Journal of Machine Learning Research 23 (2022) 1-48 Submitted 12/19; Revised 5/21; Published 2/22

Recovering shared structure from multiple networks with unknown edge distributions

Keith Levin KDLEVIN@WISC.EDU
Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706, USA

Asad Lodhia ASAD.I.LODHIA@GMAIL.COM
Unaffiliated

Elizaveta Levina ELEVINA@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109, USA

Editor: Inderjit Dhillon

Abstract

In increasingly many settings, data sets consist of multiple samples from a population of networks, with vertices aligned across networks; for example, brain connectivity networks in neuroscience. We consider the setting where the observed networks have a shared expectation, but may differ in the noise structure on their edges. Our approach exploits the shared mean structure to denoise edge-level measurements of the observed networks and estimate the underlying population-level parameters. We also explore the extent to which edge-level errors influence estimation and downstream inference. In the process, we establish a finite-sample concentration inequality for the low-rank eigenvalue truncation of a random weighted adjacency matrix, which may be of independent interest. The proposed approach is illustrated on synthetic networks and on data from an fMRI study of schizophrenia.

Keywords: Networks; Network averaging; Concentration inequalities; Spectral methods

1. Introduction

Many modern applications require simultaneous analysis of multiple networks, often with the goal of identifying structure that is shared across multiple networks. In the social sciences, this may correspond to some common underlying structure that appears, for example, in different friendship networks across high schools. In biology, one may be interested in identifying the extent to which different organisms' protein-protein interaction networks display a similar structure. In neuroscience, one may wish to identify common patterns across multiple subjects' brains in an imaging study. This last application in particular is easily abstracted to the situation where one observes a collection of independent graphs on the same vertex set, as there are well-established and widely used algorithms that map locations in individual brains onto a common atlas of so-called regions of interest (ROIs), such as the one developed by Power et al. (2011). Thus, the assumption of vertex correspondence

across graphs is especially common in multiple network analysis for neuroimaging applications; see for example Levin et al. (2017); Arroyo et al. (2019). A common approach to these problems is to treat the individual networks as independent noisy realizations of some shared structure such as a stochastic block model (Le et al., 2018) or a low-rank model (Tang et al., 2018). The goal is then to recover this underlying shared structure. The importance of analyzing brain data from multiple subjects simultaneously has spurred a particularly active line of work on this problem in neuroimaging. As a result, we primarily focus on this literature, but stress that the general problem of recovering a shared structure from multiple networks (or, more generally, multiple matrices) has applications to many other domains.

To date, most techniques for multiple-network analysis have assumed that the observed networks come from the same distribution, and typically have binary edges, a restrictive assumption in many settings. For example, in social networks, edge weights may represent the strengths of friendships, in gene expression networks, edge weights represent the extent to which pairs of genes are co-expressed (Zhang and Horvath, 2005), and in neuroimaging, edge weights represent the strength of connectivity between brain regions of interest. Substantial information is lost if these weights are truncated to binary; see, for example, Aicher et al. (2015). The shared noise distribution is also a restrictive assumption, since while we may reasonably expect that some population-level structure is shared across networks, network-level variation is likely to be heterogeneous. Specifically in neuroimaging, a lot of subject-level variation comes from head motion or deviations of the individual subject's brain from the common atlas, and there is no reason to expect these to be homogeneous. There are a number of different pipelines in use for reducing this type of noise in fMRI data (see, for example, Ciric et al., 2017, for a discussion), but all introduce artifacts of one kind or another, which we model here as potentially heterogeneous edge noise.

In this paper, we develop techniques for analyzing multiple networks without these two assumptions. In particular, we allow for weighted edges and heterogeneous noise distributions, and study how to estimate the underlying population mean. Under these conditions, the simple arithmetic mean of weighted graphs is likely to be sub-optimal, as networks with higher noise levels will contribute as much as those with less noise, and one would intuitively expect an estimate that takes noise levels into account to perform better. While there are a number of possible matrix means we might consider (see, for example, those described in Bhatia, 2007), in this paper, we focus on the case of weighted arithmetic means of networks. That is, letting $A^{(1)}, \ldots, A^{(N)} \in \mathbb{R}^{n \times n}$ be the adjacency matrices of independent graphs on the same vertex set, we are interested in estimators of the form $\sum_{s=1}^{N} \hat{w}_s A^{(s)}$, where $\{\hat{w}_s\}_{s=1}^{N}$ are non-negative, data-dependent weights summing up to 1.

There have been a number of papers written in recent years related to analysis of multiple vertex-aligned networks. Motivated by brain imaging applications similar to those discussed here, Tang et al. (2018) considered the problem of estimating a low-rank population matrix, when graphs are drawn i.i.d. from a random dot product graph model (Athreya et al., 2018), and investigated the asymptotic relative efficiency of a low-rank approximation of the sample mean of these observed graphs compared to the graph sample mean itself. Levin et al. (2017) considered the problem of analyzing multiple vertex-aligned graphs, and devised a method to compare geometric representations of graphs, typically called embeddings in the literature, for the purpose of exploratory data analysis and hypothesis testing, focused particularly

on comparing vertices across graphs. In a similar spirit, Wang et al. (2021) considered the problem of embedding multiple binary graphs whose adjacency matrices (approximately) share eigenspaces, while possibly differing in their eigenvalues. All of these papers assume binary networks and identical noise distributions on edges, in contrast to the setting we study here.

Eynard et al. (2015) developed a technique for analyzing multiple manifolds by (approximately) simultaneously diagonalizing a collection of graph Laplacians. Like our work, the technique in Eynard et al. (2015) aims to recover spectral information shared across multiple observed graphs, but differs in that the authors work with weighted similarity graphs that arise from data lying on a manifold, and derive a perturbation bound instead of applying a specific statistical or probabilistic model. The authors also require that the population graph Laplacian have a simple spectrum, an additional assumption we do not need.

A few recent papers have considered the problem of analyzing multiple networks generated from stochastic block models with the same community structure, but possibly different connection probability matrices (Tang et al., 2009; Dong et al., 2014; Han et al., 2015; Paul and Chen, 2016; Bhattacharyya and Chatterjee, 2018). Similar approaches have been developed for time-varying networks, where it is assumed that the connection probability may change over time, but community structure is constant or only slowly varying (Xu and Hero, 2014). Our setting is distinct from this line of work, since we assume a general shared structure with varying distribution of edge noise, and do not require edges to be binary.

Tang et al. (2017) considered the problem of estimating shared low-rank structure based on a sample of networks under the setting where individual edges are drawn from contaminated distributions (Huber, 1964). The paper compares the theoretical guarantees of estimates based on edge-wise (non-robust) maximum likelihood estimation, edge-wise robust maximum likelihood estimation (Ferrari and Yang, 2010), and eigenvalue truncations of both. The present work does not focus on robustness, and as a result is largely not comparable to Tang et al. (2017), although our procedures can be made robust in a similar fashion if desirable.

The remainder of this paper is organized as follows: In Section 2, we give a formal description of the problem and present necessary background. Our main results are presented in Section 3, with shows the optimality of a certain weighted network average, and Section 4, which proposes an algorithm to estimate these optimal weights from data. Section 5 briefly considers the application to community detection and estimation in the multiple-network setting. Section 6 explores the effectiveness of our method on simulated data as well as on a fMRI dataset from a study comparing schizophrenic and healthy patients. We conclude with a brief discussion in Section 7 summarizing our results and sketching directions for future work.

2. Problem Setup and Notation

Throughout this paper, we assume that we observe N undirected graphs each on n vertices with corresponding adjacency matrices, $A^{(1)}, A^{(2)}, \ldots, A^{(N)} \in \mathbb{R}^{n\times n}$. We will refer to the s-th graph and its adjacency matrix $A^{(s)}$ interchangeably. Our key assumption is that the graphs are drawn independently with shared expectation $\mathbb{E}A^{(s)} = P \in \mathbb{R}^{n\times n}$, for all $s \in [N] = \{1, 2, \ldots, N\}$. Throughout, we will denote the rank of P by $d = \operatorname{rank} P$. All

results will depend on d, but we do not make a low-rank assumption; all our results are finite sample and thus valid for any $d \leq n$. We assume that the vertices are aligned across the graphs, in the sense that the i-th vertex in graph $A^{(s)}$ is identifiable with the i-th vertex in graph $A^{(t)}$ for all $i \in [n]$ and $s, t \in [N]$. As a motivating example, consider the case where the observed graphs are obtained from fMRI neuroimaging of N patients. In such a setting, the n vertices of each graph correspond to brain regions of interest (ROIs), identified based on alignment to a common template (e.g., Power et al., 2011). This common alignment ensures that the i-th vertex in each of these connectomes corresponds to the same anatomical region, and thus this i-th vertex can be sensibly identified across graphs. If the N patients belong to a common population (e.g., shared disease status), it is reasonable to expect that these networks exhibit a shared structure. In this work, we take this shared structure to be a common expectation. Note that for the case of $N = 1$ a low-rank or another structural assumption would have to be made on P in order to enable estimation, since otherwise we would only have one observation per parameter. With $N > 1$ observed networks, this is not strictly necessary, but a low-rank or some other structural assumption would certainly enable better estimation, just like it does for the $N = 1$ case.

Crucially for the purposes of this paper, while the expectation P is shared across the graphs, we allow for each graph to exhibit different edge noise structure. That is, for each $s \in [N]$ and $i, j \in [n]$, $(A^{(s)} - P)_{ij}$ has mean 0 but otherwise arbitrary distribution $F = F_{s,ij}$, which may depend both on the subject s and the specific edge (i, j). In the motivating neuroimaging application, this corresponds to the fact that high-level anatomical and functional structure is likely common across patients, but measurement noise is likely subject-specific, and edge heterogeneity may also result from individual differences. For simplicity of notation, we allow for self-loops, i.e., treat $A^{(s)}_{ii}$ exactly the same way as the off-diagonal entries; self-loops are generally a moot point for asymptotics, since they make a negligible $O(n)$ contribution compared to the $O(n^2)$ off-diagonal entries. Throughout, we assume that all parameters, including the number of networks N, can depend on the number of vertices n, though we mostly suppress this dependence for ease of reading. We write C for a generic positive constant, not depending on n, whose value may change from one line to the next.

Before proceeding, we pause to establish notation. For an integer k, we write $[k]$ for the set $\{1, 2, \ldots, k\}$. For a vector v, we write $\|v\|$ for the Euclidean norm of v. For a matrix M, $\|M\|$ denotes the spectral norm, $\|M\|_F$ the Frobenius norm, and $\|M\|_{2,\infty}$ the $(2, \infty)$ norm, $\|M\|_{2,\infty} = \sup_{v:\|v\|=1} \|Mv\|_\infty$, where $\|v\|_\infty = \max_i |v_i|$. For a positive semidefinite matrix M, we write $\kappa(M)$ for the ratio of the largest eigenvalue of M to its largest *non-zero* eigenvalue. We use standard Landau notation to $O(\cdot), o(\cdot), \Omega(\cdot)$, and $\omega(\cdot)$ to denote growth rates. For example, $g(n) = O(f(n))$ as $n \to \infty$ means that $|g(n)| < Cf(n)$ for some constant C and all $n > n_0$, $g(n) = \Omega(f(n))$ means $g(n) > Cf(n)$, and so on. We use $\tilde{O}$ to denote growth rate up to log-factors, as in, for example, $n \log^2 n = \tilde{O}(n)$. In a slight abuse of the term, we say that an event E_n occurs *with high probability* (w.h.p.) if $\mathbb{P}[E_n^c] \leq Cn^{-(1+\epsilon)}$ for some constant $\epsilon > 0$. This definition allows us to state our results as finite-sample bounds while immediately implying asymptotic results of the form "with probability 1, event B_n occurs for at most finitely many n" by applying the Borel-Cantelli lemma.

2.1 Motivating Examples

We now present a few examples that satisfy our model assumptions, in order of increasing generality. In all cases, the question is how to optimally recover the underlying shared expectation P. We begin with one of the simplest possible settings under our model.

Example 1 (Normal measurement errors with subject-specific variance) *Assume that for each* $s = 1, 2, \ldots, N$, $\{(A^{(s)} - P)_{ij} : 1 \leq i \leq j \leq n\}$ *are independent* $\mathcal{N}(0, \rho_s)$.

A weaker assumption on the edge measurement errors would be to replace the specific distributional assumption that $(A^{(s)} - P)_{ij} \sim \mathcal{N}(0, \rho_s)$ with a more general tail bound assumption, such as sub-Gaussian or sub-gamma errors. We refer the reader to Appendix A for the definition and a few basic properties of sub-Gaussian and sub-gamma random variables, or to Boucheron et al. (2013) for a more substantial discussion.

Example 2 (Sub-gamma measurement errors with subject-specific parameter) *Assume that for each* $s = 1, 2, \ldots, N$, $\{(A^{(s)} - P)_{ij} : 1 \leq i \leq j \leq n\}$ *are independent, mean* 0*, sub-gamma with parameters* (ν_s, b_s).

Note that the assumptions of Example 2 do not require the edges to be identically distributed within a network. We can further relax the sub-gamma assumption to allow for edge-specific tail parameters rather than having a single tail parameter for each subject.

Example 3 (Sub-gamma errors with subject- and edge-specific parameters) *Assume that for all* $s = 1, 2, \ldots, N$, $\{(A^{(s)} - P)_{ij} : 1 \leq i \leq j \leq n\}$ *are independent, mean* 0*, and for each* $s \in [N]$ *and* $i, j \in [n]$, $(A^{(s)} - P)_{ij}$ *is sub-gamma with parameters* $(\nu_{s,ij}, b_{s,ij})$.

In all of these examples, there are several inference questions we may wish to ask. In this work, we focus on

1. Recovering the matrix P;
2. Recovering $X \in \mathbb{R}^{n \times d}$ when $P = XX^T$;
3. Recovering community memberships when P corresponds to a stochastic block model (Holland et al., 1983).

Given that the observed graphs $A^{(1)}, A^{(2)}, \ldots, A^{(N)}$ differ in their noise structure, the question arises as to how to combine these graphs to estimate P (or X or the community memberships). When the observed graphs are drawn i.i.d. from the same distribution, the sample mean $\bar{A} = N^{-1} \sum_{s=1}^{N} A^{(s)}$ is a natural estimate of P, and has been studied in this context in Tang et al. (2018) and in a related test statistic in Chen et al. (2020).

However, in our more general setting, where individual networks and/or edges may be more noisy than others, de-emphasizing noisier observations will lead to a more reliable estimate of P. We pursue this by choosing data-dependent weights $\{\hat{w}_s \geq 0 : s = 1, 2, \ldots, N\}$ with $\sum_{s=1}^{N} \hat{w}_s = 1$ so that the weighted mean estimate $\hat{A} = \sum_{s=1}^{N} \hat{w}_s A^{(s)}$ is optimal in some sense, or at least provably better than the sample mean $\bar{A}$.

Remark 1 (Positive semi-definite assumption on P) In what follows, we make the additional assumption that the expectation $P \in \mathbb{R}^{n\times n}$ is positive semi-definite, so that $P = XX^T$ for some $X \in \mathbb{R}^{n\times d}$, where $d \leq n$. This assumption can be removed using the techniques in Rubin-Delanchy et al. (2017), at the cost of added notational complexity. Thus, for ease of exposition, we confine ourselves to the case where $P = XX^T$, bearing in mind that our results can be easily extended to any P.

2.2 Recovering Spectral Structure

Throughout this paper, we will begin from the standard first step in spectral clustering (Rohe et al., 2011). Following the terminology of Sussman et al. (2012), we refer to this as adjacency spectral embedding (ASE). Given any adjacency matrix $A \in \mathbb{R}^{n\times n}$, with a rank d expectation P, write $P = XX^T = U_P S_P U_P^T \in \mathbb{R}^{n\times n}$ where $S_P \in \mathbb{R}^{d\times d}$ is diagonal with entries given by the d non-zero eigenvalues of P, and the d corresponding orthonormal eigenvectors are the columns of $U_P \in \mathbb{R}^{n\times d}$. Since $P = XX^T = XQ(XQ)^T$ for orthogonal $Q \in \mathbb{R}^{d\times d}$, the matrix X is only identifiable up to an orthogonal rotation. Thus, we take $X = U_P S_P^{1/2}$ without loss of generality. One can view the rows of $X \in \mathbb{R}^{n\times d}$ as latent positions in $\mathbb{R}^d$ associated with the vertices, with the expectation of an edge between nodes i and j given by the inner product $P_{ij} = X_i^T X_j$ of their latent positions, where $X_i \in \mathbb{R}^d$ is the i-th row of X. This view motivates the random dot product graph model (Athreya et al., 2018), in which the latent positions are first drawn i.i.d. from some underlying distribution on $\mathbb{R}^d$ and edges are generated independently conditioned on the latent positions. The natural estimate of the matrix $X = U_P S_P^{1/2}$ is then

$$\text{ASE}(A, d) = U_A S_A^{1/2} \in \mathbb{R}^{n\times d},$$

where the eigenvalues in S_A and eigenvectors in U_A now come from A rather than the unknown P. One can show that under appropriate conditions, the ASE recovers the matrix X up to an orthogonal rotation that does not affect the estimate $\hat{P} = U_A S_A U_A^T \in \mathbb{R}^{n\times n}$.

While we do not focus on the random dot product graph model in this paper, we make use of several generalizations of results initially established for that model. These results are summarized in Appendix B.1. In general, selecting the embedding dimension d (i.e., estimating the rank of P) is an interesting and challenging problem (see, e.g., Fishkind et al., 2013; Han et al., 2019), but it is not the focus of the present work, and we assume throughout that d is known. When P is rank d, one can show that under suitable growth conditions, the gap between the d-th largest eigenvalue and the $(d+1)$-th largest eigenvalue of P grows with n, and the same property holds for the adjacency matrix A. As a result, it is not unreasonable to assume that one can accurately determine the appropriate dimension d when the number of vertices n is large.

3. Methods and Theoretical Results

Generally speaking, we are interested in estimators of the form $\hat{A} = \sum_{s=1}^N \hat{w}_s A^{(s)}$, where $\{\hat{w}_s\}_{s=1}^N$ are data-dependent nonnegative weights summing to 1. In particular, we study how well the rank-d eigenvalue truncation of $\hat{A}$ approximates the true rank-d expectation

P. We measure this either by bounding the difference $\hat{A} - P$ in some matrix norm or by proving that we can successfully recover the matrix $X \in \mathbb{R}^{n \times d}$ from the estimate $\hat{A}$.

In this section, we largely consider the weights $\{w_s\}_{s=1}^N$ to be fixed, rather than data-dependent. The special case of normally-distributed edge noise is illustrative, as it suggests a certain choice of weights in the more general case (see Theorem 6). In Section 4, we will estimate these optimal weights and replace the fixed $\{w_s\}_{s=1}^N$ with data-dependent estimates $\{\hat{w}_s\}_{s=1}^N$.

3.1 Normal edges with subject-specific variance

Return for a moment to Example 1, in which $A_{i,j}^{(s)} \sim \mathcal{N}(P_{i,j}, \rho_s)$, independent for all $1 \leq s \leq N$ and $1 \leq i \leq j \leq n$. This very simple setting suggests a choice for the data-dependent weights $\{\hat{w}_s\}_{s=1}^N$ when constructing the matrix $\hat{A}$. Indeed, a natural extension of the estimator suggested by this example will turn out to be the right choice in the more general settings described in Section 2. The following proposition follows immediately from writing out the joint log-likelihood of the N observed networks and rearranging terms.

