Journal of Machine Learning Research 22 (2021) 1-49 Submitted 7/19; Revised 2/21; Published 3/21

Inference for Multiple Heterogeneous Networks with a Common Invariant Subspace

Jesús Arroyo JARROYO@TAMU.EDU
Department of Statistics, Texas A&M University
College Station, TX, 77843

Avanti Athreya DATHREY1@JHU.EDU
Department of Applied Mathematics and Statistics, Johns Hopkins University
Baltimore, MD, 21218, USA

Joshua Cape JOSHUA.CAPE@PITT.EDU
Department of Statistics, University of Pittsburgh
Pittsburgh, PA 15260, USA

Guodong Chen GCHEN35@JHU.EDU
Carey E. Priebe CEP@JHU.EDU
Department of Applied Mathematics and Statistics, Johns Hopkins University
Baltimore, MD, 21218, USA

Joshua T. Vogelstein JOVO@JHU.EDU
Department of Biomedical Engineering, Kavli Neuroscience Discovery Institute
Johns Hopkins University
Baltimore, MD, 21218, USA

Editor: Genevera Allen

Abstract

The development of models and methodology for the analysis of data from multiple heterogeneous networks is of importance both in statistical network theory and across a wide spectrum of application domains. Although single-graph analysis is well-studied, multiple graph inference is largely unexplored, in part because of the challenges inherent in appropriately modeling graph differences and yet retaining sufficient model simplicity to render estimation feasible. This paper addresses exactly this gap, by introducing a new model, the common subspace independent-edge multiple random graph model, which describes a heterogeneous collection of networks with a shared latent structure on the vertices but potentially different connectivity patterns for each graph. The model encompasses many popular network representations, including the stochastic blockmodel. The model is both flexible enough to meaningfully account for important graph differences, and tractable enough to allow for accurate inference in multiple networks. In particular, a joint spectral embedding of adjacency matrices—the multiple adjacency spectral embedding—leads to simultaneous consistent estimation of underlying parameters for each graph. Under mild additional assumptions, the estimates satisfy asymptotic normality and yield improvements for graph eigenvalue estimation. In both simulated and real data, the model and the embedding can be deployed for a number of subsequent network inference tasks, including dimensionality reduction, classification, hypothesis testing, and community detection. Specifically, when the embedding is applied to a data set of connectomes constructed through diffusion magnetic resonance imaging, the result is an accurate classification of brain scans by human subject and a meaningful determination of heterogeneity across scans of different individuals.

Keywords: multiple random graphs, spectral embeddings, community detection

©2021 Jesús Arroyo, Avanti Athreya, Joshua Cape, Guodong Chen, Carey E. Priebe and Joshua T. Vogelstein.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v22/19-558.html.

1. Introduction

Random graph inference has witnessed extraordinary growth over the last several decades (Goldenberg et al., 2009; Kolaczyk, 2017; Abbe, 2017). To date, the majority of work has focused primarily on inference for a single random graph, leaving largely unaddressed the problem of modeling the structure of multiple-graph data, including multi-layered and time-varying graphs (Kivelä et al., 2014; Boccaletti et al., 2014; Holme and Saramäki, 2012). Such multiple graph data arises naturally in a wide swath of application domains, including neuroscience (Bullmore and Sporns, 2009), biology (Han et al., 2004; Li et al., 2011), and the social sciences (Szell et al., 2010; Mucha et al., 2010).

Several existing models for multiple-graph data require strong assumptions that limit their flexibility or scalability to the size of real-world networks. Principled approaches to multiple-graph inference (see for example Tang et al. (2017a,b); Levin et al. (2017); Ginestet et al. (2017); Arroyo Relión et al. (2019); Kim and Levina (2019)) reflect this difficulty, in part precisely because of the inherent challenges in constructing a multiple-graph model that adequately captures real-world graph heterogeneity while remaining analytically tractable. The aim here is to address this gap.

In this paper, we resolve the following questions: first, can we construct a simple, flexible multiple random network model that can be used to approximate real-world data? Second, for such a model, can we leverage common network structure for accurate, scalable estimation of model parameters, while also being able to describe distributional properties of our estimates? Third, how can we use such estimators in subsequent inference tasks—for instance, multiple-network testing, community detection, or graph classification? Fourth, how well do such modeling, estimation, and testing procedures perform on simulated and real data, in comparison to current state-of-the-art techniques?

Towards this end, we present a semiparametric model for multiple graphs with aligned vertices that is based on a common invariant subspace structure between their expected adjacency matrices, but with allowance for heterogeneity within and across the graphs. The common invariant subspace has a meaningful interpretation and generalizes several existing models for multiple networks; moreover, the estimation of such a common subspace in a set of networks is an inherent part of well-known graph inference problems like community detection, vertex classification, or eigenvalue estimation.

The paper is organized as follows. We start by briefly encapsulating the contributions, and give an overview of the related work. In Section 2 and Section 3, we give a formal treatment of the model and the estimation procedure. In Section 4, we study the theoretical statistical performance of the estimator to provide a bound on the estimation error and describe the distributional properties. In Section 5, we investigate the empirical performance for estimation and testing when compared to other benchmark procedures. In Section 6, we analyze connectomic networks of human brain scans. Specifically, we show the ability of our model and estimation procedure to characterize and discern the differences in brain connectivity across subjects. In Section 7, we conclude with a discussion of open problems for future work.

1.1 Summary of Contributions

We model a collection of m random graphs $\mathcal{G}^{(1)}, \cdots, \mathcal{G}^{(m)}$ with n labeled vertices using Bernoulli adjacency matrices $\mathbf{A}^{(i)} \in \{0,1\}^{n\times n}, 1 \leq i \leq m$, with expectation of the form $\mathbb{E}[\mathbf{A}^{(i)}] = \mathbf{V}\mathbf{R}^{(i)}\mathbf{V}^\top$, where $\mathbf{R}^{(i)} \in \mathbb{R}^{d\times d}$ is called a score matrix, and $\mathbf{V} \in \mathbb{R}^{n\times d}$ is a matrix with orthonormal columns. Because the invariant subspaces defined by $\mathbf{V}$ are common to all the graphs, we call this model the *common subspace independent edge* (COSIE) random graph model. The score matrices $\{\mathbf{R}^{(i)}\}_{i=1}^m$ need not be diagonal and can vary with each graph. In particular, despite the shared matrix factor $\mathbf{V}$, each graph adjacency matrix can have a different expectation.

Now, to estimate $\mathbf{V}$ and $\mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)}$, we rely on the low-rank structure of the expectation matrices in the COSIE model. Specifically, we first spectrally decompose each of the adjacency matrices $\mathbf{A}^{(i)}$ to obtain the *adjacency spectral embedding* (ASE) (Sussman et al., 2012) $\widehat{\mathbf{X}}^{(i)}$, defined

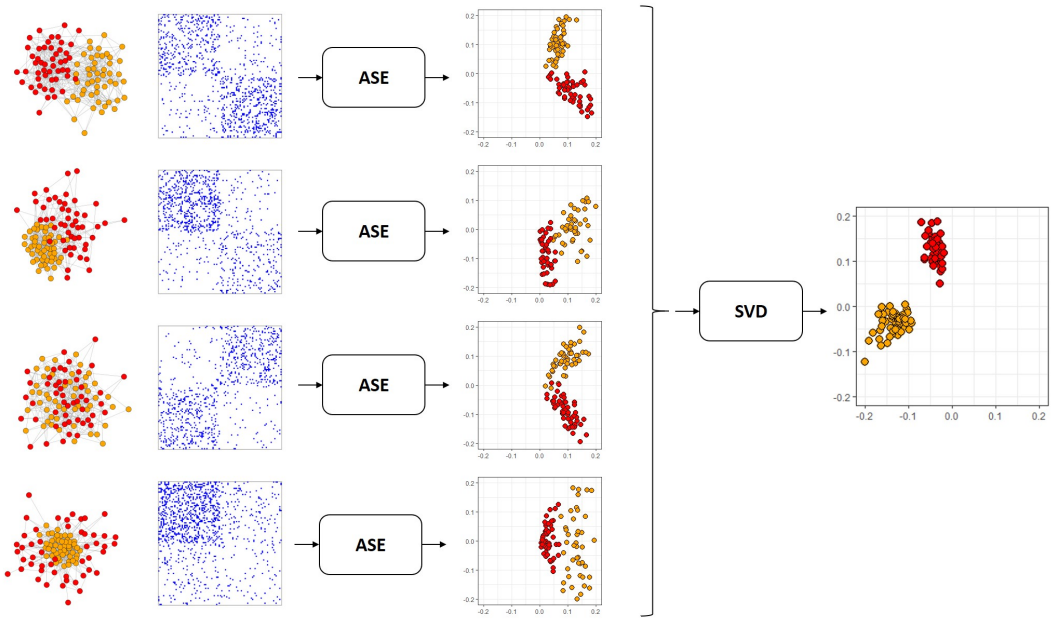


Figure 1: Graphical representation of the multiple adjacency spectral embedding algorithm (MASE, see Algorithm 1) for estimating the common invariant subspace $\mathbf{V}$ in a multilayer stochastic blockmodel (see Definition 3).

