Tail Bounds for Matrix Quadratic Forms and Bias Adjusted Spectral Clustering in Multi-layer Stochastic Block Models

Jing Lei

1Carnegie Mellon University

March 19, 2020

Abstract

We develop tail probability bounds for matrix linear combinations with matrix-valued coefficients and matrix-valued quadratic forms. These results extend well-known scalar case results such as the Hanson–Wright inequality, and matrix concentration inequalities such as the matrix Bernstein inequality. A key intermediate result is a deviation bound for matrix-valued $U$-statistics of order two and their independent sums. As an application of these probability tools in statistical inference, we establish the consistency of a novel bias-adjusted spectral clustering method in multi-layer stochastic block models with general signal structures.

1 Introduction

With the fast development of modern data acquisition technology, it is more common to have data sets with matrices as basic measurement units. One example is network data, where the data consists of a matrix recording the interaction among a set of individuals [17, 26, 13]. Another example is brain imaging, where the data matrix measures the spatiotemporal signal of brain activities [22]. In many applications, it is possible to have multiple realizations of such data matrices, such as networks measured at different time points [38, 16], and brain imaging measured with repetition under potentially different tasks and/or for different subjects [35].

An important first step in analyzing such matrix-valued data is to understand the behavior of the measurement errors, which now come in matrix form. In the simplest case, such noise will be a matrix with independent entries, and we would like to find good upper bounds for its spectral norm. Many nice and interesting results have been obtained under various settings, such as random matrix theory [3], eigenvalue perturbation and concentration theory [11, 27, 20, 18, 8], and matrix deviation inequalities [4, 36]. The matrix Bernstein
inequality and related results [32] are applicable to linear combinations of such simple noise matrices with scalar coefficients.

Some recent statistical inference problems, such as multi-layer network analysis, require more general results that are not offered by existing literature. In this paper, we extend matrix concentration inequalities in two directions. First, we provide upper bounds for linear combinations of simple 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 [14, 31] in several directions. A key intermediate step between the linear and quadratic cases is a deviation bound for matrix-valued $U$-statistics of order two.

As an application of these more general matrix concentration results, we show how they can be used to develop and understand new spectral clustering methods in multi-layer stochastic block models. The stochastic block model [15, 12] is a popular prototypical probabilistic model for network data with community structures, and has been the focus of much research activity in the past decade [7, 1]. Recently, multi-layer stochastic block models have emerged in various fields, including transportation, neural biology, internet [33], social science [28], and bioinformatics [19]. While there have been a few works on the consistency of various methods for multi-layer stochastic block models [29, 30, 6], it is unclear how the availability of multiple layers affects the hardness of estimation in general scenarios. For a network with $n$ nodes, the single layer result says it is possible to achieve consistent clustering if and only if the average node degree diverges as $n$ increases, assuming other aspects of the model are simple and fixed. For a multi-layer stochastic block model with $L$ layers, most existing theory can prove consistency if the total degree summed over all layers diverges, but under the additional assumption that the community-wise connectivity matrix of each layer is positive definite. In an effort to relax this assumption, [19] established consistency when the summed degree is at least $L^{1/2}$ up to a small poly-logarithmic factor, but with a computationally challenging least squares estimator. We show that a novel bias-adjusted spectral method can achieve consistent clustering under a similar degree requirement.

**Notation.** For a matrix $M$, $M_j$ denotes its $j$th row in the form of a column vector. $\|M\|_{q,\infty} = \max_j \|M_j\|_q$ for $q \in [1, \infty)$, and $\|M\|_\infty$ is the maximum entry-wise absolute value. When $M$ is symmetric with eigen-decomposition $\sum_j \lambda_j u_j u_j^T$, let $|M| = \sum_j |\lambda_j| u_j u_j^T$. Let $e_i$ be the $i$-th coordinate unit vector, the length of $e_i$ will depend on the context. For two symmetric matrices $A$ and $B$, $A \preceq B$ means that $B - A$ is positive semidefinite. In the statement of the theorems and their proofs, we use $C$ to denote a universal constant whose value may vary from line to line but does not depend on any of the model parameters.
We consider a sequence of independent matrices $X_1, \ldots, X_L$ with independent zero-mean entries. The goal is to provide upper bounds for operator norms of linear combinations of the form $\sum_{\ell} X_{\ell} H_{\ell}$ and quadratic forms $\sum_{\ell} X_{\ell} G_{\ell} X_{\ell}^T$, where $\{H_{\ell} : 1 \leq \ell \leq L\}, \{G_{\ell} : 1 \leq \ell \leq L\}$ are sequences of non-random matrices. Our results also cover the symmetric case where each $X_{\ell}$ has independent diagonal and upper diagonal entries.

Concentration inequalities usually require tail conditions on the entries of $X_{\ell}$. 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 $E|Y|^k \leq \frac{v^2}{2} k! R^{k-2}$ for all integers $k \geq 2$.

The Bernstein tail condition leads to concentration inequalities for sums of independent random variables [34, Chapter 2]. Since we are interested not only in linear combinations of $X_{\ell}$’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 $X_1, \ldots, X_L$. Specifically we consider the following three assumptions.

**Assumption 1.** The entries of $X_{\ell}$ are $(v_1, R_1)$-Bernstein.

**Assumption 2.** The squared entries of $X_1, \ldots, X_L$ are $(v_2, R_2)$-Bernstein.

**Assumption 2’.** The product $X_{\ell,ij} \tilde{X}_{\ell,ij}$ is $(v_2', R_2')$-Bernstein for each $(\ell, i, j)$, where $\tilde{X}_{\ell}$ is an independent copy of $X_{\ell}$.

There are two typical scenarios in which such a squared Bernstein condition in Assumption 2 holds. The first is the sub-Gaussian case: If a random variable $Y$ satisfies the sub-Gaussian condition $Ee^{Y^2/\sigma^2} \leq 2$ for some $\sigma > 0$, then a simple derivation leads to $EY^{2k} \leq 2\sigma^4 (\sigma^2)^{k-2} k!$, and hence $Y^2$ is $(4\sigma^4, \sigma^2)$-Bernstein. The second scenario is centered Bernoulli: $P(Y = 1 - p) = 1 - P(Y = -p) = p$ for some $p \in [0, 1/2]$. We have $EY^{2k} = p(1-p)^{2k} + (1-p)p^{2k} \leq p$, so that $Y^2$ is $(2p, 1)$-Bernstein. Our proof will also use the fact that if $Y^2$ is $(v_2, R_2)$-Bernstein, then the centered version $Y^2 - E(Y^2)$ is also $(v_2, R_2)$-Bernstein [37, Lemma 2].

We require Assumption 2’ in order to use a decoupling technique in establishing concentration of quadratic forms. One can show that if Assumption 2 holds then Assumption 2’ holds with $(v_2, R_2') = (v_2, R_2)$. However, when $X_{\ell,ij}$’s are centered Bernoulli random variables with parameters bounded by $p \leq 1/2$, then Assumption 2’ holds with $v_2' = 2p^2$ and $R_2' = 1$, while Assumption 2 holds with $v_2 = 2p$ and $R_2 = 1$, so that $v_2'$ can potentially be much smaller than $v_2$. For generality we will explicitly keep track of the Bernstein parameters in our results.
2.1 Linear combinations with matrix coefficients

We first present a result on linear combinations of matrices with independent and Bernstein tail entries.

**Theorem 2.1.** Let \((X_\ell : 1 \leq \ell \leq L)\) be a sequence of independent \(n \times r\) matrices with zero mean independent entries satisfying Assumption 1, and \(H_\ell\) be any sequence of \(r \times m\) non-random matrices. Then for all \(t > 0\)

\[
P\left(\left\| \sum_{\ell=1}^{L} X_\ell H_\ell \right\| \geq t \right) \leq 2(m + n) \exp\left(-\frac{t^2}{8} + R_1 \max_\ell \|H_\ell\|_{2,\infty} t\right).
\]  

(1)

A similar result holds, with \(t^2/8\) replaced by \(t^2/2\) and \(2(m + n)\) replaced by \(4(m + n)\) in (1), for symmetric \(X_\ell\)'s of size \(n \times n\) with independent \((v_1, R_1)\)-Bernstein diagonal and upper-diagonal entries and \(H_\ell\) of size \(n \times m\).

The proof of Theorem 2.1, given in Appendix A, combines the matrix Bernstein inequality [32] for symmetric matrices and a rank-one symmetric dilation trick (Lemma 2.2) to take care of the asymmetry in \(X_\ell H_\ell\).

**Definition 2 (Symmetric dilation).** For an \(n \times m\) matrix \(A\), the symmetric dilation of \(A\), denoted by \(D(A)\), is the \((n + m) \times (n + m)\) symmetric matrix

\[
D(A) = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}.
\]

The symmetric dilation is a convenient tool to reduce singular values and singular vectors of asymmetric matrices to eigenvalues and eigenvectors of symmetric matrices. See Exercise II.1.15 of [5] and Section 2.6 of [32] for example. Here we will use a special case of rank-one dilations, whose proof is elementary and omitted.

