \asymp R_1 \asymp 1$, Theorem 3 implies a high probability upper bound of $C\sqrt{\log(m+n)}(\sqrt{n}\|H\| + \|H\|_F)$, which agrees with Theorem 1.1 of Vershynin (2011). The extra $\sqrt{\log(n+m)}$ factor in our bound is because our result is a tail probability bound while Vershynin (2011) provides upper bounds on the expected value. However, in the sparse Bernoulli setting, where $\nu_1 \ll R_1 = 1$, the upper bound in Theorem 3 is better because it correctly captures the $\sqrt{\nu_1}$ factor multiplied by $\sqrt{n}\|H\| + \|H\|_F$, whereas the result in Vershynin (2011) leads to $\nu_1^{1/4}(\sqrt{n}\|H\| + \|H\|_F)$.

4.2. Matrix U-statistics and Quadratic Forms

Let

$$S = \sum_{\ell=1}^L X_\ell G_\ell X_\ell^T = \sum_{\ell=1}^L \sum_{(i,j),(i',j')} X_{\ell,ij} X_{\ell,i'j'} e_i e_{i'}^T G_{\ell,jj'} \tag{12}$$

where the summation is taken over all pairs $(i,j),(i',j') \in \{1,\dots,n\}^2$ and e_i is the canonical basis vector in $\mathbb{R}^n$ with a 1 in the ith coordinate. In this section, we will focus on the symmetric case because the bookkeeping is harder compared to the asymmetric case. The treatment for the asymmetric case is similar and the corresponding results are stated separately in Appendix B, supplementary materials for completeness.

Because X_ℓ has centered and independent diagonal and upper diagonal entries, a term in (12) has nonzero expected value only if $(i,j) = (i',j')$ or $(i,j) = (j',i')$ since this would imply $X_{\ell,ij}X_{\ell,i'j'} = X^2_{\ell,ij}$. This motivates the following decomposition of S into a quadratic component with nonzero entry-wise mean value

$$S_2 = \Big[\sum_{\ell=1}^L \sum_{1\le i<j\le n} X^2_{\ell,ij}\left(e_i e_i^T G_{\ell,jj} + e_j e_j^T G_{\ell,ii} + e_i e_j^T G_{\ell,ji} + e_j e_i^T G_{\ell,ij}\right)\Big] + \Big[\sum_{\ell=1}^L \sum_{1\le i\le n} X^2_{\ell,ii} e_i e_i^T G_{\ell,ii}\Big], \tag{13}$$

and a cross-term component with entry-wise mean-0 value

$$S_1 = S - S_2. \tag{14}$$

It is easy to check that $\mathbb{E}S_2 = \mathbb{E}S$ and $\mathbb{E}S_1 = 0$. Intuitively, the spectral norm of S_1 should be small since it is the sum of many random terms with zero mean and small correlation, which can be viewed as a U-statistic with a centered kernel function of order two. This U-statistic perspective is a key component of the analysis and will be made clearer in the proof. For a similar reason, $S_2 - \mathbb{E}S_2$ should also be small. Hence, the main contributing term in S should be the deterministic term $\mathbb{E}S_2$. To formalize this, define the following quantities,

$$\begin{aligned} \sigma_1^2 &= \textstyle\sum_{\ell=1}^{L} \|G_\ell\|^2, \\ \sigma_2 &= \max_\ell \max\left\{\|G_\ell\|_{2,\infty}, \|G_\ell^T\|_{2,\infty}\right\} \\ (\sigma_2')^2 &= \textstyle\sum_{\ell=1}^{L}\sum_{j=1}^{n} G_{\ell,jj}^2, \\ \sigma_3 &= \max_\ell \|G_\ell\|_\infty, \end{aligned}$$

where $\|\cdot\|_{2,\infty}$ is the maximum L_2-norm of each row, and $\|\cdot\|_\infty$ is the maximum entry-wise absolute value. The following theorem quantifies the random fluctuations of S_1, S_2 and S around their expectations.

Theorem 4. If $X_1, \ldots, X_L$ are independent $n \times n$ symmetric matrices with independent diagonal and upper diagonal entries satisfying Assumptions 2 and 3'. Let $G_1, \ldots, G_L$ be $n \times n$ matrices. Define $S = \sum_\ell X_\ell G_\ell X_\ell^T$ and S_1, S_2 as in (13) and (14). Then there exists a universal constant C such that with probability at least $1 - O((n+L)^{-1})$,

$$\begin{aligned} \|S_1\| \le C\Big[& v_1 n \log(L+n)\sigma_1 + \sqrt{v_1} R_1 \sqrt{Ln} \log^{3/2}(L+n)\sigma_2 \\ & + \sqrt{v_2'} \log(L+n)(\sqrt{L}\sigma_2 + \sigma_2') \\ & + (R_1^2 + R_2') \log^2(L+n)\sigma_3 \Big]. \end{aligned} \tag{15}$$

If in addition Assumption 3 holds, then with probability at least $1 - O((L+n)^{-1})$,

$$\|S_2 - \mathbb{E}S_2\| \le C\Big[\sqrt{v_2}\log(L+n)(\sqrt{L}\sigma_2 + \sigma_2') + R_2 \log(L+n)\sigma_3\Big]. \tag{16}$$

and consequently,

$$\begin{aligned} \|S - \mathbb{E}S\| \le C\Big[& v_1 n \log(L+n)\sigma_1 + \sqrt{v_1} R_1 \sqrt{Ln} \log^{3/2}(L+n)\sigma_2 \\ & + \sqrt{v_2' + v_2} \log(L+n)(\sqrt{L}\sigma_2 + \sigma_2') \\ & + (R_1^2 + R_2 + R_2') \log^2(L+n)\sigma_3 \Big]. \end{aligned} \tag{17}$$

The proof of Theorem 4 is given in Appendix B, supplementary materials, where the main effort is to control $\|S_1\|$. Unlike the linear combination case, the complicated dependence caused by the quadratic form needs to be handled by viewing S_1 as a matrix-valued U-statistic indexed by the pairs (i,j), and using a decoupling technique due to de la Peña and Montgomery-Smith (1995). This reduces the problem of bounding $\|S_1\|$ to that of bounding $\|\sum_\ell X_\ell G_\ell \widetilde{X}_\ell^T\|$, where $\widetilde{X}_1, \ldots, \widetilde{X}_L$ are iid copies of $X_1, \ldots, X_L$.