Proposition 2 *Suppose $\mathbb{E}A^{(s)} = P \in \mathbb{R}^{n \times n}$ for all $s \in [N]$, and $P = XX^T$ for some $X \in \mathbb{R}^{n \times d}$. If for all $s = 1, 2, \ldots, N$ the edges $\{(A^{(s)} - P)_{i,j} : 1 \leq i \leq j \leq n\}$ are i.i.d. normal mean 0 and known variance $\rho_s > 0$, then the maximum likelihood estimate for $X \in \mathbb{R}^{n \times d}$ (up to an orthogonal rotation) is given by $\mathrm{ASE}((\sum_{t=1}^N \rho_t^{-1})^{-1} \sum_{s=1}^N A^{(s)}/\rho_s, d)$.*

Motivated by this proposition, consider the plug-in estimator given by

$$\hat{X} = \mathrm{ASE}\left(\left(\sum_{t=1}^N \hat{\rho}_t^{-1}\right)^{-1} \sum_{s=1}^N \hat{\rho}_s^{-1} A^{(s)}, d\right), \tag{1}$$

where $\hat{\rho}_s$ is an estimate of the variance ρ_s of the edges in network s. Given P, one could naturally use the MLE for ρ_s. Since P is unknown, we estimate it as $\hat{P}^{(s)} = \hat{X}^{(s)}\hat{X}^{(s)T}$, where $\hat{X}^{(s)} = \mathrm{ASE}(A^{(s)}, d)$, and plug in for the MLE of ρ_s to obtain, for $s = 1, 2, \ldots, N$,

$$\hat{\rho}_s = \sum_{1 \leq i \leq j \leq n} \frac{2(A^{(s)} - \hat{P}^{(s)})_{i,j}^2}{n(n+1)}. \tag{2}$$

With $O(n^2)$ edges in each network, the estimates $\{\hat{\rho}_s\}_{s=1}^N$ converge to the true variances $\{\rho_s\}_{s=1}^N$ in such a way that the estimation rate of the plug-in estimator in Equation (1) matches that of the maximum-likelihood estimator in Proposition 2. The following proposition makes this claim precise.

Proposition 3 *Let $A^{(1)}, A^{(2)}, \ldots, A^{(N)}$ be independent adjacency matrices with common expectation $\mathbb{E}A^{(s)} = P = XX^T$, where $X \in \mathbb{R}^{n \times d}$, and suppose that for each $s \in [N]$, $\{(A^{(s)} - P)_{i,j} : 1 \leq i \leq j \leq n\}$ are independent $\mathcal{N}(0, \rho_s)$. Let $\mathring{X} \in \mathbb{R}^{n \times d}$ denote the maximum likelihood estimator under the assumption of known variances, as described in Proposition 2.*

Suppose that

$$\sum_{s=1}^N \rho_s^{-1} = \omega\left(\frac{n \log^2 n}{\lambda_d^2(P)}\right), \tag{3}$$

Then for all suitably large n, there exists orthogonal matrix $\mathring{V} \in \mathbb{R}^{d\times d}$ such that

$$\|\mathring{\hat{X}} - X\mathring{V}\|_{2,\infty} \leq \frac{Cd}{\lambda_d^{1/2}(P)}\left(\sum_{s=1}^{N}\rho_s^{-1}\right)^{-1/2} + \frac{Cd\kappa(P)n}{\lambda_d^{3/2}(P)}\left(\sum_{s=1}^{N}\rho_s^{-1}\right)^{-1}.$$

Further, let $\hat{X} \in \mathbb{R}^{n\times d}$ denote the estimator defined in Equation (1)*. For all suitably large n, there exists orthogonal matrix $V \in \mathbb{R}^{d\times d}$ such that with probability $1 - Cn^{-2}$,*

$$\|\hat{X} - XV\|_{2,\infty} \leq \frac{Cd}{\lambda_d^{1/2}(P)}\left(\sum_{s=1}^{N}\rho_s^{-1}\right)^{-1/2} + \frac{Cd\kappa(P)n}{\lambda_d^{3/2}(P)}\left(\sum_{s=1}^{N}\rho_s^{-1}\right)^{-1}.$$

This proposition is a special case of Theorem 9 in Section 4, and thus we delay its proof until then.

3.2 Sub-gamma edges

In settings like those in Examples 2 and 3, where there are no longer parameters controlling the noise distribution, we must resort to more general concentration inequalities. Our main tool in this setting is a generalization of a bound on the error in recovering $X = U_P S_P^{1/2} \in \mathbb{R}^{n\times d}$. Given a single adjacency matrix $A \in \mathbb{R}^{n\times n}$ with $\mathbb{E}A = P = XX^T \in \mathbb{R}^{n\times n}$, a natural estimate of $X \in \mathbb{R}^{n\times d}$ is $\mathrm{ASE}(A, d)$. The following lemma bounds the difference between this estimate and an orthogonal rotation (V in the result below) of $X = U_P S_P^{1/2}$. This bound makes no use of a particular error structure, but instead bounds the difference in terms of three different norms of $A-P$. This error term can then be bounded using standard concentration inequalities, which we will do below in the proof of Theorem 6.

Lemma 4 *Let $P = XX^T = U_P S_P U_P^T \in \mathbb{R}^{n\times n}$ be a rank-d matrix with non-zero eigenvalues $\lambda_1(P) \geq \lambda_2(P) \geq \cdots \geq \lambda_d(P) > 0$. Let $A \in \mathbb{R}^{n\times n}$ be a random symmetric matrix for which there exists a constant $c_0 \in [0, 1)$ such that with probability p_0,*

$$\|A - P\| < c_0\lambda_d(P) \tag{4}$$

for all suitably large n. Letting $\hat{X} = \mathrm{ASE}(A, d)$, for all suitably large n, there exists a random orthogonal matrix $V = V_n \in \mathbb{R}^{d\times d}$ such that with probability p_0

$$\begin{aligned}&\|\hat{X} - XV\|_{2,\infty}\\ &\quad\leq \frac{\|(A-P)U_P\|_{2,\infty}}{\lambda_d^{1/2}(P)} + \frac{C\|U_P^T(A-P)U_P\|_F}{\lambda_d^{1/2}(P)} + \frac{Cd\|A-P\|^2\kappa(P)}{\lambda_d^{3/2}(P)}.\end{aligned}$$

This lemma generalizes Theorem 18 in Lyzinski et al. (2017) and Lemma 1 in Levin et al. (2017). Details are included in Section B.1 of the Appendix. We note that recent work (Cape et al., 2019, Theorem 4.2) established a $(2, \infty)$-norm bound for eigenvector recovery, a problem related to, but fundamentally different from, the problem of recovering $X \in \mathbb{R}^{n\times d}$ considered in Lemma 4.

In order to apply Lemma 4 to the random matrix $\tilde{A} = \sum_{s=1}^{N} w_s A^{(s)}$, we need to ensure that the spectral condition in Equation (4) holds. Toward that end, the following lemma

bounds the spectral error between $\tilde{A}$ and P in terms of the weights and the sub-gamma parameters. We begin by considering the case where the weights $\{w_s\}_{s=1}^N$ are *not* data dependent. In Section 4, we will consider the case where the weights are a function of the networks $A^{(1)}, A^{(2)}, \ldots, A^{(N)}$.

Lemma 5 *Suppose that $A^{(1)}, A^{(2)}, \ldots, A^{(N)}$ are independent symmetric adjacency matrices with shared expectation $\mathbb{E}A^{(s)} = P \in \mathbb{R}^{n\times n}$, and suppose that $\{(A^{(s)} - P)_{i,j} : s \in [N], 1 \leq i \leq j \leq n\}$ are independent sub-gamma random variables with parameters $(\nu_{s,i,j}, b_{s,i,j})$. Let $w_1, w_2, \ldots, w_N \geq 0$ be non-random weights with $\sum_{s=1}^N w_s = 1$. Then with probability at least $1 - Cn^{-2}$,*

$$\left\| \tilde{A} - P \right\| \leq \frac{15\sqrt{2\eta^2}}{2} \log n,$$

where

$$\eta^2 = 2 \max_{i\in[n]} \sum_{s=1}^N \sum_{j=1}^n w_s^2 (\sqrt{2\nu_{s,i,j}} + 2b_{s,i,j})^2.$$

Lemma 5 follows from a standard matrix Bernstein inequality (Tropp, 2012). Details are provided in Appendix B.2.

While bounds for recovering X are also possible under the setting of Example 3, in which $(A^{(s)} - P)_{i,j}$ is $(\nu_{s,i,j}, b_{s,i,j})$-sub-gamma for each $s \in [N], i, j \in [n]$, the bounds are comparatively complicated functions of these parameters. For simplicity, we state the following theorem for the case where the edges in the s-th network are independent (ν_s, b_s)-sub-gamma random variables.

Theorem 6 *Under the setting of Lemma 5, with the additional condition that $P = XX^T$ for some $X \in \mathbb{R}^{n\times d}$ and $(\nu_{s,i,j}, b_{s,i,j}) = (\nu_s, b_s)$ for all $i, j \in [n]$, let $\tilde{X} = \mathrm{ASE}(\tilde{A}, d)$. Suppose that the weights $\{w_s\}_{s=1}^N$ and sub-gamma parameters $\{(\nu_s, b_s)\}_{s=1}^N$ are such that*

$$\sum_{s=1}^N w_s^2 (\nu_s + b_s^2) = o\left(\frac{\lambda_d^2(P)}{n \log^2 n} \right) \tag{5}$$

Then with probability $1 - Cn^{-2}$ there exists an orthogonal matrix $V \in \mathbb{R}^{d\times d}$ such that

$$\begin{aligned} &\|\tilde{X} - XV\|_{2,\infty} \\ &\quad \leq \frac{Cd}{\lambda_d^{1/2}(P)} \left(\sum_{s=1}^N w_s^2 (\nu_s + b_s^2) \right)^{1/2} \log n + \frac{Cdn\kappa(P)}{\lambda_d^{3/2}(P)} \left(\sum_{s=1}^N w_s^2 (\nu_s + b_s^2) \right) \log^2 n. \end{aligned}$$

This theorem follows from applying Lemma 4 with $A = \tilde{A}$, using Lemma 5 and Equation (5) to ensure that Equation (4) holds, and applying standard concentration inequalities to control the resulting bound on $\|\tilde{X} - XV\|_{2,\infty}$. Details can be found in Appendix B.2.

Remark 7 To illustrate the consequences of Theorem 6, let us consider two different settings for the sub-gamma parameters. First, consider the case when all $\{\nu_s + b_s^2\}_{s=1}^N$ are of constant order. Setting $w_s = N^{-1}$ for $s = 1, 2, \ldots, N$ then yields $\sum_s w_s^2(\nu_s + b_s^2) = O(N^{-1})$,

and Theorem 6 shows that, unsurprisingly, network averaging yields faster recovery in estimating X compared with the single-network setting, at least in the case where all of the networks have comparable levels of edge uncertainty.

Contrast this case with the setting where $(\nu_1 + b_1^2) = N^2$ and $(\nu_s + b_s^2) = 1$ for $s = 2, 3, \ldots, N$. Applying Theorem 6, if we take $w_s = N^{-1}$ for all $s \in [N]$, our estimation accuracy for X is controlled by

$$\sum_s w_s^2(\nu_s + b_s^2) = \frac{1}{N^2}\left(N^2 + N - 1\right) = O(1). \tag{6}$$

On the other hand, setting the weights to

$$w_1 = \frac{N^{-2}}{N^{-2} + N - 1} \quad \text{and} \quad w_s = \frac{1}{N^{-2} + N - 1} \quad \text{for } s = 2, 3, \ldots, N,$$

implies that the estimation rate in Theorem 6 is controlled by

$$w_1^2(\nu_1 + b_1^2) + \sum_{s=2}^N w_s^2(\nu_s + b_s^2) = \frac{1}{N^{-2} + N - 1} = O(N^{-1}).$$

Comparing this rate with the rate in Equation (6), we see that weighted network averaging can yield qualitatively better estimation in the setting where one or more networks have much higher uncertainty in their edge measurements.

3.3 Selecting network weights

When $A^{(s)}$ has sub-gamma edge noise with parameters (ν_s, b_s) common for all edges, the bound in Theorem 6 is a monotone function of the quantity $\sum_s w_s^2(\nu_s + b_s^2)$. This suggests that we choose the weights $\{w_s\}_{s=1}^N$ so as to minimize this quantity, which is achieved by taking

$$w_s = \mathring{w}_s = \frac{(\nu_s + b_s^2)^{-1}}{\sum_{t=1}^N (\nu_t + b_t^2)^{-1}} \tag{7}$$

for each $s \in [N]$. Applying Lemma 5 to $\mathring{A} = \sum_s \mathring{w}_s A^{(s)}$, and using the assumption that $(\nu_{s,i,j}, b_{s,i,j}) = (\nu_s, b_s)$ for all $s \in [N]$ and $i, j \in [n]$, we conclude that with high probability,

$$\left\|\mathring{A} - P\right\| \leq C\sqrt{\frac{n}{\sum_s (\nu_s + b_s^2)^{-1}}}\log n, \tag{8}$$

where we have used the fact that

$$\nu_s + b_s^2 \leq (\sqrt{\nu_s} + b_s)^2 \leq 2(\nu_s + b_s^2).$$

Perhaps surprisingly, when the edges are normally distributed about their expectations, this selection of weights $\{\mathring{w}_s\}_{s=1}^N$ yields the minimax optimal rate for recovering P in spectral norm. When $(A^{(s)} - P)_{i,j} \sim \mathcal{N}(0, \rho_s)$ for all $1 \leq i \leq j \leq n$, we can take the sub-gamma parameters to be $(\nu_s, b_s) = (\rho_s, 0)$ for all $s \in [N]$, and the bound in Equation (8) becomes

$$\left\|\mathring{A} - P\right\| \leq C\sqrt{\frac{n}{\sum_s \rho_s^{-1}}}\log n,$$

and this matches the minimax rate, as the following result shows. A proof can be found in the Appendix.

Theorem 8 *Let $A^{(1)}, A^{(2)}, \ldots, A^{(N)} \in \mathbb{R}^{n\times n}$ be independent symmetric adjacency matrices, with $\{(A^{(s)}_{ij} - P_{ij}) : 1 \le i \le j \le n\}$ drawn i.i.d. from a normal with mean 0 and variance ρ_s for each $s = 1, 2, \ldots, N$. Then, letting $\mathcal{S}_n = \{P \in \mathbb{R}^{n\times n} : P = P^T\}$, for all $n \ge 2$,*

$$\inf_{\hat{P}} \sup_{P\in\mathcal{S}_n} \mathbb{E}\|\hat{P} - P\| \ge C\sqrt{n}\left(\sum_{s=1}^{N} \rho_s^{-1}\right)^{-1/2},$$

where the infimum is over all estimators $\hat{P}$ of P.

4. Estimating the Sub-gamma Parameters

In the setting where each network $A^{(s)}$ has edges with shared sub-gamma parameter (ν_s, b_s), the results in Sections 3.2 and 3.3 suggested choosing our network weights according to Equation (7). Of course, in practice, we do not know the sub-gamma parameters $\{(\nu_s, b_s)\}_{s=1}^N$ and hence we must estimate them in order to obtain estimates of the optimal weights. Since $(\nu_s + b_s^2)$ is (up to a constant factor) an upper bound on the variances of the $\{(A^{(s)} - P)_{ij} : 1 \le i \le j \le n\}$, a natural estimate of $\mathring{w}_s$ is

$$\hat{w}_s = \frac{\hat{\rho}_s^{-1}}{\sum_{t=1}^{N} \hat{\rho}_t^{-1}}, \tag{9}$$

where, letting $\hat{P}^{(s)} \in \mathbb{R}^{n\times n}$ be the rank-d truncation of $A^{(s)}$ for $s = 1, 2, \ldots, N$,

$$\hat{\rho}_s = \frac{\sum_{1\le i\le j\le n}\left(A^{(s)}_{ij} - \hat{P}^{(s)}_{ij}\right)^2}{16n(n+1)}. \tag{10}$$

Comparison with Equation (2) reveals that this is, in essence, the same estimation procedure that we derived in Section 3.1, extended to the case of sub-gamma edges. The factor of 16 in the denominator comes from replacing the equality $\mathbb{E}(A^{(s)} - P)^2_{ij} = \rho_s$ with the sub-gamma moment bound (Boucheron et al., 2013, Chapter 2, Theorem 2.3)

$$\mathbb{E}(A^{(s)} - P)^2_{ij} \le 8\nu_s + 32b_s^2 \le 32(\nu_s + b_s^2). \tag{11}$$

Just as in Section 3.1, the estimated weights $\{\hat{w}_s\}_{s=1}^N$ are such that the plug-in estimate $\mathrm{ASE}(\sum_s \hat{w}_s A^{(s)}, d)$ recovers the true matrix $X \in \mathbb{R}^{n\times d}$ (up to orthogonal nonidentifiability) at the same rate as we would obtain if we knew the true sub-gamma parameters.

Theorem 9 *Suppose that $A^{(1)}, A^{(2)}, \ldots, A^{(N)}$ are independent symmetric adjacency matrices with shared expectation $\mathbb{E}A^{(s)} = P = XX^T \in \mathbb{R}^{n\times n}$, where $X \in \mathbb{R}^{n\times d}$. Suppose further that for each $s \in [N]$, $\{(A^{(s)} - P)_{i,j} : 1 \le i \le j \le n\}$ are independent sub-gamma random variables with parameters (ν_s, b_s). Let $\{\mathring{w}_s\}_{s=1}^N$ and $\{\hat{w}_s\}_{s=1}^N$ be the weights defined in Equations (7) and (9), respectively, and define the estimators*

$$\hat{X} = \mathrm{ASE}\left(\sum_{s=1}^{N} \hat{w}_s A^{(s)}, d\right), \quad \mathring{X} = \mathrm{ASE}\left(\sum_{s=1}^{N} \mathring{w}_s A^{(s)}, d\right).$$

Suppose that the parameters n, d and N grow in such a way that $d/n \leq 1$ for all suitably large n, and, some positive integer k,

$$\frac{n^{k-2}d^k\left(\log N + \log n\right)^{4k}}{N} = \Omega(1). \tag{12}$$

Suppose further that the sub-gamma parameters $\{(\nu_s, b_s)\}_{s=1}^N$ are such that

$$\left(\frac{1}{N}\sum_{s=1}^{N}(\nu_s + b_s^2)^{-1}\right)^{-1} = o\left(\frac{N\lambda_d^2(P)}{n\log^2 n}\right). \tag{13}$$

For each $s \in [N]$, define

$$\tau_s = \frac{\sum_{1\leq i\leq j\leq n}\mathbb{E}(A^{(s)} - P)_{ij}^2}{16n(n+1)}, \tag{14}$$

and suppose that

$$\lim_{n\to\infty}\frac{\sqrt{d}(\log N + \log n)^2 \max_{s\in[N]}\tau_s^{-1}(\nu_s + b_s^2)}{\sqrt{n}} = 0 \tag{15}$$

in such a way that

$$\left(\max_{s\in[N]}\tau_s^{-1}(\nu_s + b_s^2)\right)\sqrt{\sum_{s=1}^{N}\frac{\tau_s^{-1}\sum_{t=1}^{N}(\nu_t + b_t^2)^{-1}}{(\nu_s + b_s^2)^{-1}\sum_{t=1}^{N}\tau_t^{-1}}} = O\left(\frac{\sqrt{n}\log n}{\sqrt{d}\left(\log n + \log N\right)^3}\right). \tag{16}$$

Provided that

$$\sum_{s=1}^{N}\frac{(\nu_s + b_s^2)^{-1}}{\sum_t(\nu_t + b_t^2)^{-1}}\left(1 - \frac{\tau_s^{-1}}{\mathring{w}_s\sum_t\tau_t^{-1}}\right)^2 = O\left(\frac{\log n}{\sqrt{N}\left(\log n + \log N\right)}\right), \tag{17}$$

then for all suitably large n, it holds with probability $1 - O(n^{-2})$ that there exist orthogonal matrices $V, \mathring{V} \in \mathbb{R}^{d\times d}$ such that

$$\|\hat{X} - XV\|_{2,\infty} \leq \frac{Cd}{\lambda_d^{1/2}(P)}\left(\sum_{s=1}^{N}(\nu_s + b_s^2)^{-1}\right)^{-1/2} + \frac{Cd\kappa(P)n}{\lambda_d^{3/2}(P)}\left(\sum_{s=1}^{N}(\nu_s + b_s^2)^{-1}\right)^{-1}$$

and

$$\|\mathring{X} - X\mathring{V}\|_{2,\infty} \leq \frac{Cd}{\lambda_d^{1/2}(P)}\left(\sum_{s=1}^{N}(\nu_s + b_s^2)^{-1}\right)^{-1/2} + \frac{Cd\kappa(P)n}{\lambda_d^{3/2}(P)}\left(\sum_{s=1}^{N}(\nu_s + b_s^2)^{-1}\right)^{-1}.$$

That is, the plug-in estimator $\hat{X}$, based on the estimates weights $\{\hat{w}_s\}_{s=1}^N$, recovers $X \in \mathbb{R}^{n\times d}$ at the same rate as the estimator $\mathring{X}$ based on the optimal weights $\{\mathring{w}_s\}_{s=1}^N$.