**Lemma 2.2 (Rank-one dilation).** For two column vectors \(e\) and \(a\), \(D(ea^T)\) has eigen-decomposition

\[
D(ea^T) = \|e\|\|a\|\left(\frac{1}{\sqrt{2}} \begin{bmatrix} e \|e\| \\ a \|a\| \end{bmatrix} + \frac{1}{\sqrt{2}} \begin{bmatrix} e \|e\| \\ -a \|a\| \end{bmatrix} \right)^T - \frac{1}{\sqrt{2}} \begin{bmatrix} e \|e\| \\ a \|a\| \end{bmatrix} \left(\frac{1}{\sqrt{2}} \begin{bmatrix} e \|e\| \\ -a \|a\| \end{bmatrix} \right)^T
\]

and for each integer \(k \geq 2\)

\[
|D(ea^T)^k| \leq \|e\|^{k-2}\|a\|^{k-2} \begin{bmatrix} \|a\|^2ee^T \\ 0 \|e\|^2aa^T \end{bmatrix}.
\]
Remark 1. If $n = m = r = 1$, then Theorem 2.1 recovers the well known Bernstein’s inequality as a special case with a different pre-factor.

If $n \geq \min\{m, Lr\}$, then $n\|\sum_\ell H_\ell^T H_\ell\| \geq \sum_\ell\|H_\ell\|^2_F$ and the probability upper bound in Theorem 2.1 reduces to

$$
P \left[ \left\| \sum_{\ell=1}^L X_\ell H_\ell \right\| \geq t \right] \leq 2(m + n) \exp \left( -\frac{t^2/2}{v_1 n\|\sum_\ell H_\ell^T H_\ell\| + R_1 \max_\ell\|H_\ell\|_{2,\infty} t} \right).$$

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

$$
P \left[ \left\| \sum_{\ell=1}^L X_\ell H_\ell \right\| \geq t \right] \leq 2(m + n) \exp \left( -\frac{t^2/2}{v_1 \sum_\ell\|H_\ell\|^2_F + R_1 \max_\ell\|H_\ell\|_{2,\infty} t} \right).$$

Remark 2. When $L = 1$, the setting is similar to that considered in [36]. In the constant variance case (e.g., sub-Gaussian), $v_1^{1/2} \approx R_1 \approx 1$, Theorem 2.1 implies a high probability upper bound of $C \sqrt{\log(m + n)}(\sqrt{n}\|H\| + \|H\|_F)$, which agrees with Theorem 1.1 of [36]. The extra $\sqrt{\log(n + m)}$ factor in our bound is because our result is a tail probability bound while [36] provides upper bounds on the expected value. However, in the sparse Bernoulli setting, where $v_1 \ll R_1 = 1$, the upper bound in Theorem 2.1 is better because it correctly captures the $\sqrt{v_1}$ factor multiplied by $\sqrt{n}\|H\| + \|H\|_F$, whereas the result in [36] leads to $v_1^{1/4}(\sqrt{n}\|H\| + \|H\|_F)$.

2.2 Matrix U-statistics and quadratic forms

Let

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

where the summation is taken over all pairs $(i, j), (i', j') \in \{1, \ldots, n\}^2$. In this subsection 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 will be stated separately in Section 2.3.

Because $X_\ell$ has centered and independent diagonal and upper diagonal entries, a term in (4) has non-zero expected value only if $(i, j) = (i', j')$ or $(i, j) = (j', i')$ so that $X_{\ell,ij} X_{\ell,i'j'} = X_{\ell,ij}^2$. This motivates the following decomposition of $S$ into a quadratic component with nonzero entrywise mean value

$$S_2 = \sum_\ell \sum_{1 \leq i < j \leq n} X_{\ell,ij}^2 \left( e_i e_i^T G_{\ell,jj} + e_j e_j^T G_{\ell,ii} + e_i e_j^T G_{\ell,ji} + e_j e_i^T G_{\ell,ij} \right)$$

5
+ \sum_{\ell} \sum_{1 \leq i \leq n} X_{\ell,ii}^2 e_i e_i^T G_{\ell,ii}, \quad (5)

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

\[ S_1 = S - S_2. \quad (6) \]

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 indeed 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 \( \mathbb{E} S_2 \).

Define quantities

\[ \sigma_1^2 = \sum_{\ell} \|G_\ell\|^2, \]
\[ \sigma_2 = \max_{\ell} \max \{ \|G_\ell\|_{2,\infty}, \|G_\ell^T\|_{2,\infty} \} \]
\[ (\sigma_2')^2 = \sum_{\ell,j} G_{\ell,jj}^2, \]
\[ \sigma_3 = \max_{\ell} \|G_\ell\|_{\infty}. \]

**Theorem 2.3.** If \((X_\ell : 1 \leq \ell \leq L)\) are independent \( n \times n \) symmetric matrices with independent diagonal and upper diagonal entries satisfying Assumption 1 and Assumption 2'. Let \((G_\ell : 1 \leq \ell \leq L)\) be \( n \times n \) matrices. Define \( S = \sum_\ell X_\ell G_\ell X_\ell^T \) and \( S_1, S_2 \) as in (5) and (6). Then there exists a universal constant \( C \) such that with probability at least \( 1 - O((n + L)^{-1}) \),

\[ \|S_1\| \leq C \left[ v_1 n \log(L + n)\sigma_1 + \sqrt{v_1} R_1 \sqrt{nL} \log^{3/2}(L + n)\sigma_2 
+ \sqrt{v_2} \log(L + n)(\sqrt{L}\sigma_2 + \sigma_2') + (R_1^2 + R_2^2) \log^2(L + n)\sigma_3 \right], \quad (7) \]

If in addition Assumption 2 holds, then with probability at least \( 1 - O((L + n)^{-1}) \),

\[ \|S_2 - \mathbb{E} S_2\| \leq C \left[ \sqrt{v_2} \log(L + n)(\sqrt{L}\sigma_2 + \sigma_2') + R_2 \log(L + n)\sigma_3 \right]. \quad (8) \]

and consequently

\[ \|S - \mathbb{E} S\| \leq C \left[ v_1 n \log(L + n)\sigma_1 + \sqrt{v_1} R_1 \sqrt{nL} \log^{3/2}(L + n)\sigma_2 
+ \sqrt{v_2'} + v_2 \log(L + n)(\sqrt{L}\sigma_2 + \sigma_2') + (R_1^2 + R_2 + R_2') \log^2(L + n)\sigma_3 \right]. \quad (9) \]
The proof of Theorem 2.3 is given in Appendix A, where the main effort is to control \(\|S_1\|\). Unlike the linear combination case, the complex 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 [10]. This reduces the problem of bounding \(\|S_1\|\) to that of bounding \(\|\sum_\ell X_\ell G_\ell X_\ell^T\|\), where \((X_\ell : 1 \leq \ell \leq L)\) is an iid copy of \((X_\ell : 1 \leq \ell \leq L)\).

The upper bounds in Theorem 2.3 look complicated. This is because we do not make any assumption about the Bernstein parameters or the matrices \(G_\ell\). The bound can be much simplified or even improved in certain important special cases. In the sub-Gaussian case, where \(R_1 \asymp v_1^{1/2} / n \asymp v_2^{1/4}\), the first term \(v_1 n \log(L + n)\sigma_1\) dominates. This reflects the \(\sqrt{L}\) effect for sums of independent random variables. In the case \(G_\ell = G_0\) for all \(\ell\) and \(X_\ell\) are iid, we have \(\|ES\| \approx L \|X_1 G_0 X_1^T\| \asymp v_1 nL\|G_0\|\), but for the fluctuation from \(S_1\) we have \(\|S_1\| \lesssim v_1 nL\|G_0\|\) ignoring logarithmic factors. In other words, the signal is contained in \(ES_2\) which may grow linearly as \(L\), and the fluctuation from \(S_1\) only grows at a rate of \(\sqrt{L}\).

In the Bernoulli case, the situation becomes more complicated when the variance \(v_1\) is vanishing so that \(v_1 \asymp v_2 \asymp \sqrt{v_2} \ll R_1 \asymp R_2\). In the simple case of \(G_\ell = I_n\), we have \(\sigma_1 = \sqrt{L}, \sigma_2 = \sigma_3 = 1\). Thus the second term \(\sqrt{v_1} \sqrt{nL}\sigma_2\) in (7) may dominate the first term when \(nv_1 \ll 1\). In this case we also have \(\sigma_2' = \sqrt{nL}\). Therefore it is also possible that the term \(\sqrt{v_2} \sigma_2\) in (8) 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 Section 2.4.