The upper bounds in Theorem 4 look complicated. This is because we do not make any assumption about the Bernstein parameters or the matrices G_ℓ. The bound can be much simplified or even improved in certain important special cases. In the sub-Gaussian case, where $R_1 \asymp v_1^{1/2} \asymp R_2^{1/2} \asymp v_2^{1/4}$, the first term $v_1 n \log(L+n)\sigma_1$ in (15) dominates. This reflects the $L^{1/2}$ effect for sums of independent random variables. For example, in the case $G_\ell = G_0$ for all ℓ and X_ℓ are iid, we have $\|\mathbb{E}S\| \approx L\|X_1 G_0 X_1^T\| \asymp v_1 nL\|G_0\|$, but when we consider the fluctuations contributed by S_1, we have $\|S_1\| \lesssim v_1 nL^{1/2}\|G_0\|$ ignoring logarithmic factors. In other words, the signal is contained in $\mathbb{E}S_2$ whose operator norm may grow linearly as L, while the fluctuation in the operator norm of S_1 only grows at a rate of $L^{1/2}$.

Additionally, in the Bernoulli case, the situation becomes more complicated when the variance v_1 is vanishing, meaning that $v_1 \asymp v_2 \asymp (v')_2^{1/2} \ll R_1 \asymp R_2$. In the simple case of $G_\ell = I_n$, we have $\sigma_1 = L^{1/2}$, $\sigma_2 = \sigma_3 = 1$. Thus, the second term $(v_1 Ln)^{1/2}\sigma_2$ in (15) may dominate the first term when $nv_1 \ll 1$. In this case, we also have $\sigma_2' = (Ln)^{1/2}$. Therefore, it is also possible that the term $v_2^{1/2}\sigma_2$ in (16) may be large. It turns out that in such very sparse Bernoulli cases, the bound on the fluctuation term $\|S_1\|$ can be improved by a more refined and direct upper bound for $\|\sum_\ell X_\ell X_\ell^T\| = \|S\|$. The details are presented in the next section.

4.3. Sparse Bernoulli Matrices

In this section, we focus on the case where $G_\ell = I_n$ for all ℓ, and the X_ℓ's are symmetric with centered Bernoulli entries whose probability parameters are bounded by ρ. Here, ρ can be very small. In this case, Assumptions 2, 3, and 3' hold with $v_1 = v_2 = 2\rho$, $R_1 = R_2 = R_2' = 1$, $v_2' = 2\rho^2$, and the matrices G_ℓ satisfy $\sigma_1 = L^{1/2}$, $\sigma_2 = \sigma_3 = 1$, $\sigma_2' = (Ln)^{1/2}$.

Ignoring logarithmic factors, the first part of Theorem 4 becomes

$$\|S_1\| \lesssim C\left[L^{1/2} n\rho + (Ln\rho)^{1/2} + 1\right],$$

where the second term $(Ln\rho)^{1/2}$ can be dominating when $n\rho$ is small and $Ln\rho$ is large. This is suboptimal since intuitively we expect that the main variance term $L^{1/2}n\rho$ is the leading term as long as its value is large enough, which only requires $n\rho \gg L^{-1/2}$. To investigate the cause of this suboptimal bound, observe that $(Ln\rho)^{1/2}$ originates from the second term $R_1(v_1 Ln)^{1/2}\sigma_2$ in (15). Investigating the proof of Theorem 4, this term is derived by bounding $\sum_\ell \|H_\ell^T H_\ell\|$ by $\sum_\ell \|H_\ell\|^2$, which is suboptimal in this sparse Bernoulli case when applying the decoupling technique. The following result shows a sharper bound in this setting using a more refined argument.

Theorem 5. Assume $G_\ell = I_n$ for all $\ell \in \{1, \ldots, L\}$ and $X_1, \ldots, X_L$ are symmetric with centered Bernoulli entries whose parameters are bounded by ρ. If $L^{1/2}n\rho \ge C_1 \log^{1/2}(L+n)$ and $n\rho \le C_2$ for some constants C_1, C_2, then with probability at least $1 - O((n+L)^{-1})$,

$$\|S_1\| \le CL^{1/2}\rho n \log^{1/2}(L+n) \tag{18}$$

for some constant C.

The proof of Theorem 5 is given in Appendix C, supplementary materials where we modify our usage of the decoupling

technique. At a high level, the decoupling technique reduces the problem to controlling the operator norm of $\widetilde{S} = \sum_\ell X_\ell \widetilde{X}_\ell^T$ where $\widetilde{X}_\ell$ is an iid copy of X_ℓ. Instead of directly applying Theorem 3 with $H_\ell = \widetilde{X}_\ell$, we instead shift $\widetilde{X}_\ell$ back to the original Bernoulli matrix by considering $\widetilde{S} = \sum_\ell X_\ell \widetilde{A}_\ell - \sum_\ell X_\ell P_\ell$, where $\widetilde{A}_\ell$ is the original uncentered binary matrix and $P_\ell = \mathbb{E}\widetilde{A}_\ell$. Then Theorem 3 is applied to $\sum_\ell X_\ell P_\ell$ and $\sum_\ell X_\ell \widetilde{A}_\ell$ separately, where the entry-wise nonnegativity of $\widetilde{A}_\ell$ allows us to use the Perron–Frobenius theorem to obtain a sharper bound for $\|\sum_\ell \widetilde{A}_\ell^2\|$.