The proof is given in Appendix B.4. Note that the bound on $\|\mathring{X} - X\mathring{V}\|_{2,\infty}$ follows straightforwardly from Theorem 6. The analysis of $\sum_s \hat{w}_s A^{(s)}$ requires more care, since the weights $\{\hat{w}_s\}_{s=1}^N$ now depend on the observed networks.

Remark 10 The quantities $\{\tau_s^{-1}(\nu_s + b_s^2)\}_{s=1}^N$ in Equations (15), (16) and (17) are, in essence, measures of the tightness of the sub-gamma tail bounds $\mathbb{E}(A^{(s)} - P)_{i,j}^2 \leq 32(\nu_s + b_s^2)$. For example, the quantity controlled by Equation (17) is the χ^2 divergence between the distribution on $[N]$ encoded by the optimal weights $\{\mathring{w}_s : s \in [N]\}$ and the distribution given by $u_s = \tau_s^{-1} / \sum_t \tau_t^{-1}$. In the simplest case, when $\{(A^{(s)} - P)_{i,j} : 1 \leq i \leq j \leq n\}$ are i.i.d. $\mathcal{N}(0, \rho_s)$ for some $\rho_s > 0$, we have $(\nu_s, b_s) = (\rho_s, 0)$, so that $32\tau_s = \rho_s = (\nu_s + b_s^2)$ and thus $\tau_s^{-1}(\nu_s + b_s^2) = 32$ for all $s \in [N]$. The growth conditions in Equations (15), (16) and (17) are then satisfied trivially, so long as $d(\log N + \log n)^4 = o(n)$.

Remark 11 As mentioned just before Theorem 6, the more general case, in which $(A^{(s)} - P)_{i,j}$ is $(\nu_{s,i,j}, b_{s,i,j})$-sub-gamma for each $s \in [N], i, j \in [n]$, is notably more complicated to analyze than the setting where there is a single (ν_s, b_s) parameter to estimate for each network $s = 1, 2, \ldots, N$. We expect that mild structural assumptions on the matrices $[\nu_{s,i,j}]_{i,j=1}^n$ and $[b_{s,i,j}]_{i,j=1}^n$ (e.g., low rank) would yield similar estimation procedures to that described above. The technical results presented in Appendix B.2 give an indication of the cumbersome notation required to handle more general structural assumptions on the sub-gamma parameters. We leave this generalization for future work.

5. Perfect Clustering with Sub-gamma Edges

With Theorem 6 in hand, we obtain an immediate bound on the community misclassification rate in block models by an argument similar to that in Lyzinski et al. (2014).

This bound holds for *weighted* stochastic blockmodels (SBMs), in which multiple weighted graphs are drawn with a shared block structure. Weighted versions of the stochastic blockmodel have received increasing attention in recent years (see, e.g., Aicher et al., 2015). Extensions to the case of weighted edges in a multiple-network setting was recently discussed in Khim and Loh (2018), where the authors considered a network time series problem. The definition given here subsumes any variant on the weighted SBM in which, conditional on the community assignments, edge distributions are independent and obey a sub-gamma tail bound.

Definition 12 (Joint Sub-gamma SBM) Let $B \in [0,1]^{K \times K}$ and define the community membership matrix $Z \in \{0,1\}^{n \times K}$ by $Z_{ik} = 1$ if the i-th vertex belongs to community k and $Z_{ik} = 0$ otherwise. We say that random adjacency matrices $A^{(1)}, A^{(2)}, \ldots, A^{(N)} \in \mathbb{R}^{n \times n}$ are *jointly sub-gamma stochastic block model*, written

$$(A^{(1)}, A^{(2)}, \ldots, A^{(N)}) \sim \text{J-}\Gamma\text{-SBM}(n, B, Z, \{(\nu_s, b_s)\}_{s=1}^N),$$

if conditional on Z, the N adjacency matrices are independent with a common expectation $\mathbb{E}A^{(s)} = ZBZ^T$ $(s = 1, 2, \ldots, N)$, and within each adjacency matrix $A^{(s)}$, $\{A_{ij}^{(s)} : 1 \leq i \leq j \leq n\}$ are independent (ν_s, b_s)-sub-gamma random variables.

Our theoretical results from Section 3 have an immediate implication for detection and estimation of shared community structure in the joint sub-gamma SBM model. This result generalizes Theorem 6 in Lyzinski et al. (2014).

Theorem 13 *Suppose that* $(A^{(1)}, \ldots, A^{(N)})$ *are drawn from* J-Γ-SBM$(n, B, Z, \{(\nu_s, b_s)\}_{s=1}^N)$, *where* $B = YY^T \in \mathbb{R}^{K\times K}$ *is fixed with* $Y \in \mathbb{R}^{K\times d}$ *having* K *distinct rows given by* $Y_1, Y_2, \ldots, Y_K \in \mathbb{R}^d$. *Let* $\tilde{A} = \sum_{s=1}^N w_s A^{(s)}$, *where* $\{w_s\}_{s=1}^N$ *are fixed non-negative weights summing to* 1. *Let* $n_{\min} = \min_{k\in[K]} \sum_{i=1}^n Z_{ik}$ *denote the size of the smallest community. Suppose that* $\{(w_s, \nu_s, b_s)\}_{s=1}^N$ *obey the growth conditions in Equation* (5) *and that the smallest community grows as*

$$n_{\min} = \omega\left(d^2\left(\sum_{s=1}^N w_s^2(\nu_s + b_s^2)\right) + d^2\left(\sum_{s=1}^N w_s^2(\nu_s + b_s^2)\right)^2 \log^4 n\right). \tag{18}$$

Let $\tau : [n] \to [K]$ *be the true underlying assignment of vertices to communities, so that* $\tau(i) = k$ *if and only if* $Z_{ik} = 1$, *and let* $\hat{\tau} : [n] \to [K]$ *be the estimated community assignment function based on an optimal* K*-means clustering of the rows of* $\tilde{X} = \mathrm{ASE}(\tilde{A}, d)$. *Then the communities are recovered exactly almost surely, i.e., as* $n \to \infty$,

$$\mathbb{P}\left[\min_{\pi\in S_K} |\{i \in [n] : \pi(\tau(i)) \neq \hat{\tau}(i)\}| \to 0\right] = 1 \tag{19}$$

Proof The result follows from Theorem 6 and the fact that under the stochastic block model with fixed parameters, we have $\lambda_d(P) = \Omega(n)$ (see, for example, Observation 2 in Levin et al., 2017). The proof is otherwise a direct adaptation of the proof of Theorem 6 in Lyzinski et al. (2014), and we omit the details. ■

Remark 14 (Extensions of Theorem 13) This result can be generalized in two natural directions. The first would be to allow for the communication matrix B to depend on n. Generally speaking, provided the entries of B do not go to zero too quickly, the quantities $n_{\min}$ and $\lambda_d(P)$ will grow quickly enough to ensure that $\|\tilde{X} - XW\|_F \to 0$, and Theorem 13 still holds This extension is straightforward and we omit the details (though see Remark 15 below). Another generalization would be to expand the class of clustering algorithms for which the perfect recovery condition in (19) holds. For example, an argument similar to that sketched in Theorem 13 would apply equally well to another distance-based clustering method, such as K-medians, K-medoids or K-centers (Garfinkel et al., 1977). When there are K clusters, the matrix Y has K distinct rows, say, $y_1, y_2, \ldots, Y_K \in \mathbb{R}^d$, so that for all $i \in [n]$, $X_i \in \{y_1, y_2, \ldots, y_K\}$. By Theorem 6, provided the parameters grow at suitable rates, for all suitably large n, the rows of $\tilde{X}$ lie in K disjoint balls centered at the K points $\{y_1, y_2, \ldots, y_K\}$, and a clustering solution that does not place a centroid in each of these K balls can be improved upon by a solution that does. We leave it for future work to characterize the clustering algorithms that obtain this recovery guarantee and the precise growth conditions on the model parameters required for these different algorithms to succeed.

Remark 15 (Incorporating Sparsity) A standard way to incorporate sparsity in the SBM is to let $B = q_n YY^T$, where $q_n \to 0$ is a sparsity parameter. Similar to in our previous Remark, recovery of the community memberships requires, in essence, that the rows of

$\sqrt{q_n}Y$ are suitably well separated. This imposes a lower-bound on how quickly the sparsity parameter q_n can converge to zero. Specifically, the condition in Equation (5) is satisfied so long as

$$\sum_s w_s^2(\nu_s + b_s^2) = o\left(\frac{q_n^2 n}{\log^2 n}\right).$$

Since a Bernoulli with success probability q is a $(q, 1)$-sub-gamma random variable, when $N = 1$ this becomes $(q_n + 1)\log^2 n = o(q_n^2 n)$, i.e., $q_n = \omega(n^{-1/2}\log n)$. For a single network, this is a stricter requirement on the average degree than the more typical $q_n = \omega(n^{-1})\log^c n)$ for some constant $c \geq 0$. While the extension to sub-gamma edge distributions allows a much more general class of noise models, our general bounds are not necessarily tight when applied to the highly-structured setting of Bernoulli edges, where the variance is constrained by the mean. When this additional structure is present, it is, unsurprisingly, more efficient to leverage it (see, e.g., Le et al., 2018).

6. Numerical experiments

We now turn to an experimental investigation of the effect of weighted averaging. We begin with simulated data, and then turn to a neuroimaging application.

6.1 Effect of weighted averaging on estimation

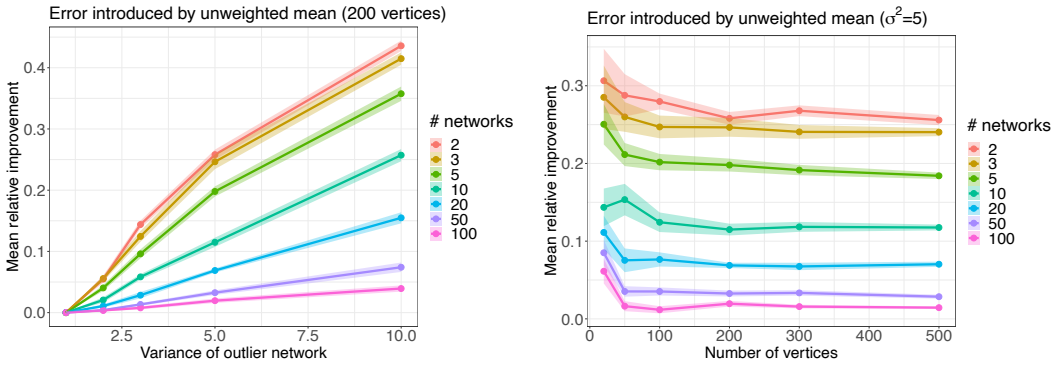
We begin by investigating the extent to which weighted averaging improves upon its unweighted counterpart in the case where we observe multiple weighted graphs with network-specific edge variances, as in Examples 1 and 2.

We consider the following simulation setup. On each trial, we generate the rows of $X \in \mathbb{R}^{n\times d}$ independently and identically distributed as $\mathcal{N}\left((1,1,1)^T, \Sigma\right)$, where

$$\Sigma = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 3 \end{bmatrix},$$

and take $P = XX^T$. Next, independently for each network $s = 1, 2, \ldots, N$, we draw edge weights $\{(A^{(s)} - P)_{ij} : 1 \leq i \leq j \leq n\}$ independently from a 0-mean Laplace distribution with variance $\sigma_s^2 > 0$. We chose this distribution because it has heavier tails than the Gaussian while still belonging to the class of sub-gamma random variables. Similar results to those presented here were also observed under Gaussian, exponential, and gamma error distributions. Without loss of generality, we take the first network to have edge variance $\sigma_1^2 \geq 1$, while all other networks have unit edge variance, so that $\sigma_s^2 = 1$ for $s > 1$. Thus, the first network is an outlier with higher edge-level variance than the other observed networks. We compare weighted and unweighted averaging, with weights estimated as described in Section 4, to obtain the weighted average $\hat{A} = \sum_{s=1}^N \hat{w}_s A^{(s)}$, and the unweighted average $\bar{A} = N^{-1}\sum_{s=1}^N A^{(s)}$. Rank-$d$ eigenvalue truncations of each of these yield estimates $\hat{P}$ and $\bar{P}$, respectively. We evaluate the weighted estimate by its relative improvement,

$$\frac{\|\bar{P} - P\| - \|\hat{P} - P\|}{\|\bar{P} - P\|}.$$



(a) (b)

Figure 1: Relative improvement of the eigenvalue truncation $\hat{P}$ of the weighted estimate $\hat{A}$ compared to its unweighted counterpart $\bar{P}$ with Laplace-distributed edge noise. Each data point is the mean of 50 independent trials. (a) Relative improvement in Frobenius norm as a function of the variance σ_1^2 of the outlier network for several values of the number of networks N, with the number of vertices $n = 200$ fixed. (b) Relative improvement as a function of the number of vertices n for different numbers of networks N, with outlier variance $\sigma_1^2 = 5$ fixed.

We can think of this quantity as a measure of the outlier's influence.

We repeat this experiment for different values of the number of vertices n, the number of networks N and the outlier variance σ_1^2, and average over 50 replications for each setting. Figure 1 summarizes the results for relative improvement measured in Frobenius norm. We also computed the relative improvement in spectral and $(2, \infty)$ matrix norms, with similar results (omitted). Figure 1a shows relative improvement as a function of the outlier variance σ_1^2, for different values of N, with $n = 200$ fixed. Figure 1b shows relative improvement as a function of n while holding the outlier variance σ_1^2 fixed, again for different values of N. Figure 1 suggests two main conclusions. First, the relative improvement is never negative, showing that even when the outlier variance is small, there is no disadvantage to using the weighted average. Second, even a single outlier with larger edge variance can significantly impact the unweighted average. Similar trends to those seen in Figure 1 apply to the error in recovering X.

6.2 Application to neuroimaging data

We briefly investigate how the choice of network average impacts downstream analyses of real data. We use the COBRE data set (Aine et al., 2017), a collection of fMRI scans from 69 healthy patients and 54 schizophrenic patients, for a total of $N = 123$ subjects. Each fMRI scan is processed to obtain a weighted graph on $n = 264$ vertices, in which each vertex represents a brain region, and edge weights capture functional connectivity, as measured by regional averages of voxel-level time series correlations. The data are processed so that the brain regions align across subjects, with the vertices corresponding to regions in the Power parcellation (Power et al., 2011).

In real data, we do not have access to the true low-rank matrix P, if such a matrix exists at all. Thus, to compare weighted and unweighted network averaging on real-world data,

we compare their impact on downstream tasks such as clustering and hypothesis testing. Even for these tasks, the ground truth is typically not known, and thus it is not possible to directly assess which method returns a better answer. Instead, we will check whether the weighted averages yield appreciably different downstream results, and point to the synthetic experiments as evidence that the weighted network average is likely the better choice.

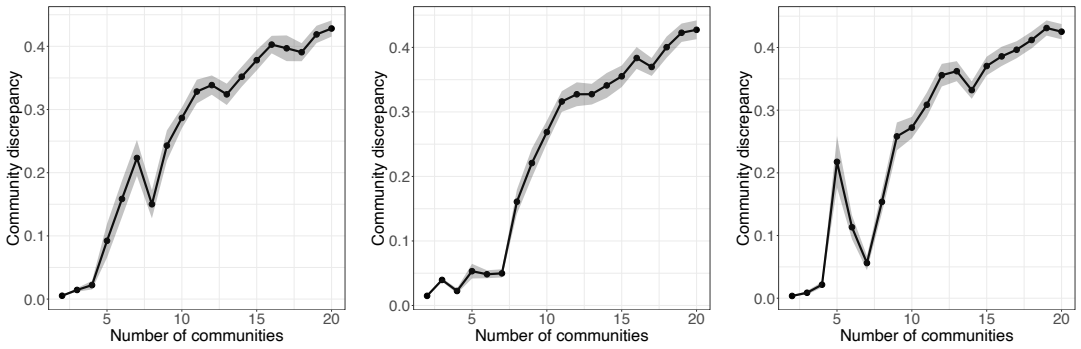
We begin by examining the effect of weighted averaging on estimated community structure. We make the assumption once again that these networks share a low-rank expectation $\mathbb{E}A^{(s)} = P = XX^T$, with $X \in \mathbb{R}^{n \times d}$. We will compare the behavior of clustering applied to the unweighted network mean $\bar{P}$ against the behavior of clustering applied to its weighted counterpart $\hat{P}$. In practice, the model rank d is unknown and must be estimated from the data. While this model selection task is important, it is not the focus of the present work, and thus instead of potentially introducing additional noise from imperfect estimation, we simply compare performance of the two estimators over a range of values of d. For each fixed model rank d, we first construct the estimate $\hat{X}^{(d)} = \mathrm{ASE}(\hat{A}, d)$, and then estimate communities by applying K-means clustering to the n rows of $\hat{X}^{(d)}$. Denote the resulting assignment of vertices to d communities by $\hat{c} \in [K]^n$, and let $\bar{c} \in [K]^n$ denote the clustering obtained by K-means applied to the rows of $\bar{X}^{(d)} = \mathrm{ASE}(\bar{A}, d)$. We measure the difference between these two assignments by the discrepancy

$$\delta(c, c') = n^{-1} \min_{\pi \in S_K} \sum_{i=1}^{n} \mathbb{I}\{c_i \neq \pi(c'_i)\}, \tag{20}$$

where S_K denotes the set of all permutations of the set $[K]$. This discrepancy measures the fraction of vertices that are assigned to different communities by c and c' after accounting for possible community relabeling. The optimization over the set of permutations in (20) can be solved using the Hungarian algorithm (Kuhn, 1955).

Figure 2 shows the discrepancy $\delta(\hat{c}, \bar{c})$ as a function of the number of communities K. For simplicity, we take the number of communities equal to the model rank d, though we note that similar patterns appear when we allow K and d to vary separately. In order to account for the possibility that the healthy and schizophrenic populations display different community structures, the three subplots of Figure 2 show the results of the community estimation experiment just described as applied only to the 69 healthy patients in the data set (subplot a), as applied only to the 54 schizophrenic patients (subplot b) and when pooling the healthy and schizophrenic patients (subplot c). Note that a similar pattern holds in all three of these cases. Each data point in the figure is the mean of 20 independent runs of K-means with random starting conditions, with shaded regions indicating two standard errors. It is clear from the plot that for a wide array of model choices, the weighted and unweighted average networks result in assigning a non-trivial fraction of the vertices to different clusters. Thus switching from unweighted to weighted averaging is likely to have considerable effects on downstream inference tasks pertaining to community structure.

The difference in task performance between these two different network averages persists even for more complicated downstream inference tasks, as we now demonstrate. The Power parcellation (Power et al., 2011) is one of many ways of assigning ROIs (i.e., nodes) to larger functional units of the brain, typically called *functional regions*. The Power parcellation assigns each of the 264 ROIs (i.e., nodes) in the COBRE data set to one of 14 different communities, corresponding to functional regions, with sizes varying between approximately



(a) (b) (c)

Figure 2: Fraction of vertices assigned to different communities when clustering the weighted and unweighted average networks based on the (a) healthy (b) schizophrenic and (c) pooled healthy and schizophrenic patients, as a function of the number of communities used. Each data point the mean of 20 independent trials, with shaded regions indicating two standard errors of the mean (randomness is present in this experiment due to starting conditions of the clustering algorithm). We see that over a broad range of model parameters (i.e., number of communities), the choice to use a weighted or unweighted network average results in different cluster assignments for between ten and forty percent of the vertices, and this pattern persists whether we pool all 123 patients or restrict our analysis to the healthy or schizophrenic patients.

