Biometrika (-), -, -, *pp*. 1–7
Printed in Great Britain -

Consistent community detection in multi-layer network data

BY JING LEI

Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.
jinglei@andrew.cmu.edu

KEHUI CHEN AND BRIAN LYNCH

Department of Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, U.S.A.
khchen@pitt.edu bcl28@pitt.edu

SUMMARY

We consider multi-layer network data where the relationships between pairs of elements are reflected in multiple modalities and may be described by multivariate or even high-dimensional vectors. Under the multi-layer stochastic block model framework, we derive consistency results for a least squares estimation of memberships. Our theorems show that, as compared to single-layer community detection, a multi-layer network provides much richer information that allows for consistent community detection from a much sparser network, with required edge density reduced by a factor of the square root of the number of layers. Moreover, the multi-layer framework can detect cohesive community structure across layers, which might be hard to detect by any single-layer or simple aggregation. Simulations and a data example are provided to support the theoretical results.

Some key words: Community detection; Consistency; Sparse network; Tensor concentration bound.

1. INTRODUCTION

A single-layer network consists of a set of n elements and a measure of pairwise interaction between them. The observed data is represented by an adjacency matrix or a more general relationship matrix $A \in \mathbb{R}^{n\times n}$, where A_{jk} $(j,k=1,\ldots,n)$ is a measure of interaction between element j and element k. Recently, many examples have shown that the relationships between different elements are reflected in multiple modalities, and that the observations may contain multivariate or even high-dimensional vectors that describe the relationship between each pair of elements. For example, in multimodal or multi-task brain connectivity studies, among a set of brain regions, one may have one source of linkage inferred from electroencephalography measures during a working memory task, and a second source of linkage inferred from resting state functional magnetic resonance imaging measures. Other examples include social networks, where the interaction between two people could be inferred from Facebook, LinkedIn, and more intimate connections such as cell phone contacts. Moreover, time-evolving networks can also be considered multi-layer network data when the set of elements remains the same over time.

A multi-layer network can be represented by a tensor object $Y \in \mathbb{R}^{m\times n\times n}$, where each layer $Y_{i\cdot\cdot}$ $(i=1,\ldots,m)$, represents a different aspect of the relationship between elements. In this paper, we will focus on multi-layer relational data with community structures, and utilize a

multi-layer stochastic block model point of view. The stochastic block model and its variants are powerful tools for modeling large networks with community structures. A single-layer stochastic block model (Holland et al., 1983) can be parametrized by (g, B). The observed adjacency matrix A satisfies $A_{jk} \sim \text{Bernoulli}(B_{g_j g_k})$ with $g \in \{1, \ldots, K\}^n$ and $B \in [0,1]^{K\times K}$. A natural extension to multi-layer stochastic block models allows the community-wise connectivity parameter B to depend on the layer. In this paper, we aim to find the overall clustering pattern of the nodes that are characterized by multiple modalities in a network structure, not the individual clustering pattern in each modality. Therefore, in our setting, the membership g is the same across layers. Allowing memberships to change in different layers could also be of interest in some applications. For example, there are works on dynamic stochastic block models where the memberships are allowed to change smoothly or follow some parametric patterns over time (Ghasemian et al., 2016; Pensky & Zhang, 2019).

In the last few years, a considerable amount of work has emerged on community detection for multi-layer networks, including the weighted modularity approach, spectral methods based on various versions of aggregation or tensor singular value decomposition, and probability model-based approaches (Tang et al., 2009; Dong et al., 2012; Kivelä et al., 2014; Xu & Hero, 2014; Han et al., 2015; Ghasemian et al., 2016; Paul & Chen, 2016, 2017; Chen & Hero, 2017; Zhang & Cao, 2017; Matias & Miele, 2017; Liu et al., 2018; Bhattacharyya & Chatterjee, 2018). Despite an explosion of disparate terminology and algorithms, there is limited work on theoretical analysis of detection limits and consistency results for the multi-layer network structure.

Han et al. (2015) studied the consistency of clustering for the same multi-layer stochastic block model studied here, but under a different asymptotic regime where the number of nodes n is fixed and the number of layers m grows to infinity. In the physics literature, Taylor et al. (2016) considered an identical edge probability matrix B across layers, in which case a signal boost can be achieved by working on the average adjacency matrix of all layers. A recent manuscript by Bhattacharyya & Chatterjee (2018) provided consistency results for spectral methods under sparse multi-layer stochastic block models, but required each layer of the connectivity matrix B to be positive definite, with the smallest singular values bounded away from zero uniformly. Paul & Chen (2017) considered other estimation methods and achieved a similar signal boost under the same positivity conditions. Pensky & Zhang (2019) considered estimating the membership for each individual layer in a dynamic stochastic block model. In the special case that memberships do not change, the method works on the average adjacency matrix and allows a similar signal boost under the positivity assumption.

The main contribution of this paper is that we propose to work with the m-layer network data directly without first averaging the adjacency matrices across layers, and the theoretical results are for general structures of B in m layers. In the proposed research, we derive a least squares estimation of memberships, and show that an m-layer network provides much richer information, allowing consistent estimation to be achieved for a sparser network, roughly by a factor of $m^{1/2}$, in each layer. The multi-layer framework only requires a well-defined block structure on the overall m layers, i.e., it is possible that none of the individual layers contain a full block structure. The theoretical analysis involves the development of a new spectral bound for tensor network data, which uses a tensor adaptation of the combinatorial approach developed in Lei & Rinaldo (2015). This new tensor concentration bound is crucial to develop the consistency result for multi-layer networks under weaker conditions on B.

2. TENSOR STOCHASTIC BLOCK MODELS

We use the symbol $\circ$ to denote the outer product of vectors. For example, if x, y, and z are vectors, then $T = x \circ y \circ z$ is the three-way tensor with $T_{ijk} = x_i y_j z_k$. For two tensors A and B, both of dimension (m, n_1, n_2), the symbol $A * B$ denotes the $m \times 1$ vector obtained by taking element-wise products of A and B, and then summing over the second and third dimensions. Finally, $\|\cdot\|^2$ denotes the sum of squares for all entries of a vector, matrix, or tensor.

A traditional single-layer stochastic block model with n nodes and K communities is parameterized by (g, B), where $g \in \{1, \ldots, K\}^n$ is a membership vector, and the $K \times K$ matrix B determines the community-wise connectivity. We also define the $n \times K$ membership matrix G such that the jth row of G is 1 in the g_jth column and 0 otherwise. The observed data A_{jk} $(j, k = 1, \ldots, n)$ has expectation $P_{jk} = B_{g_j g_k}$. The key feature of a stochastic block model is that the expectation P can be reorganized as a block-wise constant matrix by grouping nodes in the same community. In the most commonly seen Bernoulli model, A_{jk} only takes two values, 0 and 1. The edges form independently with $\mathrm{pr}(A_{jk} = 1) = P_{jk}$.

This generative model can be naturally extended to a multi-layer network. Let Y be an $m \times n \times n$ tensor, with each layer $Y_{i\cdot\cdot}$ being a random graph generated by a stochastic block model parametrized by $(g, B_{i\cdot\cdot})$. The expectation P is a $m \times n \times n$ tensor with $P_{i\cdot\cdot} = G B_{i\cdot\cdot} G^T$. In our setting, the membership vector is assumed to be common to all layers, while the connectivity parameter $B_{i\cdot\cdot}$ could be different across layers, reflecting different aspects of node interactions. Here $B_{i g_j g_k}$ denotes the g_jth row and g_kth column of the connectivity matrix for the ith layer, which equals P_{ijk}.

For example, consider a three-layer network, where each layer is generated from a three-block stochastic block model, with connectivity matrices

$$B_{1\cdot\cdot} = \begin{pmatrix} 0.6 & 0.4 & 0.4 \\ 0.4 & 0.2 & 0.2 \\ 0.4 & 0.2 & 0.2 \end{pmatrix}, \; B_{2\cdot\cdot} = \begin{pmatrix} 0.2 & 0.4 & 0.2 \\ 0.4 & 0.6 & 0.4 \\ 0.2 & 0.4 & 0.2 \end{pmatrix}, \; B_{3\cdot\cdot} = \begin{pmatrix} 0.2 & 0.2 & 0.4 \\ 0.2 & 0.2 & 0.4 \\ 0.4 & 0.4 & 0.6 \end{pmatrix}. \quad (1)$$

In this three-layer network, the ith community is more active in layer i, as reflected in the community-wise connectivity matrix $B_{i\cdot\cdot}$. From the ith layer, we can only separate the ith community from the rest. If we average the three layers to form a single-layer network, the community structure cannot be detected at all, since the average $(B_{1\cdot\cdot} + B_{2\cdot\cdot} + B_{3\cdot\cdot})/3$ is a matrix with the same value in all entries. By contrast, using the multi-layer network method developed in this paper, we are able to obtain consistent community recovery based on the tensor observation Y generated from this type of stochastic block model.

3. A LEAST SQUARES APPROACH AND ITS STATISTICAL PROPERTIES

A popular and accurate estimation method for stochastic block models is the maximum likelihood estimator, for which the consistency of membership estimation in single-layer networks has been studied by many authors, under the condition that a global maximum can be achieved (Bickel & Chen, 2009; Zhao et al., 2012; Choi et al., 2012; Amini et al., 2013; Abbe et al., 2016). Here we focus on its variant, the least squares estimator (Gao et al., 2015; Borgs et al., 2015; Chen & Lei, 2018). In practice, we have found that the least squares estimator performs at least as well as the maximum likelihood estimator. Our proposed theoretical analysis based on the least squares estimator will reveal unique features of multi-layer data.