### 2.3 The asymmetric case

Let \((X_\ell : 1 \leq \ell \leq L)\) be independent \(n \times r\) matrices with independent zero mean entries. Let \(G_\ell\) be \(r \times r\) matrices. The decomposition of the quadratic form now becomes simpler:

\[
S = \sum_\ell X_\ell G_\ell X_\ell^T = S_1 + S_2,
\]

where

\[
S_1 = \sum_{(i,j) \neq (i',j')} X_{\ell,i,j} X_{\ell,i',j'} e_i e_i^T G_{\ell,j,j'}
\]

is the mean-zero off-diagonal part and

\[
S_2 = \sum_\ell \sum_{1 \leq i \leq n, 1 \leq j \leq r} X_{\ell,i,j}^2 e_i e_i^T G_{\ell,j,j},
\]

is the diagonal part with possibly non-zero expected values on the diagonal entries.
Define
\[
\sigma_1' = \left( \sum_{\ell} \|G_{\ell}\|_F^2 \right)^{1/2}.
\]

**Theorem 2.4.** If \((X_{\ell} : 1 \leq \ell \leq L)\) are \(n \times r\) independent matrices with independent entries satisfying Assumption 1 and Assumption 2', then with probability at least \(1 - O((L + n)^{-1})\),
\[
\|S_1\| \leq C \left[ v_1 n \log(L + n)\sigma_1 + v_1 \sqrt{n} \log(L + n)\sigma_1' + \sqrt{v_1} R_1 \sqrt{nL} \log^{3/2}(L + n)\sigma_2 \\
+ \sqrt{v_2} \log(L + n)\sigma_2' + (R_1^2 + R_2') \log^2(L + n)\sigma_3 \right],
\]
for some constant \(C\). If in addition Assumption 2 holds, then with probability at least \(1 - O((L + n)^{-1})\),
\[
\|S_2 - ES_2\| \leq C \left[ \sqrt{v_2} \log(L + n)\sigma_2' + R_2 \log(L + n)\sigma_3 \right].
\]

The proof follows largely the same scheme as in the symmetric case, with two notable differences. First, in the asymmetric case \(S_2\) only has diagonal entries. So the bounds for \(S_2 - ES_2\) and \(\tilde{S}_2\) only involve \(\sigma_2'\) and \(\sigma_3\). Second, there is an additional term involving \(\sigma_1'\) in the bound of \(S_1\), which comes from the \(n\|\sum_{\ell} H_{\ell}^T H_{\ell}\|_F \vee \sum_{\ell} \|H_{\ell}\|_F^2\) term in Theorem 2.1, because in the asymmetric case it is unclear whether the maximum is achieved by the operator norm part or the Frobenius norm part.

**Remark 3.** When \(n \geq r\), we can drop the \(\sigma_1'\) term and the high probability upper bound on \(S_1\) can be reduced to
\[
\|S_1\| \leq C \left[ v_1 n \log(L + n)\sigma_1 + \sqrt{v_1} R_1 \sqrt{nL} \log^{3/2}(L + n)\sigma_2 \\
+ \sqrt{v_2} \log(L + n)\sigma_2' + (R_1^2 + R_2') \log^2(L + n)\sigma_3 \right].
\]

**Remark 4.** In the special case of \(L = 1\), \(n = 1\), and \(K\)-sub-Gaussian entries, the proof of Theorem 2.4 can be modified to show that
\[
\|S_1\| = O_P(K^2\|G\|_F),
\]
which agrees with the Hanson–Wright inequality [14, 31].

### 2.4 Sparse Bernoulli matrices

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

Ignoring logarithmic factors, the first part of Theorem 2.3 becomes

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

where the second term \( L^{1/2} n^{1/2} \rho^{1/2} \) 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 \( n \rho \gg L^{-1/2} \).

Investigating the proof of Theorem 2.3, the term \( R_1(v_1 n L)^{1/2} \sigma_2 \) in (7) comes from the bound of \( \sum_\ell \|H_\ell^T H_\ell\| \) by \( \sum_\ell \|H_\ell\|^2 \), which is suboptimal in this special case and can be improved using a more refined argument.

**Theorem 2.5.** Assume \( G_\ell = I_n \) for all \( 1 \leq \ell \leq L \) and \( (X_\ell : 1 \leq \ell \leq L) \) are symmetric with centered Bernoulli entries whose parameters are bounded by \( \rho \). If \( L^{1/2} n \rho \geq C_1 \log^{1/2}(L + n) \) and \( n \rho \leq C_2 \) for some constants \( C_1, C_2 \), then with probability at least \( 1 - O((n + L)^{-1}) \),

\[
\|S_1\| \leq C L^{1/2} \rho n \log^{1/2}(L + n)
\]

for some constant \( C \).

The proof of Theorem 2.5 is given in Appendix B. At a high level, the decoupling technique reduces the problem to controlling the operator norm of \( \tilde{S} = \sum_\ell X_\ell \tilde{X}_\ell^T \) where \( \tilde{X}_\ell \) is an iid copy of \( X_\ell \). Instead of directly applying Theorem 2.1 with \( H_\ell = \tilde{X}_\ell \), we instead shift \( \tilde{X}_\ell \) back to the original Bernoulli matrix by considering \( \tilde{S} = \sum_\ell X_\ell \tilde{A}_\ell - \sum_\ell X_\ell P_\ell \), where \( \tilde{A}_\ell \) is the original uncentered binary matrix and \( P_\ell = E \tilde{A}_\ell \). Then Theorem 2.1 is applied to \( \sum_\ell X_\ell P_\ell \) and \( \sum_\ell X_\ell \tilde{A}_\ell \) separately, where the entry-wise non-negativity of \( \tilde{A}_\ell \) allows us to use the Perron–Frobenius theorem to obtain a sharper bound for \( \|\sum_\ell \tilde{A}_\ell^2\| \).

### 3 Bias adjusted spectral clustering for multi-layer stochastic block models

We demonstrate the application of the matrix concentration inequalities in the previous section by considering spectral clustering in multi-layer stochastic block models.

A network records the interactions among a collection of individuals, such as friendship, following and linking on social media, functional connectivity among brain regions, and gene co-expression. 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
Figure 1: The connectivity matrices for each time period of the postnatal gene co-expression, with genes ordered by the estimated clusters. Tick marks denote the boundaries between the clusters.

In many applications, the interaction between individuals are recorded multiple times, resulting in multi-layer network data. Figure 1 illustrates temporal gene co-expression networks in the medial prefrontal cortex of rhesus monkeys at five different postnatal stages [23]. The medial prefrontal cortex is believed to be related to developmental brain disorders, and the subset of genes plotted here are significantly enriched for neural projection guidance and related to autism spectrum disorder. A visual inspection of the data suggests that there are roughly four groups of genes such that the genes in the same group have coherent co-expression patterns. The separation of these four groups are indicated by the axis ticks in Figure 1. But such a group partition is not obvious in each single time period. For example, in the last period labeled as “L6”, the first three groups are indistinguishable, while in the first period labeled as “L2”, the first and fourth groups are indistinguishable, as are the second and third groups.

Motivated by such multi-layer network data with a common community structure, we consider the multi-layer stochastic block model:

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

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

The inference problem is to estimate the membership vector \( \theta \) given the observed adjacency matrices \( A_1, \ldots, A_L \). We assume that the number of communities, \( K \), is known. The problem of selecting \( K \) from the data is an important problem and will not be pursued in this paper. Further discussion will be given in Section 4.

When \( L = 1 \), the community estimation problem for single layer stochastic block models is well-understood [7, 20, 1]. If \( K \) is fixed as a constant while \( n \to \infty, \rho \to 0 \) with
balanced community sizes lower and upper bounded by constant fractions of \( n \), and \( B \) is a constant matrix with distinct rows, then the community memberships can be estimated with vanishing error when \( np \to \infty \). Practical estimators include variants of spectral clustering, message passing, and likelihood-based estimators.

In the multi-layer case, consistent community recovery has been studied in some recent works. The theoretical focus is to understand how the number of layers \( L \) affects the estimation problem. [29, 6] show that consistency can be achieved if \( Ln^\rho \) diverges, but under the condition that each \( B_\ell \) is positive definite with minimum eigenvalue bounded away from zero. Such a layer-wise positivity assumption enables simple estimators based on spectral clustering of \( \sum_\ell A_\ell \), but is not plausible in examples as in Figure 1 where some layers may have zero or negative eigenvalues. To remove the positivity assumption, [19] considered a least squares estimator, and proved consistency when \( \sqrt{Ln^\rho} \) 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 can only find a local minimizer using greedy algorithms.

In the following we will motivate a spectral clustering method from the least squares perspective, and investigate its bias and the possibility of a data-driven bias reduction.

3.1 From least squares to spectral clustering

Let \( \psi : \{1, \ldots, n\} \mapsto \{1, \ldots, K\} \) be a membership vector and \( \Psi = [\Psi_1, \ldots, \Psi_k] \) be the corresponding \( n \times K \) membership matrix where each \( \Psi_k \) is an \( n \times 1 \) vector with \( \Psi_{i,k} = 1(\psi_i = k) \). Let \( I_k(\psi) = \{i : \psi_i = k\} \) and \( n_k(\psi) = |I_k(\psi)| \).