5 and 50 nodes per community. Table 1 summarizes the 14 functional regions and their purported functions. We refer to a pair of functional regions (k, ℓ), for every $k \leq \ell$ as a network cell. Thus, the $K = 14$ communities in the Power parcellation yield 105 cells. For a given parcellation, a problem of scientific interest is to identify which network cells, if any, are different in schizophrenic patients compared to the healthy controls. Such cells likely correspond to locations of functional differences between schizophrenic and healthy brains. For concreteness, consider testing the hypothesis, for each of the 105 possible cells $\{k, \ell\}$ for $1 \leq k \leq \ell \leq 14$, that the average functional connectivity within the cell is the same for the schizophrenic patients and the healthy controls. These hypotheses can be tested using either weighted or unweighted network averages over the healthy and the schizophrenic samples, which we denote $\hat{P}^{(H)}$ and $\hat{P}^{(S)}$, respectively. That is, letting C_k denote the vertices associated with the k-th functional region, we perform a two-sample t-test comparing the healthy sample cell mean $\{\hat{P}^{(H)}_{i,j} : i \in C_k, j \in C_\ell\}$ to the schizophrenic sample cell $\{\hat{P}^{(S)}_{i,j} : i \in C_k, j \in C_\ell\}$, for each pair $k \leq \ell$. Comparisons of this sort, with appropriate multiple testing correction, are common in the neuroimaging literature. Nonetheless, we are not concerned here with whether or not precisely this testing procedure is the most appropriate or most accurate method for assessing differences between the schizophrenic and healthy populations. Rather, we choose this procedure as representative of the kinds of methods typically used for comparing network populations in the literature, and our aim is to assess whether the use of weighted or unweighted averaging leads to operationally different conclusions based on the same data.

The two subplots in Figure 3 show the outcome of such a comparison. Each tile is colored according to the p-value returned by a two-sample t-test comparing the estimated

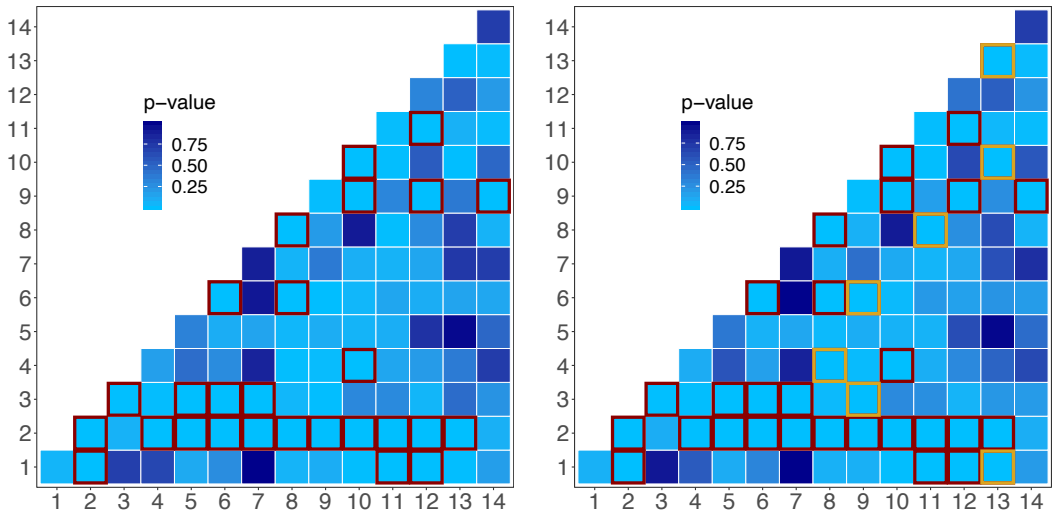
Region	Function	Nodes	Region	Function	Nodes
1	Uncertain	28	8	Visual	31
2	Sensory/somatomotor Hand	30	9	Fronto-parietal Task Control	25
3	Sensory/somatomotor Mouth	5	10	Salience	18
4	Cingulo-opercular Task Control	14	11	Subcortical	13
5	Auditory	13	12	Ventral attention	9
6	Default mode	58	13	Dorsal attention	11
7	Memory retrieval	5	14	Cerebellar	4

Table 1: Brain regions in the Power parcellation (Power et al., 2011), their functions, and the number of nodes within each functional region. The region numbers correspond to those used in Figure 3 below.

connection weights of the schizophrenic and healthy patients within the corresponding cell. Tiles highlighted by colored boxes correspond to cells for which the t-test rejected at the $\alpha = 0.01$ level after correcting for multiple comparisons via the Benjamini-Hochberg procedure. The left-hand subplot in Figure 3 shows the p-values for the unweighted test, while the right-hand subplot shows the same procedure using the weighted estimate instead of the unweighted estimate. We see that after the Benjamini-Hochberg procedure, the weighted average results in more rejections than the unweighted average, and rejects a strict superset of the cells rejected by the unweighted average. Encouragingly, the cells identified as significant by both procedures are fairly well localized, in that a few of the regions account for most of the rejected cells (e.g., region 2 alone is associated with twelve of the twenty seven cells rejected by both methods). This suggests that the differences between the healthy and schizophrenic populations are in fact localized to specific brain regions. We note that the cells rejected by the weighted procedure and accepted by the unweighted procedure are also in keeping with this localization, in that all of the seven additional cells selected by the weighted procedure are incident on at least one of regions 8, 9 or 13 (visual, fronto-parietal task control and dorsal attention, respectively).

The brain regions identified by these two procedures are consistent with existing work on the structural correlates of schizophrenia. Most strikingly, we see exceptionally strong evidence that region 2 differs between schizophrenic and healthy patients. This is in keeping with existing neuroscientific findings suggesting that somatomotor processing is altered in schizophrenic patients (see, e.g., Shinn et al., 2015; Kaufmann et al., 2015; Li et al., 2019; Hummer et al., 2020). As another example, several cells involving the default mode network (region 6) is involved in several of the cells identified by both methods. This region, which is believed to be involved in undirected thought (i.e., mind wandering), has been previously associated with schizophrenia (Bluhm et al., 2007; Whitfield-Gabrieli et al., 2009; Fox et al., 2015). It is interesting to note that this localized structure has emerged without any explicit encoding of such a structure among the cells in the testing procedure itself.

We note that the cell-level tests conducted by our two procedures outlined above are likely to be dependent, owing to the fact that each cell-level test incorporates edge-level information that is likely to be correlated within each network. The Benjamini-Yekutieli procedure, designed to control false discovery rate under such dependence, yields qualitatively similar results to those seen in Figure 3, though in that case, the unweighted procedure rejects a superset of the cells rejected by the weighted procedure. Broadly speaking, then,



(a) (b)

Figure 3: Results of the cell-level significance tests of the unweighted (left) and weighted (right) network averages. Each tile corresponds to a cell, with tiles colored according to p-values. The cells highlighted by red or yellow squares indicate those for which one or both of the weighted and the unweighted procedures rejected the null hypothesis. Cells highlighted in red are those that were rejected by both the unweighted and weighted tests after Benjamini-Hochberg correction at false discovery rate 0.01. Cells highlighted in yellow in the weighted plot (right) correspond to those rejected by the weighted procedure, but not its unweighted counterpart. All cells rejected by the unweighted procedure were also rejected by the weighted procedure. Numbering of the axes corresponds to the brain region numbering given in Table 1

the choice to use a weighted or unweighted network average has nontrivial consequences for downstream inference, in that the procedures identify different sets of cells as being implicated in the schizophrenia. On the other hand, the weighted and unweighted procedures largely agree in the cells that they identify as differing across the two populations. We conjecture that weighted network averaging will generally yield more conservative results, leading to a smaller Type I error, though our experiment just outlined shows that this need not be true uniformly. We expect that these differences will mostly affect cells that are not clear-cut in either direction, and thus the specific problem and dataset at hand will determine how much the results differ. At the same time, the non-clear cut cases are the most likely zone for new discoveries, and thus it is important to understand how different averaging choices affect the downstream analyses. Comparing the results from different averages can also be used as a measure of stability, increasing our confidence in conclusions when they agree (Yu, 2013).

7. Summary and Discussion

We have presented an approach to handling heterogeneity in edge-level noise for estimating shared structure from a collection of networks. Under the setting where edge weights are i.i.d. Gaussian within the same network, we have shown that a weighted network average

with weights proportional to estimated variances of the edges is asymptotically equivalent to the maximum-likelihood estimate in the case where the edge variances are known. We have also presented a class of estimators under weaker conditions on the tails, sub-Gaussian or sub-gamma instead of the Gaussian. While showing theoretically that these weighted estimates strictly improve upon unweighted network averages is not easily done under these weaker assumptions, synthetic experiments bear out the intuition that a weighted network average based on estimated edge variances and/or scale parameters improves upon a naïve unweighted sample mean of networks. Further, experiments on real neuroimaging data showed that the choice between weighted and unweighted network averaging has consequences for downstream inference that cannot be ignored.

A most immediate avenue for future work is to pursue a more thorough analysis of conditions under which weighted network averaging improves appreciably upon unweighted averaging. We are in the process of applying tools from random matrix theory to the multiple networks setting presented here. Further afield, considering heavy-tailed distributions of network edges is also of interest. We also believe the techniques presented in the present paper might be adapted to develop robust estimators in network settings analogous to Huber's ϵ-contamination model (Huber, 1964).

Acknowledgments

The authors acknowledge the support of the National Science Foundation, with KL and AL supported by DMS-1646108, and EL by DMS-1916222. Additional support for KL was provided by the University of Wisconsin-Madison, Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

Appendix A. sub-Gaussian and sub-gamma random variables

For the sake of completeness, we state definitions and a few relevant facts on sub-Gaussian and sub-gamma random variables; these definitions are from Boucheron et al. (2013), which can be consulted for a more thorough treatment.

Definition 16 Let Z be a random variable with $\mathbb{E}Z = 0$ and let $\psi_Z(t) = \log \mathbb{E}e^{tZ}$ denote its cumulant generating function. We say that Z is *sub-Gaussian* with variance parameter $\nu \geq 0$ if for all $t \in \mathbb{R}$, $\psi_Z(t) \leq t^2\nu/2$.

Definition 17 Let Z be a random variable with $\mathbb{E}Z = 0$ and let $\psi_Z(t) = \log \mathbb{E}e^{tZ}$ denote its cumulant generating function. Let $\nu, b \geq 0$. We say that a random variable Z is *sub-gamma* on the right tail with parameter (ν, b) if $\psi_Z(t) \leq \frac{t^2\nu}{2(1-bt)}$ for all $t < 1/b$. Similarly, we say that Z is sub-gamma on the left tail with parameter (ν, b) if $\psi_{-Z}(t) \leq \frac{t^2\nu}{2(1-bt)}$ for all $t < 1/b$. If Z is sub-gamma on both the left and the right tails with parameter (ν, b), then we say that Z is sub-gamma with parameter (ν, b), and write that Z is (ν, b)-sub-gamma.

A basic property of sub-Gaussian random variables is that their sum is sub-Gaussian: if $\{Z_i\}_{i=1}^m$ are independent sub-Gaussian variables with variance parameters ν_i, and $\{\alpha_i\}_{i=1}^m$ are real numbers, then $\sum_{i=1}^m \alpha_i Z_i$ is sub-Gaussian with variance parameter $\sum_{i=1}^m \alpha_i^2 \nu_i$.

Similarly, if $\{Z_i\}_{i=1}^m$ are independent sub-gamma variables with parameters (ν_i, b_i) for all $i \in [m]$, $\sum_{i=1}^m \alpha_i Z_i$ is sub-gamma with parameter $(\sum_{i=1}^m \alpha_i^2 \nu_i, \max_i |\alpha_i| b_i)$.

Some references use the term *sub-exponential* for the tail behavior just defined. We instead reserve this term for the special case obtained by taking $\nu = \lambda^2$, $b = 0$.

Definition 18 A random variable Z with $\mathbb{E}Z = 0$ is called sub-exponential with parameter $\lambda > 0$ if its MGF satisfies $\mathbb{E}\exp\{tZ\} \leq \exp\{t^2\lambda^2/2\}$ whenever $|t| \leq 1/\lambda$.

If a random variable Z is sub-Gaussian with parameter ν, then $Z^2 - \mathbb{E}Z^2$ is sub-exponential with parameter 16ν; see, for example, Lemma 1.12 in the lecture notes by Rigollet and Hütter (2018).

Appendix B. Proofs and Technical Results

In what follows, we prove our main results.

B.1 Proof of Lemma 4 (Concentration in $2, \infty$-norm)

Lemma 4 generalizes and extends results of Lyzinski et al. (2017) and Levin et al. (2017). To begin with, we require two technical results. The first is a modification of Proposition 16 in Lyzinski et al. (2017) to be agnostic to the growth of the spectrum of P instead of assuming a lower-bound on the growth rate of the non-zero eigenvalues of P. The proof is otherwise identical and is omitted. We note that this result is entirely deterministic, but we will apply it below when M is a random matrix with expectation P.

Proposition 19 *Let $M, P \in \mathbb{R}^{n\times n}$ with $P = XX^T$ for some $X \in \mathbb{R}^{n\times d}$. Let $P = U_P S_P U_P^T$ be the rank-d singular value decomposition of P and define $U_M \in \mathbb{R}^{n\times d}, S_M \in \mathbb{R}^{d\times d}$ so that*

$U_M S_M U_M^T$ is the rank-d eigenvalue truncation of M. Let $V_1 D V_2^T$ be the rank-d singular value decomposition of $U_P^T U_M$. Then

$$\|U_P^T U_M - V_1 V_2^T\|_F \leq \frac{d\|M-P\|^2}{\lambda_d^2(P)}.$$

Our second technical result is an adaptation of Lemma 17 in Lyzinski et al. (2017) and Lemma 4 in Levin et al. (2017), again adapted to the case where no growth assumptions on $\lambda_d(P)$ are made. The proof is a straight-foward application of Davis-Kahan style bounds (e.g., Theorem 2 in Yu et al., 2015) and a modification of the argument in Levin et al. (2017), and details are thus omitted.

Proposition 20 *Let $P \in \mathbb{R}^{n\times n}$ with $P = XX^T$ for some $X \in \mathbb{R}^{n\times d}$ and $P = U_P S_P U_P^T$ the rank-d singular value decomposition of P. Let $M \in \mathbb{R}^{n\times n}$ be random with rank-d eigenvalue truncation $U_M S_M U_M^T$, where $U_M \in \mathbb{R}^{n\times d}, S_M \in \mathbb{R}^{d\times d}$. Then*

$$\|U_M - U_P U_P^T U_M\|_F \leq \frac{C\sqrt{d}\|M-P\|}{\lambda_d(P)}. \tag{21}$$

Further, suppose that there exists a constant $c_0 \in [0,1)$ such that with probability at least p_0, $\|M-P\| \leq c_0\lambda_d(P)$. Let $V_1 D V_2^T$ be the rank-d singular value decomposition of $U_P^T U_M$, and define $V = V_1 V_2^T$. Then with probability at least p_0,

$$\|VS_M - S_P V\|_F \leq \frac{Cd\|M-P\|^2\kappa(P)}{\lambda_d(P)} + \|U_P^T(M-P)U_P\|_F \tag{22}$$

$$\|VS_M^{1/2} - S_P^{1/2}V\|_F \leq \frac{C\|VS_M - S_P V\|_F}{\lambda_d^{1/2}(P)} \quad \text{and} \tag{23}$$

$$\|VS_M^{-1/2} - S_P^{-1/2}V\|_F \leq \frac{C\|VS_M - S_P V\|_F}{\lambda_d^{3/2}(P)}. \tag{24}$$

We are now equipped to prove Lemma 4.

Proof of Lemma 4 Let $V \in \mathbb{R}^{d\times d}$ be the orthogonal matrix defined above in Proposition 20 and define the following three matrices:

$$\begin{aligned} R_1 &= U_P U_P^T U_M - U_P V, \\ R_2 &= VS_M^{1/2} - S_P^{1/2}V, \\ R_3 &= U_M - U_P U_P^T U_M + R_1 = U_M - U_P V. \end{aligned}$$

Adding and subtracting appropriate quantities,

$$\begin{aligned} U_M S_M^{1/2} - U_P S_P^{1/2} V = {} & (M-P)U_P S_P^{-1/2}V + (M-P)U_P(VS_M^{-1/2} - S_P^{-1/2}V) \\ & + U_P U_P^T(M-P)U_P V S_M^{-1/2} + R_1 S_M^{1/2} + U_P R_2 \\ & + (I - U_P U_P^T)(M-P)R_3 S_M^{-1/2}. \end{aligned}$$

Applying the triangle inequality and the fact that the Frobenius norm is an upper bound on the $(2,\infty)$-norm, we have

$$\begin{aligned}\|U_M S_M^{1/2} - U_P S_P^{1/2} V\|_{2,\infty} &\leq \|(M-P)U_P S_P^{-1/2} V\|_{2,\infty} \\ &+ \|(M-P)U_P(VS_M^{-1/2} - S_P^{-1/2}V)\|_F + \|U_P U_P^T (M-P)U_P V S_M^{-1/2}\|_F \\ &+ \|(I - U_P U_P^T)(M-P)R_3 S_M^{-1/2}\|_F + \|R_1 S_M^{1/2}\|_F + \|U_P R_2\|_F.\end{aligned} \tag{25}$$

We will bound each of these summands in turn. Firstly, by definition of the spectral and $(2,\infty)$ norms,

$$\|(M-P)U_P S_P^{-1/2} V\|_{2,\infty} \leq \|(M-P)U_P\|_{2,\infty}\|S_P^{-1/2}\| \leq \frac{\|(M-P)U_P\|_{2,\infty}}{\lambda_d^{1/2}(P)}. \tag{26}$$

The second term on the right-hand side of (25) is bounded as

$$\begin{aligned}\|(M-P)U_P(VS_M^{-1/2} - S_P^{-1/2}V)\|_F &\leq \|M-P\|\|U_P\|\|VS_M^{-1/2} - S_P^{-1/2}V\|_F \\ &= \|M-P\|\|VS_M^{-1/2} - S_P^{-1/2}V\|_F,\end{aligned}$$

from which Proposition 20 implies that with probability at least p_0,

$$\|(M-P)U_P(VS_M^{-1/2} - S_P^{-1/2}V)\|_F \leq \frac{C\|M-P\|\|VS_M - S_P V\|_F}{\lambda_d^{3/2}(P)}. \tag{27}$$

By the assumption in Equation (4), with probability at least p_0,

$$\|S_M^{-1/2}\| \leq (\lambda_d(P) - \|M-P\|)^{-1/2} \leq C\lambda_d^{-1/2}(P). \tag{28}$$

Thus, the third term on the right-hand side of (25) satisfies

$$\begin{aligned}\|U_P U_P^T (M-P)U_P V S_M^{-1/2}\|_F &\leq \|U_P\|\|U_P^T(M-P)U_P\|_F\|V\|\|S_M^{-1/2}\| \\ &\leq \|U_P^T(M-P)U_P\|_F\|S_M^{-1/2}\| \\ &\leq \frac{\|U_P^T(M-P)U_P\|_F}{\sqrt{\lambda_d(P) - \|M-P\|}} \leq \frac{C\|U_P^T(M-P)U_P\|_F}{\lambda_d^{1/2}(P)},\end{aligned} \tag{29}$$

where the penultimate inequality follows from Equation (28) and the last inequality holds with probability at least p_0 by the assumption in Equation (4).

Considering the fourth term on the right-hand side of (25), recall that $R_3 = U_M - U_P V$, so that by adding and subtracting appropriate quantities we have

$$\begin{aligned}&(I - U_P U_P^T)(M-P)R_3 S_M^{-1/2} \\ &\quad= (I - U_P U_P^T)(M-P)(U_M - U_P U_P^T U_M)S_M^{-1/2} \\ &\qquad+ (I - U_P U_P^T)(M-P)((I - U_P U_P^T)(M-P) - U_P V)S_M^{-1/2}.\end{aligned} \tag{30}$$

The former of these quantities is bounded as

$$\begin{aligned}\|(I - U_P U_P^T)(M - P)(U_M - U_P U_P^T U_M) S_M^{-1/2}\|_F \\ \leq \|I - U_P U_P^T\| \|M - P\| \|U_M - U_P U_P^T U_M\|_F \|S_M^{-1/2}\| \leq \frac{C\sqrt{d}\|M - P\|^2}{\lambda_d^{3/2}(P)}\end{aligned}$$

where we have used Equations (21) and (28) to obtain the second inequality. The second term on the right-hand side of (30) is bounded by Proposition 19 and Equation (28) as

$$\begin{aligned}\|(I - U_P U_P^T)(M - P)((I - U_P U_P^T)(M - P) - U_P V) S_M^{-1/2}\|_F \\ \leq \|I - U_P U_P^T\| \|M - P\| \|U_P\| \|U_P^T U_M - V\|_F \|S_M^{-1/2}\| \leq \frac{Cd\|M - P\|^3}{\lambda_d^{5/2}(P)},\end{aligned}$$

Thus, combining the above two displays with Equation (30),

$$\begin{aligned}\|(I - U_P U_P^T)(M - P) R_3 S_M^{-1/2}\|_F \\ \leq \frac{C\sqrt{d}\|M - P\|^2(\lambda_d(P) + \sqrt{d}\|M - P\|)}{\lambda_d^{5/2}(P)} \leq \frac{Cd\|M - P\|^2}{\lambda_d^{3/2}(P)},\end{aligned} \tag{31}$$

where the last inequality holds with probability at least p_0 by the assumption in Equation (4).