For multi-layer network data, given an observation $Y \in \mathbb{R}^{m\times n\times n}$ and the number of communities K, the least squares estimator is

$$(\hat{g}, \hat{B}) = \arg\min_{h,\tilde{B}} \sum_{i=1}^{m} \omega_i \sum_{1\le j\ne l\le n} (Y_{ijl} - \tilde{B}_{ih_jh_l})^2, \tag{2}$$

where $\omega = (\omega_1, \ldots, \omega_m)$ are user-defined weights for each layer. The minimization is over all possible $h \in \{1, \ldots, K\}^n$ and all possible $\tilde{B} \in \mathbb{R}^{m\times K\times K}$. At first sight, there are a large number of parameters to estimate, but some derivation reveals that the optimal $\tilde{B}$ is uniquely determined given a membership vector. This is analogous to the profile likelihood perspective used in Stephens (2000); Bickel & Chen (2009).

In the following, we present consistency results for the optimal solution to problem (2). For a fixed community vector h, the optimization problem (2) over $\tilde{B}$ is a simple least squares fit, and the optimal $\tilde{B}$ is obtained by averaging the corresponding entries of Y according to the membership given by h. After profiling out $\tilde{B}$ in (2), the optimization problem essentially involves finding a partition h such that the within-group residual sum of squares is minimized when the entries of each layer of Y are partitioned into $K(K+1)/2$ blocks according to h. Thus, the well-known total variance decomposition implies that the profiled optimization problem over h is equivalent to maximizing the between-group sum of squares,

$$\begin{aligned} f(h; Y) = &\sum_{k=1}^{K} \binom{n_k(h)}{2} \left\| \frac{Y * (\omega \circ H_k \circ H_k)}{n_k(h)\{n_k(h)-1\}} \right\|^2 \\ &+ \sum_{1\le j<k\le K} n_j(h) n_k(h) \left\| \frac{Y * (\omega \circ H_j \circ H_k)}{n_j(h) n_k(h)} \right\|^2, \end{aligned} \tag{3}$$

where H is the membership matrix corresponding to the membership vector h, i.e., the ith row of H is 1 at the location h_i and zero otherwise. We let H_k be the kth column of H and $n_k(h) = \|H_k\|_1$.

In the rest of the paper, we use $\omega_i = 1$ $(i = 1, \ldots, m)$. The theory can be easily extended to the more general case of unequal ω_i's. To accurately describe the community recovery error in terms of the network sparsity and community separation, we adopt the following settings, which we state as assumptions:

Assumption 1. Network sparsity: $B = \rho_n B^0$, where ρ_n controls the overall network sparsity and the entries of B^0 are of constant order with maximum entry equal to 1.

Assumption 2. Community separation: $\delta^2 = \min_{1\le j\ne j'\le K} \|B^0_{\cdot j\cdot} - B^0_{\cdot j'\cdot}\|^2 > 0$.

Assumption 3. Assume $m \le cn$ for some constant c.

Assumption 1 is common in the literature of network community detection. Assumption 2 is a minimum requirement for K-community structure. Assumption 3 is made merely for presentational simplicity in the main theorem. The analysis covers the case of larger m as well; see Remark 1 at the end of this section for further details.

Our theoretical analysis of the least squares estimator in (2) consists of three main steps, which we outline next as Lemma 1 through Lemma 3. The key technical component is a new spectral concentration bound for tensor data, which may be of independent interest. We state the tensor concentration theorem at the end of this section. All proofs are given in the online supplementary material.

LEMMA 1. *Under Assumption* 2*,* $f(h;P)$ *is uniquely maximized by* $h = g$*, up to a label permutation.*

This lemma ensures parameter identification as the population optimum.

LEMMA 2. *Under Assumptions* 1 *and* 3*, with probability tending to one as* $n \to \infty$*, we have, for some universal constant* c_1*,*

$$\sup_{h\in\{1,\dots,K\}^n} |f(h;Y) - f(h;P)| \le c_1\kappa_n,$$

where $\kappa_n = K(\log n)\{(n\rho_n) \vee \log n\}^{1/2}\left[n\rho_n m^{1/2} + K(\log n)\{(n\rho_n) \vee \log n\}^{1/2}\right]$.

This lemma gives a uniform upper bound for the sampling errors. This is a main technical contribution of this paper. The proof of Lemma 2 relies on Theorem 2 stated at the end of this section, which extends recent spectral bounds for single-layer network data (Lei & Rinaldo, 2015; Chin et al., 2015; Le et al., 2017). The proof technique is a tensor adaptation of the combinatorial approach developed in Feige & Ofek (2005) and Lei & Rinaldo (2015).

Next we need to analyze the population optimality gap for incorrect membership vectors. Let g be the true membership and h be any membership vector. For $k = 1,\dots,K$ define $C_k = \{1 \le j \le n : g_j = k\}$ and for any other membership vector h define $\hat{C}_k = \{1 \le j \le n : h_j = k\}$. Define $e_h(k,l)$ as the number of nodes labeled as k in h and l in g:

$$e_h(k,l) = |C_l \cap \hat{C}_k|\,.$$

For $k = 1,...,K$ and membership vector h, let $e_h(k)$ be the second largest value in $\{e_h(k,l) : 1 \le l \le K\}$. Define

$$\eta_h = \max_k e_h(k)/n_{\min}\,,$$

where $n_{\min}$ is the smallest community size in g.

Intuitively, the largest value in $\{e_h(k,l) : 1 \le l \le K\}$ corresponds to the correctly clustered nodes, so $(K-1)e_h(k)$ can be viewed as an upper bound on the number of incorrectly clustered nodes in $\hat{C}_k$. When η_h is small, each of the true communities C_l must assign a majority proportion of its nodes to some $\hat{C}_{k(l)}$, and the mapping from l to $k(l)$ must be one-to-one since otherwise η_h will be large. Therefore, our basic strategy of proving consistency of membership estimation is to show that $\eta_h = o_P(1)$ if h is an optimal solution to (3).

LEMMA 3. *Under Assumption* 2*, there exists a universal constant* c_2 *such that*

$$f(g;P) - f(h;P) \ge c_2\eta_h n_{\min}^2\rho_n^2\delta^2 K^{-1}.$$

Combining the above three lemmas, we have the following main result:

THEOREM 1 (MAIN THEOREM). *Let* h *be a solution to the least squares problem. Then, under Assumptions* 1*,* 2*, and* 3*, there exists a universal constant* $C > 0$ *such that the following statements hold with probability tending to one:*

In the sparse case, where $n\rho_n < \log n$*,*

$$\eta_h \le CK\left(\frac{n}{n_{\min}}\right)^2\left(\frac{m}{\delta^2}\right)\left\{\frac{K(\log n)^{3/2}}{n\rho_n m^{1/2}}\right\}\left\{1 + \frac{K(\log n)^{3/2}}{n\rho_n m^{1/2}}\right\}.$$

In the moderately dense case, where $n\rho_n \geq \log n$,

$$\eta_h \leq CK \left(\frac{n}{n_{\min}}\right)^2 \left(\frac{m}{\delta^2}\right) \left\{\frac{K \log n}{(n\rho_n m)^{1/2}}\right\} \left\{1 + \frac{K \log n}{(n\rho_n m)^{1/2}}\right\}.$$

A prototypical case in the study of stochastic block models is the balanced community case, where $K = O(1)$ and $n_{\min} \asymp n$. Moreover, it is natural to assume that $\delta^2 \asymp m$. This is the case if a constant fraction of layers in the multi-layer stochastic block model exhibits the same scale of between-community connectivity difference, or more precisely, if there is a constant fraction of i's in $\{1, ..., m\}$ such that $\max_{j,j'} \|B^0_{ij\cdot} - B^0_{ij'\cdot}\| \geq c$ for some positive constant c. In particular, if the layers of B^0 are generated independently from a non-degenerate distribution, we have $\delta^2 \asymp m$ with high probability when m is large.

The state-of-the-art one-layer stochastic block model result requires $\rho_n n \to \infty$ for consistent community recovery. Under the standard assumptions made above, in the sparse case where $n\rho_n < \log n$, we only require $n\rho_n m^{1/2}/(\log n)^{3/2} \to \infty$ to guarantee a vanishing proportion of mis-clustered nodes. Roughly speaking, the m-layer least squares estimator combines the signal from all layers and enhances the signal strength by a factor of $m^{1/2}$.

Finally we introduce the key technical component in our proof: the tensor concentration result. Let A be an $n \times n \times m$ tensor, with each layer $A_{\cdot\cdot l}$ being an inhomogeneous Erdős–Rényi random graph with expectation P. The maximum entry of P is of order ρ_n. For presentational simplicity we assume that $m = n$. The more general case can be treated by padding the tensor with zeros when $m < n$. The case of $m > n$ is discussed in Remark 1 below.

THEOREM 2. *For any* $(x, y, z) \in \mathbb{S}^n \times \mathbb{S}^n \times \mathbb{S}^n$, *let* $W = A - P$. *We have* $|\langle W, x \circ y \circ z\rangle| \leq c(\log n)\{(n\rho_n) \vee \log n\}^{1/2}$ *for some universal constant* c *with all but vanishing probability as* $n \to \infty$.

Remark 1. When $m > n$ the tensor spectral concentration bound in Theorem 2 becomes $(\log m)\{(m\rho_n) \vee \log m\}^{1/2}$. When $m > n$ but $m = O(n^2)$, using the same analysis as in the proof of Theorem 1, the least squares estimator can achieve consistent community estimation when $n\rho_n$ grows faster than $(\log n)^{3/2}/m^{1/2}$, which is still an $m^{1/2}$ improvement over the density requirement for single-layer networks. A larger m beyond the order of n^2 can further reduce the required sparsity level ρ_n, but the rate of signal boost is no longer $m^{1/2}$. This regime is of less practical interest, because $\rho_n \gg n^{-2}$ is a minimum requirement for each layer to have at least one edge.