5. Further Simulation Study

In the following simulation study, we show that bias-adjusting sum of squared adjacency matrices constructed in (7) has a measurable impact on the downstream spectral clustering accuracy, and that our method performs favorably against other competing methods. This builds upon the simulation initially shown in Section 2.2.

Data-generating process. We design the following simulation setting to highlight the importance of bias adjustment for $\sum_\ell A_\ell^2$. We consider $n = 500$ nodes per network across $K = 3$ communities, with imbalanced sizes $n_1 = 200$, $n_2 = 50$, and $n_3 = 250$. We construct two edge-probability matrices that share the same eigenvectors,

$$W = \begin{bmatrix} 1/2 & 1/2 & -\sqrt{2}/2 \\ 1/2 & 1/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 & 0 \end{bmatrix}. \tag{19}$$

The two edge-probability matrices are

$$B^{(1)} = W \begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.4 \end{bmatrix} W^\top \approx \begin{bmatrix} 0.62 & 0.22 & 0.46 \\ 0.22 & 0.62 & 0.46 \\ 0.46 & 0.46 & 0.85 \end{bmatrix},$$

$$B^{(2)} = W \begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & -0.4 \end{bmatrix} W^\top \approx \begin{bmatrix} 0.22 & 0.62 & 0.46 \\ 0.62 & 0.22 & 0.46 \\ 0.46 & 0.46 & 0.85 \end{bmatrix}.$$

We then generate $L = 100$ layers of adjacency matrices, where each layer is drawn by setting the edge-probability matrices $B_\ell = \rho B^{(1)}$ for $\ell \in \{1, \ldots, L/2\}$ and $B_\ell = \rho B^{(2)}$ for $\ell \in \{L/2+1, \ldots, L\}$. Using this, we generate the adjacency matrices via (1), with ρ varying from 0.025 to 0.2.

We choose this particular simulation setting for two reasons. First, the first two eigenvectors in W are not sufficient to distinguish between the first two communities. Hence, methods based on $\sum_\ell A_\ell$ are not expected to perform well since the third eigen-component cancels out in the summation. Second, the average degrees among the three communities are drastically different, which are 251ρ, 191ρ, and 327ρ, respectively. This means the variability of degree matrix D_ℓ's diagonal entries will be high, helping demonstrating the effect of our method's bias adjustment.

Methods we consider. We consider the following four ways to aggregate information across all L layers, three of which were used earlier in Figure 2: (a) the sum of adjacency matrices without squaring (i.e., considering $M = \sum_\ell A_\ell$, "Sum"), (b) the sum of squared adjacency matrices (i.e., considering $M = \sum_\ell A_\ell^2$, "SoS"), (c) our proposed bias-adjusted sum of squared adjacency matrices (i.e., considering (7), or equivalently $M = \sum_\ell A_\ell^2$ and then zeroing out the diagonal entries, "SoS-Debias"), and (d) column-wise concatenating the adjacency matrices together, specifically, considering

$$M = \begin{bmatrix} A_1 & A_2 & \cdots & A_L \end{bmatrix} \in \mathbb{R}^{n \times (Ln)}.$$

(i.e., "Tensor matricization"). This method is commonly-used in the tensor literature (see, e.g., Zhang and Xia 2018), where the L adjacency matrices can be viewed as a $n \times n \times L$ tensor, and the column-wise concatenation converts the tensor into a matrix. Then, using one of the four construction of the aggregated matrix M, we then apply spectral clustering onto M, meaning we first compute the matrix containing the leading K left singular vectors of M and perform K-means on its rows.

Additionally, we consider two methods that developed in Paul and Chen (2020) called Linked Matrix Factorization (i.e., "LMF") and Co-regularized Spectral Clustering (i.e., "Co-reg"). These two methods fall outside the framework of the four methods discussed above. Instead, they use optimization procedures designed with different so-called fusion techniques to solve for an appropriate low-dimensional embedding shared among all L layers, and then perform K-means clustering on its rows.

Results. The results shown in Figure 3 demonstrate that bias-adjusting the diagonal entries of $\sum_\ell A_\ell^2$ has a noticeable impact on the clustering accuracy. Using the aforementioned simulation setting and methods, we vary ρ from 0.025 to 0.2 in 15 equally spaced values, and compare the methods for each setting of ρ across 100 trials by measuring the average Hamming distance (i.e., $n^{-1}d(\widehat{\theta}, \theta)$ defined in (2)) between the true memberships in θ and the estimated membership $\widehat{\theta}$. We observe phenomenons in Figure 3 which all agree with our intuition and theoretical results. Specifically, summing the adjacency matrices hinders our ability to cluster the nodes due to the cancelation of positive and negative eigenvalues (green squares), and the diagonal bias induced by squaring the adjacency matrices has a profound effect in the range of $\rho \in [0.08, 0.17]$, which our bias-adjusted sum-of-squared method removes (purple diamonds verses blue circles). We also see that our bias-adjusted sum-of-squared method out-performs Linked Matrix Factorization (red circles) and Coregularized Spectral Clustering (gray squares). While the LMF method and Coreg method show some improvements over the Sum and SoS methods, respectively, they still behave qualitatively similar. This observation suggests that these two methods may have similar difficulty in aggregating layers without positivity or removing the diagonal bias.