The least squares estimator of [19] seeks to minimize the within block sum of squares.

\[
\hat{\theta} = \arg\min_{\psi} \sum_{\ell=1}^L \sum_{1 \leq i < j \leq n} (A_{\ell,ij} - \hat{B}_{\ell,\psi_i,\psi_j}(\psi))^2 \tag{14}
\]

where

\[
\hat{B}_{\ell,k,l}(\psi) = \begin{cases} 
\frac{\sum_{i,j \in I_k(\psi)} A_{\ell,ij}}{n_k(\psi)(n_k(\psi)-1)} & k = l \\
\frac{\sum_{i \in I_k(\psi), j \in I_l(\psi)} A_{\ell,ij}}{n_k(\psi)n_l(\psi)} & k \neq l
\end{cases}
\]

is the sample mean estimate of \( B_\ell \) under a given membership vector \( \psi \).

If we accept the approximation \( n_k(\psi)(n_k(\psi) - 1) \approx n_k^2(\psi) \), and multiply the least squares objective function (14) by 2, using the total variance decomposition, the objective function becomes

\[
\max_{\psi} \sum_{\ell=1}^L \sum_{1 \leq k,l \leq K} \frac{(\Psi_k^T A_\ell \Psi_l)^2}{n_k(\psi)n_l(\psi)},
\]
which is equivalent to

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

where \( \tilde{\Psi} = [\tilde{\Psi}_1, \ldots, \tilde{\Psi}_K] \) with \( \tilde{\Psi}_k = \Psi_k / \sqrt{n_k(\psi)} \) is the column normalized version of \( \Psi \).

Now \( \tilde{\Psi} \) is orthonormal: \( \tilde{\Psi}^T \tilde{\Psi} = I_K \). For any orthonormal matrix \( U \in \mathbb{R}^{n \times K} \) and symmetric \( n \times n \) matrix \( A \) we have

$$\left\| U^T A U \right\|_F^2 = \text{tr}(U^T A U U^T A U) \leq \text{tr}(U^T A^2 U) .$$

The right hand side of the above inequality is maximized by the leading \( K \) eigenvectors of \( A \) (eigenvalues ordered by absolute value). For this \( U \), the inequality becomes equality. Under the multi-layer stochastic block model, the expected values \( (P_\ell = \mathbb{E} A_\ell : 1 \leq \ell \leq L) \) share roughly the same leading principal subspace as determined by the common community structure. So such a \( U \) should be close to a solution to the original problem.

Therefore, a relaxation of the approximate version of the original problem is

$$\max_{U \in \mathbb{R}^{n \times K}} \left\{ U^T \left( \sum_{\ell=1}^L A_\ell^2 \right) U \right\}, \quad (15)$$

which is a standard spectral problem. The community estimation is then obtained by applying a clustering algorithm to the rows of \( \hat{U} \), a solution to (15).

### 3.2 The necessity of bias adjustment

Let \( P_\ell = \mathbb{E} A_\ell \), so that \( P_\ell \) is the matrix obtained by zeroing out diagonal entries of \( \tilde{P}_\ell = \rho \Theta B_\ell \Theta^T \). Let \( X_\ell = A_\ell - P_\ell \). Then

$$\sum_\ell A_\ell^2 = \sum_\ell P_\ell^2 + \sum_\ell (X_\ell P_\ell + P_\ell X_\ell) + S , \quad (16)$$

where \( S = \sum_\ell X_\ell^2 \). The first term is the signal term, with each summand close to \( \tilde{P}_\ell^2 = \Theta B_\ell^2 \Theta^T \), and will add up over the layers, because each \( B_\ell^2 \) is positive semi-definite. The second is a mean 0 noise term, which can be controlled using Theorem 2.1. The third term \( S = \sum_\ell X_\ell^2 \) is a squared error term and will also add up over the layers, which will likely introduce some bias.
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:

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

These two matrices are chosen such that spectral clustering applied to either the sum of the original un-squared adjacency matrices or the sum of squared adjacency matrices without bias adjustment would be sub-optimal or inconsistent in the very sparse regime. Our simulation uses $n = 200$ nodes with 100 in each community. We set $L = 30$ and each $B_t$ is randomly and independently chosen from $B^{(1)}$ and $B^{(2)}$ with equal probability, and use five different values of $\rho$ between 0.02 and 0.06. For each value of $\rho$ we repeat the experiment 20 times and apply spectral clustering to three matrices: (1) the average adjacency matrix, (2) the sum of squared adjacency matrices, and (3) a bias-adjusted sum of squared adjacency matrices, which will be introduced in the next subsection. The plotted numbers are average proportion of mis-clustered nodes. By construction the average adjacency matrix has only one significant eigen-component and the result is very sensitive to the number of eigenvectors used for spectral clustering. When we use two eigenvectors, the performance is rather poor as reported in Figure 2. It is also easy to generate cases in which the average adjacency matrix carries no signal at all. The sum of squared adjacency
matrices does carry signal but needs the density to be high in order to overcome the bias. The bias-adjusted sum of square adjacency matrices performs the best.

### 3.3 Bias-adjusted sum of squared adjacency spectral clustering

From (16) we see that the squared error term $S$ has positive expected value and hence may cause systematic bias in the principal subspace of $\sum \mathbf{A}_\ell^2$. Using the further decomposition of $S$ as in (5) and (6) with $G_\ell = I_n$, we see that the non-zero expected value comes from $S_2$, which is a diagonal matrix with

\[(S_2)_{ii} = \sum_\ell \sum_j X_{\ell,ij}^2\]

\[= \sum_\ell \sum_j P_{\ell,ij}^2 1(A_{\ell,ij} = 0) + (1 - P_{\ell,ij})^2 1(A_{\ell,ij} = 1)\]

\[\leq Ln \max_{\ell,ij} P_{\ell,ij}^2 + \sum_\ell d_{\ell,i}\]  

where $d_{\ell,i} = \sum_j A_{\ell,ij}$. The expected value of $\sum_\ell d_{\ell,i}$ is $\sum_{\ell,j} P_{\ell,ij} \propto Ln \max_{\ell,ij} P_{\ell,ij}$. 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 $(S_2)_{ii}$.

A key observation is that $\sum_\ell d_{\ell,i}$ can be computed from the data, so we can remove it to reduce the bias. Therefore, we arrive at the following bias-adjusted spectral clustering algorithm.

Let $D_\ell$ be the diagonal matrix consisting of the degrees of $A_\ell$: $(D_\ell)_{ii} = d_{\ell,i}$. The bias-adjusted sum of squared adjacency matrix is

\[S_0 = \sum_\ell (A_\ell^2 - D_\ell) .\]  

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 (18).

### 3.4 Analysis of the bias-adjusted spectral clustering

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 jointly across $M$ layers. To simplify the discussion, we focus on the following setting.

**Assumption 3.** 1. The number of communities $K$ is fixed and community sizes are balanced: There exists constant $c$ such that each community size is in $[c^{-1}n/K, cn/K]$. 


2. 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 $[0,1]$. The minimum eigenvalue of $\sum_{\ell} B_{\ell,0}^2$ is at least $cL$ for some constant $c > 0$.

Part 1 simplifies the effect of the community sizes and the number of communities. This setting has been well-studied in the SBM literature for $L = 1$ [20]. Part 2 puts the focus on the effect of the overall edge density parameter $\rho$, and requires a linear growth of the aggregated squared connectivity matrices, which is much less restrictive than the layer-wise positivity assumption. In the asymptotic regime that $n \to \infty$ and $\rho \to 0$, it is known that consistent community estimation is possible when $n\rho \to \infty$ when $L = 1$. In the multilayer setting when $L \to \infty$ one should expect a lower requirement on overall density when we have more layers, as we aggregate information across layers.

**Theorem 3.1.** Under Assumption 3, if $L^{1/2} n \rho \geq C_1 \log^{1/2}(L + n)$ and $n \rho \leq C_2$ 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 the bias-adjusted sum of squared adjacency matrices $S_0$ in (18) correctly estimates the membership of all but a $C \left( \frac{1}{n} + \frac{\log(L + n)}{Ln^2 \rho^2} \right)$ proportion of nodes for some constant $C$ with probability at least $1 - O((L + n)^{-1})$.

An immediate consequence of Theorem 3.1 is the Hamming distance consistency of the bias-adjusted sum of squares spectral clustering, provided that $L^{1/2} n \rho / \sqrt{\log(L + n)} \to \infty$.

The condition $n \rho \lesssim 1$ is used for notational simplicity. Our analysis can be modified to cover other regimes such as $\rho \to \infty$ with more complicated bookkeeping. However, this is less interesting since it is well-known that consistent community estimation is possible in this regime even if $L = 1$. The condition $L^{1/2} n \rho \gg \log^{1/2}(L + n)$ is required in order for the error bound in Theorem 3.1 to imply consistency, and is suitable for the linear squared signal accumulation assumed in part 2 of Assumption 3. If we assume a different speed of accumulation, this requirement needs to be changed accordingly.