as $\widehat{\mathbf{X}}^{(i)} = \widehat{\mathbf{V}}^{(i)}|\widehat{\mathbf{D}}^{(i)}|^{1/2}$, where $\widehat{\mathbf{D}}^{(i)}$ is the $d \times d$ diagonal matrix of the top d eigenvalues of $\mathbf{A}^{(i)}$, sorted by magnitude, and $\widehat{\mathbf{V}}^{(i)}$ the $n \times d$ matrix of associated orthonormal eigenvectors. We leverage the common subspace structure and use these eigenvector estimates $\widehat{\mathbf{V}}^{(i)}$ to obtain an improved estimate of the true common subspace $\mathbf{V}$. We use $\widehat{\mathbf{V}}^{(i)}$ to build the *multiple adjacency spectral embedding* (MASE), as follows. We define the $n \times md$ matrix $\widehat{\mathbf{U}}$ by $\widehat{\mathbf{U}} = (\widehat{\mathbf{V}}^{(1)}, \cdots, \widehat{\mathbf{V}}^{(m)})$ and we let $\widehat{\mathbf{V}}$ be the matrix of the top d leading left singular vectors of $\widehat{\mathbf{U}}$. Figure 1 provides a graphical illustration of the MASE procedure for a collection of stochastic blockmodel graphs. We set $\widehat{\mathbf{R}}^{(i)} = \widehat{\mathbf{V}}^{\top}\mathbf{A}^{(i)}\widehat{\mathbf{V}}$. The MASE algorithm, then, outputs $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{R}}^{(i)}$, for $1 \leq i \leq m$.

The main contributions of the paper are listed as follows.

1. We introduce a flexible, tractable model for multiple random graphs, the *common subspace independent edge* (COSIE) model, that retains enough homogeneity—via the common subspace—for ease of estimation, and permits sufficient heterogeneity—via the potentially distinct score matrices $\mathbf{R}^{(i)}$—to model important collections of different graphs. The COSIE model is appropriate for modeling a variety of real network phenomena. Indeed, the ubiquitous stochastic blockmodel (SBM) and its extension to multilayer networks (Holland et al., 1983) is a special case of the COSIE model (Proposition 4). COSIE also provides a generalization to the multiple-graph setting of latent position models such as the *random dot product graph* (Young and Scheinerman, 2007; Athreya et al., 2017).

2. The estimates obtained by MASE provide consistent estimation for the common subspace matrix $\mathbf{V}$ and the score matrices $\mathbf{R}^{(i)}$. In particular, Theorem 7 shows that under mild

assumptions on the sparsity of the graphs, there exists a constant C such that for all $n \geq n_0$, we have

$$\mathbb{E}\left[\min_{\mathbf{W}\in\mathcal{O}_d} \|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F\right] \leq C\left(\sqrt{\frac{1}{nm}} + \frac{1}{n}\right),$$

where $\mathcal{O}_d$ represents the set of $d \times d$ orthogonal matrices. This result employs bounds introduced by Fan et al. (2019) for subspace estimation in distributed covariance matrices. Furthermore, under delocalization assumptions on $\mathbf{V}$, we can show that the MASE estimates $\widehat{\mathbf{R}}^{(i)}$ are asymptotically normal as the size of the graph n increases, with a bias term that decays as m increases. Namely, there exists a sequence of orthogonal matrices $\mathbf{W}$ such that

$$\frac{1}{\sigma_{i,k,l}}(\widehat{\mathbf{R}}^{(i)} - \mathbf{W}^\top\mathbf{R}^{(i)}\mathbf{W} + \mathbf{H})_{kl} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$ and for $k, l \in [d]$, where $\mathbf{H}$ is a random matrix that satisfies $\mathbb{E}[\|\mathbf{H}\|_F] = O(\frac{d}{\sqrt{m}})$, $\sigma^2_{i,k,l} = O(1)$, and $\mathcal{N}(0,1)$ is the standard normal distribution. The norm of the score matrices typically grows with the size of the graphs, and in particular, our method requires that $\|\mathbf{R}^{(i)}\|_F = \omega(\log n)$. The asymptotic distribution result then shows that the estimates of the score matrices obtained by MASE are accurate, allowing their use in subsequent inference tasks. These results and simulation evidence (see Figure 2 in Section 5) also suggest that the multiplicity of graphs in a COSIE can reduce the bias in eigenvalue estimation as observed in Tang (2018).

3. The embedding method MASE can be successfully deployed for a number of subsequent network inference tasks, including dimensionality reduction, classification, hypothesis testing and community detection, and it performs competitively with respect to (and in some cases significantly better than) a number of existing procedures. In Section 4.1 we show the ability of the method to perform consistent community detection in multilayer stochastic blockmodels. Section 5 examines the empirical performance of MASE in a number of graph inference tasks. In particular, hypothesis testing on populations of graphs is a relatively nascent area in which methodologically sound procedures for graph comparison are scarce, and we demonstrate that COSIE provides foundations for the critical (and wide-open) question of multi-sample graph hypothesis testing, whereby we can use the estimates to build principled tests to compare different networks.

4. COSIE and MASE are robust to the challenges of real data, and provide domain-relevant insights for real data analysis. In particular, COSIE and MASE can be deployed effectively in a pressing real-data problem: that of identifying differences in brain connectivity in medical imaging on human subjects. In Section 6, we use COSIE to model a collection of brain networks, specifically data from the HNU1 study (Zuo et al., 2014) consisting of diffusion magnetic resonance imaging (dMRI), and show that the MASE procedure elucidates differences across heterogeneous networks but also recognizes important similarities, and provides a rigorous statistical framework within which biologically relevant differences can be assessed.

1.2 Related Work

Latent space approaches for the multiple graph data setting have been presented before, but they either tend to limit the heterogeneity in the distributions across the graphs, or include a large number of parameters, which complicate the scalability and interpretation. Levin et al. (2017) presents a method to estimate the latent positions for a set of graphs. Although the method can in principle obtain different latent positions for each graph that do not require any further Procrustes alignment, the method is only studied under a joint RDPG model which assumes that the latent positions of all the graphs are the same. Wang et al. (2019c) introduced a semiparametric model for graphs

that is able to effectively handle heterogeneous distributions. However, their model usually requires a larger number of parameters to represent the same distributions than COSIE, which complicates the interpretation, and the non-convexity of the problem makes estimation more difficult. In fact, to represent a multilayer SBM with K communities this model requires to embed the graphs in a latent space of dimension $O(K^2)$, compared to $O(K)$ for COSIE (see Proposition 4). Tensor factorizations (Zhang and Xia, 2018) provide a way to model the structure of a collection of matrices, but current approaches that are based on tensor decompositions or other matrix factorizations (Zhang et al., 2019; Wang et al., 2019b,a) present challenges on estimation due to the non-convexity of the problem, and interpretation because of the large number of parameters. In particular, the model introduced in (Wang et al., 2019b) has a similar factorization structure to COSIE, but the non-convexity of the likelihood complicates the tractability of the model. More recently, Nielsen and Witten (2018) studied a multiple RDPG model that is a special case of the model of Wang et al. (2019c) by imposing further identifiability constraints. These constraints limit the ability of their model to represent heterogeneous distributions, and, in fact, it is not possible to obtain equivalent statements to Proposition 4 under this model. Bayesian formulations of this problem have also been introduced (Durante et al., 2017; Durante and Dunson, 2018), but computational methods to fit these models limit their applicability to much smaller graphs, with vertices numbering only in the dozens.

In the single graph setting, the adjacency spectral embedding of a graph, $\widehat{\mathbf{X}}$, is a consistent, asymptotically normal estimate of underlying graph parameters for the RDPG model (Sussman et al., 2012; Lyzinski et al., 2014; Athreya et al., 2016). In Levin et al. (2017), it is shown that such a spectral embedding can be profitably deployed for estimation and testing in multiple independent graphs, when all the graphs are sampled from the same probability matrix. Here, the COSIE model provides the ground for studying a multiple heterogeneous graph setting, in which each graph is sampled from a different probability matrix.