Intuition behind results. We provide additional intuition behind the results shown in Figure 3 by visualizing the impact of the diagonal terms on the overall spectrum and quantifying the loss of population signal due to the bias.

First, we demonstrate in Figure 4 that the third leading eigenvalue of $\sum_\ell A_\ell^2$ when $\rho = 0.15$ is indistinguishable from the remaining bulk "noise" eigenvalues if the diagonal bias is not removed (left), but becomes well-separated if so (middle). Recall by construction (19), all three eigenvectors are needed for recovering the communities. Hence, if the third eigenvalue of $\sum_\ell A_\ell^2$ is indistinguishable from fourth through last eigenvalues (i.e., the "noise"), then we should expect many nodes to be mis-clustered. This is exactly what Figure 4(left) shows, where the third eigenvalue (denoted by the left-most red vertical

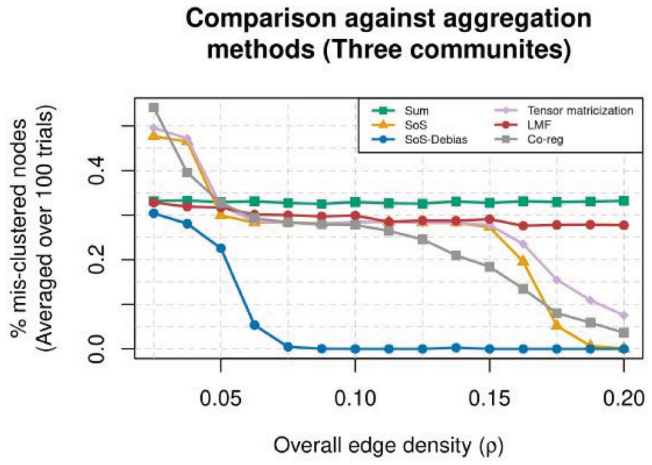

Figure 3. The average proportion of mis-clustered nodes for eight methods (measured via Hamming distance $n^{-1}d(\widehat{\theta},\theta)$ shown in (2), averaged over 100 trials), with $n = 500$ with three unequally sized communities among overall edge densities ranging from $\rho \in [0.025, 0.2]$ and $L = 100$ layers. Six methods' performance are shown: "Sum" (green squares), "SoS" (orange triangles), "Bias-adjusted SoS" (blue circles), "Tensor matricization" (purple diamonds), "LMF" (red circles), and "Co-reg" (gray squares).

line) is not separated from the remaining eigenvalues. However, when we appropriately bias-adjust $\sum_\ell A_\ell^2$ via (7), then Figure 4(middle) shows that the third eigenvalue is now well-separated from the remaining eigenvalues. This demonstrates the importance of bias-adjustment for community estimation in this regime of ρ.

Next, in Figure 4(right), we show that this lack-of-separation between the third eigenvalue and the noise can be observed on the population level. Specifically, we show that the population counterpart of $\sum_\ell A_\ell^2$ has considerable diagonal bias that makes the accurate estimation of the third eigenvector nearly impossible when ρ is too small. To show this, for a particular value of ρ, recall from our theory that the population counterpart of $\sum_\ell A_\ell^2$ is

$$\sum_{\ell=1}^{L}(P_\ell^2 + \widetilde{D}_\ell), \text{ for a diagonal matrix } \widetilde{D}_\ell \text{ where}$$

$$\widetilde{D}_{\ell,ii} = \sum_{j=1}^{n} P_{\ell,ij} \text{ for } 1 \le i \le n,$$

and $P_\ell = \mathbb{E}A_\ell$. Let $\lambda_1, \ldots, \lambda_n$ denote the n eigenvalues of the above matrix, dependent on ρ. We then plot $(\lambda_3 - \lambda_4)/\lambda_4$ against ρ in Figure 4 (right). This plot demonstrates that when ρ is too small, the diagonal entries (represented by $\widetilde{D}_\ell$'s) add a disproportionally large amount of bias that makes it impossible to accurately distinguish between the third and fourth eigenvectors. Additionally, the raise in the eigengap in Figure 4(right) at $\rho = 0.15$ corresponds to when "SoS" starts to improve in Figure 3 (orange triangles). This means starting at $\rho = 0.15$, the effect of the diagonal bias starts to diminish, and at larger values of ρ, the sum of squared adjacency matrices contains accurate information for community estimation (both with and without bias adjustment). We report additional results in Appendix E, supplementary materials, where we report the time needed for each method, visualize the lack of concentration in the nodes' degrees in sparse graphs and its effect on the spectral embedding, and also report that the qualitative trends in Figure 3 remain the same when we either consider the varying-membership setting (described in Corollary 2) or an additional variant of spectral clustering where the eigenvectors are reweighted by its corresponding eigenvalues.

6. Data Application: Gene Coexpression Patterns in Developing Monkey Brain

We analyze the microarray dataset of developing rhesus monkeys' tissue from the medial prefrontal cortex introduced in Section 1 that was originally collected in Bakken et al. (2016) to demonstrate the utility of our bias-adjusted sum-of-squared spectral clustering method. As described in other work that analyze this data (Liu et al. 2018; Lei, Chen, and Lynch 2019), this is a suitable dataset to analyze as other work have well-documented that the gene coexpression patterns in monkeys' tissue from this brain region change dramatically over development. Specifically, the data from Bakken et al. (2016) consists of the gene co-expression network of 10 different developmental times (starting from 40 days in the embryo to 48 months after birth) derived from microarray data, where each of the developmental time points corresponds to post-mortem tissue samples of multiple unique rhesus monkeys. With this data, we aim to show that our bias-adjusted sum-of-squared spectral clustering method produces insightful gene communities.