The proof of Theorem 3.1 is given in Appendix C. The main effort is to establish a refined operator norm bound for $S - \sum_{\ell} D_{\ell}$ where $S = \sum_{\ell} X_{\ell}^2$ is the sum of squared noise matrix in (16), and the refined operator norm bound for $S_1$ in Section 2.4 plays an important role. Once the operator norm bound is established, the clustering consistency follows from a standard analysis of the $k$-means algorithm (Lemma C.1).
4 Discussion

An important theoretical question in the study of stochastic block models is the critical threshold for community recovery. 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 recovery accuracy when the density and/or separation are above this threshold, as well as impossibility for community recovery below this threshold. For single-layer stochastic block models, this problem has been studied by many authors, such as [24, 2, 39, 25]. The case of multi-layer stochastic block models is much less clear, especially for generally structured layers. The upper bounds proved in [29, 6] imply achievability of vanishing error proportion when $L\rho \rightarrow \infty$ under a layer-wise positivity assumption. Our results requires a stronger $L^{1/2}n\rho/\sqrt{\log n} \rightarrow \infty$ condition, but without layer-wise positivity. 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 3.1 seems to suggest a positive answer. A rigorous claim will require a formal lower bound analysis, where the simplified constructions such as that in [39] cannot work, since it does not reflect the additional hardness brought to the estimation problem by unknown layer-wise structures.

The consistency result for multi-layer stochastic block models 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 [9, 21]. The probability tools developed in this paper, such as Theorem 2.1, Theorem 2.3, Theorem 2.4, and Theorem 2.5, may be useful for other statistical inference problems involving matrix-valued measurements and noise. For example, in dynamic networks where the network parameters change smoothly over time, one may use nonparametric kernel smoothing techniques [30] and the matrix concentration inequalities developed in this paper to control the aggregated noise and perhaps obtain more refined analysis.

A Proofs for general concentration results

Proof of Theorem 2.1. We will prove the asymmetric case first. The symmetric case follows by consider upper and lower diagonal of $X_{\ell}$ separately and use union bound.

First consider the case of a single pair of $(X, H)$, where $X$ is $n \times r$ with independent $(v_1, R_1)$-Bernstein entries, and $H$ is $r \times m$. Then

$$XH = \sum_{1 \leq i \leq n, 1 \leq j \leq r} X_{ij}e_iH^T_j.$$
By Lemma 2.2 we have
\[
\left| \mathbb{E} \left[ \mathcal{D}(X_{i,j}e_iH_j^T) \right]^k \right| \leq \mathbb{E}|X_{i,j}|^k \left| \mathbb{E} \left[ \mathcal{D}(e_iH_j^T) \right]^k \right| \\
\leq (v_1/2)^k! R_{k-2}^k \|H_j\|^{k-2} \left[ \begin{array}{cc} \|H_j\|^2e_ie_i^T & 0 \\ 0 & H_jH_j^T \end{array} \right]
\]

Now take the sum over \(i, j\).
\[
\sum_{1 \leq i \leq n, 1 \leq j \leq r} \left[ \|H_j\|^2e_ie_i^T \begin{array}{c} 0 \\ H_jH_j^T \end{array} \right] \leq \sum_{\ell} \left[ \begin{array}{cc} \|H_{\ell}\|^2I_n & 0 \\ 0 & nH^T H \end{array} \right].
\]

By Theorem 6.2 of [32], we have
\[
\mathbb{P}(\| \mathcal{D}(XH) \| \geq t) \leq 2(m + n) \exp \left( -\frac{t^2/2}{v_1(n\|H^T H\| \vee \|H\|^2_{F}) + R_1\|H\|_{2,\infty}} \right).
\]

The proof for the case of sum \(\sum_{\ell} X_{\ell}H_{\ell}\) follows by modifying the above argument where the summation in (19) takes another outer layer of summation over \(\ell\) and becomes
\[
\sum_{\ell} \sum_{1 \leq i \leq n, 1 \leq j \leq r} \left[ \|H_j\|^2e_ie_i^T \begin{array}{c} 0 \\ H_jH_j^T \end{array} \right] \leq \sum_{\ell} \left[ \begin{array}{cc} \|H_{\ell}\|^2I_n & 0 \\ 0 & n\sum_{\ell} H_{\ell}^T H_{\ell} \end{array} \right].
\]

The proof for the symmetric case, let \(X_{\ell}^{(u)}\) be the diagonal and upper-diagonal part of \(X_{\ell}\), and \(X_{\ell}^{(l)} = X_{\ell} - X_{\ell}^{(u)}\). The claim follows by upper bounding \(\mathbb{P}(\| \sum_{\ell} X_{\ell}^{(u)} H_{\ell}\| \geq t/2)\) and \(\mathbb{P}(\| \sum_{\ell} X_{\ell}^{(l)} H_{\ell}\| \geq t/2)\) using the asymmetric result, and combining with union bound.

**Proof of Theorem 2.3.** The proof uses decoupling. Let \(\tilde{X}_{\ell}\) be an independent copy of \(X_{\ell}\). Define
\[
\tilde{S} = \sum_{\ell} X_{\ell}G_{\ell}\tilde{X}_{\ell}^T,
\]
\[
\tilde{S}_2 = \sum_{\ell} \sum_{1 \leq i < j \leq n} X_{\ell,ij}\tilde{X}_{\ell,ij} \left( e_ie_i^T G_{\ell,ij} + e_je_j^T G_{\ell,ji} + e_ie_i^T G_{\ell,ji} + e_je_j^T G_{\ell,ij} \right) \\
+ \sum_{\ell} \sum_{1 \leq i \leq n} X_{\ell,ii}\tilde{X}_{\ell,ii}e_ie_i^T G_{\ell,ii},
\]
and
\[
\tilde{S}_1 = \tilde{S} - \tilde{S}_2.
\]
Now we expand $S_1$:

\[ S_1 = S - S_2 \]

\[ = \sum_{\ell} \sum_{\substack{1 \leq i < j \leq n \\{i,j\} \neq \{i',j'\}}} X_{\ell,ij} X_{\ell,i'j'} \left( e_i e_{i'}^T G_{\ell,jj'} + e_j e_{j'}^T G_{\ell,ii'} + e_i e_{i'}^T G_{\ell,ji'} + e_j e_{j'}^T G_{\ell,ij'} \right) \]

\[ + \sum_{\ell} \sum_{\substack{1 \leq i < j \leq n \\{i,j\} \neq \{i',j'\}}} X_{\ell,ii} X_{\ell,i'j'} (e_i e_{i'}^T G_{\ell,ii'} + e_i e_{i'}^T G_{\ell,ii'}) \]

\[ + \sum_{\ell} \sum_{\substack{1 \leq i < j \leq n \\{i,j\} \neq \{i',j'\}}} X_{\ell,ij} X_{\ell,i'j'} (e_i e_{i'}^T G_{\ell,ii'} + e_j e_{j'}^T G_{\ell,ii'}) \]

\[ + \sum_{\ell} \sum_{\substack{1 \leq i < j \leq n \\{i,j\} \neq \{i',j'\}}} X_{\ell,ii} X_{\ell,i'j'} e_i e_{i'} G_{\ell,ii'} , \]

which can be viewed as a matrix-valued U-statistic defined on the vectors $(X_{\ell,ij} : 1 \leq \ell \leq L)$ indexed by pairs $(i,j)$ such that $1 \leq i \leq j \leq n$. Using the de-coupling inequality (Theorem 1 of [10]) we have

\[ \mathbb{P}(\|S_1\| \geq t) \leq C_2 \mathbb{P}(\|\tilde{S}_1\| \geq t/C_2) \]  

for some universal constant $C_2$ and all $t > 0$.

The plan is to control $\|\tilde{S}_1\|$ by $\|\tilde{S}_1\| \leq \|\tilde{S}\| + \|\tilde{S}_2\|$.

**Step 1: Controlling $\tilde{S}$.** Let $H_\ell = G_\ell \tilde{X}_\ell$. In order to apply Theorem 2.1 to control $\tilde{S}_2 = \sum_\ell X_\ell H_\ell$ conditioning on $H_\ell$, we need to upper bound

\[ \left\| \sum_\ell H_\ell^T H_\ell \right\| = \left\| \sum_\ell \tilde{X}_\ell G_\ell^T G_\ell \tilde{X}_\ell^T \right\| \]

and

\[ \max_\ell \|H_\ell\|_{2,\infty} = \max_{i,j} \|\tilde{X}_\ell G_{\ell,j}\|. \]

We first consider $\|\sum_\ell H_\ell^T H_\ell\|$. With high probability over $\tilde{X}_\ell$, we have

\[ \left\| \sum_\ell H_\ell^T H_\ell \right\| = \left\| \sum_\ell \tilde{X}_\ell G_\ell^T G_\ell \tilde{X}_\ell^T \right\| \leq \sum_\ell \|\tilde{X}_\ell G_\ell^T G_\ell \tilde{X}_\ell^T\| \]
\[ \sum_{\ell} \|X_\ell G_\ell^T\|^2 \]
\[ \leq \sum_{\ell} \left[ v_1 n \log(L + n) \|G_\ell\|^2 + R_1^2 \|G_\ell^T\|^2_{2,\infty} \log^2(L + n) \right] \]
\[ = v_1 n \log(L + n) \sum_{\ell} \|G_\ell\|^2 + R_1^2 \log^2(L + n) \sum_{\ell} \|G_\ell^T\|^2_{2,\infty} \]
\[ \leq v_1 n \log(L + n) \sigma_1^2 + R_1^2 \log^2(L + n) \sigma_2^2, \]

where the fourth line follows from applying Theorem 2.1 to each individual $X_\ell G_\ell^T$ with union bound over $\ell$ and the fact that the entries of $X_\ell$ are $(v_1, R_1)$-Bernstein.