The method for estimating the common invariant subspace is related to other similar approaches in the literature. Crainiceanu et al. (2011) proposed a population singular value decomposition method for representing a sample of rectangular matrices, and use it to study arrays of images. In a different work, Fan et al. (2019) introduced a method for estimating the principal components of distributed data. Their method computes the leading eigenvectors of the covariance matrix on each server, and then obtains the leading eigenvectors of the average of the subspace projections, which is shown to converge to the solution of PCA with the same rate as if the data were not distributed. Both methods correspond to the unscaled ASE version of MASE. In our case, we show that this particular way of estimating the parameters of the COSIE model results in a consistent estimator of the common invariant subspace, and asymptotically normal estimators of the individual parameters.

The MASE algorithm can be used to extend spectral clustering to the multiple-graph setting when all the graphs share the same community structure but possibly different connectivity probabilities. This community detection method is scalable and can effectively handle the heterogeneity of the graphs. Other spectral approaches for this problem rely in averaging the adjacency matrices (Han et al., 2015; Bhattacharyya and Chatterjee, 2018; Paul and Chen, 2020) or a transformation of them (Bhattacharyya and Chatterjee, 2017), which requires stronger assumptions on the connection probabilities matrices across the graphs, or can increase the computational cost. Other approaches for this problem, including likelihood and modularity maximization (Han et al., 2015; Peixoto, 2015; Mucha et al., 2010; Paul and Chen, 2020) require to solve large combinatorial problems, which usually limits their scalability.

2. Common Subspace Independent-edge Random Graphs

We consider a sample of m graphs $\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(m)}$, with $\mathcal{G}^{(i)} = (\mathcal{V}, \mathcal{E}^{(i)})$, where $\mathcal{V} = \{1, \ldots, n\}$ denotes a set of n labeled vertices common to all the graphs, and $\mathcal{E}^{(i)} \subset \mathcal{V} \times \mathcal{V}$ is the set of edges corresponding to graph i. Assume for simplicity that the graphs are undirected with no self-loops (it is worth emphasizing however, that the results of this paper can be easily extended to allow for

loops, directed or weighted graphs). For each graph $\mathcal{G}^{(i)}$, denote by $\mathbf{A}^{(i)}$ to the $n \times n$ adjacency matrix that represents the edges; the matrix $\mathbf{A}^{(i)}$ is binary, symmetric and hollow, and $\mathbf{A}^{(i)}_{uv} = 1$ if $(u, v) \in \mathcal{E}^{(i)}$.

To model the graphs we consider an independent-edge random graph framework, in which the edges of a graph $\mathbf{A}$ are conditionally independent given a probability matrix $\mathbf{P} \in [0, 1]^{n \times n}$, so that each entry of $\mathbf{P}_{uv}$ denotes the probability of a Bernoulli random variable representing an edge between the corresponding vertices u and v, and hence, the probability of observing a graph is given by

$$\mathbb{P}(\mathbf{A}) = \prod_{u>v} \mathbf{P}_{uv}^{\mathbf{A}_{uv}} (1 - \mathbf{P}_{uv})^{1-\mathbf{A}_{uv}}.$$

We also write $\mathbf{P} = \mathbb{E}[\mathbf{A}]$. This framework encompasses many popular statistical models for graphs (Holland et al., 1983; Hoff et al., 2002; Bickel and Chen, 2009) which differ in the way the structure of the matrix $\mathbf{P}$ is defined. In the single graph setting, imposing some structure in the matrix $\mathbf{P}$ is necessary to make the estimation problem feasible, as there is only one graph observation. Under this framework, we can characterize the distribution of a sample of graphs $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}$ by their expectations $\mathbf{P}^{(1)}, \ldots, \mathbf{P}^{(m)}$. As our goal is to model a heterogenous population of graphs with possible different distributions, we do not assume that the expected matrices are all equal, so as in the single graph setting, further assumptions on the structure of these matrices are necessary.

To introduce our model, we start by reviewing some models for a single graph that motivate our approach. One of such models is the *random dot product graph* (RDPG) defined below.

Definition 1 (Random dot product graph model (Young and Scheinerman, 2007))**.** *Let* $\mathbf{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times d}$ *be a matrix such that the inner product of any two rows satisfy* $0 \leq X_u^\top X_v \leq 1$*. We say that a random adjacency matrix* $\mathbf{A}$ *is distributed as a random dot product graph with latent positions* $\mathbf{X}$*, and write* $\mathbf{A} \sim \text{RDPG}(\mathbf{X})$*, if the conditional distribution of* $\mathbf{A}$ *given* $\mathbf{X}$ *is*

$$\mathbb{P}(\mathbf{A}) = \prod_{u>v} (X_u^\top X_v)^{\mathbf{A}_{uv}} (1 - X_u^\top X_v)^{1-\mathbf{A}_{uv}}.$$

The RDPG is a type of latent space model (Hoff et al., 2002) in which the rows of $\mathbf{X}$ correspond to latent positions of the vertices, and the probability of an edge between a pair of vertices is proportional to the angle between their corresponding latent vectors, which gives an appealing interpretation of the matrix of latent positions $\mathbf{X}$. Under the RDPG model, the matrix of edge probabilities satisfies $\mathbf{P} = \mathbf{X}\mathbf{X}^\top$, so the RDPG framework contains the class of independent-edge graph distributions for which $\mathbf{P}$ is a positive semidefinite matrix with rank at most d. More recently, Rubin-Delanchy et al. (2017) introduces the generalized RDPG (GRDPG) model, which extends this framework to the whole class of matrices $\mathbf{P}$ with rank at most d, by introducing a diagonal matrix $\mathbf{I}(p, q)$ to the model, such that $\mathbf{I}(p, q)$ is of size $d \times d$ and there are p diagonal entries equal to 1 and q entries equal to -1. Given $\mathbf{X}$ and $\mathbf{I}(p, q)$, the edge probabilities are modeled as $\mathbf{P} = \mathbf{X}\mathbf{I}(p, q)\mathbf{X}^\top$, which keeps the interpretation of the latent space while extending the class of graphs that can be represented within this formulation.

The stochastic blockmodel (SBM) (Holland et al., 1983) is a particular example of a RDPG in which there are only $K < n$ different rows in $\mathbf{X}$, aiming to model community structure in a graph. In a SBM, vertices are partitioned into K different groups, so each vertex $u \in \mathcal{V}$ has a corresponding label $z_u \in \{1, \ldots, K\}$. The probability of an edge appearing between two vertices only depends on their corresponding labels, and these probabilities are encoded in a matrix $\mathbf{B} \in \mathbb{R}^{K \times K}$ such that

$$\mathbb{P}(\mathbf{A}_{uv} = 1) = \mathbf{P}_{uv} = \mathbf{B}_{z_u z_v}.$$

To write the SBM as a RDPG, let $\mathbf{Z} \in \{0, 1\}^{n \times K}$ be a binary matrix that denotes the community memberships of the vertices, with $\mathbf{Z}_{uk} = 1$ if $z_u = k$, and 0 otherwise. Then, the edge probability matrix of the SBM is

$$\mathbf{P} = \mathbf{Z}\mathbf{B}\mathbf{Z}^\top, \tag{1}$$

and denote $\mathbf{A} \sim \text{SBM}(\mathbf{Z}, \mathbf{B})$. Let $\mathbf{B} = \mathbf{W}\mathbf{D}\mathbf{W}^\top$ be the eigendecomposition of $\mathbf{B}$. From this representation, it is easy to see that if we write $\mathbf{X} = \mathbf{Z}\mathbf{W}|\mathbf{D}|^{1/2}$, then the distribution of a SBM corresponds to a GRPDG graph with latent positions $\mathbf{X}$. In particular, if $\mathbf{B}$ is a positive semidefinite matrix, this is also a RDPG. Multiple extensions to the SBM have been proposed in order to produce a more realistic and flexible model, including degree heterogeneity (Karrer and Newman, 2011), multiple community membership (Airoldi et al., 2007), or hierarchical partitions (Lyzinski et al., 2016). These extensions usually fall within the same framework of Equation (1) by placing different constraints on the matrix $\mathbf{Z}$, and hence they can also be studied within the RDPG model.

In defining a model for multiple graphs, we adopt a low-rank assumption on the expected adjacency matrices, as the RDPG model and all the other special cases do. To leverage the information of multiple graphs, a common structure among them is necessary. We thus assume that all the expected adjacency matrices of the independent edge graphs share a common invariant subspace, but allow each individual matrix to be different within that subspace.