Preprocessing procedure. The microarray dataset from Bakken et al. (2016) contains $n = 9173$ genes measured among many samples across the $L = 10$ layers, which we preprocess into 10 adjacency matrices in the following way in line with other work like Langfelder and Horvath (2008). We use these specific set of n genes, following the analysis in Liu et al. (2018), since they map to the human genome. First, for each layer $\ell \in \{1, \ldots, L\}$, we construct the Pearson correlation matrix. Then, we convert each correlation matrix into adjacency matrix by hard-thresholding at 0.72 in absolute value, resulting in 10 adjacency matrices $A_1, \ldots, A_L$. We choose this particular threshold since it yields sparse and scale-free networks that have many disjoint connected components individually but have one connected component after aggregation, as reported in Appendix F, supplementary materials. Lastly, we remove all the genes corresponding to nodes whose total degree across all 10 layers is less than 90. This value is chosen since the median total degree among all nodes that do not have any neighbors in five or more of the layers (i.e., a degree of zero in more than half the layers) is 89. In the end, we have 10 adjacency matrices $A_1, \ldots, A_L \in \{0,1\}^{7836\times 7836}$, each representing a network corresponding to 7836 genes. We note that the above procedure of transforming correlation matrices into adjacency matrices is unlikely to procedure networks that severely violate the layer-wise positivity assumption commonly required by other methods—this hypothetically could happen if many pairs of genes display high negative correlations, but this is not typical in genomic data. Nonetheless, we are interested in what insights the bias-adjusted sum-of-squared spectral clustering method can reveal for this dataset.

Results and interpretation The following results show that bias-adjusted sum-of-squared spectral clustering finds meaningful gene communities. Prior to using our method, we select

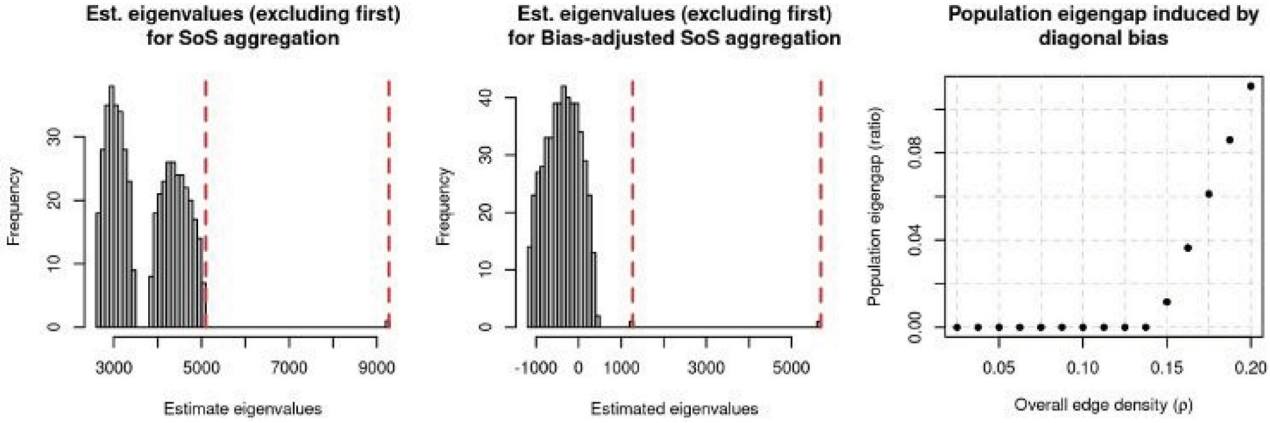

Figure 4. (Left): For one realization of $A_1, \ldots, A_L$ given the setup described in the simulation with $\rho = 0.15$, a histogram of all 500 eigenvalues of $\sum_\ell A_\ell^2$, where the red vertical dashed lines denote the second and third eigenvalues. (The first eigenvalue is too large to be shown.) (Middle): Similar to the left plot, but showing the 500 eigenvalues of the bias-adjusted variant of $\sum_\ell A_\ell^2$ (i.e., setting the diagonal to be all 0's). (Right): The population eigengap $(\lambda_3 - \lambda_4)/\lambda_3$ computed from $\sum_\ell P_\ell^2 + \widetilde{D}_\ell$ for varying values of ρ.