Finally, we bound the last two summands on the right-hand side of (25). Recall that $R_1 = U_P U_P^T U_M - U_P V = U_P(U_P^T U_M - V)$, Proposition 19 yields

$$\|R_1\|_F \leq \|U_P\| \|U_P^T U_M - V\|_F \leq \frac{Cd\|M - P\|^2}{\lambda_d^2(P)},$$

whence by the assumption in Equation (4),

$$\|R_1 S_M^{1/2}\|_F \leq \|R_1\|_F \|S_M^{1/2}\| \leq \frac{Cd\|M - P\|^2 \lambda_1^{1/2}(P)}{\lambda_d^2(P)}. \tag{32}$$

Recalling $R_2 = V S_M^{1/2} - S_P^{1/2} V$, Proposition 20 implies that

$$\|R_2\|_F \leq \frac{C\|V S_M - S_P V\|_F}{\lambda_d^{1/2}(P)}. \tag{33}$$

Applying this bound along with Equations (26), (27), (29), (31) and (32) to Equation (25), collecting terms and making use of the assumption in Equation (4) once more,

$$\begin{aligned}\|U_M S_M^{1/2} - U_P S_P^{1/2} V\|_{2,\infty} \leq \frac{\|(M - P)U_P\|_{2,\infty}}{\lambda_d^{1/2}(P)} + \frac{C\|V S_M - S_P V\|_F}{\lambda_d^{1/2}(P)} \\ + \frac{C\|U_P^T(M - P)U_P\|_F}{\lambda_d^{1/2}(P)} + \frac{Cd\|M - P\|^2}{\lambda_d^{3/2}(P)}\left(1 + \kappa^{1/2}(P)\right).\end{aligned}$$

Applying Proposition 20 to $\|VS_M - S_P V\|_F$,

$$\|U_M S_M^{1/2} - U_P S_P^{1/2} V\|_{2,\infty} \leq \frac{\|(M-P)U_P\|_{2,\infty}}{\lambda_d^{1/2}(P)} + \frac{C\|U_P^T(M-P)U_P\|_F}{\lambda_d^{1/2}(P)} + \frac{Cd\|M-P\|^2}{\lambda_d^{3/2}(P)}\left(1+\kappa^{1/2}(P)+\kappa(P)\right).$$

Noting that $\kappa(P) \geq 1$ completes the proof. ■

B.2 Proof of Theorem 6 (Concentration under sub-gamma assumptions)

Using the results presented above, we are now able to prove Lemma 5 and Theorem 6. Our proof of Theorem 6 relies on applying Lemma 4 to the case of $M = \tilde{A}$. This in turn requires a bound on the spectral norm error that is central to Lemma 4. This spectral norm error is controlled by Lemma 5, the proof of which relies on the following matrix Bernstein bound.

Theorem 21 (Tropp (2012) Theorem 6.2) *Let $\{Z_k\}$ be a finite sequence of independent, random, self-adjoint n-by-n matrices each satisfying $\mathbb{E}Z_k = 0$ and $\mathbb{E}Z_k^p \preceq p!R^{p-2}M_k^2/2$ for $p = 2, 3, \ldots$, where $R \in \mathbb{R}$ and $\{M_k\} \subset \mathbb{R}^{n\times n}$ are deterministic matrices and $\preceq$ denotes the semidefinite ordering, Define $\eta^2 = \|\sum_k M_k^2\|$. Then for all $t \geq 0$,*

$$\mathbb{P}\left[\left\|\sum_k Z_k\right\| \geq t\right] \leq n\exp\left\{\frac{-t^2}{2\eta^2 + 2Rt}\right\}.$$

Proof of Lemma 5 To apply Theorem 21, we first decompose $\sum_{s=1}^N w_s A^{(s)} - P$ into a sum of independent zero-mean symmetric matrices. Letting $\{e_1, e_2, \ldots, e_n\}$ denote the standard basis vectors, define, for all $1 \leq i \leq j \leq n$ the n-by-n matrices

$$\mathbf{B}_{i,j} = \begin{cases} e_i e_j^T + e_j e_i^T & \text{if } 1 \leq i < j \leq n, \\ e_i e_i^T & \text{if } i = j, \end{cases} \tag{34}$$

and the n-by-n random matrices $\mathbf{Z}_{s,i,j} = (w_s A^{(s)} - P)_{i,j}\mathbf{B}_{i,j}$ for all $1 \leq i \leq j \leq n$ and all $s \in [N]$. We bold these matrices in this proof to remind the reader that indexing by s, i, j is specifying one of $O(Nn^2)$ matrices, rather than a matrix entry. The set $\{\mathbf{Z}_{s,i,j} : s \in [N], 1 \leq i \leq j \leq n\}$ sum to $\sum_s w_s A^{(s)} - P$, and satisfy the symmetry, independence and unbiasedness assumptions required of Theorem 21. By the assumption that $(A^{(s)} - P)_{i,j}$ is $(\nu_{s,i,j}, b_{s,i,j})$-sub-gamma, we have that $w_s(A^{(s)} - P)_{i,j}$ is $(w_s^2\nu_{s,i,j}, w_s b_{s,i,j})$-sub-gamma. A moment-bounding argument similar to that in Theorem 2.3 of Boucheron et al. (2013)

yields that

$$\begin{aligned}
\mathbb{E}|w_s(A^{(s)}-P)_{i,j}|^p &= \int_0^\infty pu^{p-1}\mathbb{P}[|w_s(A^{(s)}-P)|_{i,j}>u]du \\
&\le 2p\int_0^\infty \left(\sqrt{2w_s^2\nu_{s,i,j}t}+w_sb_{s,i,j}t\right)^{p-1}\frac{e^{-t}(w_s\sqrt{2\nu_{s,i,j}t}+2w_sb_{s,i,j}t)}{2t}dt \\
&\le 2^{p-1}pw_s^p\int_0^\infty \left[(2\nu_{s,i,j})^{p/2}t^{p/2-1}+(2b_{s,i,j})^pt^{p-1}\right]e^{-t}dt \\
&= 2^{p-1}pw_s^p[(2\nu_{s,i,j})^{p/2}\Gamma(p/2)+(2b_{s,i,j})^p\Gamma(p)] \\
&\le 2^pp!\,w_s^p(\sqrt{2\nu_{s,i,j}}+2b_{s,i,j})^p.
\end{aligned}$$

Thus, defining $\eta_{s,i,j}=2(\sqrt{2\nu_{s,i,j}}+2b_{s,i,j})$ and $R=\max_{s,i,j}\eta_{s,i,j}$, we have

$$\mathbb{E}\mathbf{Z}_{s,i,j}^p=\mathbb{E}\left[w_s(A^{(s)}-P)_{i,j}\right]^p\mathbf{B}_{i,j}^p\preceq p!w_s^p\eta_{s,i,j}^p\mathbf{B}_{i,j}^p\preceq\frac{p!}{2}R^{p-2}(\sqrt{2}w_s\eta_{s,i,j}\mathbf{B}_{i,j})^2,$$

since $\mathbf{B}_{i,j}^p=\mathbf{B}_{i,j}^{2-\lfloor p/2\rfloor}\preceq\mathbf{B}_{i,j}^2$. Taking the bounding matrices M_k in Theorem 21 to be $\{\sqrt{2}w_s\eta_{s,i,j}\mathbf{B}_{i,j}:1\le s\le N,1\le i\le j\le n\}$, η in Theorem 21 becomes

$$\begin{aligned}
\eta^2 &= \left\|\sum_{s=1}^N w_s^2\sum_{1\le i\le j\le n}2\eta_{s,i,j}^2\mathbf{B}_{i,j}^2\right\| \\
&= 2\left\|\sum_{s=1}^N\sum_{i=1}^n w_s^2\eta_{s,i,i}^2e_ie_i^T+\sum_{s=1}^N\sum_{1\le i<j\le n}w_s^2\eta_{s,i,j}^2(e_ie_i^T+e_je_j^T)\right\| \\
&= 2\max_i\sum_{s=1}^N\sum_{j=1}^n w_s^2\eta_{s,i,j}^2.
\end{aligned}$$

Applying Theorem 21, we have

$$\mathbb{P}\left[\left\|\sum_{s=1}^N\sum_{1\le i\le j\le n}\mathbf{Z}_{s,i,j}\right\|\ge t\right]\le n\exp\left\{\frac{-t^2/2}{\eta^2+Rt}\right\}.$$

Taking $t=3(\sqrt{2\eta^2}+3R)\log n$ is enough to ensure that $t^2\ge 6(\eta^2+Rt)\log n$, so that

$$\mathbb{P}\left[\left\|\sum_{s=1}^N\sum_{1\le i\le j\le n}\mathbf{Z}_{s,i,j}\right\|\ge 3(\sqrt{2\eta^2}+3R)\log n\right]\le n^{-2}.$$

Note that we have

$$R=\max_{i,j,s}\eta_{s,i,j}w_s=\sqrt{\max_{i,j,s}(\eta_{s,i,j}w_s)^2}\le\sqrt{\max_{i\in[n]}\sum_{j=1}^n\sum_{s=1}^N(\eta_{s,i,j}w_s)^2}=\sqrt{\eta^2/2},$$

so our upper bound can be further simplified to

$$\left\| \sum_{s=1}^{N} w_s A^{(s)} - P \right\| \leq \frac{15\sqrt{2\eta^2}}{2} \log n,$$

which completes the proof. ■

Recall that our aim in proving Theorem 6 is to establish a bound on the $(2,\infty)$-norm of $\tilde{X} - XV$, for a suitably-chosen orthogonal matrix $V \in \mathbb{R}^{d\times d}$, where $\tilde{A} = \sum_{s=1}^{N} w_s A^{(s)}$ and $\tilde{X} = \text{ASE}(\tilde{A}, d)$. By Lemma 4, in order to bound this norm, it suffices to control $\|\tilde{A} - P\|$, $\|U_P^T(\tilde{A} - P)U_P\|_F$, $\|(\tilde{A} - P)U_P\|_{2,\infty}$ and the spectrum of P. Lemma 5 controls the first of these terms, while Propositions 22 and 23 control the Frobenius and $(2,\infty)$-norms, respectively.

Proposition 22 *Let $\{A^{(s)}\}_{s=1}^{N}$ be independent and for all $s = 1, 2, \ldots, N$, $\{(A^{(s)} - P)_{i,j} : 1 \leq i \leq j \leq n\}$ are independent and $(A^{(s)} - P)_{i,j}$ is $(\nu_{s,i,j}, b_{s,i,j})$-sub-gamma. Assume that $P = \mathbb{E}A^{(s)}$ is rank d with $d = O(n)$, and let $U_{j,k}$ denote the (j,k)-element of the matrix $U_P \in \mathbb{R}^{n\times d}$ of eigenvectors of P with non-zero eigenvalues. Define for all $k, \ell \in [d]$ the quantities*

$$\bar{\nu}_{k,\ell} = \sum_{s=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n} w_s^2 U_{i,k}^2 U_{j,\ell}^2 \nu_{s,i,j}, \qquad \bar{b}_{k,\ell} = \max_{s\in[N], i,j\in[n]} w_s |U_{i,k}U_{j,\ell}| b_{s,i,j}.$$

Then for any fixed collection of weights $\{w_s\}_{s=1}^{N}$ with $w_s \geq 0$ and $\sum_{s=1}^{N} w_s = 1$,

$$\left\| U_P^T \left(\sum_{s=1}^{N} w_s A^{(s)} - P \right) U_P \right\|_F \leq 2 \left(\sum_{k=1}^{d}\sum_{\ell=1}^{d} (\sqrt{2\bar{\nu}_{k,\ell}} + 2\bar{b}_{k,\ell})^2 \right)^{1/2} \log n$$

Further, when the top eigenvectors of P delocalize so that $|U_{i,k}| \leq Cn^{-1/2}$ for all $i \in [n], k \in [d]$, it holds with probability at least $1 - Cn^{-2}$ that

$$\left\| U_P^T \left(\sum_{s=1}^{N} w_s A^{(s)} - P \right) U_P \right\|_F \leq \frac{Cd}{n} \left(\sum_{s=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n} w_s^2 \nu_{s,i,j} + 2 \max_{s\in[N],i,j\in[n]} w_s^2 b_{s,i,j}^2 \right)^{1/2} \log n.$$

Proof Fix $k, \ell \in [d]$ and consider $[U_P^T(\sum_{s=1}^{N} w_s A^{(s)} - P)U_P]_{k,\ell}$. Then

$$\left[U_P^T \left(\sum_{s=1}^{N} w_s A^{(s)} - P \right) U_P \right]_{k,\ell} = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{s=1}^{N} w_s (A^{(s)} - P)_{i,j} U_{i,k} U_{j,\ell}, \tag{35}$$

is a sum of independent mean-zero random variables. Since $(A^{(s)} - P)_{i,j}$ is $(\nu_{s,i,j}, b_{s,i,j})$-sub-gamma, $w_s(A^{(s)} - P)_{i,j}U_{i,k}U_{j,\ell}$ is $(w_s^2 U_{i,k}^2 U_{j,\ell}^2 \nu_{s,i,j}, w_s|U_{i,k}U_{j,\ell}|b_{s,i,j})$-sub-gamma Applying a standard Bernstein inequality to the quantity in Equation (35) thus yields that

$$\mathbb{P}\left[|U_P^T(A - P)U_P|_{k,\ell} > t \right] \leq 2\exp\left\{ -\frac{\bar{\nu}_{k,\ell}}{\bar{b}_{k,\ell}^2} \left(1 + \frac{\bar{b}_{k,\ell} t}{\bar{\nu}_{k,\ell}} - \sqrt{1 + \frac{2\bar{b}_{k,\ell} t}{\bar{\nu}_{k,\ell}}} \right) \right\},$$

where

$$\bar{\nu}_{k,\ell} = \sum_{s=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n} w_s^2 U_{i,k}^2 U_{j,\ell}^2 \nu_{s,i,j}, \qquad \bar{b}_{k,\ell} = \max_{s\in[N],i,j\in[n]} w_s |U_{i,k}U_{j,\ell}| b_{s,i,j}.$$

Taking $t = 2(\sqrt{2\bar{\nu}_{k,\ell}} + 2\bar{b}_{k,\ell})\log n$ is enough to ensure that $|U_P^T(A-P)U_P|_{k,\ell} > t$ with probability at most $2n^{-4}$. A union bound over all $k, \ell \in [d]$ yields

$$\|U_P^T(A-P)U_P\|_F^2 \leq 4\sum_{k=1}^{d}\sum_{\ell=1}^{d}(\sqrt{2\bar{\nu}_{k,\ell}} + 2\bar{b}_{k,\ell})^2 \log^2 n, \tag{36}$$

with probability at least $1 - 2d^2n^{-4} \geq 1 - Cn^{-2}$, owing to our assumption that $d = O(n)$. Taking square roots in Equation (36) yields the first claim.

When the eigenvectors of P delocalize, we have

$$\bar{\nu}_{k,\ell} = \sum_{s=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n} w_s^2 U_{i,k}^2 U_{j,\ell}^2 \nu_{s,i,j} \leq \frac{C^2}{n^2}\sum_{s=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n} w_s^2 \nu_{s,i,j},$$

and

$$\bar{b}_{k,\ell} = \max_{s\in[N],\ i,j\in[n]} w_s |U_{i,k}U_{j,\ell}| b_{s,i,j} \leq \frac{C \max_{s\in[N],\ i,j\in[n]} w_s b_{s,i,j}}{n},$$

whence we have

$$\begin{aligned}
\|U_P^T(A-P)U_P\|_F^2 &\leq C^2 \sum_{k=1}^{d}\sum_{\ell=1}^{d} n^{-2}\left(\sqrt{2\sum_{s=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n} w_s^2\nu_{s,i,j}} + 2\max_{s,i,j} w_s b_{s,i,j}\right)^2 \log^2 n \\
&\leq C^2\left(4d^2n^{-2}\sum_{s,i,j} w_s^2\nu_{s,i,j} + 8d^2n^{-2}\max_{s,i,j} w_s^2 b_{s,i,j}^2\right)\log^2 n \\
&= \frac{d^2C^2}{n^2}\left(\sum_{s,i,j} w_s^2\nu_{s,i,j} + 2\max_{s,i,j} w_s^2 b_{s,i,j}^2\right)\log^2 n,
\end{aligned}$$

and taking square roots completes the proof. ∎

Proposition 23 *Let $\{A^{(s)}\}_{s=1}^N$ be independent and for all $s = 1, 2, \ldots, N$, $\{(A^{(s)} - P)_{i,j} : 1 \leq i \leq j \leq n\}$ are independent and $(A^{(s)} - P)_{i,j}$ is $(\nu_{s,i,j}, b_{s,i,j})$-sub-gamma. Assume that $P = \mathbb{E}A^{(s)}$ is rank d with $d = O(n)$, and let $U_{j,k}$ denote the (j,k)-element of the matrix $U_P \in \mathbb{R}^{n\times d}$ of eigenvectors of P with non-zero eigenvalues. Define for all $i \in [n]$ and $k \in [d]$ the quantities*

$$\nu_{i,k}^* = \sum_{s=1}^{N}\sum_{j=1}^{n} w_s^2 U_{j,k}^2 \nu_{s,i,j}, \qquad b_{i,k}^* = \max_{s\in[N],j\in[n]} w_s |U_{j,k}| b_{s,i,j}.$$

Then with probability at least $1 - Cn^{-2}$,

$$\|(\tilde{A} - P)U_P\|_{2,\infty} \leq 2 \max_{i \in [n]} \left(\sum_{k=1}^{d} \left(\sqrt{2\nu^*_{i,k}} + 2b^*_{i,k} \right)^2 \right)^{1/2} \log n$$

If, in addition, the top eigenvectors of P *delocalize,*

$$\|(A - P)U_P\|_{2,\infty} \leq \frac{Cd^{1/2}}{\sqrt{n}} \max_{i \in [n]} \left(\sum_{s=1}^{N} \sum_{j=1}^{n} w_s^2 \nu_{s,i,j} + \max_{s \in [N], j \in [n]} w_s^2 b^2_{s,i,j} \right)^{1/2} \log n$$

also with probability at least $1 - Cn^{-2}$,

Proof Fixing $i \in [n]$, the vector formed by i-th row of $(\sum_{s=1}^{N} w_s A^{(s)} - P)U_P$, which we denote by $[(\sum_{s=1}^{N} w_s A^{(s)} - P)U_P]_i \in \mathbb{R}^d$, satisfies

$$\left\| \left[\left(\sum_{s=1}^{N} w_s A^{(s)} - P \right) U_P \right]_i \right\|^2 = \sum_{k=1}^{d} \left(\sum_{s=1}^{N} \sum_{j=1}^{n} w_s (A^{(s)} - P)_{i,j} U_{j,k} \right)^2 .$$

Fix some $k \in [d]$, and consider $Z_{i,k} = \sum_{j=1}^{n} \sum_{s=1}^{N} w_s (A^{(s)} - P)_{i,j} U_{j,k}$, which is a sum of nN independent zero-mean random variables. By our assumptions, $w_s U_{j,k} (A^{(s)} - P)_{i,j}$ is $(w_s^2 U^2_{j,k} \nu_{s,i,j}, w_s |U_{j,k}| b_{s,i,j})$-sub-gamma and $Z_{i,k}$ is $(\nu^*_{i,k}, b^*_{i,k})$-sub-gamma. Writing $\nu^*_{i,k} = \sum_{s=1}^{N} \sum_{j=1}^{n} w_s^2 U^2_{j,k} \nu_{s,i,j}$, A standard Bernstein inequality yields

$$\mathbb{P}\left[|Z_{i,k}| > t \right] \leq 2 \exp \left\{ - \left(\frac{t}{b^*_{i,k}} + \frac{\nu^*_{i,k}}{(b^*_{i,k})^2} - \frac{\nu^*_{i,k}}{(b^*_{i,k})^2} \sqrt{1 + 2t b^*_{i,k} / \nu^*_{i,k}} \right) \right\} .$$

Choosing $t = 2 \left(\sqrt{2\nu^*_{i,k}} + 2b^*_{i,k} \right) \log n$ suffices to ensure that the event

$$\left\{ |Z_{i,k}| > 2 \left(\sqrt{2\nu^*_{i,k}} + 2b^*_{i,k} \right) \log n \right\}$$

occurs with probability at most $2n^{-4}$. A union bound over all d entries of the vector $[(\sum_{s=1}^{N} w_s A^{(s)} - P)U_P]_i$ yields that with probability at least $1 - 2dn^{-4}$,

$$\left\| \left[\left(\sum_{s=1}^{N} w_s A^{(s)} - P \right) U_P \right]_i \right\|^2 \leq 4 \sum_{k=1}^{d} \left(\sqrt{2\nu^*_{i,k}} + 2b^*_{i,k} \right)^2 \log^2 n,$$

and another union bound over all $i \in [n]$ yields that with probability at least $1 - 2dn^{-3}$,

$$\left\| \left(\sum_{s=1}^{N} w_s A^{(s)} - P \right) U_P \right\|_{2,\infty} \leq 2 \max_{i \in [n]} \left(\sum_{k=1}^{d} \left(\sqrt{2\nu^*_{i,k}} + 2b^*_{i,k} \right)^2 \right)^{1/2} \log n. \tag{37}$$

Since $d = O(n)$ by assumption, this event holds with probability at least $1 - Cn^{-2}$, as we set out to prove.