Definition 2 (Common Subspace Independent Edge graphs)**.** *Let $\mathbf{V} = (V_1, \dots, V_n)^\top \in \mathbb{R}^{n \times d}$ be a matrix with orthonormal columns, and $\mathbf{R}^{(1)}, \dots, \mathbf{R}^{(m)} \in \mathbb{R}^{d \times d}$ be symmetric matrices such that $0 \leq V_u^\top \mathbf{R}^{(i)} V_v \leq 1$ for all $u, v \in [n]$, $i \in [m]$. We say that the random adjacency matrices $\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(m)}$ are jointly distributed according to the* common subspace independent-edge *graph model with bounded rank d and parameters $\mathbf{V}$ and $\mathbf{R}^{(1)}, \dots, \mathbf{R}^{(m)}$ if for each $i = 1, \dots, m$, given $\mathbf{V}$ and $\mathbf{R}^{(i)}$ the entries of each $\mathbf{A}^{(i)}$ are independent and distributed according to*

$$\mathbb{P}(\mathbf{A}^{(i)}) = \prod_{u<v} (V_u^\top \mathbf{R}^{(i)} V_v)^{\mathbf{A}^{(i)}_{uv}} (1 - V_u^\top \mathbf{R}^{(i)} V_v)^{1 - \mathbf{A}^{(i)}_{uv}}.$$

We denote by $(\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(m)}) \sim \text{COSIE}\left(\mathbf{V}; \mathbf{R}^{(1)}, \dots, \mathbf{R}^{(m)}\right)$ to the joint model for the adjacency matrices.

Under the COSIE graph model, the expected adjacency matrices of the graphs share a common invariant subspace $\mathbf{V}$ that can be interpreted, upon scaling by the score matrices, as the common joint latent positions of the m graphs. This invariant subspace can only be identified up to an orthogonal transformation, so the interpretation of the latent positions is preserved. Note that each graph is marginally distributed as a GRDPG. The score matrices $\mathbf{R}^{(i)}$ can be expressed as $\mathbf{R}^{(i)} = \mathbf{W}^{(i)}\mathbf{D}^{(i)}(\mathbf{W}^{(i)})^\top$ as the eigendecomposition of $\mathbf{R}^{(i)}$, such that $\mathbf{W}^{(i)}$ is a orthogonal matrix of size $d \times d$, and $\mathbf{D}^{(i)}$ is a diagonal matrix containing the eigenvalues. Then, the corresponding latent positions of graph i in the GRDPG model are given by $\mathbf{X}^{(i)} = \mathbf{V}\mathbf{W}^{(i)}|\mathbf{D}^{(i)}|^{1/2}$, with $(|\mathbf{D}|)_{uv} = |\mathbf{D}_{uv}|$.

The model also introduces individual score matrices $\mathbf{R}^{(1)}, \dots, \mathbf{R}^{(m)}$ that control the connection probabilities between the edges of each graph. When an $\mathbf{R}^{(i)}$ is diagonal, its entries contain the eigenvalues of the corresponding probability matrix $\mathbf{P}^{(i)}$, but in general $\mathbf{R}^{(i)}$ does not have to be diagonal. Formal identifiability conditions of the model are discussed in Section 2.1.

The COSIE model can capture any distribution of multiple independent-edge graphs with aligned vertices when the model is equipped with a distribution for the score matrices and d is sufficiently large. That is, for any set of probability matrices $\mathbf{P}^{(1)}, \dots, \mathbf{P}^{(m)}$, there exist an embedding dimension $d \leq n$ such that those matrices can be represented with the COSIE model. Indeed, it is trivial to note that if $d = n$, we can set $\mathbf{V} = \mathbf{I}$ and $\mathbf{R}^{(i)} = \mathbf{P}^{(i)}$. However, for many classes of interest, the embedding dimension d necessary to exactly or approximately represent the graphs is usually much smaller than n. In those cases, the COSIE model can effectively reduce the dimensionality of the problem from $O(mn^2)$ different parameters to only $O(nd + md^2)$.

To motivate our model, we consider the *multilayer stochastic blockmodel* for multiple graphs, originally introduced by Holland et al. (1983). In this model, the community labels of the vertices remain fixed across all the graphs in the population, but the connection probability between and within communities can be different on each graph. The parsimony and simplicity of the model, while maintining heterogeneity across the graphs, has allowed its use in different statistical tasks,

including community detection (Han et al., 2015), multiple graph inference (Pavlović et al., 2020; Kim and Levina, 2019) and modelling time-varying networks (Matias and Miele, 2017). In brain networks, communities are usually in agreement with functional brain systems, which are common across individuals (Power et al., 2011). Formally, the model is defined as follows.

Definition 3 (Multilayer stochastic blockmodel (Holland et al., 1983)). *Let $\mathbf{Z} \in \{0,1\}^{n\times K}$ be a matrix such that $\sum_{k=1}^{K} \mathbf{Z}_{uk} = 1$ for each $u \in [n]$, and $\mathbf{B}^{(1)}, \dots, \mathbf{B}^{(m)} \in [0,1]^{K\times K}$ be symmetric matrices. The random adjacency matrices $\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(m)}$ are jointly distributed as a multilayer SBM, denoted by $(\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(m)}) \sim \text{SBM}\left(\mathbf{Z}; \mathbf{B}^{(1)}, \dots, \mathbf{B}^{(i)}\right)$, if each $\mathbf{A}^{(i)}$ is independently distributed as $\text{SBM}\left(\mathbf{Z}, \mathbf{B}^{(i)}\right)$.*

The next proposition formalizes our intuition statement about the semiparametric aspect of the model. Namely, the COSIE graph model can represent a multilayer SBM with K communities using a dimension d that is at most K. The proof can be found in the appendix.

Proposition 4. *Suppose that $\mathbf{Z} \in \{0,1\}^{n\times K}$ and $\mathbf{B}^{(1)}, \dots, \mathbf{B}^{(m)} \in [0,1]^{K\times K}$ are the parameters of the multilayer SBM. Then, for some $d \leq K$, there exists a matrix with orthonormal columns $\mathbf{V} \in \mathbb{R}^{n\times d}$ and symmetric matrices $\mathbf{R}^{(1)}, \dots, \mathbf{R}^{(m)} \in \mathbb{R}^{d\times d}$ such that*

$$\mathbf{Z}\mathbf{B}^{(i)}\mathbf{Z}^\top = \mathbf{V}\mathbf{R}^{(i)}\mathbf{V}^\top.$$

The previous result shows that the multilayer SBM is a special case of the COSIE model. Conversely, if $\mathbf{V} \in \mathbb{R}^{n\times d}$ is the invariant subspace of the COSIE model, and $\mathbf{V}$ has only K different rows, then the COSIE model is equivalent to the multilayer SBM. Furthermore, several extensions of the single-graph SBM can be translated directly into the multilayer setting, and are also contained within the COSIE model. By allowing the matrix $\mathbf{Z}$ to have more than one non-zero value on each row, overlapping memberships can be incorporated (Latouche et al., 2011), and if the rows of $\mathbf{Z}$ are nonnegative real numbers such that $\sum_{k=1}^{K} \mathbf{Z}_{uk} = 1$ then we obtain an extension of the mixed membership model (Airoldi et al., 2007), which can further incorporate degree heterogeneity by multiplying the rows of $\mathbf{Z}$ by a constant (Zhang et al., 2020). More broadly, if the rows of $\mathbf{V}$ in a COSIE graph are characterized by a hierarchical structure, then the graphs correspond to a multilayer extension of the hierarchical SBM (Lyzinski et al., 2016). Some of these extensions have not been presented before, and hence our work provides an avenue for studying these models, which are interesting in various applications.

2.1 Identifiability

In the COSIE model, note that any orthogonal transformation $\mathbf{W} \in \mathbb{R}^{d\times d}$ of the parameters keeps the probability matrix of the model unchanged. Indeed, observe that

$$\mathbf{P}^{(i)} = \mathbf{V}\mathbf{R}^{(i)}\mathbf{V}^\top = (\mathbf{V}\mathbf{W})(\mathbf{W}^\top\mathbf{R}^{(i)}\mathbf{W})(\mathbf{V}\mathbf{W})^\top,$$

and therefore the most we can hope is that the parameters of the model are identifiable within the equivalence class

$$\mathcal{L}(\mathbf{V}, \{\mathbf{R}^{(i)}\}_{i=1}^m) = \{\mathbf{U}, \{\mathbf{S}^{(i)}\}_{i=1}^m |\ \mathbf{U} = \mathbf{V}\mathbf{W}, \mathbf{S}^{(i)} = \mathbf{W}^\top\mathbf{R}^{(i)}\mathbf{W}, i \in [m] \text{ for some } \mathbf{W} \in \mathcal{O}_d\}.$$

This non-identifiability is unavoidable in many latent space models, including the RDPG. However, with multiple graphs the situation is more nuanced, as we do not require the probability matrix of each graph to have the same rank. Note that Definition 2 does not restrict the matrices $\mathbf{R}^{(1)}, \dots, \mathbf{R}^{(m)}$ to be full rank, so the rank of each individual graph can be smaller than the dimension of the joint model.