4. NUMERICAL EXPERIMENTS

4·1. *Algorithm*

The theoretical results developed above pertain to a global optimum of problem (2). In practice, how to achieve or approximate the global optimum remains a challenging and interesting algorithmic question. A full search over K^n possible labelings takes exponential computing time. Greedy search methods such as the label switching algorithm (Bickel & Chen, 2009) and the tabu search (Zhao et al., 2012) have been used in network clustering. The algorithm that we used in our numerical experiments can be viewed as a label-switching method with batch updates, and can also be viewed as an adaptation of Lloyd's algorithm for k-means clustering to network data. The algorithm proceeds as follows:

(i) Initialize by k-means on n slices of the data, where the jth data point is a column slice $Y_{\cdot\cdot j}$, viewed as a vector of length mn.

(ii) Assume that the current iteration starts with a membership vector g^{old}. Find a new community vector g^{new} where

$$g_j^{new} = \arg\min_{k\in\{1,\ldots,K\}} \sum_{i=1}^{m} \omega_i \sum_{l\neq j} \{Y_{ijl} - B^{old}_{ikg_l^{old}}\}^2.$$

(iii) Compute B^{new}:

$$B^{new}_{ikk'} = \frac{\sum_{j\neq l} Y_{ijl} \mathbb{1}_{\{g_j^{new}=k\}} \mathbb{1}_{\{g_l^{new}=k'\}}}{\sum_{j\neq l} \mathbb{1}_{\{g_j^{new}=k\}} \mathbb{1}_{\{g_l^{new}=k'\}}}.$$

(iv) Compute the least squares loss function with respect to g^{new} and B^{new}, and update with $g^{old} \leftarrow g^{new}$ and $B^{old} \leftarrow B^{new}$ if the loss function reduces.

(v) Repeat steps ii-iv until the objective function cannot be further reduced.

Lloyd's algorithm is arguably the most popular approach for k-means clustering, due to its simplicity and fast convergence. It is not guaranteed to converge to a local minimum though, if the local minimum is defined as a partition of the data where moving any single point to a different cluster increases the objective function. Arthur & Vassilvitskii (2007) showed that Lloyd's algorithm combined with a good starting point instead of a purely random start can provide an accurate approximate solution with small optimality gap. Here we initialize our algorithm by k-means on slices of data, known as marginal clustering, which has been proved to be a very good initial point in the co-clustering literature (Anagnostopoulos et al., 2008). In our numerical studies, we repeat the algorithm three times and retain the choice with the smallest objective value, and use weights $\omega_i \equiv 1$. The algorithm performs very satisfactorily in our simulations shown in Section 4·2.

4·2. *Simulations*

In this section, we illustrate the performance of our proposed method using simulations where we generate multi-layer networks $Y \in \mathbb{R}^{m\times n\times n}$ given membership vector g and community connectivity tensor B. We compare our multi-layer clustering method to a single-layer method based on the same least squares criterion as in (2), either applied to only the first layer of Y, or to the average over all layers of Y. In addition, we compare our method with spectral clustering applied to the average of the layers. In each simulation trial, given a membership matrix G and a B with each layer symmetric, the upper triangular entries of $Y_{i\cdot\cdot}$ are independently generated as Bernoulli random variables with probabilities given by $P_{i\cdot\cdot} = GB_{i\cdot\cdot}G^T$. The nodes are divided into clusters such that the number of nodes in each of the first $K-1$ communities is $\lfloor n/K \rfloor$, and the Kth community contains the remaining nodes. When we instead used unequal community sizes, our method performed similarly.

In Simulation I, the entries of B are randomly generated in each trial. To do this, we first generate $B_0 \in \mathbb{R}^{m\times K\times K}$, where the upper triangular and diagonal entries of each layer $B_{0,i\cdot\cdot}$ are generated independently from Uniform$(0, 0\cdot 5)$, and the lower triangular entries are set equal to their corresponding upper triangular entries. We then set $B = rB_0$, where $r \leq 1$ is a preselected positive parameter that controls sparsity. If a layer of B_0 has Kth singular value less than a preset cutoff value, that layer is regenerated to ensure that we have well-formed K-block structures.

We consider $n \in \{50, 100, 200\}$ as the number of nodes, $m = 2^j$ $(j = 1, \ldots, 6)$ as the number of layers, $K \in \{2, 3, 4\}$ as the number of communities, and $r \in \{0\cdot 05, 0\cdot 1, 0\cdot 25, 0\cdot 5, 1\}$ as the level of sparsity. For each combination of values (n, m, r, K), we run 100 simulation trials. The proportion of nodes correctly clustered by a given method is averaged over all trials. In Figures 1 and 2 we plot the success rates against different values of r, and we only show the

results for $m = 2$ and $m = 8$, respectively, as the success rates for relatively dense networks are close to 1. The performance for $m = 4$ is in between these and we omit it to save space. In Figure 3 we show the success rate of our method against different values of $m \in \{2, 4, 8, 16, 32, 64\}$, for relatively sparse networks with $r \in \{0 \cdot 05, 0 \cdot 1, 0 \cdot 25\}$. Here, we see that sufficiently large m and n can allow for nearly 100% correct assignment, even for small r. This supports the results of Theorem 1 and the discussion following it. The performance of our multi-layer method is clearly superior to that of the single-layer methods. All of the methods show improved performance as r increases, n increases, or K decreases. The performance of our multi-layer method also improves as m increases, while the performance of the single-layer methods remains relatively static with m. We note that under the layer-wise positivity assumption, the spectral method on the average of the layers has been proven to have superior performance in Taylor et al. (2016); Paul & Chen (2017); Bhattacharyya & Chatterjee (2018). However, in our simulations, the $B_{i\cdot\cdot}$ are generated randomly without any imposed positivity, so there can be signal cancellation due to averaging. Moreover, in our simulations, the spectral method on the average of the layers does not work as well as the least squares method on the average of the layers. The reason for this is that our B matrices are randomly generated and the kth singular value of the averaged layers could be very small. Spectral methods do not work well in these scenarios.

In Simulation II, B_0 is assigned a constant value over all trials by using $m = K = 3$ and defining each layer of B_0 as in (1). In this case, the average over all layers of $B = rB_0$ is a constant matrix. Each individual layer can only identify 2 unique clusters, although there are 3 clusters when all layers are considered. Figure 4 shows simulation results in this scenario for the three least squares methods considered above, calculating the proportion of nodes correctly identified over 100 simulation trials. As before, our multi-layer method performs the best, increasing toward 100% correct clustering as r and n increase. As expected, the averaging method performs poorly irrespective of r and n. Using only the first layer results in a performance that improves slightly with r and n, but as expected never approaches 100% accuracy.

4·3. *Application to gene network study*

In this section, we apply our method to a gene network dataset. The data have been described and used by Liu et al. (2018), and include gene co-expression data in the medial prefrontal cortex from studies of rhesus monkeys at different stages of development. For the prenatal period, they consider a 6-layer network corresponding to 6 age categories, labeled E40 to E120 to indicate the number of embryonic days of age. For the postnatal period, they consider a 5-layer network corresponding to 5 layers within the medial prefrontal cortex, labeled L2 to L6. Studies of the medial prefrontal cortex have been used to understand developmental brain disorders, and Liu et al. (2018) make special note of sets of genes that are significantly enriched for neural projection guidance, which has been shown to be related to autism spectrum disorder. These genes are marked in red in their Figures 5 and S10. We focus on the set of neural projection guidance-enriched genes, which results in $n = 154$ nodes for the prenatal network, and $n = 117$ nodes for the postnatal network. The networks were constructed by soft thresholding the sample correlation computed from 423 samples from several groups, and we refer to Liu et al. (2018) for details of the calculation.

Our clustering results are visualized for the prenatal and postnatal data in Figures 5 and 6, respectively, in which each layer's connectivity matrix is ordered according to the results of the multi-layer clustering. Here $K = 4$ clusters are used based on visual inspection of the communities of red genes in Figures 5 and S10 of Liu et al. (2018). In both the prenatal and postnatal networks, the individual layers show the 4 clusters grouped in different ways, giving partial views of the overall clustering and revealing the advantage of our method in finding a cohesive portrait

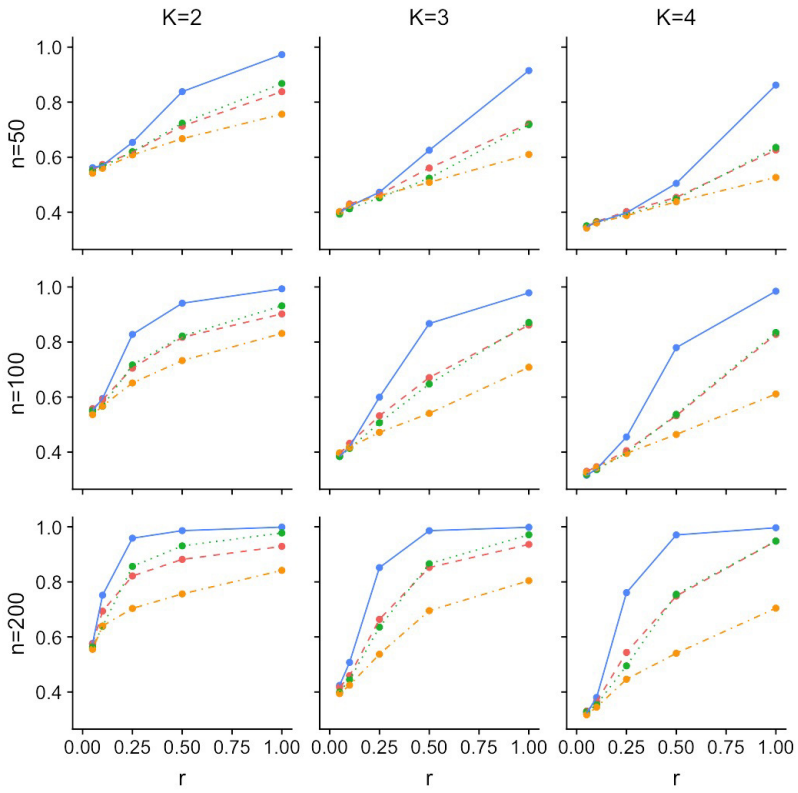

Fig. 1. Simulation I: Proportion of nodes correctly assigned for $m = 2$ layers and $r \in \{0 \cdot 05, 0 \cdot 1, 0 \cdot 25, 0 \cdot 5, 1\}$. The blue solid line is for the multi-layer method, the red dashed line is for the single-layer least squares method applied to the average of the layers, the green dotted line is for the single-layer least squares method applied to the first layer, and the orange dotted and dashed line is for the spectral method on the average of the layers.