If the top eigenvectors of P delocalize, then we have

$$\nu^*_{i,k} \leq \frac{C^2}{n} \sum_{s=1}^{N} \sum_{j=1}^{n} w_s^2 \nu_{s,i,j}, \qquad b^*_{i,k} \leq \frac{C \max_{s\in[N],j\in[n]} w_s b_{s,i,j}}{\sqrt{n}},$$

and we can strengthen Equation (37) to

$$\|(A-P)U_P\|_{2,\infty} \leq \frac{Cd^{1/2}}{\sqrt{n}} \max_{i\in[n]} \left(\sum_{s=1}^{N} \sum_{j=1}^{n} w_s^2 \nu_{s,i,j} + \max_{s\in[N],j\in[n]} w_s^2 b_{s,i,j}^2 \right)^{1/2},$$

also with probability at least $1 - Cn^{-2}$, again since $d = O(n)$ by assumption. ■

We are now ready to prove Theorem 6.

Proof of Theorem 6 By Lemma 5, with probability at least $1 - Cn^{-2}$,

$$\|\tilde{A} - P\| \leq 15 \sqrt{n \sum_{s=1}^{N} w_s^2 (\sqrt{2\nu_s} + b_s)^2 \log n} \leq C \sqrt{n \sum_{s=1}^{N} w_s^2 (\nu_s + b_s^2) \log n}. \tag{38}$$

The assumption in Equation (5) ensures that for suitably large n, $\|\tilde{A} - P\| \leq \lambda_d(P)/2$, so that Equation (4) of Lemma 4 is satisfied with high probability. The result will now follow immediately from Lemma 4, provided we can bound both $\|U_P^T(\tilde{A} - P)U_P\|_F$ and $\|(\tilde{A} - P)U_P\|_{2,\infty}$, also with high probability.

Proposition 22 implies that with probability at least $1 - Cn^{-2}$,

$$\|U_P^T(\tilde{A} - P)U_P\|_F \leq Cd \left(\sum_{s=1}^{N} w_s^2 (\nu_s + b_s^2) \right)^{1/2} \log n,$$

and Proposition 23 implies that with probability at least $1 - Cn^{-2}$,

$$\|(\tilde{A} - P)U_P\|_{2,\infty} \leq C\sqrt{d} \left(\sum_{s=1}^{N} w_s^2 (\nu_s + b_s^2) \right)^{1/2} \log n.$$

Plugging the above two bounds, along with Equation (38), into Lemma 4, since $d \geq 1$, we have

$$\begin{aligned} &\|U_{\tilde{A}} S_{\tilde{A}}^{1/2} - U_P S_P^{1/2} W\|_{2,\infty} \\ &\qquad \leq \frac{Cd}{\lambda_d^{1/2}(P)} \left(\sum_{s=1}^{N} w_s^2 (\nu_s + b_s^2) \right)^{1/2} \log n + \frac{Cd\kappa(P)n}{\lambda_d^{3/2}(P)} \left(\sum_{s=1}^{N} w_s^2 (\nu_s + b_s^2) \right) \log^2 n, \end{aligned}$$

which completes the proof. ■

B.3 Proof of Theorem 8 (Minimax estimation in spectral norm)

In this section, we establish Theorem 8, showing that the weighted network average $\sum_s \mathring{w}_s A^{(s)}$ recovers P at the minimax rate when the edge errors in each network are drawn from a network-specific normal distribution. Our main tool is Theorem 2.7 in Tsybakov (2009), which we restate here, adapted to the present setting.

Theorem 24 (Tsybakov (2009) Theorem 2.7) *Let Θ be a parameter space endowed with a distance $\delta(\cdot,\cdot)$ and containing $M \geq 1$ elements $\theta_0, \theta_1, \ldots, \theta_M$, such that, for $s > 0$,*

i For all $0 \leq i < j \leq M$, $\delta(\theta_i, \theta_j) \geq 2s$

ii Writing P_j for the distribution induced by θ_j, $P_j \ll P_0$ for all $j = 1, 2, \ldots, M$, and

$$\frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(P_j, P_0) \leq \alpha \log M$$

for some $\alpha \in (0, 1/8)$.

Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}\delta(\hat{\theta}, \theta) \geq c_\alpha s,$$

where the infimum is over all estimators and c_α is a constant depending only on α.

Proof of Theorem 8 We will apply Theorem 24 to a suitably-chosen collection of symmetric n-by-n matrices, under the distance

$$\delta(P^{(1)}, P^{(2)}) = \|P^{(1)} - P^{(2)}\|,$$

where $P^{(1)}, P^{(2)} \in \mathbb{R}^{n\times n}$. Letting $\mathcal{S}_n$ denote the set of symmetric n-by-n matrices, any $P \in \mathcal{S}_n$ induces a distribution on $(\mathbb{R}^{n\times n})^N$ given by

$$A_{ij}^{(s)} \overset{\text{ind.}}{\sim} \mathcal{N}(P_{ij}, \rho_s), \quad 1 \leq i \leq j \leq n, s = 1, 2, \ldots, N,$$

Making use of the independence structure, the Kullback-Leibler divergence between the distributions induced by two symmetric matrices $P^{(1)}, P^{(2)} \in \mathcal{S}_n$ is given by

$$\mathrm{KL}(P^{(1)}, P^{(2)}) = \sum_{1\leq i\leq j\leq n} \sum_{s=1}^{N} \mathrm{KL}\left(\mathcal{N}(P_{ij}^{(1)}, \rho_s), \mathcal{N}(P_{ij}^{(2)}, \rho_s)\right). \tag{39}$$

The KL-divergence betwen two normal distributions Q_1, Q_2 with means μ_1 and μ_2 and common variance $\sigma^2 > 0$ is given by (letting $Z_1 \sim Q_1$)

$$\mathrm{KL}\left(Q_1, Q_2\right) = \mathbb{E}Q_1(Z_1) \log \frac{Q_1(Z_1)}{Q_2(Z_1)} = \frac{(\mu_1 - \mu_2)^2}{\sigma^2}.$$

Applying this identity to Equation (39),

$$\mathrm{KL}(P^{(1)}, P^{(2)}) = \sum_{1\leq i\leq j\leq n} \sum_{s=1}^{N} \frac{\left(P_{ij}^{(1)} - P_{ij}^{(2)}\right)^2}{\rho_s} \leq \|P^{(1)} - P^{(2)}\|_F^2 \sum_{s=1}^{N} \rho_s^{-1}. \tag{40}$$

To each $\xi \in \{0,1\}^n$, associate a symmetric matrix $M^{(\xi)} \in \mathbb{R}^{n\times n}$ given by

$$M_{ij}^{(\xi)} = \begin{cases} \xi_j & \text{if } 1 = i \le j \le n, \xi_j = 1 \\ 0 & \text{if } 1 < i \le j \le n \\ M_{ji}^{(\xi)} & \text{if } 1 \le j < i \le n \end{cases}$$

Let $\mathcal{M}_{n,d} \subseteq \{0,1\}^n$ be a collection such that for all $\xi, \xi' \in \mathcal{M}_{n,d}$ with $\xi \neq \xi'$, the Hamming distance $d_H(\xi, \xi')$ between ξ and ξ' is at least d, for some $d \le n$ to be specified below. Letting $\eta > 0$, for each $\xi \in \mathcal{M}_{n,d}$, construct

$$P^{(\xi)} = \frac{\eta}{\sqrt{d}} M^{(\xi)}.$$

Recall that for a matrix $P \in \mathbb{R}^{n\times n}$, the spectral norm is lower-bounded by the maximum Euclidean norm of any column of P. That is, letting $e_1, e_2, \ldots, e_n \in \mathbb{R}^n$ denote the standard basis vectors,

$$\|P\| \ge \max_{i\in[n]} \|Pe_i\| = \max_{i\in[n]} \|P_{\cdot,i}\|.$$

Then for any $\xi, \xi' \in \mathcal{M}_{n,d}$ distinct,

$$\left\|P^{(\xi)} - P^{(\xi')}\right\|^2 \ge \sum_{j=1}^{n} \left(P_{j,1}^{(\xi)} - P_{j,1}^{(\xi')}\right)^2 = \frac{\eta^2 d_H(\xi,\xi')}{d}.$$

By definition of $\mathcal{M}_{n,d}$, $d_H(\xi, \xi') \ge d$, whence for all $\xi, \xi' \in \mathcal{M}_{n,d}$ distinct,

$$\left\|P^{(\xi)} - P^{(\xi')}\right\| \ge \eta. \tag{41}$$

In addition to the matrices $\{P^{(\xi)} : \xi \in \{0,1\}^n\}$, define the matrix $P^{(0)} \in \mathbb{R}^{n\times n}$ by

$$P_{ij}^{(0)} = \begin{cases} \eta & \text{if } i = j = n \\ 0 & \text{otherwise.} \end{cases}$$

By construction, for all $\xi \in \mathcal{M}_{n,d}$,

$$\|P^{(\xi)} - P^{(0)}\| \ge \eta \left(1 + \frac{1}{d}\right)^{1/2} \ge \eta.$$

Thus, the collection of matrices $\left\{P^{(t)} : t \in \mathcal{M}_{n,d} \cup \{0\}\right\}$ obeys Condition 24 in Theorem 24 with $s = \eta/2$.

Similarly, for all $\xi \in \mathcal{M}_{n,d}$, writing $|\xi|$ to denote the number of non-zero entries in ξ,

$$\left\|P^{(\xi)} - P^{(0)}\right\|_F^2 = \eta^2 + \left(P_{1,1}^{(\xi)}\right)^2 + 2\sum_{j=2}^{n} \left(P_{1,j}^{(\xi)}\right)^2 \le \eta^2 + \frac{2|\xi|\eta^2}{d} \le \eta^2 \left(1 + \frac{2n}{d}\right), \tag{42}$$

where we have used the trivial upper bound $|\xi| \le n$. Applying Equation (42) to Equation (40), it follows that

$$\frac{1}{|\mathcal{M}_{n,d}|} \sum_{\xi\in\mathcal{M}_{n,d}} \mathrm{KL}\left(P^{(\xi)}, P^{(0)}\right) \le \eta^2 \left(1 + \frac{2n}{d}\right) \left(\sum_{s=1}^{N} \rho_s^{-1}\right). \tag{43}$$

By the Gilbert-Varshamov bound (see, e.g., Jiang and Vardy, 2004, and citations therein) we can choose $\mathcal{M}_{n,d}$ so that

$$|\mathcal{M}_{n,d}| \geq \frac{2^n}{\sum_{j=0}^{d-1} \binom{n}{j}}.$$

Letting $H(p) = -p \log_2 p - (1-p)\log_2(1-p) \in [0,1]$ denote the binary entropy function, recall the inequality (Ash, 1990, Chapter 4) $\sum_{j=0}^{d-1} \binom{n}{j} \leq 2^{nH\left(\frac{d-1}{n}\right)}$, valid for $d-1 < n/2$. Taking logarithms, we have

$$\log |\mathcal{M}_{n,d}| \geq \log \frac{2^n}{2^{nH((d-1)/n)}} = n\left(1 - H\left(\frac{d-1}{n}\right)\right)\log 2. \tag{44}$$

Choosing $d = \lfloor n/3 \rfloor + 1$ is enough to ensure that for all $n \geq 2$ n/d is bounded by a constant while $H((d-1)/n)$ is bounded away from 1, so that

$$\left(1 + \frac{2n}{d}\right) \leq C\left(1 - H\left(\frac{d-1}{n}\right)\right)\log 2$$

for some constant $C > 0$ not depending on n. Applying this bound to Equation (43),

$$\frac{1}{|\mathcal{M}_{n,d}|} \sum_{\xi \in \mathcal{M}_{n,d}} \mathrm{KL}\left(P^{(\xi)}, P^{(0)}\right) \leq C\left(1 - H\left(\frac{1}{c_d}\right)\right)\eta^2 \left(\sum_{s=1}^{N} \rho_s^{-1}\right). \tag{45}$$

Combining Equation (44) with Equation (45),

$$\frac{1}{|\mathcal{M}_{n,d}|} \sum_{\xi \in \mathcal{M}_{n,d}} \mathrm{KL}\left(P^{(\xi)}, P^{(0)}\right) \leq \frac{C\eta^2}{n} \left(\sum_{s=1}^{N} \rho_s^{-1}\right) \log |\mathcal{M}_n|. \tag{46}$$

Choosing

$$\eta = C_1 \sqrt{n} \left(\sum_{s=1}^{N} \rho_s^{-1}\right)^{-1/2}$$

for suitably small constant $C_1 > 0$ ensures that

$$C\eta^2 \leq \alpha n \left(\sum_{s=1}^{N} \rho_s^{-1}\right)^{-1}$$

for some $\alpha \in (0, 1/8)$. Plugging this into Equation (46), it follows that

$$\frac{1}{|\mathcal{M}_{n,d}|} \sum_{\xi \in \mathcal{M}_{n,d}} \mathrm{KL}\left(P^{(\xi)}, P^{(0)}\right) \leq \alpha \log |\mathcal{M}_n|,$$

and Condition 24 of Theorem 24 is satisfied with $s/2 = \eta$. Applying this theorem, we conclude that

$$\inf_{\hat{P}} \sup_{P \in \mathcal{S}_n} \mathbb{E}\|\hat{P} - P\| \geq C\sqrt{n} \left(\sum_{s=1}^{N} \rho_s^{-1}\right)^{-1/2},$$

completing the proof. ■

B.4 Proof of Theorem 9 (Estimating the sub-gamma weights)

In this section, we provide a proof of Theorem 9. We remind the reader that in this section, $\{\mathring{w}_s\}_{s=1}^N$ are non-negative weights, summing to 1, defined by

$$\mathring{w}_s = \frac{(\nu_s + b_s^2)^{-1}}{\sum_{t=1}^N (\nu_t + b_t^2)^{-1}} \tag{47}$$

for each $s \in [N]$, where (ν_s, b_s) is the sub-gamma parameter for the edges in the s-th network. We further remind the reader that we estimate these weights by

$$\hat{w}_s = \frac{\hat{\rho}_s^{-1}}{\sum_{t=1}^n \hat{\rho}_t^{-1}}, \tag{48}$$

where for each $s \in [N]$, letting $\hat{P}^{(s)}$ denote the rank-d eigenvalue truncation of $A^{(s)}$,

$$\hat{\rho}_s = \frac{\sum_{1 \le i \le j \le n} (A^{(s)} - \hat{P}^{(s)})_{i,j}^2}{16n(n+1)}. \tag{49}$$

For ease of notation, define

$$\tilde{\rho}_s = \frac{\sum_{1 \le i \le j \le n} (A^{(s)} - P)_{i,j}^2}{16n(n+1)}, \tag{50}$$

noting that by definition of τ_s in Equation (14), it follows from a basic property of sub-gamma random variables that

$$\tau_s = \mathbb{E}\tilde{\rho}_s \le \frac{8\nu_s + 32b_s^2}{32} \le \nu_s + b_s^2. \tag{51}$$

The following technical results will prove useful in our main proof.

Proposition 25 *Suppose the networks $A^{(1)}, A^{(2)}, \ldots, A^{(N)}$ are independent and for each $s = 1, 2, \ldots, N$ the edges $\{(A^{(s)} - P)_{i,j} : 1 \le i \le j \le n\}$ are independent (ν_s, b_s)-sub-gamma random variables. With probability at least $1 - Cn^{-2}$ it holds for all $s \in [N]$ that*

$$\|A^{(s)} - P\|_F \le Cn(\nu_s + b_s^2)^{1/2}(\log N + \log n).$$

Proof By a standard tail inequality for sub-gamma random variables, for all $t \ge 0$,

$$\mathbb{P}\left[(A^{(s)} - P)_{i,j}^2 > (\sqrt{2\nu_s t} + b_s t)^2\right] = \mathbb{P}\left[|A_{i,j}^{(s)} - P_{i,j}| > \sqrt{2\nu_s t} + b_s t\right] \le C\exp(-t).$$

Setting $t = 4\log n + \log N$, taking a union bound over all $\{(i,j) : 1 \le i \le j \le n\}$ and upper bounding $(\sqrt{2\nu_s t} + b_s t)^2 \le Ct^2(\nu_s + b_s^2)$ since $t \ge 1$, it holds with probability at least $1 - CN^{-1}n^{-2}$ that for all $1 \le i \le j \le n$,

$$(A^{(s)} - P)_{i,j}^2 \le C(\nu_s + b_s^2)(\log N + \log n)^2.$$

Summing over $1 \le i < j \le n$, $\|A^{(s)} - P\|_F^2 \le Cn^2(\nu_s + b_s^2)(\log N + \log n)^2$ with probability at least $1 - CN^{-1}n^{-2}$. A union bound over $s \in [N]$ yields the result. ■

Proposition 26 *Suppose that the networks $A^{(1)}, A^{(2)}, \ldots, A^{(N)}$ are independent and for each $s = 1, 2, \ldots, N$ the edges $\{(A^{(s)} - P)_{i,j} : 1 \leq i \leq j \leq n\}$ are independent (ν_s, b_s)-sub-gamma random variables. Let $\hat{\rho}_s$ and $\tilde{\rho}_s$ be as defined in Equations (49) and (50), respectively, for all $s \in [N]$, and suppose that $d \leq n$ for all suitably large n. With probability at least $1 - Cn^{-2}$, it holds for all $s \in [N]$ that*

$$|\hat{\rho}_s - \tilde{\rho}_s| \leq \frac{C\sqrt{d}(\nu_s + b_s^2)(\log N + \log n)^2}{\sqrt{n}}.$$