Now we turn to $\max_\ell \|H_\ell\|_{2,\infty}$. Applying Theorem 2.1 to $X_\ell G_{\ell,j}$ and taking union bound over $j, \ell$ we get, with high probability
\[ \max_{\ell,j} \|H_\ell\|_{2,\infty} \lesssim \sqrt{v_1} \sqrt{n \log(L + n)} \max_\ell \|G_\ell\|_{2,\infty} + R_1 \max_\ell \|G_\ell\|_{\infty} \log(L + n) \]
\[ \lesssim \sqrt{v_1} \sqrt{n \log(L + n)} \sigma_2 + R_1 \log(L + n) \sigma_3. \]

Conditioning on the intersection of these two events above, applying Theorem 2.1 we conclude with high probability
\[ \|\hat{S}\| \lesssim \sqrt{v_1} \sqrt{n \log(L + n)} \left[ v_1 n \log(L + n) \sigma_1^2 + R_1^2 \log^2(L + n) \sigma_2^2 \right]^{1/2} \]
\[ + R_1 \log(L + n) \left[ \sqrt{v_1} \sqrt{n \log(L + n)} \sigma_2 + R_1 \log(L + n) \sigma_3 \right] \]
\[ \lesssim v_1 n \log(L + n) \sigma_1 + \sqrt{v_1} R_1 \sqrt{n \log(L + n)} \log^{3/2}(L + n) \sigma_2 + R_1^2 \log^2(L + n) \sigma_3. \]

**Step 2:** $\hat{S}_2$. Let $Z_{\ell,ij} = X_{\ell,ij} X_{\ell,ij}^T$. By construction, the off-diagonal part of $\hat{S}_2$ is
\[ \sum_{\ell} \sum_{1 \leq i < j \leq n} Z_{\ell,ij} (e_i e_j^T G_{\ell,ij} + e_j e_i^T G_{\ell,ij}). \]

Consider the first component $\sum_{\ell} \sum_{1 \leq i < j \leq n} Z_{\ell,ij} G_{\ell,ij} e_i e_j^T$. Lemma 2.2 implies that
\[ \mathbb{E} \left[ \|D(Z_{\ell,ij} G_{\ell,ji} e_i e_j^T)\|^k \right] \preceq \frac{v_2^2 G_{\ell,ij}^2}{2} |R_2 G_{\ell,ji}|^{k-2} \left[ \begin{array}{cc} e_i e_i^T & 0 \\ 0 & e_j e_j^T \end{array} \right] \]
provided that $Z_{\ell,ij}$’s are $(v_2', R_2')$-Bernstein.

Summing over $(\ell, i, j)$ we obtain
\[ \sum_{\ell,i<j} \mathbb{E} \left[ \|D(Z_{\ell,ij} G_{\ell,ji} e_i e_j^T)\|^k \right] \preceq \frac{v_2'}{2} |R_2'\max_{\ell,i,j} G_{\ell,ji}|^{k-2} \max_{i,j} \sum_{\ell,i} G_{\ell,ji}^2 \max_{\ell,i} \sum_{\ell,j} G_{\ell,ji}^2 \right] I_{2n} \]
\[
\frac{\nu'}{2} |R'_2 \sigma_3|^{k-2} L \sigma_2^2 I_{2n}.
\]

Then with high probability the off-diagonal part of \(\tilde{S}_2\) is bounded by
\[
\sqrt{\nu'_2 L \log(L+n) \sigma_2 + R'_2 \log(L+n) \sigma_3}.
\]  

(23)

For the diagonal part of \(\tilde{S}_2\), the \(i\)th diagonal entry is
\[
\sum_{\ell} \sum_j Z_{\ell,ij} G_{\ell, jj}.
\]

Then the operator norm of diagonal part of \(\tilde{S}_2\) is bounded by its maximum entry, which is further bounded by, using standard Bernstein’s inequality
\[
\sqrt{\nu'_2 \log(L+n) \left( \sum_{\ell,j} G_{\ell, jj}^2 \right)^{1/2}} + R'_2 \max_{\ell,j} |G_{\ell, jj}| \log(L+n)
\]
\[
\leq \sqrt{\nu'_2 \log(L+n) \sigma_2^2 + R'_2 \log(L+n) \sigma_3}.
\]  

(24)

Now (7) follows by combining (22), (23), and (24) together with the de-coupling inequality (20).

The claim regarding \(S - ES\) only requires an additional bound on \(\|S_2 - ES_2\|\), which can be obtained using an identical to that of \(\tilde{S}_2\) with \((v_2, R_2)\) replacing \((v'_2, R'_2)\). ■

**Proof of Theorem 2.4.** Define \(\tilde{S}, \tilde{S}_1, \tilde{S}_2\) accordingly. It is easy to check that \(\tilde{S}_2\) only has diagonal entries and can be bounded by the same technique as in the symmetric case where
\[
\|\tilde{S}_2\| \lesssim \sqrt{\nu'_2 \log(L+n) \sigma_2^2 + R'_2 \log(L+n) \sigma_3}.
\]  

(25)

with high probability.

For \(\tilde{S}\), let \(H_\ell = G_\ell \tilde{X}_\ell^T\), then with high probability
\[
\|H_\ell\| \lesssim \sqrt{\nu_1 \log(L+n)} \left( \sqrt{n} \|G_\ell\| \lor \|G_\ell\|_F \right) + R_1 \log(L+n) \|G_\ell\|_2,\infty.
\]

and
\[
\left\| \sum_\ell H_\ell^T H_\ell \right\| \lesssim \nu_1 \log(L+n) \left( n \sigma_1^2 + (\sigma_1')^2 \right) + R_1^2 L \log^2(L+n) \sigma_2^2,
\]  

(26)

The rest of the proof is the same as that of Theorem 2.3. ■
B  Proofs for the sparse Bernoulli case

The proof of Theorem 2.5 follows a similar idea to that of Theorem 2.3, which uses decoupling and reduces the problem to a linear combination in the form of $\sum_{\ell} X_{\ell} H_{\ell}$. The proof here uses a refinement in constructing $H_{\ell}$ and controlling $\|\sum_{\ell} H_{\ell}^T H_{\ell}\|$ using properties of Bernoulli random variables. The refinement involves carefully bounding the degrees of $A_{\ell}$, as well as $\sum_{\ell} A_{\ell}^2$, which is provided in Lemma B.1.

**Lemma B.1.** Let $A_1, \ldots, A_L$ be independent adjacency matrices generated by a multi-layer stochastic block model satisfying the condition of Theorem 3.1. The following holds with probability at least $1 - O((L + n)^{-1})$ for some universal constant $C$:

1. $\max_{\ell,i} d_{\ell,i} \leq C \log(L + n)$.
2. $\max_i \sum_{\ell} d_{\ell,i} \leq CLn\rho$.
3. $\sum_{\ell,i} d_{\ell,i} \leq CLn^2\rho$.
4. $\|\sum_{\ell} A_{\ell}^2\| \leq CLn\rho$.

**Proof.** Parts 1 follows from direct application of Bernstein’s inequality and union bound:

$$\mathbb{P}[d_{\ell,i} - n\rho \geq t] \leq \exp \left(- \frac{t^2/2}{4n\rho + 2t} \right),$$

and use the assumption that $n\rho \leq C_2 \log n$.

For part 2, $\sum_{\ell} d_{\ell,i}$ has expected value at most $Ln\rho$. To control the deviation, Bernstein’s inequality implies that

$$\mathbb{P} \left[ \sum_{\ell} (d_{\ell,i} - \mathbb{E}d_{\ell,i}) \geq t \right] \leq \exp \left(- \frac{t^2/2}{4\rho nL + 2t} \right)$$

with probability at least $1 - O((L + n)^{-1})$

$$\max_i \sum_{\ell} (d_{\ell,i} - \mathbb{E}d_{\ell,i}) \leq C \left[ \rho^{1/2}n^{1/2}L^{1/2} \log^{1/2}(L + n) + \log(L + n) \right]$$

$$\leq C\rho^{1/2}n^{1/2}L^{1/2} \log^{1/2}(L + n).$$

For part 3, first we have $\mathbb{E} \sum_{\ell,i} d_{\ell,i} \leq Ln^2\rho$, and the deviation satisfies

$$\mathbb{P} \left[ \sum_{\ell,i} X_{\ell,ij} \geq t \right] \leq \exp \left(- \frac{t^2/2}{16\rho n^2L + 4t} \right).$$
The claim follows from the assumption $\rho n^2 L \gtrsim \rho^{1/2} n L^{1/2} \sqrt{\log(L + n) + \log(L + n)}$.