of the communities. In Figure 5, the connectivity matrix for the first layer E40 seems to differentiate 2 groups of genes, including the first cluster and the combined next 3 clusters. The second layer E50 shows 3 groups, including the first cluster, the second 2 clusters, and the fourth cluster. In the third layer E70, clusters 1 and 3 appear distinct, while clusters 2 and 4 look like they could be grouped. In E80, E90, and E120, the last 2 clusters seem to be grouped, the first cluster seems to be either distinct or grouped with the last 2, and the second cluster is distinct. The postnatal network analysis, shown in Figure 6, reveals similar phenomena. Each of the layers gives partial or weak information about the overall clustering, and combining them gives a stronger signal that captures the structure of the gene clustering over all layers.

5. Discussion

The greedy algorithm works very well in numerical experiments, but the rigorous theoretical analysis of the approximate solutions is still an open question, even in single-layer network data analysis. The authors believe this would be an interesting and challenging topic for future study. As pointed out in the introduction, there are also various other estimators for multi-layer network

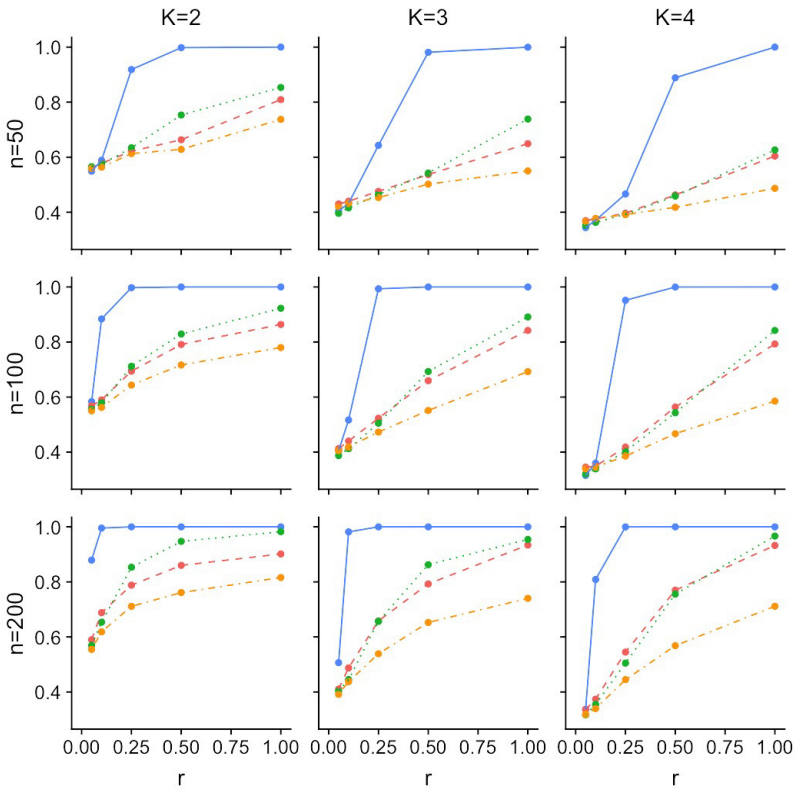

Fig. 2. Simulation I: Proportion of nodes correctly assigned for $m = 8$ layers and $r \in \{0 \cdot 05, 0 \cdot 1, 0 \cdot 25, 0 \cdot 5, 1\}$. The blue solid line is for the multi-layer method, the red dashed line is for the single-layer least squares method applied to the average of the layers, the green dotted line is for the single-layer least squares method applied to the first layer, and the orange dotted and dashed line is for the spectral method on the average of the layers.

data proposed in the literature, especially tensor spectral methods. The theoretical analysis of these methods, and subsequent comparisons, are also of interest.

Our model can also be considered through the perspective of modeling non-binary interactions, extending the traditional single-layer network that records binary interactions among nodes. There are other types of pairwise interactions considered in the literature, such as the categorical interaction in Lelarge et al. (2015) and the continuous-valued interaction in Xu et al. (2017). It would be interesting to investigate and understand these models in a unified framework.

ACKNOWLEDGEMENT

The authors are grateful for the comments the reviewers and editors have provided to improve the paper. The authors would like to thank Dr. Kathryn Roeder and Dr. Fuchen Liu for providing the gene network data and helpful discussion. Jing Lei's research is partially supported by the U.S. National Science Foundation grant DMS-1553884. Kehui Chen and Brian Lynch's research is partially supported by the U.S. National Science Foundation grant DMS-1612458.

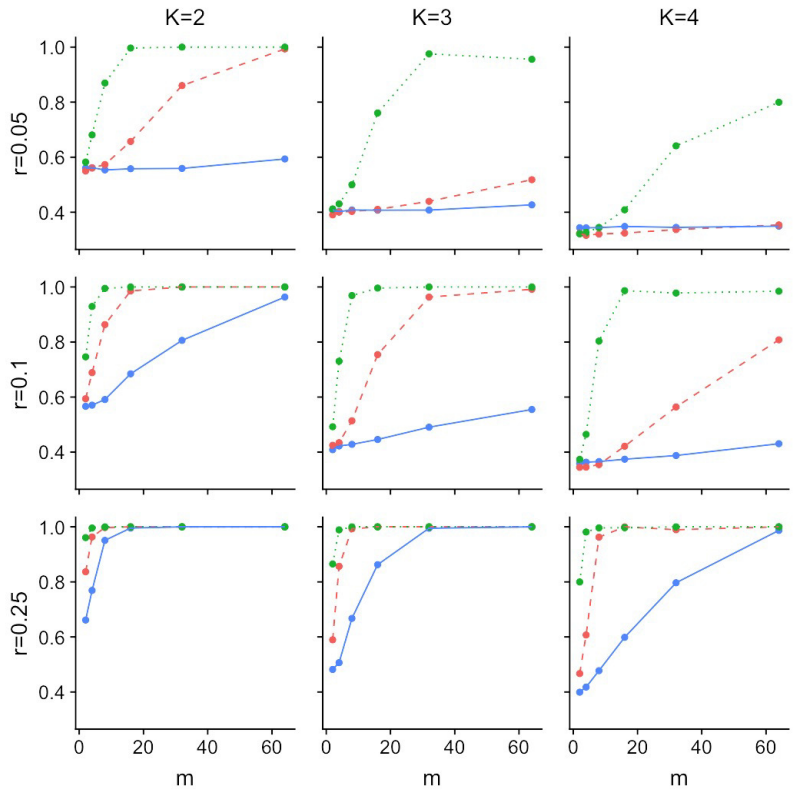

Fig. 3. Simulation I: Proportion of nodes correctly assigned by the multi-layer method, for $m \in \{2, 4, 8, 16, 32, 64\}$ and $r \in \{0 \cdot 05, 0 \cdot 1, 0 \cdot 25\}$. The blue solid line, red dashed line, and green dotted line correspond to $n = 50$, 100, and 200, respectively.

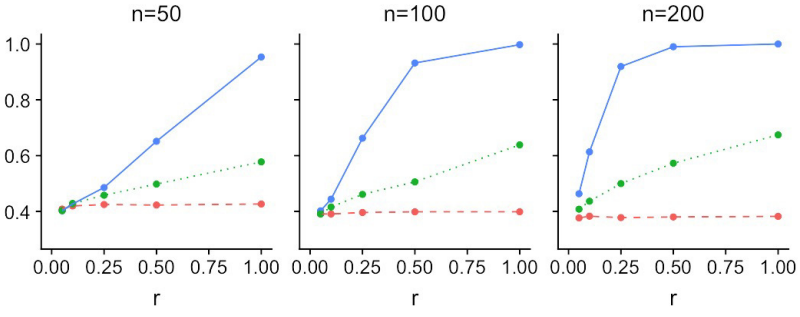

Fig. 4. Simulation II: Proportion of nodes correctly assigned for $r \in \{0 \cdot 05, 0 \cdot 1, 0 \cdot 25, 0 \cdot 5, 1\}$. The blue solid line is for the multi-layer method, the red dashed line is for the single-layer method applied to the average of the layers, and the green dotted line is for the single-layer method applied to the first layer.