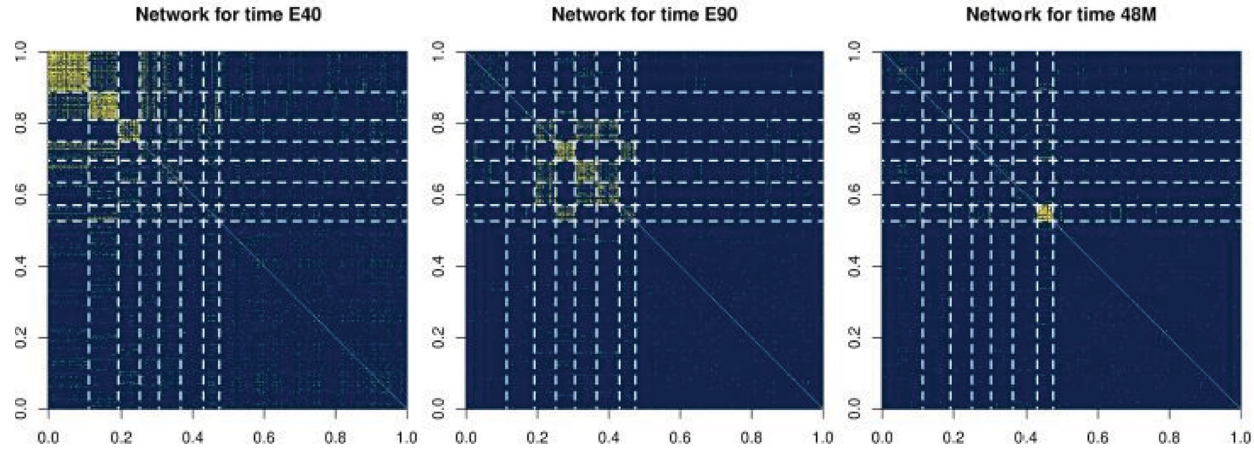

Figure 5. Three of 10 adjacency matrices where the genes are ordered according to the estimated $K = 8$ communities. Blue pixels correspond to the absence of an edge between the corresponding genes in A_ℓ's, while yellow pixels correspond to an edge. The dashed white lines denote the separation among the $K = 8$ gene communities. The adjacency matrices shown in Figure 1 correspond to the same three developmental times (from left to right), and are formed by selecting only the genes in Communities 1, 4, 5, and 7.

the dimensionality and number of communities to be $K = 8$ based on a scree plot of the singular values of the bias-adjusted variant of $\sum_\ell A_\ell^2$. We perform our bias-adjusted spectral clustering on this matrix with $K = 8$, and visualize three out of the 10 adjacency matrices using the estimated communities in Figure 5 (which are the full adjacency matrices corresponding to the three adjacency matrices shown in Figure 1). We see that as development occurs from 40 days in the embryo to 48 months after birth, there are different gene communities that are most-connected. This visually demonstrates different biological processes in brain tissue that are most active at different stages of development. Labeling the communities 1–8 from top left to bottom right, our results show that starting at 40 days in the embryo, Community 1 is highly coordinated (i.e., densely connected), and ending at 48 months after birth, Community 7 is highly coordinated. All the genes in Community 8 are sparsely connected throughout all 10 adjacency matrices, suggesting that these genes are not strongly correlated with many other genes throughout development.

To interpret these $K = 8$ communities, we perform a gene ontology analysis, using the `cluster-Profiler::enrichGO` function on the gene annotation in the Bioconductor package `org.Mmu.eg.db` to analyze the scientific interpretation of each of the K communities of genes within rhesus monkeys. Table 1 shows the results. We see the first seven communities are highly enriched for cell processes closely related to brain development—we can interpret Figure 5 and Table 1 together as which biological systems are most active in a coordinated fash-

Table 1. Gene ontology of the estimated $K = 8$ communities of genes.

Community	Description	GO ID	*p*-value
1	RNA splicing	GO:0008380	1.07×10^{-11}
2	Nuclear transport	GO:0051169	3.15×10^{-5}
3	Neuron development	GO:0048666	2.08×10^{-8}
4	Chromosome segregation	GO:0007059	1.31×10^{-8}
5	Neuron projection development	GO:0031175	1.51×10^{-5}
6	Regulation of transporter activity	GO:0032409	5.68×10^{-6}
7	Anchoring junction	GO:0070161	8.86×10^{-5}
8	None		

NOTE: Here, "GO" denotes the gene ontology ID, and "*p*-value" denotes the Fisher's exact test to denote an enrichment (i.e., significance or over-representation) of a particular GO for the genes in said community compared to all other genes.

ion at different developmental stage. Since genes in the eighth community are sparsely connected across all developmental time and is not enriched for any cell processes, we infer that these genes are unlikely to be coordinated to drive any process related to brain development. Together, these results demonstrate that the bias-adjusted sum-of-squared spectral clustering is able to find meaningful gene communities. Visualizations of all 10 adjacency matrices, beyond those shown in Figure 5 and explicit reporting of the edge densities, as well as stability analyses that demonstrate how the results vary when different tuning parameters are used, are included in Appendix F, supplementary materials.

7. Discussion

While we establish community estimation consistency in this article, there are two major additional theoretical directions we hope our results will help shed light into for future work. First, an important theoretical question in the study of stochastic block models is the critical threshold for community estimation. This involves finding a critical rate of the overall edge density and/or the separation between rows of $B_{\ell,0}$, and proving achievability of certain community estimation accuracy when the density and/or separation are above this threshold, as well as impossibility for nontrivial community recovery below this threshold. For single-layer SBMs, this problem has been studied by many authors, such as Massoulié (2014), Abbe and Sandon (2015), Zhang and Zhou (2016), and Mossel, Neeman, and Sly (2018). The case of multi-layer SBMs is much less clear, especially for generally structured layers. The upper bounds proved in Paul and Chen (2020) and Bhattacharyya and Chatterjee (2018) imply achievability of vanishing error proportion when $Ln\rho \to \infty$ under a layer-wise positivity assumption. Our results requires a stronger $L^{1/2}n\rho/\log^{1/2}(L+n) \to \infty$ condition, but does not require a layer-wise positivity assumption. Ignoring logarithmic factors, is a rate of $L^{1/2}$ the right price to pay for not having the layer-wise positivity assumption? The error analysis in the proof of Theorem 1 seems to suggest a positive answer, but a rigorous claim will require a formal lower bound analysis. We note that the simplified constructions such as that in Zhang and Zhou (2016) designed for single-layer SBMs are unlikely to work, since they do not reflect the additional hardness brought to the estimation problem by unknown layer-wise structures.

Second, the consistency result for multi-layer SBMs also makes it possible to extend other inference tools developed for single-layer data to multi-layer data. One such example is model selection and cross-validation (Chen and Lei 2018; Li, Levina, and Zhu 2020). The probability tools developed in this article, such as Theorems 3 and 4 and Theorem B.2, supplementary materials, may be useful for other statistical inference problems involving matrix-valued measurements and noise. For example, our theoretical analyses could refine the theoretical analyses for multilayer graphs that go beyond SBMs, such as degree-corrected SBMs or random dot-product graphs in general (Nielsen and Witten 2018; Arroyo et al. 2021). Alternatively, in dynamic networks where the network parameters change smoothly over time, one may use nonparametric kernel smoothing techniques in Pensky and Zhang (2019) and the matrix concentration inequalities developed in this article to control the aggregated noise and perhaps obtain more refined analysis in those settings.