For part 4, first decompose
$$\sum_{\ell} A^2_{\ell} = S_{1A} + S_{2A},$$
where $S_{2A}$ is the diagonal part of $\sum_{\ell} A^2_{\ell}$, with
$$(S_{2A})_{ii} = \sum_{\ell} d_{\ell, i}.$$ Use part 2, we have with high probability
$$\|S_{2A}\| = \max_i (S_{2A})_{ii} \leq CN\rho.$$ (27)

For the off-diagonal part $S_{1A} = \sum_{\ell} A^2_{\ell} - S_{2A}$, we can obtain a high probability bound using decoupling. Let $\tilde{S}_{1A}$ be the corresponding version of $S_{1A}$ for $\sum_{\ell} A_\ell \tilde{A}_\ell$. For a matrix $M$, let $\|M\|_{1,\infty} = \max_i \sum_j |M_{ij}|$ be the maximum row-wise $\ell_1$ norm. Using symmetric dilation, Perron-Frobenius theorem and non-negativity of $A_\ell$, $\tilde{A}_\ell$ we have
$$\|\tilde{S}_{1A}\| \leq \max \left\{ \|\tilde{S}_{1A}\|_{1,\infty}, \|\tilde{S}_{1A}^T\|_{1,\infty} \right\} \leq \max \left\{ \left\| \sum_{\ell} A_{\ell} \tilde{A}_{\ell} \right\|_{1,\infty}, \left\| \sum_{\ell} \tilde{A}_{\ell} A_{\ell} \right\|_{1,\infty} \right\}.$$ By symmetry, it suffices to upper bound the maximum row sum of $\sum_{\ell} A_{\ell} \tilde{A}_{\ell}$. The sum of the $i$th row is
$$\sum_{\ell,j,m} A_{\ell,im} \tilde{A}_{\ell,jm} = \sum_{\ell,m} A_{\ell,im} \tilde{d}_{\ell,m}$$
whose expected value is upper bounded by $\rho^2 n^2 L$.

Conditioning on the event that $\max_{\ell,m} \tilde{d}_{\ell,m} \leq C \log(L + n)$ and $\max_i \sum_{\ell} \tilde{d}_{\ell,i} \leq CLn\rho$, the mean deviation
$$\sum_{\ell} \sum_m X_{\ell,im} \tilde{d}_{\ell,m}$$
can be bounded Bernstein’s inequality
$$\mathbb{P} \left[ \sum_{\ell,m} X_{\ell,im} \tilde{d}_{\ell,m} \geq t \mid \tilde{A}_\ell \right] \leq \exp \left( -\frac{t^2/2}{4\rho \sum_{\ell,m} \tilde{d}_{\ell,m}^2 + 2t \max_{\ell,m} \tilde{d}_{\ell,m}} \right) \leq \exp \left( -\frac{t^2/2}{4C\rho \log(L + n) \sum_{\ell,m} \tilde{d}_{\ell,m} + 2Ct \log(L + n)} \right)$$
\[
\leq \exp \left( -\frac{t^2}{4C\log(L+n)\rho^2n^2L + 2Ct \log(L+n)} \right).
\]

Using union bound over \(i\) we conclude that with probability at least \(1 - O((L+n)^{-1})\)
\[
\max_i \sum_{\ell,m} X_{\ell,m} \tilde{d}_{\ell,m} \leq C\rho nL^{1/2} \log(L+n) \leq C\rho^2 n^2 L.
\]

Therefore we proved that with high probability \(\|\tilde{S}_{1,A}\| \leq C\rho^2 n^2 L\). Combining this with (27) we have with high probability
\[
\left\| \sum_{\ell} A_{\ell}^2 \right\| \leq \|S_{1,A}\| + \|S_{2,A}\| \leq C\rho n L. \quad \blacksquare
\]

**Proof of Theorem 2.5.** By the sparse Bernoulli assumption, \(X_\ell\) satisfy Assumption 1 with \((v_1, R_1) = (2\rho, 1)\) and Assumption 2’ with \((v'_2, R'_2) = (2\rho^2, 1)\).

First using the decoupling argument we reduce the problem to controlling \(\tilde{S}\) and \(\tilde{S}_2\) respectively.

For \(\tilde{S}_2\), it is easy to verify that \(\tilde{S}_2\) is a diagonal matrix with
\[
(\tilde{S}_2)_{ii} = \sum_{\ell,j} X_{\ell,ij} \tilde{X}_{\ell,ij},
\]
which is a sum of \(nL\) independent zero-mean, \((2\rho^2, 1)\)-Bernstein random variables. Using Bernstein’s inequality and union bound over \(i\), we have with probability at least \(1 - (L+n)^{-1}\)
\[
\|\tilde{S}_2\| = \max_i \left| (\tilde{S}_2)_{ii} \right| \leq C\rho n^{1/2} L^{1/2} \sqrt{\log(L+n)}.
\]  

(28)

Now we turn to \(\tilde{S}\). Recall that \(X_\ell = A_\ell - P_\ell\), where \(A_\ell\) consists of the uncentered versions of the corresponding entries of \(X_\ell\), and \(P_\ell = \mathbb{E}A_\ell\).

\[
\tilde{S} = \sum_{\ell} X_\ell \tilde{X}_\ell = \sum_{\ell} X_\ell \tilde{A}_\ell - \sum_{\ell} X_\ell P_\ell
\]

Using Theorem 2.1 and the fact that \(\|P_\ell\| \leq n\rho\) and \(\|P_\ell\|_{2,\infty} \leq \rho n^{1/2}\), we have with probability at least \(1 - ((L+n)^{-1})\) and universal constant \(C\)
\[
\sum_{\ell} X_\ell P_\ell \leq C \left[ \rho^{3/2} n^{3/2} L^{1/2} \sqrt{\log(L+n)} + \rho n^{1/2} \log(L+n) \right]
\]
\[
\leq C\rho^{3/2} n^{3/2} L^{1/2} \sqrt{\log(L+n)}. 
\]

(29)
Now we focus on $\sum_\ell X_\ell \tilde{A}_\ell$ by conditioning on $\tilde{A}_\ell$. By Lemma B.1, with high probability $\|\sum_\ell \tilde{A}_\ell^2\| \leq C\rho n L$ and $\max_\ell \|\tilde{A}_\ell\|_{2,\infty} = \max_\ell \tilde{d}_{\ell,i}^{3/2} \leq C \log^{1/2}(L + n)$. Applying Theorem 2.1 conditioning on this event we have with high probability

$$
\left\| \sum_\ell X_\ell \tilde{A}_\ell \right\| \leq C \left[ \rho^{1/2} n^{1/2} (\rho n L \log(L + n))^{1/2} + \log^{3/2}(L + n) \right]
\leq C \rho n L^{1/2} \log^{1/2}(L + n).
$$

Combining (29) and (30) we obtain with high probability

$$
\|\tilde{S}\| \leq C \rho n L^{1/2} \log^{1/2}(L + n).
$$

The claimed bound holds for $\tilde{S}_1$ by combining (28) and (31), and hence holds for $S_1$ by de-coupling.

C Proof of consistency of bias-adjusted spectral clustering

The plan is to decompose the matrix $S_0$ into the sum of a signal term and a noise term, where the signal term has a leading principal subspace with perfect clustering, and then apply matrix perturbation results (the Davis-Kahan sin $\Theta$ theorem) combined with a standard error analysis of the k-means algorithm. We first introduce some notation a preliminary result for the k-means problem.

Given an $n \times d$ matrix $\hat{U}$, the K-means problem is an optimization problem

$$
\min_{\Theta, X} \|\hat{U} - \Theta X\|_F^2
$$

where the minimization is over all $\Theta \in \{0, 1\}^{n \times K}$ with each row has exactly one “1”, and all $X \in \mathbb{R}^{K \times d}$. We say a pair $(\hat{\Theta}, \hat{X})$ is an $(1 + \epsilon)$-approximate solution if its objective function value is no larger than $(1 + \epsilon)$ times the optimal value.

**Lemma C.1** (Simplified from Lemma 5.3 of [20]). Let $U$ be an $n \times d$ matrix with $K$ distinct rows with minimum pairwise Euclidean norm separation $\gamma$. Let $\hat{U}$ be another $n \times d$ matrix and $(\hat{\Theta}, \hat{X})$ be an $(1 + \epsilon)$-approximate solution to K-means problem with input $\hat{U}$, then the number of errors in $\hat{\Theta}$ as an estimate of the row clusters of $U$ is no larger than

$$
C_\epsilon \|\hat{U} - U\|_F^2 \gamma^{-2}
$$

for some constant $C_\epsilon$ depending only on $\epsilon$. 