The following proposition characterizes the identifiability of the model. Recall that for a given matrix $\mathbf{R} \in \mathbb{R}^{d\times d}$ with singular value decomposition $\mathbf{R} = \mathbf{U}_1\mathbf{S}\mathbf{U}_2$, where $\mathbf{U}_1, \mathbf{U}_2 \in \mathbb{R}^{d\times d}$ are orthogonal matrices and $\mathbf{S}$ is a non-negative diagonal matrix, the spectral norm is given by $\|\mathbf{R}\| := \max_{i\in[d]} \mathbf{S}_{ii}$, and the Frobenius norm is defined as $\|\mathbf{R}\|_F := \left(\sum_{i=1}^d \sum_{j=1}^d \mathbf{R}_{ij}^2\right)^{1/2}$. The proof of the next result is given on the appendix.

Proposition 5 (Model identifiability.)**.** *Let $\mathbf{V} \in \mathbb{R}^{n\times d}$ be a matrix with orthonormal columns and $\mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)}$ be symmetric matrices such that these are the parameters of the bounded rank d COSIE model.*

a) For any for any pair of indices $i, j \in [m]$, the pairwise spectral distance $\|\mathbf{R}^{(i)} - \mathbf{R}^{(j)}\|$ and the Frobenius distance $\|\mathbf{R}^{(i)} - \mathbf{R}^{(j)}\|_F$ are identifiable.

b) Define $\widetilde{\mathbf{R}} = \left(\mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)}\right) \in \mathbb{R}^{d\times dm}$. If the matrix $\widetilde{\mathbf{R}}$ has a full rank, then $\mathbf{V}$ is identifiable up to an orthogonal transformation.

c) Given $\mathbf{V}$, the matrices $\mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)}$ are identifiable.

The previous results present identifiability characterizations at different levels. At the weakest level, the first part of Proposition 5 ensures that even if the parameters are not identifiable, the Frobenius or spectral distances between the score matrices of the graphs are unique. This property allows the use of distance-based methods for any subsequent inference in multiple graph problems, such as multidimensional scaling, k-means or k-nearest neighbors; some examples are shown in Section 5 and 6.

Proposition 5 also provides an identifiability condition for the invariant subspace $\mathbf{V}$. The matrices $\mathbf{R}^{(i)}$ do not need to have the same rank, but the joint model may require a larger dimension than the ranks of each individual graph to represent all the graphs, which is given by the rank of $\mathbf{R}$. To illustrate this scenario, consider the multilayer SBM with 3 communities. Let $\mathbf{B}^{(1)}, \mathbf{B}^{(2)} \in \mathbb{R}^{3\times 3}$ be real matrices, $\mathbf{Z} \in \{0,1\}^{n\times 3}$ a community membership matrix, and $a > b$ some constants so that

$$\mathbf{B}^{(1)} = \begin{pmatrix} a & b & b \\ b & a & a \\ b & a & a \end{pmatrix}, \qquad \mathbf{B}^{(2)} = \begin{pmatrix} a & a & b \\ a & a & b \\ b & b & a \end{pmatrix}.$$

Although the joint model contains three communities, each matrix $\mathbf{B}^{(i)}$ is rank 2, so each graph individually only contains two communities. However, since the concatenated matrix whose columns are given by $\left(\mathbf{B}^{(1)}, \mathbf{B}^{(2)}\right)$ has full rank (i.e rank 3), the three communities of the model can be identified in the joint model, and thus this can be represented as a rank-3 COSIE model.

The last part of Proposition 5 shows that the individual parameters of a graph $\mathbf{R}^{(i)}$ are only identifiable with respect to a given basis of the eigenspace, and hence, all the interpretations that can be derived from depend on $\mathbf{V}$. Some instances of the COSIE model, including the multilayer SBM, provide specific characterizations of $\mathbf{R}^{(i)}$ that can facilitate further interpretation of the parameters.

3. Fitting COSIE by Spectral Embedding of Multiple Adjacency Matrices

This section presents a method to fit the model provided in Definition 2 to a sample of m adjacency matrices $(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}) \sim \text{COSIE}\left(\mathbf{V}; \mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)}\right)$. Fitting the model requires an estimator $\hat{\mathbf{V}}$ for the common subspace $\mathbf{V}$, for which we present a spectral approach. Given an estimated common subspace, we estimate the individual score matrices $\mathbf{R}^{(i)}$ for each graph by least squares, which has a simple solution in out setting. In all of our analysis, we assume that the dimension of the model d is known in advance, but we present a method to estimate it in practice.

We start by defining the *adjacency spectral embedding* (ASE) of an adjacency matrix $\mathbf{A}$, which is a standard tool for estimating the latent positions of a RDPG.

Definition 6 (Adjacency spectral embedding (Sussman et al., 2012; Rubin-Delanchy et al., 2017))**.** *For an adjacency matrix* $\mathbf{A}$*, let* $\mathbf{A} = \widehat{\mathbf{V}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^\top + \widehat{\mathbf{V}}_\perp\widehat{\mathbf{D}}_\perp\widehat{\mathbf{V}}_\perp^\top$ *be the eigendecomposition of* $\mathbf{A}$ *such that* $(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}_\perp)$ *is the* $n \times n$ *orthogonal matrix of eigenvectors, with* $\widehat{\mathbf{V}} \in \mathbb{R}^{n\times d}$*,* $\widehat{\mathbf{V}}_\perp \in \mathbb{R}^{n\times(n-d)}$*, and* $\widehat{\mathbf{D}}$ *is a diagonal matrix containing the* d *largest eigenvalues in magnitude. The* scaled adjacency spectral embedding *of* $\mathbf{A}$ *is defined as* $\widehat{\mathbf{X}} = \widehat{\mathbf{V}}|\widehat{\mathbf{D}}|^{1/2}$*. We refer to* $\widehat{\mathbf{V}}$ *as the* unscaled adjacency spectral embedding*, or simply as the leading eigenvectors of* $\mathbf{A}$*.*

The ASE provides a consistent and asymptotically normal estimator of the corresponding latent positions under the RDPG and GRDPG models as the number of vertices n increases (Sussman et al., 2014; Athreya et al., 2017; Rubin-Delanchy et al., 2017). Therefore, under the COSIE model for which $\mathbf{P} = \mathbf{V}\mathbf{R}\mathbf{V}^\top$, the matrix $\widehat{\mathbf{V}}$ obtained by ASE is a consistent estimator of $\mathbf{V}$, up to an orthogonal transformation, provided that the corresponding $\mathbf{R}$ has full rank. This suggests that given a sample of adjacency matrices, one can simply use the unscaled ASE of any single graph to obtain an estimator of $\mathbf{V}$. However, this method is not leveraging the information about $\mathbf{V}$ on all the graphs.

To give an intuition in how to fit the joint model for the data, we first consider working with the expected probability matrix of each graph $\mathbf{P}^{(i)}$. For simplicity in all our analysis here, we assume that all the matrices $\mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)}$ have full rank, so each matrix $\mathbf{P}^{(i)}$ has exactly d non-zero eigenvalues. Let $\mathbf{V}^{(i)}$ and $\mathbf{X}^{(i)}$ be the unscaled and scaled ASE of the matrix $\mathbf{P}^{(i)}$. Then

$$\begin{aligned}\mathbf{V}^{(i)} &= \mathbf{V}\mathbf{W}^{(i)},\\ \mathbf{X}^{(i)} &= \mathbf{V}\mathbf{W}^{(i)}|\mathbf{D}^{(i)}|^{1/2},\end{aligned}$$

for some orthogonal matrix $\mathbf{W}^{(i)} \in \mathbb{R}^{d\times d}$, and a diagonal matrix $\mathbf{D}^{(i)}$ containing the eigenvalues of $\mathbf{P}^{(i)}$. Note that the matrices $\mathbf{W}^{(i)}$ are possibly different for each graph, which is problematic when we want to leverage the information of all the graphs. Consider the matrix

$$\mathbf{U} = \left(\mathbf{V}^{(1)} \ \cdots \ \mathbf{V}^{(m)}\right),$$

that is formed by concatenating the unscaled ASEs of each graph, resulting in matrices of size $n \times (dm)$. Alternatively, consider the same matrix but concatenating the scaled ASEs, given by

$$\mathbf{U}' = \left(\mathbf{X}^{(1)} \ \cdots \ \mathbf{X}^{(m)}\right).$$

Note that both matrices $\mathbf{U}$ and $\mathbf{U}'$ are rank d, and the d left singular vectors of either of them corresponds to the common subspace $\mathbf{V}$, up to some orthogonal transformation. The SVD step effectively aligns the multiple ASEs to a common spectral embedding. In the scaled ASE, the SVD also eliminates the effect of the eigenvalues of each graph, which are nuisance parameters in the common subspace estimation problem.