Supplementary Materials

The online supplementary file contains technical proofs of main theorems, and further details about data access, simulation study, real data analysis.

Acknowledgments

We thank Kathryn Roeder, Fuchen Liu, and Xuran Wang for providing the data and code used in Liu et al. (2018) to preprocess the data. We also thank the editor, the associate editor, and two anonymous reviewers for their helpful suggestions to improve the manuscript.

Funding

Jing Lei's research is partially funded by NSF grant DMS-2015492 and NIH grant R01 MH123184.

References

Abbe, E. (2017), "Community Detection and Stochastic Block Models: Recent Developments," *The Journal of Machine Learning Research*, 18, 6446–6531. [3]

Abbe, E., and Sandon, C. (2015), "Community Detection in General Stochastic Block Models: Fundamental Limits and Efficient Algorithms for Recovery," *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, 670–688. [1,12]

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), "Mixed Membership Stochastic Blockmodels," *The Journal of Machine Learning Research*, 9, 1981–2014. [1]

Arroyo, J., Athreya, A., Cape, J., Chen, G., Priebe, C. E., and Vogelstein, J. T. (2021), "Inference for Multiple Heterogeneous Networks with a Common Invariant Subspace," *Journal of Machine Learning Research*, 22, 1–49. [12]

Bai, Z., and Silverstein, J. W. (2010), *Spectral Analysis of Large Dimensional Random Matrices* (Vol. 20), New York: Springer. [3]

Bakken, T. E., Miller, J. A., Ding, S.-L., Sunkin, S. M., Smith, K. A., Ng, L., Szafer, A., Dalley, R. A., Royall, J. J., Lemon, T., et al. (2016), "A Comprehensive Transcriptional Map of Primate Brain Development," *Nature*, 535, 367–375. [1,10]

Bandeira, A. S., and Van Handel, R. (2016), "Sharp Nonasymptotic Bounds on the Norm of Random Matrices with Independent Entries," *The Annals of Probability*, 44, 2479–2506. [3]

Bhattacharyya, S., and Chatterjee, S. (2018), "Spectral Clustering for Multiple Sparse Networks: I," arXiv preprint arXiv:1805.10594. [2,3,12]

Bickel, P. J., and Chen, A. (2009), "A Nonparametric View of Network Models and Newman–Girvan and Other Modularities," *Proceedings of the National Academy of Sciences*, 106, 21068–21073. [1,3]

Cape, J., Tang, M., and Priebe, C. E. (2017), "The Kato–Temple Inequality and Eigenvalue Concentration with Applications to Graph Inference," *Electronic Journal of Statistics*, 11, 3954–3978. [3]

Chen, K., and Lei, J. (2018), "Network Cross-validation for Determining the Number of Communities in Network data," *Journal of the American Statistical Association*, 113, 241–251. [12]

de la Peña, V. H., and Montgomery-Smith, S. J. (1995), "Decoupling Inequalities for the Tail Probabilities of Multivariate U-Statistics," *The Annals of Probability*, 23, 806–816. [8]

Dong, X., Frossard, P., Vandergheynst, P., and Nefedov, N. (2012), "Clustering with Multi-layer Graphs: A Spectral Perspective," *IEEE Transactions on Signal Processing*, 60, 5820–5831. [1]

Feige, U. and Ofek, E. (2005), "Spectral Techniques Applied to Sparse Random Graphs," *Random Structures & Algorithms*, 27, 251–275. [3]

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010), "A Survey of Statistical Network Models," *Foundations and Trends® in Machine Learning*, 2, 129–233. [1]

Han, Q., Xu, K., and Airoldi, E. (2015), "Consistent Estimation of Dynamic and Multi-Layer Block Models," in *International Conference on Machine Learning*, pp. 1511–1520. [1]

Hanson, D. L., and Wright, F. T. (1971), "A Bound on Tail Probabilities for Quadratic Forms in Independent Random Variables," *The Annals of Mathematical Statistics*, 42, 1079–1083. [3]

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983), "Stochastic Blockmodels: First Steps," *Social Networks*, 5, 109–137. [1]

Jin, J. (2015), "Fast Community Detection by SCORE," *Annals of Statistics*, 43, 57–89. [1]

Karrer, B., and Newman, M. E. (2011), "Stochastic Nlockmodels and Community Structure in Networks," *Physical Review E*, 83, 016107. [1]

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. (2014), "Multilayer Networks," *Journal of Complex Networks*, 2, 203–271. [1]

Kolaczyk, E. D. (2009), *Statistical Analysis of Network Data*, New York: Springer. [1]

Langfelder, P., and Horvath, S. (2008), "WGCNA: An R Package for Weighted Correlation Network Analysis," *BMC Bioinformatics*, 9, 559. [10]

Latouche, P., Birmele, E., and Ambroise, C. (2012), "Variational Bayesian Inference and Complexity Control for Stochastic Block Models," *Statistical Modelling*, 12, 93–115. [1]

Le, C. M., Levina, E., and Vershynin, R. (2017), "Concentration and Regularization of Random Graphs," *Random Structures & Algorithms*, 51, 538–561. [3]

Lei, J. (2018), "Network Representation Using Graph Root Distributions," arXiv preprint arXiv:1802.09684. [3]

Lei, J., Chen, K., and Lynch, B. (2019), "Consistent Community Detection in Multi-Layer Network Data," *Biometrika*, 107, 61–73. [2,3,10]