Proof By definition,

$$\begin{aligned}
|\hat{\rho}_s - \tilde{\rho}_s| &= \frac{\left|\sum_{1 \leq i \leq j \leq n} (A^{(s)} - \hat{P}^{(s)})_{i,j}^2 - (A^{(s)} - P)_{i,j}^2\right|}{16n(n+1)} \\
&\leq \frac{\sum_{1 \leq i \leq j \leq n} (\hat{P}^{(s)} - P)_{i,j}^2}{16n(n+1)} + \frac{\sum_{1 \leq i \leq j \leq n} \left|(A^{(s)} - P)_{i,j}(\hat{P}^{(s)} - P)_{i,j}\right|}{16n(n+1)} \\
&\leq \frac{C\|\hat{P}^{(s)} - P\|_F^2}{n^2} + \frac{C\|A^{(s)} - P\|_F \|\hat{P}^{(s)} - P\|_F}{n^2},
\end{aligned}$$

where the second bound follows from an application of the Cauchy-Schwarz inequality. Since $\hat{P}^{(s)}$ is the rank-d truncation of $A^{(s)}$, and both $\hat{P}^{(s)}$ and P are rank-d,

$$\|\hat{P}^{(s)} - P\|_F^2 \leq Cd\|A^{(s)} - P\|^2 \leq Cdn(\nu_s + b_s^2)(\log N + \log n)^2,$$

where the second inequality holds with high probability simultaneously for all $s \in [N]$ by applying Lemma 5 to $\|A^{(s)} - P\|$ separately for each $s \in [N]$ followed by a union bound argument similar to that given in the previous proof. Taking square roots in the above display and applying Proposition 25 to bound $\|A^{(s)} - P\|_F$, we conclude that

$$\begin{aligned}
|\hat{\rho}_s - \tilde{\rho}_s| &\leq \frac{C\|\hat{P}^{(s)} - P\|_F^2}{n^2} + \frac{C\|A^{(s)} - P\|_F \|\hat{P}^{(s)} - P\|_F}{n^2} \\
&\leq \frac{Cd(\nu_s + b_s^2)(\log N + \log n)^2}{n} + \frac{C\sqrt{d}(\nu_s + b_s^2)(\log N + \log n)^2}{\sqrt{n}} \\
&\leq C(\nu_s + b_s^2)\frac{\sqrt{d}(\log N + \log n)}{\sqrt{n}},
\end{aligned}$$

where we have used the assumption that $d/n \leq 1$ for all suitably large n. ■

Proposition 27 *Suppose that the networks $A^{(1)}, A^{(2)}, \ldots, A^{(N)}$ are independent and that for each $s = 1, 2, \ldots, N$, the edges $\{(A^{(s)} - P)_{i,j} : 1 \leq i \leq j \leq n\}$ are independent (ν_s, b_s)-sub-gamma random variables. Let $\hat{\rho}_s$ and $\tilde{\rho}_s$ be as defined in Equations (49) and (50), respectively, for all $s \in [N]$ and suppose that the growth assumption in Equation (12) from Theorem 9 holds. That is, there exists a positive integer k such that*

$$\frac{n^{k-2}d^k (\log N + \log n)^{4k}}{N} = \Omega(1). \tag{52}$$

Then with probability $1 - O(n^{-2})$*, it holds for all* $s \in [N]$ *that*

$$|\tilde{\rho}_s - \tau_s| \leq \frac{\sqrt{d}(\nu_s + b_s^2)(\log N + \log n)^2}{\sqrt{n}}.$$

Proof By definition,

$$\tilde{\rho}_s - \tau_s = \sum_{1 \leq i \leq j \leq n} \frac{(A^{(s)} - P)_{i,j}^2 - \mathbb{E}(A^{(s)} - P)_{i,j}^2}{16n(n+1)},$$

is a sum of independent mean-0 random variables. By Chebyshev's inequality,

$$\mathbb{P}\left[|\tilde{\rho}_s - \tau_s| > t\right] \leq \frac{C\mathbb{E}\left[\sum_{1 \leq i \leq j \leq n}(A^{(s)} - P)_{i,j}^2 - \mathbb{E}(A^{(s)} - P)_{i,j}^2\right]^\ell}{n^{2\ell} t^\ell} \tag{53}$$

for any $t > 0$ and any even integer $\ell \geq 2$. We will bound the expectation on the right-hand side via a standard counting argument.

For ease of notation, let $Z_{i,j} = (A^{(s)} - P)_{i,j}^2 - \mathbb{E}(A^{(s)} - P)_{i,j}^2$, so that

$$\tilde{\rho}_s - \mathbb{E}\tilde{\rho}_s = \frac{\sum_{1 \leq i \leq j \leq n} Z_{i,j}}{16n(n+1)}.$$

The $\{Z_{i,j} : 1 \leq i \leq j \leq n\}$ are independent mean-0, and since $(A^{(s)} - P)_{i,j}$ are (ν_s, b_s)-sub-gamma, all moments of $Z_{i,j}$ exist, with

$$\mathbb{E}Z_{i,j}^k \leq \mathbb{E}(A^{(s)} - P)_{i,j}^{2k} \leq C_k(\nu_s + b_s^2)^k,$$

where C_k is a constant depending on k but not on any other parameters. Let $\vec{i}$ denote an ℓ-tuple of numbers from $[n]$, $\vec{i} = (i_1, i_2, \ldots, i_\ell)$ with $i_a \in [n]$ for all $a \in [\ell]$. Write $\vec{i} \leq \vec{j}$ to mean that $i_a \leq j_a$ for all $a = 1, 2, \ldots, \ell$. Defining the set $\mathcal{I}_{n,\ell} = \{(\vec{i}, \vec{j}) : \vec{i} \leq \vec{j}\}$, we have

$$\mathbb{E}\left|\sum_{1 \leq i \leq j \leq n} Z_{i,j}\right|^\ell = \sum_{(\vec{i},\vec{j}) \in \mathcal{I}_{n,\ell}} \mathbb{E}\prod_{k=1}^{\ell} Z_{i_k,j_k}.$$

We can identify each $(\vec{i}, \vec{j}) \in \mathcal{I}_{n,\ell}$ with an ℓ-tuple of pairs $((i_1, j_1), (i_2, j_2), \ldots, (i_\ell, j_\ell))$. For $1 \leq i \leq j \leq n$, let $m_{(i,j)}(\vec{i}, \vec{j})$ denote the number of times that the pair (i, j) appears in the ℓ-tuple of pairs $(\vec{i}, \vec{j})$. That is,

$$m_{(i,j)}(\vec{i}, \vec{j}) = |\{a \in [\ell] : i_a = i, j_a = j\}|.$$

We can rewrite our expectation of interest as

$$\mathbb{E}\left|\sum_{1 \leq i \leq j \leq n} Z_{i,j}\right|^\ell = \sum_{(\vec{i},\vec{j}) \in \mathcal{I}_{n,\ell}} \mathbb{E}\prod_{(i,j) \in (\vec{i},\vec{j})} Z_{i,j}^{m_{(i,j)}(\vec{i},\vec{j})}, \tag{54}$$

and we see immediately that by independence of the $Z_{i,j}$, if $m_{(i_a,j_a)}(\vec{i},\vec{j}) = 1$ for some $a \in [\ell]$, then the corresponding term in the sum has

$$\mathbb{E}\prod_{k=1}^{\ell} Z_{i_k,j_k} = \mathbb{E}\prod_{(i,j)\in S(\vec{i},\vec{j})} Z_{i,j}^{m_{(i,j)}(\vec{i},\vec{j})} = 0.$$

Since $\{Z_{i,j} : 1 \le i \le j \le n\}$ are independent (ν_s, b_s)-sub-gamma, Hölder's inequality implies that for any $(\vec{i},\vec{j}) \in \mathcal{I}_{n,\ell}$, letting $E(\vec{i},\vec{j}) = \{(i_a, j_a) : a \in [\ell]\}$,

$$\mathbb{E}\prod_{k=1}^{\ell} Z_{i_k,j_k} = \prod_{e\in E(\vec{i},\vec{j})} \mathbb{E}Z_e^{m_e(\vec{i},\vec{j})} \le \prod_{e\in E(\vec{i},\vec{j})} \left(\mathbb{E}Z_e^{\ell}\right)^{\frac{m_e(\vec{i},\vec{j})}{\ell}} \le C_\ell(\nu_s + b_s^2)^\ell. \tag{55}$$

We identify each $(\vec{i},\vec{j}) \in \mathcal{I}_{n,\ell}$ with a partition of $[\ell]$ in the following way: for $a, b \in [\ell]$, take $a \sim b$ if and only if $(i_a, j_a) = (i_b, j_b)$. Under this identification, the nonzero elements of the sum on the right-hand side of Equation (54) are precisely those whose corresponding partition of $[\ell]$ has no singleton parts. Thus, to bound the expectation on the left-hand side of Equation (54), it will suffice to count how many such partitions correspond to non-zero terms in the right-hand sum, and apply the bound in Equation (55) to those terms.

Let $\mathcal{P}_\ell^+$ denote the set of all partitions of $[\ell]$ having no singleton part. Any $\pi \in \mathcal{P}_\ell^+$ can correspond to at most $(n(n+1)/2)^{\ell/2}$ pairs $(\vec{i},\vec{j}) \in \mathcal{I}_{n,\ell}$, since we must associate each part of π with some $(i,j) \in [n]$ satisfying $i \le j$, and each $\pi \in \mathcal{P}_\ell^+$ has at most $\ell/2$ parts. Thus, we can can bound

$$\mathbb{E}\left|\sum_{1\le i\le j\le n} Z_{i,j}\right|^\ell = \sum_{\pi\in\mathcal{P}_\ell^+} \mathbb{E}\prod_{(i,j)\in(\vec{i},\vec{j})} Z_{i,j}^{m_{(i,j)}(\vec{i},\vec{j})} \le \frac{C_\ell|\mathcal{P}_\ell^+|n^\ell(\nu_s+b_s^2)^\ell}{2^{\ell/2}} \le C_\ell n^\ell(\nu_s+b_s^2)^\ell,$$

where we have used the fact that $|\mathcal{P}_\ell^+| \le 2^\ell$ and we have gathered all constants, possibly depending on ℓ but not on n, into C_ℓ. Plugging this back into Equation (53), we conclude that for each $s \in [N]$,

$$\mathbb{P}\left[|\tilde{\rho}_s - \tau_s| > t_s\right] \le \frac{\mathbb{E}\left|\tilde{\rho}_s - \tau_s\right|^\ell}{t_s^\ell} \le \frac{Cn^\ell(\nu_s+b_s^2)^\ell}{n^{2\ell}t_s^\ell} = \frac{C(\nu_s+b_s^2)^\ell}{n^\ell t_s^\ell}.$$

Setting

$$t_s = \frac{\sqrt{d}(\nu_s+b_s^2)(\log N + \log n)^2}{\sqrt{n}},$$

we have

$$\mathbb{P}\left[|\tilde{\rho}_s - \tau_s| > t_s\right] \le \frac{C(\nu_s+b_s^2)^\ell}{n^\ell t_s^\ell} = \frac{C}{(nd)^{\ell/2}(\log N + \log n)^{2\ell}},$$

and a union bound over $s \in [N]$ yields

$$\mathbb{P}\left[\exists s \in [N] : |\tilde{\rho}_s - \tau_s| > t_s\right] \le \frac{CN}{(nd)^{\ell/2}(\log N + \log n)^{2\ell}}.$$

Taking $\ell = 2k$ for a positive integer k chosen in accordance with our growth assumption in Equation (52), it follows that

$$\mathbb{P}\left[\exists s \in [N] : |\tilde{\rho}_s - \tau_s| > t_s\right] = O(n^{-2}),$$

which completes the proof. ■

Proposition 28 *Let $A^{(1)}, A^{(2)}, \ldots, A^{(N)}$ be independent random networks with shared expectation $P = XX^T \in \mathbb{R}^{n\times n}$, where $X \in \mathbb{R}^{n\times d}$, and suppose that for all $s \in [N]$, $\{(A^{(s)} - P)_{i,j} : 1 \le i \le j \le n\}$ are independent (ν_s, b_s)-sub-gamma random variables. Let $\{\hat{w}_s\}_{s=1}^N$ and $\{\tau_s\}_{s=1}^N$ be as in Equations (48) and (51), respectively, and define*

$$u_s = \frac{\tau_s^{-1}}{\sum_{t=1}^N \tau_t^{-1}} \tag{56}$$

for each $s \in [N]$. Under the same growth assumptions as Theorem 9, for all suitably large n, it holds with probability $1 - O(n^{-2})$ that

$$\sum_{s=1}^N |\hat{w}_s - u_s|(\nu_s + b_s^2)^{1/2} \le \frac{C\log n}{\log n + \log N}\left(\sum_{s=1}^N (\nu_s + b_s^2)^{-1}\right)^{-1/2}.$$

Proof Applying the triangle inequality followed by Propositions 26 and 27, with probability $1 - O(n^{-2})$ it holds for all $s \in [N]$ that

$$|\hat{\rho}_s - \tau_s| \le |\hat{\rho}_s - \tilde{\rho}_s| + |\tilde{\rho}_s - \tau_s| \le \frac{C\sqrt{d}(\nu_s + b_s^2)(\log N + \log n)^2}{\sqrt{n}} = C(\nu_s + b_s^2)\gamma_n, \tag{57}$$

where, for ease of notation, we let

$$\gamma_n = \frac{\sqrt{d}(\log N + \log n)^2}{\sqrt{n}}.$$

Thus, for any $s \in [N]$, and n suitably large,

$$\begin{aligned} |\hat{\rho}_s^{-1} - \tau_s^{-1}| = \frac{|\hat{\rho}_s - \tau_s|}{\tau_s\hat{\rho}_s} &\le \frac{C(\nu_s + b_s^2)\gamma_n}{\tau_s\hat{\rho}_s} \le \frac{C(\nu_s + b_s^2)\gamma_n}{\tau_s^2\left(1 - \tau_s^{-1}|\hat{\rho}_s - \tau_s|\right)} \\ &\le \frac{C(\nu_s + b_s^2)\gamma_n}{\tau_s^2\left(1 - C\tau_s^{-1}(\nu_s + b_s^2)\gamma_n\right)}, \end{aligned} \tag{58}$$

where the inequalities follow from successive application of Equation (57).

Expanding the definitions of $\hat{w}_s$ and u_s,

$$\begin{aligned} \sum_{s=1}^N |\hat{w}_s - u_s|(\nu_s + b_s^2)^{1/2} &= \sum_{s=1}^N \left|\frac{\hat{\rho}_s^{-1}}{\sum_t \hat{\rho}_t^{-1}} - \frac{\tau_s^{-1}}{\sum_t \tau_t^{-1}}\right|(\nu_s + b_s^2)^{1/2} \\ &\le \sum_{s=1}^N \frac{\left|\hat{\rho}_s^{-1} - \tau_s^{-1}\right|(\nu_s + b_s^2)^{1/2}}{\sum_t \tau_t^{-1}} + \sum_{s=1}^N \frac{\hat{\rho}_s^{-1}(\nu_s + b_s^2)^{1/2}\sum_t \left|\hat{\rho}_t^{-1} - \tau_t^{-1}\right|}{\left(\sum_t \hat{\rho}_t^{-1}\right)\left(\sum_t \tau_t^{-1}\right)}. \end{aligned}$$

Applying the bound in Equation (58) and using our assumption in Equation (15) that $\tau_s^{-1}(\nu_s + b_s^2)\gamma_n = o(1)$ uniformly over $s \in [N]$,

$$\begin{aligned}\sum_{s=1}^{N} &|\hat{w}_s - u_s|(\nu_s + b_s^2)^{1/2} \\ &\leq \sum_{s=1}^{N} \frac{C\gamma_n \tau_s^{-2}(\nu_s + b_s^2)(\nu_s + b_s^2)^{1/2}}{\sum_t \tau_t^{-1}} + \sum_{s=1}^{N}\sum_{t=1}^{N} \frac{C\gamma_n \tau_t^{-2}(\nu_t + b_t^2)\tau_s^{-1}(\nu_s + b_s^2)^{1/2}}{\left(\sum_t \hat{\rho}_t^{-1}\right)\left(\sum_t \tau_t^{-1}\right)}.\end{aligned} \tag{59}$$

Considering the first of these two sums, we have

$$\sum_{s=1}^{N} \frac{\gamma_n \tau_s^{-1}(\nu_s + b_s^2)^{3/2}}{\tau_s \sum_t \tau_t^{-1}} \leq \left(\gamma_n \max_s \tau_s^{-1}(\nu_s + b_s^2)\right) \sum_{s=1}^{N} \frac{\tau_s^{-1}(\nu_s + b_s^2)^{1/2}}{\sum_t \tau_t^{-1}}. \tag{60}$$

By concavity of the square root function,

$$\sum_{s=1}^{N} \frac{\tau_s^{-1}\sqrt{(\nu_s + b_s^2)\sum_t (\nu_t + b_t^2)^{-1}}}{\sum_t \tau_t^{-1}} \leq \left(\sum_{s=1}^{N} \frac{\tau_s^{-1}\sum_t (\nu_t + b_t^2)^{-1}}{(\nu_s + b_s^2)^{-1}\sum_t \tau_t^{-1}}\right)^{1/2} = \sqrt{\sum_{s=1}^{N} \frac{u_s}{\mathring{w}_s}}, \tag{61}$$

where we have used the definitions of $\{u_s\}_{s=1}^N$ and $\{\mathring{w}_s\}_{s=1}^N$ in Equations (56) and (47), respectively. Rearranging and applying this bound to Equation (60),

$$\sum_{s=1}^{N} \frac{\gamma_n \tau_s^{-2}(\nu_s + b_s^2)^{3/2}}{\sum_t \tau_t^{-1}} \leq \gamma_n \left(\max_s \tau_s^{-1}(\nu_s + b_s^2)\right) \left(\frac{\sum_{s=1}^N u_s/\mathring{w}_s}{\sum_{s=1}^N (\nu_s + b_s^2)^{-1}}\right)^{1/2}$$

Applying our growth assumption from Equation (16), we conclude that

$$\sum_{s=1}^{N} \frac{\gamma_n \tau_s^{-1}(\nu_s + b_s^2)^{3/2}}{\tau_s \sum_t \tau_t^{-1}} \leq \left(\sum_{s=1}^{N} (\nu_s + b_s^2)^{-1}\right)^{-1/2} \frac{\log n}{\log n + \log N}. \tag{62}$$

Turning to the second sum on the right-hand side of Equation (59), another application of Equation (58) to bound $\hat{\rho}_s^{-1}$, followed by Equation (15), implies that

$$\begin{aligned}\sum_{s=1}^{N}\sum_{t=1}^{N} \frac{\gamma_n(\nu_t + b_t^2)(\nu_s + b_s^2)^{1/2}}{\tau_s \tau_t^2 \left(\sum_t \hat{\rho}_t^{-1}\right)\left(\sum_t \tau_t^{-1}\right)} &\leq C\gamma_n \left(\max_s \frac{(\nu_s + b_s^2)}{\tau_s}\right) \sum_{s=1}^{N}\sum_{t=1}^{N} \frac{\tau_t^{-1}\tau_s^{-1}(\nu_s + b_s^2)^{1/2}}{\left(\sum_t \tau_t^{-1}\right)^2} \\ &= C\gamma_n \left(\max_s \frac{(\nu_s + b_s^2)}{\tau_s}\right) \sum_{s=1}^{N} \frac{\tau_s^{-1}(\nu_s + b_s^2)^{1/2}}{\sum_t \tau_t^{-1}}.\end{aligned}$$

Applying Equation (61) once again, we have

$$\sum_{s=1}^{N}\sum_{t=1}^{N} \frac{\gamma_n(\nu_t + b_t^2)(\nu_s + b_s^2)^{1/2}}{\tau_s \tau_t^2 \left(\sum_t \hat{\rho}_t^{-1}\right)\left(\sum_t \tau_t^{-1}\right)} \leq C\gamma_n \left(\max_s \frac{(\nu_s + b_s^2)}{\tau_s}\right) \left(\frac{\sum_{s=1}^N u_s/\mathring{w}_s}{\sum_t (\nu_t + b_t^2)^{-1}}\right)^{1/2}.$$

Once again applying our growth assumption in Equation (16), we have

$$\sum_{s=1}^{N}\sum_{t=1}^{N}\frac{\gamma_n(\nu_t+b_t^2)(\nu_s+b_s^2)^{1/2}}{\tau_s\tau_t^2\left(\sum_t\hat{\rho}_t^{-1}\right)\left(\sum_t\tau_t^{-1}\right)}\leq C\left(\sum_{s=1}^{N}(\nu_s+b_s^2)^{-1}\right)^{-1/2}\frac{\log n}{\log n+\log N}.$$

Applying this bound and Equation (62) to Equation (59),

$$\sum_{s=1}^{N}|\hat{w}_s-u_s|(\nu_s+b_s^2)^{1/2}\leq C\left(\sum_{s=1}^{N}(\nu_s+b_s^2)^{-1}\right)^{-1/2}\frac{\log n}{\log n+\log N},$$

completing the proof. ■

We are now ready to prove Theorem 9.