In practice, we do not have access to the expected value of the matrices $\mathbf{P}^{(i)}$, but we use the procedure described above with the adjacency matrices to obtain an estimator $\widehat{\mathbf{V}}$ of the common subspace. Given $\widehat{\mathbf{V}}$, we proceed to estimate the individual parameters of the graphs by minimizing a least squares function,

$$\widehat{\mathbf{R}}^{(i)} = \underset{\mathbf{R}\in\mathbb{R}^{d\times d}}{\operatorname{argmin}} \|\mathbf{A}^{(i)} - \widehat{\mathbf{V}}\mathbf{R}\widehat{\mathbf{V}}^\top\|_F^2 = \underset{\mathbf{R}\in\mathbb{R}^{d\times d}}{\operatorname{argmin}} \|\mathbf{R} - \widehat{\mathbf{V}}^\top\mathbf{A}^{(i)}\widehat{\mathbf{V}}\|_F^2,$$

and so, the estimator has a closed form given by

$$\widehat{\mathbf{R}}^{(i)} = \widehat{\mathbf{V}}^\top\mathbf{A}^{(i)}\widehat{\mathbf{V}}.$$

Once these parameters are known, one can also estimate the matrix of edge probabilities as

$$\widehat{\mathbf{P}}^{(i)} = \widehat{\mathbf{V}}\widehat{\mathbf{R}}^{(i)}\widehat{\mathbf{V}}^\top.$$

We call this procedure the *multiple adjacency spectral embedding* (MASE), and it is summarized in Algorithm 1.

Algorithm 1 Multiple adjacency spectral embedding (MASE)

Input: Sample of graphs $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}$; embedding dimensions d and $\{d_i\}_{i=1}^m$.

1. For each $i \in [m]$, obtain the adjacency spectral embedding of $\mathbf{A}^{(i)}$ on d_i dimensions, and denote it by $\widehat{\mathbf{V}}^{(i)} \in \mathbb{R}^{n \times d_i}$.
2. Let $\widehat{\mathbf{U}} = \left(\widehat{\mathbf{V}}^{(1)} \ \cdots \ \widehat{\mathbf{V}}^{(m)}\right)$ be the $n \times (\sum_{i=1}^m d_i)$ matrix of concatenated spectral embeddings.
3. Define $\widehat{\mathbf{V}} \in \mathbf{R}^{n \times d}$ as the matrix containing the d leading left singular values of $\widehat{\mathbf{U}}$.
4. For each $i \in [m]$, set $\widehat{\mathbf{R}}^{(i)} = \widehat{\mathbf{V}}^\top \mathbf{A}^{(i)} \widehat{\mathbf{V}}$.

Output: $\widehat{\mathbf{V}}, \{\widehat{\mathbf{R}}^{(i)}\}_{i=1}^m$.

The MASE method presented for estimating the parameters of the COSIE model only relies on singular value decompositions, and hence it is simple and computationally scalable. This method extends the ideas of a single spectral graph embedding to a multiple graph setting. The first step of Algorithm 1, which is typically the major burden of the method, can be carried in parallel. The estimation of the score matrices only requires to know an estimate for $\mathbf{V}$, and hence allows to easily compute an out-of-sample-embedding for a new graph $\mathbf{A}^{(m+1)}$ without having to update all the parameter estimates.

Figure 1 shows a graphical representation of MASE for estimating the common subspace of a set of four graphs. The graphs correspond to the multilayer SBM with two communities, and the connection matrices $\mathbf{B}^{(i)}$ are selected in a way that the graphs have an heterogeneous structure within the sample. Note that after step 1 of Algorithm 1 the latent positions obtained by ASE are slightly rotated relatively to each graph. After estimating the common subspace by SVD, a common set of latent positions is found, which look tightly clustered within their community, showing that MASE is able to leverage the information of all the graphs in this example.

The first step of Algorithm 1 corresponds to a separate ASE of each graph. We stated the algorithm using the unscaled version of the ASE, and later in Section 4, we show that this version of the method is consistent in estimating the common subspace $\mathbf{V}$. In practice, this step can be replaced with other spectral embeddings of $\mathbf{A}^{(i)}$, including the scaled version described above. Note that the scaled ASE also makes use of the eigenvalues, and thus it puts more weight onto the columns of the matrix $\widehat{\mathbf{U}}'$ that correspond to the eigenvectors associated with the largest eigenvalues, which can be convenient, especially if the dimension d is overestimated. Other spectral embeddings that also aim to estimate the eigenspace of the adjacency matrices, including the Laplacian spectral embedding or a regularized version of it (Priebe et al., 2019; Le et al., 2017), might be preferred in some circumstances, but we do not explore this direction further.

3.1 Choice of Embedding Dimension

In practice, the dimension of the common subspace d is usually unknown. Moreover, each individual $\mathbf{R}^{(i)}$ might not be full rank, and so estimating a different embedding dimension d_i for each graph might be necessary. The actual values of d and $\{d_i\}$ correspond to the ranks of $\mathbf{U}$ and $\{\mathbf{P}^{(i)}\}$ respectively, and so they can be approximated by estimating the ranks of $\{\widehat{\mathbf{A}}^{(i)}\}$ and $\widehat{\mathbf{U}}$. The scree plot method, which consists in looking for an elbow in the plot of ordered singular values of a

matrices, provides an estimator of these quantities, and can be automatically performed using the method proposed by Zhu and Ghodsi (2006). We use this method to fit the model to real data in Section 6.

4. Theoretical Results

In this section, we study the statistical performance of Algorithm 1 in estimating the parameters of the COSIE model when a sample of graphs is given. We first study the expected error in estimating the common subspace, and show that this error decreases as a function of the number of graphs. This result demonstrates that our method is able to leverage the information of the multiple graphs to improve the estimation of $\mathbf{V}$.

To study the estimation error of our method, we consider a sequence of parameters of the COSIE model $\{\mathbf{R}^{(1,n)}, \dots, \mathbf{R}^{(m,n)} \in \mathbb{R}^{d\times d}, \mathbf{V}^{(n)} \in \mathbb{R}^{n\times d}\}_{n=n_0}^{\infty}$, and present non-asymptotic error bounds on the estimation of $\mathbf{V}^{(n)}$, as well as a result on the asymptotic distribution of the estimators of the score matrices as n goes to infinity. The magnitude of the entries of the parameters typically changes with n, and in order to obtain consistent estimators we will require that $\|\mathbf{R}^{(i,n)}\| \to \infty$. This is a natural assumption considering the fact that $\mathbf{R}^{(i,n)}$ contains the eigenvalues of the expected adjacency matrix of a graph, and requiring these eigenvalues to grow as n increases is necessary in order to control the error of the ASE (Athreya et al., 2017). In the following, we omit the dependence of the parameters in n to ease the notation. To simplify the analysis, we also assume that all the score matrices $\mathbf{R}^{(1)}, \dots, \mathbf{R}^{(m)}$ have full rank.

Our next theorem introduces a bound on the expected error in estimating the common subspace of the COSIE model, which is the basis of our theoretical results. The proof of this result relies on bounds from Fan et al. (2019) for studying the eigenvector estimation error in distributed PCA. However, our setting presents some substantial differences, including the distribution of the data and the fact that the expected adjacency matrices are not jointly diagonalizable. Moreover, the scaled version of the algorithm adds the complication of working with eigenvalues, and extending the theory to this case is more challenging. We thus work with the unscaled version of the ASE in Algorithm 1. Before stating the theorem, we introduce some notation. For a given square symmetric matrix $\mathbf{M} \in \mathbb{R}^{r\times r}$, denote by $\lambda_1(\mathbf{M}) \geq \ldots \geq \lambda_r(\mathbf{M})$ to the ordered eigenvalues of $\mathbf{M}$, and define $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ as the smallest and largest eigenvalue in magnitude of $\mathbf{M}$, and $\delta(\mathbf{M}) = \max_{u\in[r]} \sum_{v=1}^{r} \mathbf{M}_{uv}$ the largest sum of the rows of $\mathbf{M}$. Given a pair of sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, denote by $a_n \lesssim b_n$ if there exists some constant $C > 0, n_0 > 0$ such that $a_n \leq Cb_n$ for all $n \geq n_0$.