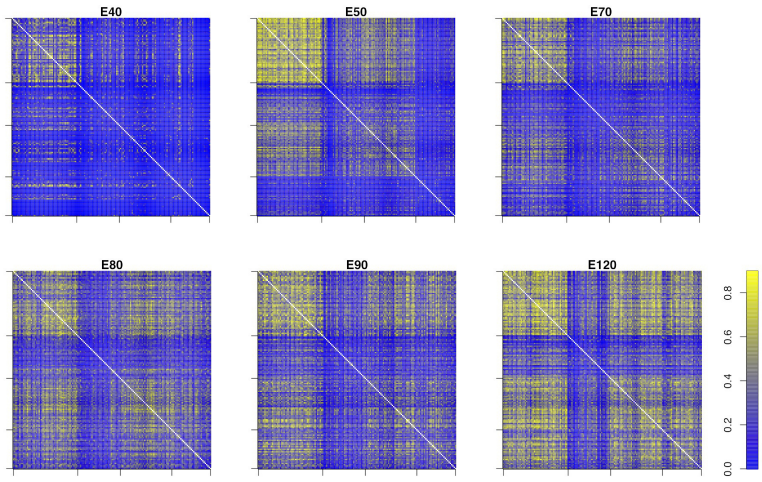

Fig. 5. The connectivity matrices for each layer of the pre-natal data, with genes ordered by the clusters. Tick marks denote the boundaries between the clusters.

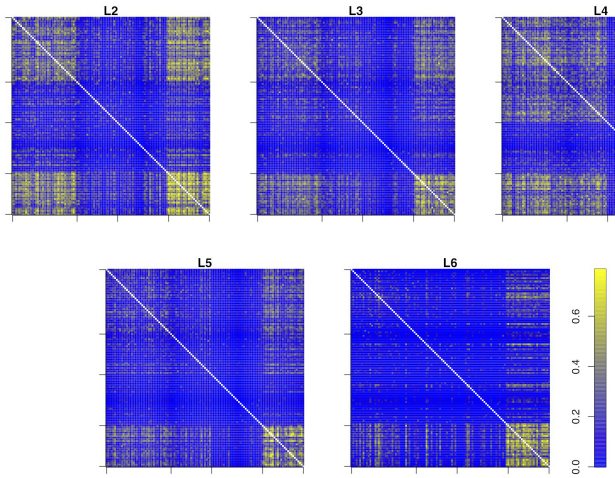

Fig. 6. The connectivity matrices for each layer of the post-natal data, with genes ordered by the clusters. Tick marks denote the boundaries between the clusters.

REFERENCES

ABBE, E., BANDEIRA, A. S. & HALL, G. (2016). Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory* **62**, 471–487.

AMINI, A. A., CHEN, A., BICKEL, P. J. & LEVINA, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics* **41**, 2097–2122.

ANAGNOSTOPOULOS, A., DASGUPTA, A. & KUMAR, R. (2008). Approximation algorithms for co-clustering. In *PODS*.

ARTHUR, D. & VASSILVITSKII, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics.

BHATTACHARYYA, S. & CHATTERJEE, S. (2018). Spectral clustering for multiple sparse networks: I. *arXiv preprint arXiv:1805.10594* .

BICKEL, P. J. & 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.

BORGS, C., CHAYES, J. & SMITH, A. (2015). Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems*.

CHEN, K. & LEI, J. (2018). Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association* **113**, 241–251.

CHEN, P.-Y. & HERO, A. O. (2017). Multilayer spectral graph clustering via convex layer aggregation: Theory and algorithms. *IEEE Transactions on Signal and Information Processing over Networks* **3**, 553–567.

CHIN, P., RAO, A. & VU, V. (2015). Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference on Learning Theory*.

CHOI, D. S., WOLFE, P. J. & AIROLDI, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99**, 273–284.

DONG, X., FROSSARD, P., VANDERGHEYNST, P. & NEFEDOV, N. (2012). Clustering with multi-layer graphs: A spectral perspective. *IEEE Trans. Signal Processing* **60**, 5820–5831.

FEIGE, U. & OFEK, E. (2005). Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms* **27**, 251–275.

GAO, C., LU, Y. & ZHOU, H. H. (2015). Rate-optimal graphon estimation. *The Annals of Statistics* **43**, 2624–2652.

GHASEMIAN, A., ZHANG, P., CLAUSET, A., MOORE, C. & PEEL, L. (2016). Detectability thresholds and optimal algorithms for community structure in dynamic networks. *Physical Review X* **6**, 031005.

HAN, Q., XU, K. & AIROLDI, E. (2015). Consistent estimation of dynamic and multi-layer block models. In *International Conference on Machine Learning*.

HOLLAND, P. W., LASKEY, K. B. & LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social networks* **5**, 109–137.

KIVELÄ, M., ARENAS, A., BARTHELEMY, M., GLEESON, J. P., MORENO, Y. & PORTER, M. A. (2014). Multilayer networks. *Journal of Complex Networks* **2**, 203–271.

LE, C. M., LEVINA, E. & VERSHYNIN, R. (2017). Concentration and regularization of random graphs. *Random Structures & Algorithms* **51**, 538–561.

LEI, J. & RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics* **43**, 215–237.

LELARGE, M., MASSOULIÉ, L. & XU, J. (2015). Reconstruction in the labelled stochastic block model. *IEEE Transactions on Network Science and Engineering* **2**, 152–163.

LIU, F., CHOI, D., XIE, L. & ROEDER, K. (2018). Global spectral clustering in dynamic networks. *Proceedings of the National Academy of Sciences* , 201718449.

MATIAS, C. & MIELE, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 1119–1141.

PAUL, S. & CHEN, Y. (2016). Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electronic Journal of Statistics* **10**, 3807–3870.

PAUL, S. & CHEN, Y. (2017). Consistency of community detection in multi-layer networks using spectral and matrix factorization methods. *arXiv preprint arXiv:1704.07353* .

PENSKY, M. & ZHANG, T. (2019). Spectral clustering in the dynamic stochastic block model. *Electronic Journal of Statistics* **13**, 678–709.

STEPHENS, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**, 795–809.

TANG, W., LU, Z. & DHILLON, I. S. (2009). Clustering with multiple graphs. In *International Conference on Data Mining (ICDM)*. IEEE.

TAYLOR, D., SHAI, S., STANLEY, N. & MUCHA, P. J. (2016). Enhanced detectability of community structure in multilayer networks through layer aggregation. *Physical review letters* **116**, 228301.

XU, K. S. & HERO, A. O. (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing* **8**, 552–562.

XU, M., JOG, V. & LOH, P.-L. (2017). Optimal rates for community estimation in the weighted stochastic block model. *arXiv preprint arXiv:1706.01175* .

ZHANG, J. & 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.

ZHAO, Y., LEVINA, E. & ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics* **40**, 2266–2292.

Supplementary Material for "Consistent community detection in multi-layer network data"

In this supplementary material we prove technical results in the article "Consistent community detection in multi-layer network data" by Lei, Chen, and Lynch. All references, as well as theorem and equation numbering, refer to the original paper.

Proof of Lemmas and Main Theorem

Proof of Lemma 1. The discussion preceding (3) implies that maximizing $f(h;P)$ over h is equivalent to minimizing the profiled version of (2) over all possible h with Y replaced by P. By construction of P, the objective function in (2) equals zero when $h=g$. When h, g correspond to different partitions, there exist $1\le j<l\le n$ such that $h_j=h_l$ but $g_j\neq g_l$. Then there exist $1\le i\le m$ and $1\le k\le n$ such that $P_{ijk}\neq P_{ilk}$, so that $P_{ijk}-\tilde{B}_{ih_jh_k}$ and $P_{ilk}-\tilde{B}_{ih_lh_k}$ cannot both be zero. So the objective function (2) must be strictly positive. Therefore if h achieves the minimum of (2), then $h_j=h_l$ implies $g_j=g_l$. Given that h contains at most K groups, we conclude that $h=g$ up to label permutation. □

Proof of Lemma 2. A little rearranging gives

$$2f(h;Y)=\sum_{k=1}^{K}\frac{\|Y*(\omega\circ H_k\circ H_k)\|^2}{n_k(h)\{n_k(h)-1\}}+\sum_{k\neq l}\frac{\|Y*(\omega\circ H_k\circ H_l)\|^2}{n_k(h)n_l(h)}.$$

We only focus on $\omega_i=1$ $(i=1,\dots,m)$. For the diagonal blocks, if $n_k(h)\le 1$ for some k, then $Y*(\omega\circ H_k\circ H_k)=0=P*(\omega\circ H_k\circ H_k)$ so we can focus on the case $n_k(h)\ge 2$. Then by the Cauchy-Schwartz inequality,

$$\begin{aligned}
&\frac{\|Y*(\omega\circ H_k\circ H_k)\|^2}{n_k(h)\{n_k(h)-1\}}-\frac{\|P*(\omega\circ H_k\circ H_k)\|^2}{n_k(h)\{n_k(h)-1\}}\\
=&\frac{\langle 2P*(\omega\circ H_k\circ H_k)+(Y-P)*(\omega\circ H_k\circ H_k),\ (Y-P)*(\omega\circ H_k\circ H_k)\rangle}{n_k(h)\{n_k(h)-1\}}\\
\lesssim&\frac{n_k^2(h)m^{1/2}p_{\max}n_k(h)\log n\{(np_{\max})\vee\log n\}^{1/2}+n_k^2(h)(\log n)^2\{(np_{\max})\vee\log n\}}{n_k^2(h)}\\
=&n_k(h)m^{1/2}p_{\max}\log n\{(np_{\max})\vee\log n\}^{1/2}+(\log n)^2\{(np_{\max})\vee\log n\},
\end{aligned}$$

where the third line uses $\|P*(\omega\circ H_k\circ H_k)\|\lesssim n_k^2(h)m^{1/2}p_{\max}$ and $\|(Y-P)*(\omega\circ H_k\circ H_k)\|\lesssim n_k(h)\log n\{(np_{\max})\vee\log n\}^{1/2}$ with high probability by Theorem 2. To see the latter inequality, recall that $w=(1,...,1)^T$,

$$\begin{aligned}
\|(Y-P)*(\omega\circ H_k\circ H_k)\|=&\max_{x\in\mathbb{R}^m:\|x\|=1}\langle(Y-P)*(\omega\circ H_k\circ H_k),x\rangle\\
=&\max_{x\in\mathbb{R}^m:\|x\|=1}\langle Y-P,x\circ H_k\circ H_k\rangle\\
=&n_k\max_{x\in\mathbb{R}^m:\|x\|=1}\langle Y-P,x\circ\frac{H_k}{\sqrt{n_k}}\circ\frac{H_k}{\sqrt{n_k}}\rangle\\
\le&n_k\max_{x\in\mathbb{R}^m:\|x\|=1;y,z\in\mathbb{R}^n,\|y\|=\|z\|=1}\langle Y-P,x\circ y\circ z\rangle,
\end{aligned}$$

and the rest follows from Theorem 2.