Proof of Theorem 9 We first establish the bound on $\|\mathring{X}-X\mathring{V}\|_{2,\infty}$. Defining $\mathring{A}=\sum_w \mathring{w}A^{(s)}$, we have $\mathring{X}=\mathrm{ASE}(\mathring{A},d)$. Setting $w_s=\mathring{w}_s$ for $s\in[N]$ in Theorem 6, Equation (5) holds by virtue of our assumption in Equation (13), and thus

$$\|\mathring{X}-X\mathring{V}\|_{2,\infty}\leq\frac{Cd\log n}{\lambda^{1/2}(P)\sqrt{\sum_{s=1}^{N}(\nu_s+b_s^2)^{-1}}}+\frac{Cdn\kappa(P)\log^2 n}{\lambda_d^{3/2}(P)\sum_{s=1}^{N}(\nu_s+b_s^2)^{-1}}. \tag{63}$$

To prove the corresponding bound on $\|\hat{X}-XV\|_{2,\infty}$, write $\hat{A}=\sum_s\hat{w}_sA^{(s)}$ for ease of notation, where $\{\hat{w}_s\}_{s=1}^N$ are as in Equation (48). So long as for some $c_0\in[0,1)$,

$$\left\|\hat{A}-P\right\|<c_0\lambda_d(P)\quad\text{ eventually,} \tag{64}$$

Lemma 4 implies that

$$\begin{aligned}&\|\hat{X}-XV\|_{2,\infty}\\&\quad\leq\frac{\left\|\left(\hat{A}-P\right)U_P\right\|_{2,\infty}}{\lambda_d^{1/2}(P)}+\frac{C\left\|U_P^T\left(\hat{A}-P\right)U_P\right\|_F}{\lambda_d^{1/2}(P)}+\frac{Cd\kappa(P)\left\|\hat{A}-P\right\|^2}{\lambda_d^{3/2}(P)}\end{aligned} \tag{65}$$

We will assume for now that Equation (64) holds, and we will bound each of the norms on the right-hand side of Equation (65) in turn.

Noting that $\sum_s(\hat{w}_s-\mathring{w}_s)P=0$ and applying the triangle inequality,

$$\left\|\left(\hat{A}-P\right)U_P\right\|_{2,\infty}\leq\sum_{s=1}^{N}|\hat{w}_s-\mathring{w}_s|\left\|(A^{(s)}-P)U_P\right\|_{2,\infty}+\left\|\left(\mathring{A}-P\right)U_P\right\|_{2,\infty}. \tag{66}$$

By a slight adaptation of the arguments used to prove the non-delocalized version of Proposition 23, it holds with probability $1-O(n^{-2})$ that for all $s\in[N]$,

$$\left\|(A^{(s)}-P)U_P\right\|_{2,\infty}\leq C\sqrt{d}(\nu_s+b_s^2)^{1/2}(\log N+\log n), \tag{67}$$

where the $\log N$ term is included to permit a union bound over all $s \in [N]$. Similarly, the non-delocalized version of Proposition 23 implies that with probability at least $1 - O(n^{-2})$,

$$\left\| \left(\mathring{A} - P \right) U_P \right\|_{2,\infty} \leq \sqrt{d} \left(\sum_{s=1}^{N} \mathring{w}_s^2 (\nu_s + b_s^2) \right)^{1/2} \log n = \sqrt{d} \left(\sum_{s=1}^{N} (\nu_s + b_s^2)^{-1} \right)^{-1/2} \log n,$$

where we have plugged in the definition of $\mathring{w}_s$. Applying this and Equation (67) to Equation (66), it holds with probability at least $1 - O(n^{-2})$ that

$$\begin{aligned} &\left\| \left(\hat{A} - P \right) U_P \right\|_{2,\infty} \\ &\qquad \leq C\sqrt{d} \sum_{s=1}^{N} |\hat{w}_s - \mathring{w}_s| \, (\nu_s + b_s^2)^{1/2} (\log N + \log n) + \frac{C\sqrt{d} \log n}{\sqrt{\sum_{s=1}^{N} (\nu_s + b_s^2)^{-1}}} \end{aligned} \tag{68}$$

By a similar argument, this time using the non-delocalized version of Proposition 22 applied to $A^{(1)}, A^{(2)}, \ldots, A^{(N)}$ and $\mathring{A}$, it holds with probability $1 - O(n^{-2})$ that

$$\begin{aligned} &\left\| U_P^T \left(\hat{A} - P \right) U_P \right\|_F \\ &\qquad \leq Cd \sum_{s=1}^{N} |\hat{w}_s - \mathring{w}_s| \, (\nu_s + b_s^2)^{1/2} (\log N + \log n) + \frac{Cd \log n}{\sqrt{\sum_{s=1}^{N} (\nu_s + b_s^2)^{-1}}}. \end{aligned} \tag{69}$$

Finally, applying Lemma 5 to $\mathring{A}$ and with $N = 1$ to each of $A^{(1)}, A^{(2)}, \ldots, A^{(N)}$ separately, it holds with probability at least $1 - O(n^{-2})$ that

$$\left\| \mathring{A} - P \right\| \leq C \left(\sum_{s=1}^{N} \mathring{w}_s^2 (\nu_s + b_s^2) \right)^{1/2} \sqrt{n} \log n = \frac{C\sqrt{n} \log n}{\sqrt{\sum_{s=1}^{N} (\nu_s + b_s^2)^{-1}}}$$

and for all $s = 1, 2, \ldots, N$,

$$\left\| A^{(s)} - P \right\| \leq C (\nu_s + b_s^2)^{1/2} \sqrt{n} \left(\log n + \log N \right),$$

with the $\log N$ terms again included to allow a union bound over all $s \in [N]$. Applying the triangle inequality,

$$\left\| \hat{A} - P \right\| \leq C\sqrt{n} \left(\sum_{s=1}^{N} |\hat{w}_s - \mathring{w}_s| \, (\nu_s + b_s^2)^{1/2} \left(\log n + \log N \right) + \frac{\log n}{\sqrt{\sum_{s=1}^{N} (\nu_s + b_s^2)^{-1}}} \right) \tag{70}$$

Plugging the bounds in Equations (68), (69) and (70) into Equation (65), we have that with probability $1 - O(n^{-2})$,

$$\begin{aligned}
&\|\hat{X} - XV\|_{2,\infty} \\
&\quad \leq \frac{C\sqrt{d}}{\lambda_d^{1/2}(P)} \left(\sum_{s=1}^{N} |\hat{w}_s - \mathring{w}_s| \, (\nu_s + b_s^2)^{1/2} (\log N + \log n) + \frac{\log n}{\sqrt{\sum_{s=1}^{N} (\nu_s + b_s^2)^{-1}}} \right) \\
&\qquad + \frac{Cd}{\lambda_d^{1/2}(P)} \left(\sum_{s=1}^{N} |\hat{w}_s - \mathring{w}_s| \, (\nu_s + b_s^2)^{1/2} (\log N + \log n) + \frac{\log n}{\sqrt{\sum_{s=1}^{N} (\nu_s + b_s^2)^{-1}}} \right) \\
&\qquad + \frac{Cdn\kappa(P)}{\lambda_d^{3/2}(P)} \left(\sum_{s=1}^{N} |\hat{w}_s - \mathring{w}_s| \, (\nu_s + b_s^2)^{1/2} \, (\log n + \log N) + \frac{\log n}{\sqrt{\sum_{s=1}^{N} (\nu_s + b_s^2)^{-1}}} \right)^2 .
\end{aligned}$$

Collecting terms,

$$\begin{aligned}
&\|\hat{X} - XV\|_{2,\infty} \\
&\quad \leq \frac{Cd}{\lambda_d^{1/2}(P)} \left(\sum_{s=1}^{N} |\hat{w}_s - \mathring{w}_s| \, (\nu_s + b_s^2)^{1/2} (\log N + \log n) + \frac{\log n}{\sqrt{\sum_{s=1}^{N} (\nu_s + b_s^2)^{-1}}} \right) \\
&\qquad + \frac{Cdn\kappa(P)}{\lambda_d^{3/2}(P)} \left(\sum_{s=1}^{N} |\hat{w}_s - \mathring{w}_s| \, (\nu_s + b_s^2)^{1/2} \, (\log n + \log N) + \frac{\log n}{\sqrt{\sum_{s=1}^{N} (\nu_s + b_s^2)^{-1}}} \right)^2 .
\end{aligned}$$

Comparing this bound with that in Equation (63), it will suffice for us to show that for all suitably large n,

$$\sum_{s=1}^{N} |\hat{w}_s - \mathring{w}_s| \, (\nu_s + b_s^2)^{1/2} (\log N + \log n) \leq C \frac{\log n}{\sqrt{\sum_{s=1}^{N} (\nu_s + b_s^2)^{-1}}}. \tag{71}$$

Note that applying Equation (71) to Equation (70), followed by our growth assumption in Equation (13), will imply that

$$\|\hat{A} - P\| \leq \frac{C\sqrt{n} \log n}{\sqrt{\sum_{s=1}^{N} (\nu_s + b_s^2)^{-1}}} = o\left(\lambda_d(P)\right),$$

so that Equation (64) holds eventually. Thus, our proof will be complete once we establish Equation (71).

Applying the triangle inequality,

$$\sum_{s=1}^{N} |\hat{w}_s - \mathring{w}_s| \, (\nu_s + b_s^2)^{1/2} \leq \sum_{s=1}^{N} |\hat{w}_s - u_s| \, (\nu_s + b_s^2)^{1/2} + \sum_{s=1}^{N} |u_s - \mathring{w}_s| \, (\nu_s + b_s^2)^{1/2}. \tag{72}$$

By Proposition 28, with probability $1 - O(n^{-2})$, the first of these two sums is bounded by

$$\sum_{s=1}^{N} |\hat{w}_s - u_s| \, (\nu_s + b_s^2)^{1/2} \leq \frac{C \log n}{\log n + \log N} \left(\sum_{s=1}^{N} (\nu_s + b_s^2)^{-1} \right)^{-1/2}, \tag{73}$$

The second sum on the right-hand side of Equation (72) is bounded by

$$\begin{aligned} \sum_{s=1}^{N} |u_s - \mathring{w}_s| \, (\nu_s + b_s^2)^{1/2} &= \sum_{s=1}^{N} \mathring{w}_s \left| 1 - \frac{u_s}{\mathring{w}_t} \right| (\nu_s + b_s^2)^{1/2} \\ &\leq \sqrt{\sum_{s=1}^{N} \mathring{w}_s \left(1 - \frac{u_s}{\mathring{w}_s} \right)^2} \sqrt{\sum_{s=1}^{N} \mathring{w}_s (\nu_s + b_s^2)}, \end{aligned}$$

where the inequality follows from Cauchy-Schwarz. Noting that

$$\mathring{w}_s (\nu_s + b_s^2) = \frac{1}{\sum_t (\nu_t + b_t^2)^{-1}},$$

it follows that

$$\sum_{s=1}^{N} |u_s - \mathring{w}_s| \, (\nu_s + b_s^2)^{1/2} \leq \sqrt{\frac{N \sum_{s=1}^{N} \mathring{w}_s \left(1 - \frac{u_s}{\mathring{w}_s} \right)^2}{\sum_t (\nu_t + b_t^2)^{-1}}}.$$

Applying this and Equation (73) to Equation (72), we have

$$\sum_{s=1}^{N} |\hat{w}_s - \mathring{w}_s| \, (\nu_s + b_s^2)^{1/2} \leq \frac{1}{\sqrt{\sum_{s=1}^{N} (\nu_s + b_s^2)^{-1}}} \left(\frac{C \log n}{\log n + \log N} + \sqrt{N \sum_{s=1}^{N} \mathring{w}_s \left(1 - \frac{u_s}{\mathring{w}_s} \right)} \right).$$

Applying the growth assumption in Equation (17) yields

$$\sum_{s=1}^{N} |\hat{w}_s - \mathring{w}_s| \, (\nu_s + b_s^2)^{1/2} \leq \frac{C \log n}{(\log n + \log N) \sqrt{\sum_{s=1}^{N} (\nu_s + b_s^2)^{-1}}}.$$

Multiplying by $(\log n + \log N)$ establishes Equation (71), completing the proof. ■

References

C. Aicher, A. Z. Jacobs, and A. Clauset. Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2):221–248, 2015.

C. J. Aine, H. J. Bockholt, J. R. Bustillo, J. M. Cañive, A. Caprihan, C. Gasparovic, F. M. Hanlon, J. M. Houck, R. E. Jung, J. Lauriello, J. Liu, A. R. Mayer, N. I. Perrone-Bizzozero, S. Posse, J. M. Stephen, J. A. Turner, V. P. Clark, and Vince D. Calhoun. Multimodal Neuroimaging in Schizophrenia: Description and Dissemination. *Neuroinformatics*, 15(4):343–364, 2017.

J. Arroyo, D. Kessler, E. Levina, and S. F. Taylor. Network classification with applications to brain connectomics. *Annals of Applied Statistics*, 13(3):1648–1677, 2019.

R. B. Ash. *Information Theory*. Dover Publications, 1990.

A. Athreya, D. E. Fishkind, K. Levin, V. Lyzinski, Y. Park, Y. Qin, D. L. Sussman, M. Tang, J. T. Vogelstein, and C. E. Priebe. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92, 2018.

R. Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007.

S. Bhattacharyya and S. Chatterjee. Spectral clustering for multiple sparse networks: I. *arXiv:1805.10594*, 2018.

R. L. Bluhm, J. Miller, R. A. Lanius, E. A. Osuch, K. Boksman, R. W. Neufeld, J. Théberge, B. Schaefer, and P. Williamson. Spontaneous low-frequency fluctuations in the BOLD signal in schizophrenic patients: anomalies in the default network. *Schizophrenia Bulletin*, 33(4):1004–1012, 2007.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

J. Cape, M. Tang, and C. E. Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, 47 (5):2405–2439, 2019.

L. Chen, N. Josephs, L. Lin, J. Zhou, and E. D. Kolaczyk. A spectral-based framework for hypothesis testing in populations of networks. *arXiv:2011.12416*, 2020.

R. Ciric, D. H. Wolf, J. D. Power, D. R. Roalf, G. L. Baum, K. Ruparel, R. T. Shinohara, M. A. Elliott, S. B. Eickhoff, C. Davatzikos, R. C. Gur amd R. E. Gur, D. S. Bassett, and T. D. Satterthwaite. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage*, 154: 174–187, 2017.

X. Dong, P. Frossard, P. Vandergheynst, and N. Nefedov. Clustering on multi-layer graphs via subspace analysis on grassmann manifolds. *IEEE Transactions on Signal Processing*, 62(4):905–918, 2014.

D. Eynard, A. Kovnatsky, M. M. Bronstein, K. Glashoff, and A. M. Bronstein. Multimodal manifold analysis by simultaneous diagonalization of laplacians. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(12):2505–2517, 2015.

D. Ferrari and Y. Yang. Maximum lq-likelihood estimation. *The Annals of Statistics*, 38 (2):753–783, 2010.

D. E. Fishkind, D. L. Sussman, M. Tang, J. Vogelstein, and C. E. Priebe. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Application*, 34(1):23–39, 2013.

K. C. Fox, R. N. Spreng, M. Ellamil, J. R. Andrews-Hanna, and K. Christoff. The wandering brain: meta-analysis of functional neuroimaging studies of mind-wandering and related spontaneous thought processes. *Neuroimage*, 111:611–621, 2015.

R. S. Garfinkel, A. W. Neebe, and M. R. Rao. The m-center problem: Minimax facility location. *Management Science*, 23(10):1133–1142, 1977.

Q. Han, K. S. Xu, and E. M. Airoldi. Consistent estimation of dynamic and multi-layer block models. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1511–1520, 2015.

X. Han, Q. Yang, and Y. Fan. Universal rank inference via residual subsampling with application to large networks. *arXiv:1912.11583*, 2019.

P. W Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: first steps. *Social Networks*, 5:109–137, 1983.

P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35:73–101, 1964.

T. A. Hummer, M. G. Yung, J. Go n, S. K. Conroy, M. M. Francis, N. F. Mehdiyoun, and A. Breier. Functional network connectivity in early-stage schizophrenia. *Schizophrenia Research*, 218:107–115, 2020.

T. Jiang and A. Vardy. Asymptotic improvement of the gilbert-varshamov bound on the size of binary codes. *IEEE Transactions on Information Theory*, 50(8), 2004.

T. Kaufmann, K. C. Skåtun, D. Alnæs, N. T. Doan, E. P. Duff, S. Tønnesen, E. Roussos, T. Ueland, S. R. Aminoff, T. V. Lagerberg, I. Agartz, I. S. Melle, S. M. Smith, O. A. Andreassen, and L. T. Westlye. Disintegration of sensorimotor brain networks in schizophrenia. *Schizophrenia Bulletin*, 41(6):1326–1335, 2015.

J. Khim and P.-L. Loh. A theory of maximum likelihood for weighted infection graphs. *arXiv:1806.05273*, 2018.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.

C. M. Le, K. Levin, and E. Levina. Estimating a network from multiple noisy realizations. *Electronic Journal of Statistics*, 12(2):4697–4740, 2018.

K. Levin, A. Athreya, M. Tang, V. Lyzinski, and C. E. Priebe. A central limit theorem for an omnibus embedding of random dot product graphs. *arXiv:1705.09355*, 2017.

S. Li, N. Hu, W. Zhang, B. Tao, J. Dai, Y. Gong, Y. Tan, D. Cai, and S. Lui. Dysconnectivity of multiple brain networks in schizophrenia: A meta-analysis of resting-state functional connectivity. *Frontiers in Psychiatry*, 10(482), 2019.

V. Lyzinski, D. L. Sussman, M. Tang, A. Athreya, and C. E. Priebe. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic Journal of Statistics*, 8(2):2905–2922, 2014.

V. Lyzinski, M. Tang, A. Athreya, Y. Park, and C. E. Priebe. Community detection and classification in hierarchical stochastic blockmodels. *IEEE Transactions in Network Science and Engineering*, 2017.

S. Paul and Y. Chen. Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electronic Journal of Statistics*, 10:3807–3870, 2016.

J. D. Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, A. C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar, et al. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.

P. Rigollet and J.-C. Hütter. High dimensional statistics lecture notes. Accessed May, 2018, 2018. URL `http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf`.

K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

P. Rubin-Delanchy, C. E. Priebe, M. Tang, and J. Cape. A statistical interpretation of spectral embedding: the generalised random dot product graph. *arXiv 1709.05506*, 2017.

A. K. Shinn, J. T. Baker, K. E. Lewandowski, D. Öngür, and B. M. Cohen. Aberrant cerebellar connectivity in motor and association networks in schizophrenia. *Frontiers in Human Neuroscience*, 9(134), 2015.

D. L. Sussman, M. Tang, D. E. Fishkind, and C. E. Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107:1119–1128, 2012.

R. Tang, M. Tang, J. T. Vogelstein, and C. E. Priebe. Robust estimation from multiple graphs under gross error contamination. *arXiv:1707.03487*, 2017.

R. Tang, M. Ketcha, A. Badea, E. D. Calabrese, D. S. Margulies, J. T. Vogelstein, C. E. Priebe, and D. L. Sussman. Connectome smoothing via low-rank approximations. *IEEE Transactions on Medical Imaging*, 38(6):1446–1456, 2018.

W. Tang, Z. Lu, and I. S. Dhillon. Clustering with multiple graphs. In *Proceedings of the 9th IEEE International Conference on Data Mining (ICDM)*, pages 1016–1021, 2009.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

S. Wang, J. Arroyo, J. T. Vogelstein, and C. E. Priebe. Joint embedding of graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1324–1336, 2021.

S. Whitfield-Gabrieli, H. W. Thermenos, S. Milanovic, M. T. Tsuang, S. V. Faraone, R. W. McCarley, M. E. Shenton, A. I. Green, A. Nieto-Castanon, P. LaViolette, J. Wojcik, J. D. Gabrieli, and L. J. Seidman. Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia. *Proceedings of the National Academy of Sciences*, 106(4):1279–1284, 2009.

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

B. Yu. Stability. *Bernoulli*, 19(4):1484–1500, 2013.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102:315–323, 2015.

B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1):1–45, 2005.