Theorem 7. *Let $\mathbf{R}^{(1)}, \dots, \mathbf{R}^{(m)}$ be a collection of full rank symmetric matrices of size $d \times d$, $\mathbf{V} \in \mathbb{R}^{n\times d}$ a matrix with orthonormal columns, and $(\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(m)}) \sim \text{COSIE}\left(\mathbf{V}, \mathbf{R}^{(1)}, \dots, \mathbf{R}^{(m)}\right)$ a sample of m random adjacency matrices, and set $\mathbf{P}^{(i)} = \mathbf{V}\mathbf{R}^{(i)}\mathbf{V}^{\top}$. Define*

$$\varepsilon = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \frac{\delta(\mathbf{P}^{(i)})}{\lambda_{\min}^2(\mathbf{R}^{(i)})}}. \tag{2}$$

Let $\widehat{\mathbf{V}}$ be the estimator of $\mathbf{V}$ obtained by Algorithm 1. Suppose that $\min_{i\in[m]} \delta(\mathbf{P}^{(i)}) = \omega(\log n)$ and $\varepsilon = o(1)$ as $n \to \infty$. Then,

$$\mathbb{E}\left[\min_{\mathbf{W}\in\mathcal{O}_d} \|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F\right] \lesssim \sqrt{\frac{d}{m}}\varepsilon + \sqrt{d}\varepsilon^2. \tag{3}$$

The above theorem requires two conditions on the expected adjacency matrices, that control the estimation error of each ASE for a single graph. First, the largest expected vertex degree $\delta(\mathbf{P}^{(i)})$

needs to grow at rate $\omega(\log n)$ to ensure the concentration in spectral norm of the adjacency matrix to its expectation (see Theorem 20 of Athreya et al. (2017)). Second, the condition $\varepsilon = o(1)$ is required to control the average estimation error of each ASE. For this condition to hold, it is enough to have that $\delta(\mathbf{P}^{(i)}) = o\left(\lambda^2_{\min}(\mathbf{R}^{(i)})\right)$ for each graph $i \in [m]$. In particular, observe that

$$\delta(\mathbf{P}^{(i)}) = \|\mathbf{P}^{(i)}\|_1 \leq \sqrt{n}\lambda_{\max}(\mathbf{R}^{(i)}),$$

so it is sufficient that $\max_{i\in[m]} \frac{\lambda_{\max}(\mathbf{R}^{(i)})}{\lambda^2_{\min}(\mathbf{R}^{(i)})} = o\left(\frac{1}{\sqrt{n}}\right)$. Specific conditions for the multilayer SBM that relate the community sizes and the eigenvalues of the connection probability matrices are presented in Section 4.1.

Theorem 7 bounds the expected error in estimating the subspace of $\mathbf{V}$ by the sum of two terms in Equation (3). To understand these terms, note that the error can be partitioned as

$$\begin{aligned}\min_{\mathbf{W}\in\mathcal{O}_d} \|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F &\leq \|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top\|_F \\ &\leq \left\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top\right\|_F + \left\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top\right\|_F, \end{aligned} \tag{4}$$

where $\widetilde{\mathbf{V}}$ is the matrix containing the d leading eigenvectors of $\frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\widehat{\mathbf{V}}^{(i)}(\widehat{\mathbf{V}}^{(i)})^\top\right]$. The first term of the right hand side of Equation (4) measures the variance of the estimated projection matrix, which corresponds to the first term in Equation (3). Thus, this term converges to zero as the number of graphs increases. The estimated projection is not unbiased, and therefore the second term of Equation 4 is always positive, but its order is usually smaller than the variance. This bound is a function of ε, which quantifies the average error in estimating $\mathbf{V}$ by using the leading eigenvectors of $\mathbf{A}^{(i)}$ on each network. Note that the inclusion of an orthogonal transformation $\mathbf{W}$ has to do with the fact that the common subspace is only identifiable up to such transformation.

The above result establishes that when the sample size m is small ($m \lesssim 1/\varepsilon$), the variance term dominates the error as observed in the left hand side of Equation (3), but as the sample size increases this term vanishes. The second term is the bias of estimating $\mathbf{V}$ by using the leading eigenvectors of each network, and this term dominates when the sample size is large. These type of results are common in settings when a global estimator is obtained from the average of local estimators (Lee et al., 2017; Arroyo and Hou, 2016; Fan et al., 2019).

To illustrate the result of Theorem 7, consider the following example on the Erdös-Rényi (ER) model (Erdős and Rényi, 1959; Gilbert, 1959). In the ER random graph model $G(n,p)$, the edges between any pair of the n vertices are formed independently with a constant probability p. We consider a sample of adjacency matrices $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}$, each of them having a different edge probability $p_1, \ldots, p_m$ so that $\mathbf{A}^{(i)} \sim G(n, p_i)$. By defining $\mathbf{V} = \frac{1}{\sqrt{n}}\mathbf{1}_n$, where $\mathbf{1}_n$ is the n-dimensional vector with all entries equal to one, we can represent these graphs under the COSIE model by setting $\mathbf{P}^{(i)} = p_i n \mathbf{V}\mathbf{V}^\top$. Hence, $\delta(\mathbf{P}^{(i)}) = p_i n$ and

$$\varepsilon = \sqrt{\frac{1}{nm}\sum_{i=1}^m \frac{1}{p_i}}.$$

Therefore, if $p_i = \omega(\log(n)/n)$, Theorem 7 implies that

$$\mathbb{E}\left[\min_{\mathbf{W}\in\mathcal{O}_d} \|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F\right] \lesssim \sqrt{\frac{1}{nm^2}\sum_{i=1}^m \frac{1}{p_i}} + \frac{1}{nm}\sum_{i=1}^m \frac{1}{p_i}$$

In the dense regime when $p_i = \Omega(1)$, the error of estimating the common subspace is of order $O\left(\frac{1}{\sqrt{nm}} + \frac{1}{n}\right)$. An alternative estimator for the common subspace $\mathbf{V}$ is given by the ASE of the

mean adjacency matrix, ASE $\left(\frac{1}{m}\sum_{i=1}^{m}\mathbf{A}^{(i)}\right)$, for which the expected subspace estimation error is $O\left(\frac{1}{\sqrt{mn}}\right)$ (Tang et al., 2019). When $m \lesssim n$, the error rate of both estimators coincide. However, the estimator of Tang et al. (2019) is only valid when all the graphs have the same expected adjacency matrices, and can perform poorly in heterogeneous populations of graphs (see Section 5). The ER model described above can be regarded as a special case of the multilayer SBM with only one community, which we consider next.

4.1 Perfect Recovery in Community Detection

Estimation of the common subspace in the multilayer SBM is of particular interest due to its relation with the community detection problem (Girvan and Newman, 2002; Abbe, 2017). In the multilayer SBM with K communities, the matrix of the common subspace basis $\mathbf{V}$ has K different rows, each of them corresponding to a different community. An accurate estimator $\widehat{\mathbf{V}}$ will then reflect this structure, and when the community labels of the vertices are not known, clustering the rows of $\widehat{\mathbf{V}}$ into K groups can reveal these labels. Different clustering procedures can be used for this goal, and in this section we focus on K-means clustering defined next. Suppose that $\mathcal{Z}$ is the set of valid membership matrices for n vertices and K communities, that is,

$$\mathcal{Z} = \left\{ \mathbf{Z} \in \{0,1\}^{n\times K} : \sum_{v=1}^{K} \mathbf{Z}_{uv} = 1,\ u \in [n] \right\}.$$

The community assignments are obtained by solving the K-means problem

$$(\widehat{\mathbf{Z}}, \widehat{\mathbf{C}}) = \underset{\widetilde{\mathbf{Z}}\in\mathcal{Z}, \widetilde{\mathbf{C}}\in\mathbb{R}^{K\times K}}{\operatorname{argmin}} \|\widetilde{\mathbf{Z}}\widetilde{\mathbf{C}} - \widehat{\mathbf{V}}\|_F. \tag{5}$$

The matrix $\widehat{\mathbf{Z}}$ thus contains the estimated community assignments, and $\widehat{\mathbf{C}}$ their corresponding centroids.

To understand the performance in community detection of the procedure described above, consider a multilayer SBM with parameters $\mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(m)} \in [0,1]^{K\times K}$ and $\mathbf{Z} \in \{0,1\}^{n\times K}$. Denote the community sizes by $n_1, \ldots, n_K$, such that $\sum_{j=1}^{k} n_j = n$, and define the quantities $\mathbf{\Xi} = \operatorname{diag}(n_1, \ldots, n_K)$, $n_{\min} = \min_{j\in[K]} n_j$, $n_{\max} = \max_{j\in[K]} n_j$ and $\mathbf{P}^{(i)} = \mathbf{Z}\mathbf{B}^{(i)}\mathbf{Z}^\top$. To obtain a consistent estimator of the common invariant subspace, we use the following assumption.