For the off-diagonal blocks, similarly,

$$\begin{aligned}
&\frac{\|Y*(\omega\circ H_k\circ H_l)\|^2}{n_k(h)n_l(h)}-\frac{\|P*(\omega\circ H_k\circ H_l)\|^2}{n_k(h)n_l(h)}\\
=&\frac{\langle 2P*(\omega\circ H_k\circ H_l)+(Y-P)*(\omega\circ H_k\circ H_l),\ (A-P)*(\omega\circ H_k\circ H_l)\rangle}{n_k(h)n_l(h)}\\
\lesssim&\{n_k(h)n_l(h)\}^{1/2}p_{\max}\log n\{m(np_{\max}\vee\log n)\}^{1/2}+(\log n)^2(np_{\max}\vee\log n).
\end{aligned}$$

Summing over k,l we have

$$\begin{aligned}
|f(h;A)-f(h;P)|\lesssim&p_{\max}\log n\{m(np_{\max}\vee\log n)\}^{1/2}\sum_{k,l=1,\dots,K}(n_kn_l)^{1/2}+K^2(\log n)^2(np_{\max}\vee\log n)\\
=&p_{\max}\log n\{m(np_{\max}\vee\log n)\}^{1/2}\left(\sum_{k=1}^K n_k^{1/2}\right)^2+K^2(\log n)^2(np_{\max}\vee\log n)\\
\le&Knp_{\max}\log n\{m(np_{\max}\vee\log n)\}^{1/2}+K^2(\log n)^2(np_{\max}\vee\log n)\equiv\kappa_n.
\end{aligned}$$

□

Proof of Lemma 3. Let $h\neq g$ be another membership vector. Let $C_k=\{1\le i\le n: g_i=k\}$ be the set of nodes in the kth true cluster, and $\hat{C}_k=\{1\le i\le n: h_i=k\}$ be the set of nodes in the kth h-cluster. Then by definition $e_h(k,l)=|\hat{C}_k\cap C_l|$. When $h\neq g$, there exists (k,l,l') such that $l\neq l'$, $e_h(k,l)>0$, and $e_h(k,l')>0$. Without loss of generality, assume $k=1$, $l=1$, $l'=2$, and $e_h(1,1)\ge e_h(1,2)$.

By assumption there exists k such that $\|B_{\cdot 1k}-B_{\cdot 2k}\|_2\ge\delta$. Given this k, there exists an l such that $e_h(l,k)\ge n_k/K$.

Let m_1 be the number of distinct node pairs in $(\hat{C}_1\times\hat{C}_l)\cap(C_1\times C_k)$. Let m_2 be the number of unique node pairs in $(\hat{C}_1\times\hat{C}_l)\cap(C_2\times C_k)$. We have $m_1\gtrsim e_h(1,1)n_k/K$ and $m_2\gtrsim e_h(1,2)n_k/K$.

Then the within-cluster variance under h is at least $\{m_1m_2/(m_1+m_2)\}(\rho_n\delta)^2$ and we have the following optimality gap:

$$f(g,P)-f(h,P)\ge\frac{m_1m_2}{m_1+m_2}(\rho_n\delta)^2\gtrsim(\rho_n\delta)^2e_h(1,2)n_k/K.$$

The claim follows from $e_h(1,2)\ge\eta_h n_{\min}$ and $n_k\ge n_{\min}$. □

Proof of Theorem 1. Combining Lemma 3 and Lemma 2, we have, with high probability,

$$\eta_h\le\frac{K\{f(g;P)-f(h;P)\}}{c_2n_{\min}^2\delta^2}\le\frac{2Kc_1\kappa_n}{c_2n_{\min}^2\rho_n^2\delta^2}. \tag{S.1}$$

In the sparse case, under the assumption $n\rho_n\le\log n$, we have $\kappa_n=K(\log n)^{3/2}\left\{n\rho_nm^{1/2}+K(\log n)^{3/2}\right\}$. Plugging κ_n in (S.1) we have

$$\eta_h\le\frac{2cK^2n(\log n)^{3/2}\rho_nm^{1/2}}{n_{\min}^2\rho_n^2\delta^2}\left\{1+\frac{K(\log n)^{3/2}}{n\rho_nm^{1/2}}\right\}=cK\frac{n^2m}{n_{\min}^2\delta^2}\left\{\frac{K(\log n)^{3/2}}{n\rho_nm^{1/2}}\right\}\left\{1+\frac{K(\log n)^{3/2}}{n\rho_nm^{1/2}}\right\}.$$

Similarly, the claim in the moderately dense case follows by plugging in $\kappa_n=K\log n(n\rho_n)^{1/2}\left\{n\rho_nm^{1/2}+K\log n(n\rho_n)^{1/2}\right\}$ in (S.1). □

TENSOR CONCENTRATION BOUND

Proof of Theorem 2. Fix $\delta\in(0,1)$; for example, we can take $\delta=1/2$. Let T be the intersection of the n-dimensional unit ball and points whose coordinates are grid points of length $\delta/n^{1/2}$. For each $u\in\mathbb{R}^n$ such that $\|u\|\le 1-\delta$, the cube of side length $\delta/n^{1/2}$ centered at u is contained in T. It then follows that

$(1-\delta)\mathbb{S}^n \subseteq \text{convhull}(T)$. As a consequence, for $(x,y,z) \in \mathbb{S}^n \times \mathbb{S}^n \times \mathbb{S}^n$,

$$\begin{aligned}|\langle W, x\circ y\circ z\rangle| =& (1-\delta)^{-3}\,|\langle W, (1-\delta)x\circ(1-\delta)y\circ(1-\delta)z\rangle| \\ \leq& (1-\delta)^{-3} \sup_{(x,y,z)\in T\times T\times T} |\langle W, x\circ y\circ z\rangle|\,.\end{aligned}$$

Therefore, we only need to deal with the vectors in $T\times T\times T$.

Let $d = (n\rho)\vee \log n$ (we use ρ for ρ_n).

For each $(x,y,z)\in T\times T\times T$, consider index sets $\mathcal{L}_{x,y,z} = \left\{(i,j,l) : |x_i y_j z_l| \leq d^{1/2}/n\right\}$. These are what we call the light triplets of (i,j,l).

For $(x,y,z)\in T\times T\times T$, let $u_{ijl} = x_i y_j z_l 1(|x_i y_j z_l| \leq d^{1/2}/n) + x_j y_i z_l 1(|x_j y_i z_l| \leq d^{1/2}/n)$. Again using Bernstein's inequality and the fact that $\sum_{i<j,l} u_{ijl}^2 \leq 2$ we have

$$\text{pr}\left(\left|\sum_{(i,j,l)\in\mathcal{L}_{x,y,z}} w_{ijl}u_{ijl}\right| \geq cd^{1/2}\right) \leq 2\exp\left\{-\frac{\frac{1}{2}c^2 d}{\rho\sum_{i<j,l}u_{ijl}^2 + \left(\frac{2d^{1/2}}{3n}\right)cd^{1/2}}\right\} \leq \quad 2\exp\left\{-\frac{c^2 n}{4+(4/3)c}\right\}.$$

According to Proposition S1, we have $|T|\times|T|\times|T| \leq e^{n3\log(9/\delta)}$. Thus one can pick a constant c large enough so that union bound yields

$$\text{pr}\left(\sup_{(x,y,z)\in T^3}\left|\sum_{(i,j,l)\in\mathcal{L}_{x,y,z}} w_{ijl}u_{ijl}\right| \geq cd^{1/2}\right) \leq 2n^{-1}.$$

Next we control the triplets in $\mathcal{L}_{x,y,z}^c$. By symmetry, it suffices to control the triplets with positive coordinates, $\bar{\mathcal{L}}_1 = \{(i,j,l) : x_i>0, y_j>0, z_l>0\}$. The other cases can be controlled using the same technique, and only differ in the constant factor of the bound. The required result is provided in Lemma 4 and the proof is complete. □

LEMMA 4 (HEAVY TRIPLET BOUND). *For any given $c>0$, there exists a constant C depending only on c such that*

$$\sup_{x,y,z\in T^3}\left|\sum_{(i,j,l)\in\bar{\mathcal{L}}_1} x_i y_j z_l w_{ijl}\right| \leq Cd^{1/2}\log n,$$

with probability at least $1-2n^{-c}$.

Proof of Lemma 4. We start with some notation:

Let $I_1 = \{i : \delta/n^{1/2} \leq x_i \leq 2\delta/n^{1/2}\}$, $I_s = \{i : 2^{s-1}\delta/n^{1/2} < x_i \leq 2^s\delta/n^{1/2}\}$ for $s = 2, ..., \lceil \log_2(n^{1/2}/\delta)\rceil$. Define J_t, L_u similarly.

Let $e(I,J,L)$ be the number of distinct edges between I and J on layers indexed in L.

Let $\mu(I,J,L) = E\{e(I,J,L)\}$, $\bar{\mu}(I,J,L) = |I||J||L|d/n$.