24
Proof of Theorem 3.1. Let $Q_\ell = \rho \Theta B_{\ell,0} \Theta^T$ then $P_\ell = Q_\ell - \text{diag}(Q_\ell)$ and

$$\sum_\ell P_\ell^2 = \sum_\ell Q_\ell^2 + \|\text{diag}(Q_\ell)\|^2 - Q_\ell \text{diag}(Q_\ell) - \text{diag}(Q_\ell) Q_\ell$$

$$= \sum_\ell Q_\ell^2 + E_1,$$

where $E_1 = \sum_\ell \|\text{diag}(Q_\ell)\|^2 - Q_\ell \text{diag}(Q_\ell) - \text{diag}(Q_\ell) Q_\ell$.

Define

$$E_2 = \sum_\ell X_\ell P_\ell + P_\ell X_\ell,$$

$$E_3 = S_2 - \sum_\ell D_\ell,$$

$$E_4 = S_1,$$

where $S_1, S_2$ are defined as in (5) and (6) with $G_\ell = I_n$. By the definition of $S_0$ in (18) and the decomposition (16) we have

$$S_0 = \sum_\ell Q_\ell^2 + E_1 + E_2 + E_3 + E_4.$$

Let $\Theta = \tilde{\Theta} \Lambda$ where $\Lambda$ is a $K \times K$ diagonal matrix with $k$th diagonal entry being the $\ell_2$ norm of the $k$th column of $\Theta$. Then $\Theta$ is orthonormal. By the balanced community size assumption the minimum eigenvalue of $\Lambda$ is lower bounded by $cn^{1/2}$ for some constant $c$ (recall that $K$ is assumed to be a constant). The signal term

$$\sum_\ell Q_\ell^2 = \rho^2 \tilde{\Theta}^T \Lambda \left( \sum_\ell B_{\ell,0} \Lambda^2 B_{\ell,0} \right) \Lambda \tilde{\Theta} \geq c n \rho^2 \tilde{\Theta}^T \Lambda \left( \sum_\ell B_{\ell,0} B_{\ell,0} \right) \Lambda \tilde{\Theta} \geq c L n \rho^2 \tilde{\Theta}^T \Lambda^2 \tilde{\Theta} \geq c L n \rho^2 \tilde{\Theta}^T \tilde{\Theta},$$

where we used $\Lambda^2 \succeq cn^2 I_K$ and Assumption 3. Eq. (32) implies that the matrix $\sum_\ell Q_\ell^2$ is rank $K$ and the leading eigen-space is spanned by the columns of $\tilde{\Theta}$. Eq. (33) implies the smallest non-zero eigenvalue of $\sum_\ell Q_\ell^2$ is lower bounded by $c L n \rho^2$.

The first bias term $E_1$ is non-random and satisfies $\|E_1\| \leq L n \rho^2$. 

25
For the noise term $E_2$, applying Theorem 2.1 with $H_\ell = P_\ell$ and realizing that $\|P_\ell\| \leq n\rho$ and $\|P_\ell\|_{2,\infty} \leq \sqrt{n}\rho$ we have with high probability

$$\|E_2\|_{\infty} \leq CL^{1/2}n^{1/2}\rho^{3/2}\log^{1/2}(L + n).$$

(34)

and consequently

$$\|E_2\| \leq CL^{1/2}n^{3/2}\rho^{3/2}\log^{1/2}(L + n).$$

(35)

For $E_3 = S_2 - \sum_\ell D_\ell$, the decomposition (17) implies that $\|S_2 - \sum_\ell D_\ell\|$ can be upper bounded deterministically by $Ln\rho^2$.

Next we control $E_4 = S_1$ by refining the result of Theorem 2.3 using properties of Bernoulli random variables. The details are given in Theorem 2.5 and we have

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

(36)

with high probability. Thus

$$\frac{\|E_1 + E_2 + E_3 + E_4\|}{\lambda_K(\sum_\ell Q_\ell^2)} \leq C\frac{Ln\rho^2 + L^{1/2}n\rho \log^{1/2}(L + n)}{Ln^2\rho^2} \leq C + \frac{C\log^{1/2}(L + n)}{L^{1/2}n\rho},$$

where $\lambda_K(\sum_\ell Q_\ell^2)$ is the $K$th and smallest non-zero eigenvalue of $\sum_\ell Q_\ell^2$. Let $U, \hat{U}$ be the $n \times K$ matrices consisting of the leading eigenvectors of $\sum_\ell Q_\ell^2$ and $S_0$, respectively. By the Davis-Kahan sin $\Theta$ theorem, we have $\|\hat{U} - U\|_F \leq \sqrt{K}\|\hat{U} - U\| \leq \frac{\sqrt{K}\|E_1 + E_2 + E_3 + E_4\|}{\lambda_K(\sum_\ell Q_\ell^2) - \|E_1 + E_2 + E_3 + E_4\|} \lesssim n^{-1} + \log^{1/2}(L + n)/(L^{1/2}n\rho)$. The rest proof follows from Lemma C.1 because part 1 of Assumption 3 implies that the minimum separation of two distinct rows in $U$ is at least $C/\sqrt{n}$ for some constant $C$.

**References**

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

[2] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *arXiv preprint arXiv:1503.00609*, 2015.

[3] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.
[4] Afonso S Bandeira and Ramon Van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, 44(4):2479–2506, 2016.

[5] Rajendra Bhatia. *Matrix Analysis*. Springer-Verlag, 1997.

[6] Sharmodeep Bhattacharyya and Shirshendu Chatterjee. Spectral clustering for multiple sparse networks: I. *arXiv preprint arXiv:1805.10594*, 2018.

[7] Peter J Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[8] Joshua Cape, Minh Tang, and Carey E Priebe. The kato–temple inequality and eigenvalue concentration with applications to graph inference. *Electronic Journal of Statistics*, 11(2):3954–3978, 2017.

[9] Kehui Chen and Jing Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 113(521):241–251, 2018.

[10] Victor H de la Peña and Stephen J Montgomery-Smith. Decoupling inequalities for the tail probabilities of multivariate u-statistics. *The Annals of Probability*, pages 806–816, 1995.

[11] Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

[12] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[13] Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

[14] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

[15] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[16] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.

[17] Eric D Kolaczyk. *Statistical analysis of network data*. Springer, 2009.
[18] Can M Le, Elizaveta Levina, and Roman Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 51(3):538–561, 2017.

[19] Jing Lei, Kehui Chen, and Brian Lynch. Consistent community detection in multi-layer network data. *Biometrika*, 2019.

[20] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

[21] Tianxi Li, Elizaveta Levina, and Ji Zhu. Network cross-validation by edge sampling. *Biometrika*, 2020, to appear.

[22] Martin A Lindquist. The statistical analysis of fmri data. *Statistical Science*, 23(4):439–464, 2008.

[23] Fuchen Liu, David Choi, Lu Xie, and Kathryn Roeder. Global spectral clustering in dynamic networks. *Proceedings of the National Academy of Sciences*, 115(5):927–932, 2018.

[24] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703, 2014.

[25] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018.

[26] Mark Newman. *Networks: an introduction*. Oxford University Press, 2009.

[27] Sean O’Rourke, Van Vu, and Ke Wang. Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra and its Applications*, 540:26–59, 2018.

[28] Subhadeep Paul and Yuguo Chen. Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electronic Journal of Statistics*, 10(2):3807–3870, 2016.

[29] Subhadeep Paul and Yuguo Chen. Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *arXiv preprint arXiv:1704.07353*, 2017.

[30] Marianna Pensky and Teng Zhang. Spectral clustering in the dynamic stochastic block model. *Electronic Journal of Statistics*, 13(1):678–709, 2019.

[31] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[32] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.
[33] Toni Valles-Catala, Francesco A Massucci, Roger Guimera, and Marta Sales-Pardo. Multilayer stochastic block models reveal the multilayer structure of complex networks. Physical Review X, 6(1):011036, 2016.

[34] A. W. van der Vaart and J. A. Wellner. Weak Convergence and Empirical Processes. Springer-Verlag, 1996.

[35] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, and Wu-Minn HCP Consortium. The wu-minn human connectome project: an overview. Neuroimage, 80:62–79, 2013.

[36] Roman Vershynin. Spectral norm of products of random and deterministic matrices. Probability theory and related fields, 150(3-4):471–509, 2011.

[37] Tengyao Wang, Quentin Berthet, and Yaniv Plan. Average-case hardness of rip certification. In Advances in Neural Information Processing Systems, pages 3819–3827, 2016.

[38] Kevin S Xu and Alfred O Hero. Dynamic stochastic blockmodels for time-evolving social networks. IEEE Journal of Selected Topics in Signal Processing, 8(4):552–562, 2014.

[39] Anderson Y Zhang and Harrison H Zhou. Minimax rates of community detection in stochastic block models. The Annals of Statistics, 44(5):2252–2280, 2016.