Lei, J., and Rinaldo, A. (2015), "Consistency of Spectral Clustering in Stochastic Block Models," *The Annals of Statistics*, 43, 215–237. [1,3,5]

Li, T., Levina, E., and Zhu, J. (2020), "Network Cross-Validation by Edge Sampling," *Biometrika*, 107, 257–276. [12]

Litvak, N., and Van Der Hofstad, R. (2013), "Uncovering Disassortativity in Large Scale-Free Networks," *Physical Review E*, 87, 022801. [3]

Liu, F., Choi, D., Xie, L., and Roeder, K. (2018), "Global Spectral Clustering in Dynamic Networks," *Proceedings of the National Academy of Sciences*, 115, 927–932. [10,12]

Massoulié, L. (2014), "Community Detection Thresholds and the Weak Ramanujan Property," in *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, pp. 694–703. [12]

Matias, C., and Miele, V. (2017), "Statistical Clustering of Temporal Networks Through a Dynamic Stochastic Block Model," *Journal of the Royal Statistical Society*, Series B, 79, 1119–1141. [1]

McSherry, F. (2001), "Spectral Partitioning of Random Graphs," in *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science*, pp. 529–537. IEEE. [1]

Mossel, E., Neeman, J., and Sly, A. (2018), "A Proof of the Block Model Threshold Conjecture," *Combinatorica*, 38, 665–708. [12]

Ndaoud, M. (2018), "Sharp Optimal Recovery in the Two Component Gaussian Mixture Model," arXiv preprint arXiv:1812.08078. [3]

Newman, M. (2009), *Networks: An Introduction*, Oxford: Oxford University Press. [1]

Newman, M. E. (2002), "Assortative Mixing in Networks," *Physical Review Letters*, 89, 208701. [3]

Nielsen, A. M., and Witten, D. (2018), "The Multiple Random Dot Product Graph Model," arXiv preprint arXiv:1811.12172. [12]

O'Rourke, S., Vu, V., and Wang, K. (2018), "Random Perturbation of Low Rank Matrices: Improving Classical Bounds," *Linear Algebra and its Applications*, 540, 26–59. [3]

Paul, A., Cai, Y., Atwal, G. S., and Huang, Z. J. (2012), "Developmental Coordination of Gene Expression between Synaptic Partners during GABAergic Circuit Assembly in Cerebellar Cortex," *Frontiers in Neural Circuits*, 6, 37. [2]

Paul, S., and Chen, Y. (2020), "Spectral and Matrix Factorization Methods for Consistent Community Detection in Multi-Layer Networks," *The Annals of Statistics*, 48, 230–250. [2,3,9,12]

Peixoto, T. P. (2013), "Parsimonious Module Inference in Large Networks," *Physical Review Letters*, 110, 148701. [1]

Pensky, M., and Zhang, T. (2019), "Spectral Clustering in the Dynamic Stochastic Block Model," *Electronic Journal of Statistics*, 13, 678–709. [2,12]

Rohe, K., Chatterjee, S., and Yu, B. (2011), "Spectral Clustering and the High-Dimensional Stochastic Block Model," *The Annals of Statistics*, 39, 1878–1915. [1]

Rudelson, M., and Vershynin, R. (2013), "Hanson-Wright Inequality and Sub-Gaussian Concentration," *Electronic Communications in Probability*, 18, 1–9. [3]

Székely, G. J., and Rizzo, M. L. (2014), "Partial Distance Correlation with Methods for Dissimilarities," *The Annals of Statistics*, 42, 2382–2412. [3]

Tang, W., Lu, Z., and Dhillon, I. S. (2009), "Clustering with Multiple Graphs," in *International Conference on Data Mining (ICDM)*, IEEE, pp. 1016–1021. [1]

Tropp, J. A. (2012), "User-Friendly Tail Bounds for Sums of Random Matrices," *Foundations of Computational Mathematics*, 12, 389–434. [3,7]

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer-Verlag. [6]

Vershynin, R. (2011), "Spectral Norm of Products of Random and Deterministic Matrices," *Probability Theory and Related Fields*, 150, 471–509. [3,7]

Wang, T., Berthet, Q., and Plan, Y. (2016), "Average-Case Hardness of RIP Certification," in *Advances in Neural Information Processing Systems*, pp. 3819–3827. [7]

Werling, D. M., Pochareddy, S., Choi, J., An, J.-Y., Sheppard, B., Peng, M., Li, Z., Dastmalchi, C., Santpere, G., Sousa, A. M., et al. (2020), "Whole-Genome and RNA Sequencing Reveal Variation and Transcriptomic Coordination in the Developing Human Prefrontal Cortex," *Cell Reports*, 31, 107489. [2]

Xu, K. S., and Hero, A. O. (2014), "Dynamic Stochastic Blockmodels for Time-Evolving Social Networks," *IEEE Journal of Selected Topics in Signal Processing*, 8, 552–562. [1]

Zhang, A., and Xia, D. (2018), "Tensor SVD: Statistical and Computational Limits," *IEEE Transactions on Information Theory*, 64, 7311–7338. [9]

Zhang, A. R., Cai, T. T., and Wu, Y. (2022), "Heteroskedastic PCA: Algorithm, Optimality, and Applications," *The Annals of Statistics*, 50, 53–80. [3]

Zhang, A. Y., and Zhou, H. H. (2016), "Minimax Rates of Community Detection in Stochastic Block Models," *The Annals of Statistics*, 44, 2252–2280. [12]

Zhang, J., and Cao, J. (2017), "Finding Common Modules in a Time-Varying Network with Application to the Drosophila Melanogaster Gene Regulation Network," *Journal of the American Statistical Association*, 112, 994–1008. [1]