Let $\lambda_{stu} = e(I_s,J_t,L_u)/\bar{\mu}_{stu}$.

Let $\alpha_s = |I_s|2^{2s}/n$, $\beta_t = |J_t|2^{2t}/n$, $\gamma_u = |L_u|2^{2u}/n$, $\sigma_{stu} = \lambda_{stu}d^{1/2}n^{1/2}2^{-(s+t+u)}$.

Then

$$\begin{aligned}\sum_{(i,j,l)\in\bar{\mathcal{L}}_1} x_i y_j z_l a_{ijl} \leq& 2\sum_{(s,t,u):2^{s+t+u}\geq d^{1/2}n^{1/2}} e(I_s,J_t,L_u)\frac{2^s\delta}{n^{1/2}}\frac{2^t\delta}{n^{1/2}}\frac{2^u\delta}{n^{1/2}} \\ =& 2\delta^3 d^{1/2}\sum_{(s,t,u):2^{s+t+u}\geq d^{1/2}n^{1/2}} \sigma_{stu}\alpha_s\beta_t\gamma_u.\end{aligned}$$

Now we split the triplets (s,t,u) under consideration into eighteen categories. Let $\mathcal{C} = \{(s,t,u) : 2^{s+t+u} \geq d^{1/2}n^{1/2}, |I_s|, |L_u| \leq |J_t|\}$, and define the following:

$\mathcal{C}_1 = \{(s,t,u)\in\mathcal{C} : \sigma_{stu} \leq 1\}$.

$\mathcal{C}_2 = \{(s,t,u) \in \mathcal{C}\backslash\mathcal{C}_1 : \lambda_{stu} \le ec_2\}$.
$\mathcal{C}_3 = \{(s,t,u) \in \mathcal{C}\backslash(\mathcal{C}_1 \cup \mathcal{C}_2) : 2^{s+u} \ge d^{1/2}n^{1/2}2^t\}$.
$\mathcal{C}_4 = \{(s,t,u) \in \mathcal{C}\backslash(\mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3) : \log\lambda_{stu} > (1/4)\{2t\log 2 + \log(1/\beta_t)\}\}$.
$\mathcal{C}_5 = \{(s,t,u) \in \mathcal{C}\backslash(\mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3 \cup \mathcal{C}_4) : 2t\log 2 \ge \log(1/\beta_t)\}$.
$\mathcal{C}_6 = \{(s,t,u) \in \mathcal{C}\backslash(\mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3 \cup \mathcal{C}_4 \cup \mathcal{C}_5)\}$.

The other twelve categories can be defined using a similar partition of

$$\mathcal{C}' = \{(s,t,u) : 2^{s+t+u} \ge d^{1/2}n^{1/2}, |J_t|, |L_u| \le |I_s|\},$$
$$\mathcal{C}'' = \{(s,t,u) : 2^{s+t+u} \ge d^{1/2}n^{1/2}, |I_s|, |J_t| \le |L_u|\},$$

by rotating the roles of (I,s,α), (J,t,β), (L,u,γ). These sets are not disjoint. Our argument is still valid as the overlap only makes the sum larger.

We now analyze separately each of the first six cases. Towards that end, we will repeatedly make use of the following simple facts:

$$\sum_s \alpha_s \le \sum_i |2x_i/\delta|^2 \le 4\delta^{-2}, \quad \sum_t \beta_t \le 4\delta^{-2}, \quad \sum_u \gamma_u \le 4\delta^{-2}.$$

Triplets in $\mathcal{C}_1$: In this case we get the bound

$$\sum_{(s,t,u)} \alpha_s\beta_t\gamma_u\sigma_{stu}1\{(s,t,u) \in \mathcal{C}_1\} \le \sum_{(s,t,u)} \alpha_s\beta_t\gamma_u 1\{(s,t) \in \mathcal{C}_1\}$$
$$\le \sum_{(s,t,u)} \alpha_s\beta_t\gamma_u = 4\|x\|_2^2\delta^{-2}4\|y\|_2^2\delta^{-2}4\|z\|_2^2\delta^{-2} \le 64\delta^{-6}.$$

Triplets in $\mathcal{C}_2$: In this case

$$\sigma_{stu} = \lambda_{stu}d^{1/2}n^{1/2}2^{-(s+t+u)} \le \lambda_{stu} \le ec_2.$$

Therefore,

$$\sum_{(s,t)} \alpha_s\beta_t\sigma_{stu}1\{(s,t) \in \mathcal{C}_2\} \le \sum_{(s,t,u)} \alpha_s\beta_t\gamma_u ec_2 1\{(s,t) \in \mathcal{C}_2\}$$
$$\le ec_2 \sum_{(s,t,u)} \alpha_s\beta_t\gamma_u \le ec_2 64\delta^{-6}.$$

Triplets in $\mathcal{C}_3$: In this case $2^{s-t+u} \ge d^{1/2}n^{1/2}$. Also by the bounded degree lemma (Lemma 6), we have $e(I_s, J_t, L_u) \le c_1|I_s||L_u|d$, and hence $\lambda_{stu} \le c_1 n/|J_t|$. Thus,

$$\sum_{(s,t,u)} \alpha_s\beta_t\gamma_u\sigma_{stu}1\{(s,t,u) \in \mathcal{C}_3\} = \sum_{s,u} \alpha_s\gamma_u \sum_t \beta_t\sigma_{stu}1\{(s,t,u) \in \mathcal{C}_3\}$$
$$= \sum_{s,u} \alpha_s\gamma_u \sum_t |J_t|\frac{2^{2t}}{n}\lambda_{stu}d^{1/2}n^{1/2}2^{-(s+t+u)}1\{(s,t,u) \in \mathcal{C}_3\}$$
$$\le \sum_{s,u} \alpha_s\gamma_u \sum_t |J_t|\frac{2^{2t}}{n}\frac{c_1 n}{|J_t|}d^{1/2}n^{1/2}2^{-(s+t+u)}1\{(s,t,u) \in \mathcal{C}_3\}$$
$$\le c_1 \sum_{s,u} \alpha_s\gamma_u \sum_t \frac{d^{1/2}n^{1/2}}{2^{s-t+u}}1\{(s,t,u) \in \mathcal{C}_3\} \le 2c_1 \sum_{s,u} \alpha_s\gamma_u \le 32c_1\delta^{-4},$$

where the first inequality uses $\lambda_{stu} \le c_1 n/|J_t|$, the second inequality follows from the definition of $\mathcal{C}_3$, and the third inequality follows from the fact that the nonzero summands over t are all bounded by 1 and form a geometric sequence.

In order to bound the triplets in $\mathcal{C}_4$, $\mathcal{C}_5$, and $\mathcal{C}_6$, we will rely on the second case described in the bounded discrepancy lemma (Lemma 5), which we can rewrite in an equivalent form as

$$\lambda_{stu}|I_s||J_t||L_u|\frac{d}{n}\log\lambda_{stu} \le c_3|J_t|\log\frac{2^{2t}}{|J_t|}.$$

Rearranging, this is in turn equivalent to

$$\sigma_{stu}\alpha_s\gamma_u\log\lambda_{stu} \le c_3\frac{2^{s-t+u}}{d^{1/2}n^{1/2}}\left(2t\log 2+\log\beta_t^{-1}\right). \tag{S.2}$$

Triplets in $\mathcal{C}_4$: The inequality $\log\lambda_{stu} > (1/4)\{2t\log 2+\log(1/\beta_t)\}$ and (S.2) imply that $\alpha_s\gamma_u\sigma_{stu} \le 4c_3 2^{s-t+u}/(d^{1/2}n^{1/2})$. Then

$$\begin{aligned}\sum_{(s,t,u)}\alpha_s\beta_t\gamma_u\sigma_{stu}1\{(s,t,u)\in\mathcal{C}_4\} &= \sum_t\beta_t\sum_{s,u}\alpha_s\gamma_u\sigma_{stu}1\{(s,t)\in\mathcal{C}_4\}\\ \le 4c_3\sum_t\beta_t\sum_{s,u}2^{s-t+u}/(d^{1/2}n^{1/2})1\{(s,t,u)\in\mathcal{C}_4\} &\le 8c_3\log n\sum_t\beta_t \le 32c_3\delta^{-2}\log n,\end{aligned}$$

where the first inequality uses the property of $\mathcal{C}_4$ which implies that $\alpha_s\gamma_u\sigma_{stu} \le 4c_3 2^{s-t+u}/(d^{1/2}n^{1/2})$, and the second inequality follows from the fact that for each u the nonzero summand over s is a geometric sequence bounded by 1 because $(s,t,u)\notin\mathcal{C}_3$, and the number of distinct u values is bounded by $\log n$ when n is large enough.

Triplets in $\mathcal{C}_5$: In this case we have $2t\log 2 \ge \log\beta_t^{-1}$. Also because $(s,t,u)\notin\mathcal{C}_4$, we have $\log\lambda_{stu} \le 4^{-1}(2t\log 2+\log\beta_t^{-1}) \le t\log 2$. Thus $\lambda_{stu}\le 2^t$. On the other hand, because $(s,t,u)\notin\mathcal{C}_1$, $1\le\sigma_{stu} = \lambda_{stu}d^{1/2}n^{1/2}2^{-(s+t+u)} \le d^{1/2}n^{1/2}2^{-s-u}$. Thus $2^{s+u}\le d^{1/2}n^{1/2}$.

Because $(s,t,u)\notin\mathcal{C}_2$, we have $\log\lambda_{stu}\ge 1$. Combining with $2t\log 2\ge\log\beta_t^{-1}$, (S.2) implies that

$$\sigma_{stu}\alpha_s\gamma_u \le c_3\frac{2^{s-t+u}}{d^{1/2}n^{1/2}}4t\log 2,$$

$$\begin{aligned}\sum_{(s,t,u)}\alpha_s\beta_t\gamma_u\sigma_{stu}1\{(s,t,u)\in\mathcal{C}_5\} &= \sum_t\beta_t\sum_{s,u}\alpha_s\gamma_u\sigma_{stu}1\{(s,t,u)\in\mathcal{C}_5\}\\ &\le \sum_t\beta_t\sum_{s,u}c_3\frac{2^{s-t+u}}{d^{1/2}n^{1/2}}4t(\log 2)1\{(s,t,u)\in\mathcal{C}_5\}\\ &\le 4c_3\log 2\sum_t\beta_t t2^{-t}\sum_{s,u}\frac{2^{s+u}}{d^{1/2}n^{1/2}}1\{(s,t,u)\in\mathcal{C}_5\}\\ &\le 4c_3\log 2(\log n)\sum_t\beta_t \le 16c_3\delta^{-2}\log n\,,\end{aligned}$$

where the third inequality holds for the same reason as the double sum over (s,u) in the case of $\mathcal{C}_4$.

Triplets in $\mathcal{C}_6$: We have $2t\log 2<\log\beta_t^{-1}$. Because $(s,t,u)\notin\mathcal{C}_4$, we have $\log\lambda_{stu}\le(1/2)\log\beta_t^{-1}\le\log\beta_t^{-1}$ where the last inequality comes from the fact that $\log\lambda_{stu}\ge 1$ because $(s,t,u)\notin\mathcal{C}_2$. Thus,

$$\begin{aligned}\sum_{(s,t,u)}\alpha_s\beta_t\gamma_u\sigma_{stu}1\{(s,t,u)\in\mathcal{C}_6\} &= \sum_{s,u}\alpha_s\gamma_u\sum_t\beta_t\lambda_{stu}d^{1/2}n^{1/2}2^{-(s+t+u)}1\{(s,t,u)\in\mathcal{C}_6\}\\ &\le \sum_{s,u}\alpha_s\gamma_u\sum_t d^{1/2}n^{1/2}2^{-(s+t+u)}1\{(s,t,u)\in\mathcal{C}_6\}\\ &\le 2\sum_{s,u}\alpha_s\gamma_u \le 32\delta^{-4}.\end{aligned}$$

□

LEMMA 5. *There exist universal constants c_2, c_3 such that with probability at least $1-4n^{-1}$ for all triplets (I,J,L) at least one of the following holds:*

$$\frac{e(I,J,L)}{\bar{\mu}(I,J,L)} \le ec_2,$$
$$e(I,J,L)\log\frac{e(I,J,L)}{\bar{\mu}(I,J,L)} \le c_3\max(|I|,|J|,|L|)\log\frac{n}{\max(|I|,|J|,|L|)}.$$

Proof. Let $d_{ij\cdot}=\sum_{l=1}^n A_{ijl}$, $d_{i\cdot l}=\sum_{j=1}^n A_{ijl}$. We confine the argument to the event $\max_{i,j,l}\max(d_{i\cdot l},d_{ij\cdot})\le c_1 d$, which has probability at least $1-n^{-1}$ for large enough constant c_1, by Bernstein's inequality and union bound.

Consider the case $|J|=\max(|I|,|J|,|L|)$. The other cases are similar.

If $|J|\ge n/e$, then by the bounded degree lemma (Lemma 6) we have $e(I,J,L)\le |I||L|c_1 d$ and hence $e(I,J,L)/\bar{\mu}(I,J,L)\le |I||L|c_1 d/(|I||J||L|d/n)\le c_1 e$.

Now assume $|J|<n/e$. Let $k\ge 8$ be a number to be specified later. By a deviation bound for sums of Bernoulli random variables we have

$$\mathrm{pr}\{e(I,J,L)\ge k\bar{\mu}(I,J,L)\}\le \exp\left\{-\frac{1}{2}(k\log k)\bar{\mu}\right\}.$$

For a given number $c_3>0$, define $t(I,J,L)$ as the unique value of t such that $t\log t=\{c_3|J|/\bar{\mu}(I,J,L)\}\log(n/|J|)$. Let $k(I,J,L)=\max\{8,t(I,J,L)\}$.

Then,

$$\begin{aligned}\mathrm{pr}\left\{e(I,J,L)\ge k(I,J,L)\bar{\mu}(I,J,L)\right\} &\le \exp\left\{-\frac{1}{2}\bar{\mu}(I,J,L)k(I,J,L)\log k(I,J,L)\right\}\\ &\le \exp\left(-\frac{1}{2}c_3|J|\log\frac{n}{|J|}\right).\end{aligned}$$

Therefore, the probability that there exists (I,J,L) such that $|I|,|L|\le|J|\le n/e, e(I,J,L)\ge k(I,J,L)\bar{\mu}(I,J,L)$ is less than or equal to

$$\begin{aligned}&\sum_{(I,J,L):|I|,|L|\le|J|\le n/e}\exp\left(-\frac{1}{2}c_3|J|\log\frac{n}{|J|}\right)\\ \le&\sum_{(h,g,m):1\le h,m\le g\le n/e}\ \sum_{(I,J,L):|I|=h,|J|=g,|L|=m}\exp\left(-\frac{1}{2}c_3 g\log\frac{n}{g}\right)\\ =&\sum_{(h,g,m):1\le h,m\le g\le n/e}\binom{n}{h}\binom{n}{g}\binom{n}{m}\exp\left(-\frac{1}{2}c_3 g\log\frac{n}{g}\right)\\ \le&\sum_{(h,g,m):1\le h,m\le g\le n/e}\left(\frac{ne}{h}\right)^h\left(\frac{ne}{g}\right)^g\left(\frac{ne}{m}\right)^m\exp\left(-\frac{1}{2}c_3 g\log\frac{n}{g}\right)\\ =&\sum_{(h,g,m):1\le h,m\le g\le n/e}\exp\left(-\frac{1}{2}c_3 g\log\frac{n}{g}+h\log\frac{n}{h}+h+g\log\frac{n}{g}+g+m\log\frac{n}{m}+m\right)\\ \le&\sum_{(h,g,m):1\le h,m\le g\le n/e}\exp\left(-\frac{1}{2}c_3 g\log\frac{n}{g}+3g\log\frac{n}{g}+3g\right)\\ \le&\sum_{(h,g,m):1\le h,m\le g\le n/e}\exp\left(-\frac{1}{2}(c_3-12)g\log\frac{n}{g}\right)\le\sum_{(h,g,m):1\le h,m\le g\le n/e}n^{-\frac{1}{2}(c_3-12)}\le n^{-\frac{1}{2}(c_3-18)},\end{aligned}$$

where the inequalities repeatedly use the assumption that $h,m\le g\le n/e$ and the fact that $t\log(n/t)$ is increasing on $[1,n/e]$.

As a result, with probability at least $1 - n^{-(1/2)(c_3-18)}$, we have $e(I,J,L) \le k(I,J,L)\bar{\mu}(I,J,L)$ for all $|I|, |L| \le |J| \le n/e$. As a final step, we further divide the set of triplets (I,J,L) satisfying $|I|, |L| \le |J| \le n/e$ into two groups by the value of $k(I,J,L)$. For the triplets for which $k(I,J,L) = 8$, we get

$$e(I,J,L) \le k(I,J,L)\bar{\mu}(I,J,L) = 8\bar{\mu}(I,J,L).$$

For all the other triplets $k(I,J,L) = t(I,J,L) > 8$, and we have $e(I,J,L)/\bar{\mu}(I,J,L) \le t(I,J,L)$. Thus

$$\frac{e(I,J,L)}{\bar{\mu}(I,J,L)} \log \frac{e(I,J,L)}{\bar{\mu}(I,J,L)} \le t(I,J,L) \log t(I,J,L) = \frac{c_3|J|}{\bar{\mu}(I,J,L)} \log \frac{n}{|J|},$$

which implies that

$$e(I,J,L) \log \frac{e(I,J,L)}{\bar{\mu}(I,J,L)} \le c_3|J| \log \frac{n}{|J|}.$$

The desired claim follows by letting $c_2 = \max(c_1, 8)$, and $c_3 = 20$. □

PROPOSITION S1. $|T| < e^{n \log(9/\delta)}$.

Proof. For each point in T, consider the ℓ_∞ ball of side length $\delta/n^{1/3}$ centering at that point. These balls are disjoint, have volume $\delta^n n^{-n/3}$ and diameter $\delta n^{1/6}$, and are hence inside the ball of radius $1 + \delta n^{1/6}$. Therefore $|T|$ equals the number of all such ℓ_∞ balls, which is no more than, by argument of volume and Stirling's formula,

$$\begin{aligned}\frac{(1+\delta n^{1/6})^n \pi^{n/2}}{(1+n/2)!(\delta^n n^{-n/3})} \le & \frac{(1+\delta)^n n^{6/n} \pi^{n/2}}{(2\pi)^{1/2}(1+n/2)^{3/2+n/2} e^{-1-n/2} \delta^n n^{-n/3}} \\ < & (1+\delta^{-1})^n (2\pi e)^{n/2} < e^{n \log(9/\delta)}.\end{aligned}$$

□

LEMMA 6 (BOUNDED DEGREE). *Under the assumption of Theorem* 2*, there exists a universal constant* $c_1 > 0$ *such that with probability at least* $1 - n^{-1}$

$$\sup_{i,l} \sum_j A_{ijl} \le c_1 d\,.$$

Proof. The proof follows from a standard application of Bernstein's inequality and union bound, which is identical to the corresponding lemma in Lei & Rinaldo (2015). □

[*Received* 2 *January* 2017. *Editorial decision on* 1 *April* 2017]