Assumption 1. *There exist some absolute constants $\kappa \in (0,1], \gamma > 0$ such that*

$$\frac{1}{m}\sum_{i=1}^{m} \frac{\sqrt{K}\left(\sum_{j=1}^{K} n_j^2\right)^{1/2}}{n_{\min}^{2-\kappa}} \frac{\lambda_1(\mathbf{B}^{(i)})}{\lambda_{\min}^2(\mathbf{B}^{(i)})} \le \gamma.$$

The previous assumption requires a control on ratios of the community sizes and the eigenvalues of the connectivity matrices. When all communities have comparable sizes (i.e., $n_{\min} \ge c n_{\max}$ for some absolute constant $c > 0$), then a sufficient condition is that the smallest eigenvalue of each $\mathbf{B}^{(i)}$ approaches zero at a rate

$$\min_{i\in[m]} |\lambda_{\min}(\mathbf{B}^{(i)})| = \Omega\left(\frac{\sqrt{K}\lambda_1^{1/2}(\mathbf{B}^{(i)})}{n^{1/2-\kappa/2}}\right).$$

On the other hand, if the smallest eigenvalue of each $\mathbf{B}^{(i)}$ is bounded away from zero by an absolute constant, then the size of the largest community can be at most $n_{\max} = O\left(\frac{n_{\min}^{2-\kappa}}{K}\right)$.

The next Corollary is a consequence of Theorem 7 for the multilayer SBM setting. The proof is given on the Appendix.

Corollary 8. *Consider a sample of m adjacency matrices from the multilayer SBM*

$$(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}) \sim \text{SBM}\left(\mathbf{Z}; \mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(m)}\right),$$

and let $\widehat{\mathbf{V}}$ the estimated subspace from Algorithm 1. Suppose that $n_{\min} = \omega(1)$, $\delta(\mathbf{P}^{(i)}) = \omega(\log n)$, and that Assumption 1 holds. Then, for sufficiently large n,

$$\mathbb{E}\left[\min_{\mathbf{W}\in\mathcal{O}_d} \|\widehat{\mathbf{V}} - \mathbf{Z}\boldsymbol{\Xi}^{-1/2}\mathbf{W}\|_F\right] \lesssim \sqrt{\frac{\gamma K}{m n_{\min}^{\kappa}}} + \frac{\gamma\sqrt{K}}{n_{\min}^{\kappa}}. \tag{6}$$

The above corollary presents sufficient conditions for consistent estimation of the common invariant subspace, which contains the information of the community labels encoded in the matrix $\mathbf{Z}$. The corollary requires the size of the smallest community to grow as n increases, which is something commonly assumed in the literature (Rohe et al., 2011; Lyzinski et al., 2014).

The following result provides a bound on the expected number of misclustered vertices after clustering the rows of $\widehat{\mathbf{V}}$ according to Equation (5). Due to the non-identifiability of the community labels, these are recovered up to a permutation $\mathbf{Q} \in \mathcal{P}_K$, where $\mathcal{P}_K$ is the set of $K \times K$ permutation matrices.

Theorem 9. *Suppose that $\widehat{\mathbf{Z}}$ is the membership matrix obtained by Equation* (5)*. Under the conditions of Corollary 8, there exist some absolute constants $\gamma > 0$, $\kappa \in (0, 1]$ such that the expected number of misclustered vertices satisfies*

$$\mathbb{E}\left[\min_{\mathbf{Q}\in\mathcal{P}_K} \|\widehat{\mathbf{Z}} - \mathbf{Z}\mathbf{Q}\|_F\right] \lesssim \sqrt{\frac{\gamma K n_{\max}}{m n_{\min}^{\kappa}}} + \frac{\gamma\sqrt{K n_{\max}}}{n_{\min}^{\kappa}}. \tag{7}$$

Similar to the estimation error of the common subspace in Theorem 7, the bound on the expected clustering error depends on two terms in Equation (7). When the number of graphs m increases, the first term vanishes. The second term also decays to zero as long as $K n_{\max} = o(n_{\min}^{2\kappa})$, and therefore the algorithm recovers the community labels correctly as n goes to infinity, provided that $m = \Omega(n_{min}^{\kappa})$. On the other hand, our theory does not guarantee perfect community recovery for $m = O(1)$ because the first term in Equation (7) has at least a constant order. We remark that this fact is a technical consequence of our analysis of subspace estimation error, and it is analogous to similar results obtained in spectral clustering on a single graph (Rohe et al., 2011). In contrast, previous work by Lyzinski et al. (2014) shows, via a careful analysis of the row-wise error in subspace estimation, that spectral methods can produce perfect recovery in community detection in the single-graph setting. Developing a similar result in our setting requires further investigation that we leave as future work.

4.2 Asymptotic Normality of the Estimated Score Matrices

Now, we study the asymptotic distribution of the individual estimates of the score matrices. We show not only that the entries of $\widehat{\mathbf{R}}^{(i)}$ accurately approximate the entries of $\mathbf{R}^{(i)}$ up to some orthogonal transformation, but that under delocalization assumptions on $\mathbf{V}$, the entries $\widehat{\mathbf{R}}_{kl}^{(i)}$, where $1 \leq k \leq d, 1 \leq l \leq d$, exhibit asymptotic normality. In effect, we prove a central limit with a bias term, but this bias term decays as the number of graphs m increases. The central limit theorem depends, in essence, on the ability to estimate the common subspace $\mathbf{V}$ to sufficient precision.

To make such central limit theorem precise, we require the following delocalization and edge variance assumptions.

Assumption 2 (Delocalization of $\mathbf{V}$)**.** *There exist constants $c_1, c_2 > 0$, and an orthogonal matrix $\mathbf{W} \in \mathcal{O}_d$ such that each entry of $\mathbf{VW}$ satisfies*

$$\frac{c_1}{\sqrt{n}} < (\mathbf{VW})_{kl} < \frac{c_2}{\sqrt{n}}, \qquad \forall k \in [n], l \in [d].$$

The delocalization assumption requires that the score matrix $\mathbf{R}^{(i)}$ influence the connectivity of enough edges in the graph. Because the common subspace $\mathbf{V}$ is invariant to orthogonal transformations, the assumption only requires the existence of a particular orthogonal matrix $\mathbf{W}$ that satisfies the entrywise inequalities. This delocalization assumption is satisfied by a wide variety of graphs; for example, if we consider $\mathbf{V}$ as the eigenvectors of an Erdős-Rényi graph, its entries are all of order $1/\sqrt{n}$, and similarly if $\mathbf{V}$ are the eigenvectors of the probability matrix of a stochastic blockmodel all of whose block sizes grow linearly with n (see Proposition 10 below).

Assumption 3 (Edge variance)**.** *The sum of the variance of the edges satisfies for all $i \in [m]$*

$$s^2(\mathbf{P}^{(i)}) := \sum_{s=1}^{n}\sum_{t=1}^{n} \mathbf{P}^{(i)}_{st}(1-\mathbf{P}^{(i)}_{st}) = \omega(1), \qquad \textit{and}$$

$$\sum_{s=1}^{n}\sum_{t=1}^{n} \mathbf{P}^{(i)}_{st}(1-\mathbf{P}^{(i)}_{st})(n\mathbf{V}_{sk}\mathbf{V}_{tl} + n\mathbf{V}_{tk}\mathbf{V}_{sl})^2 = \omega(1), \qquad \textit{for } k \neq l.$$

Assumptions 2 and 3 hold for a wide variety of graphs. In particular, when the community sizes of an SBM are balanced and the connection probabilities do not decay very fast, the next proposition shows that these assumptions hold. The proof is included on the Appendix.

Proposition 10. *Let $\mathbf{Z} \in \{0,1\}^{n\times K}$ be a membership matrix with community sizes $n_1, \ldots, n_K$, and $\mathbf{B}^{(i)} \in [0,1]^{K\times K}$ a connectivity matrix .*

a) If $\min_{k\in[K]} n_k = \Omega(n)$, then Assumption 2 holds for the common invariant subspace of a multilayer SBM with parameter $\mathbf{Z}$.

b) If in addition to a), $\min_{u,v\in[K],i\in[m]} \mathbf{B}^{(i)}_{uv}(1-\mathbf{B}^{(i)}_{uv}) = \omega(1/n^2)$, then Assumption 3 also holds.

The score matrices are symmetric, so we focus on their upper triangular and diagonal elements, and for a symmetric matrix $\mathbf{R} \in \mathbb{R}^{d\times d}$, we define $\mathrm{vec}(\mathbf{R}) \in \mathbb{R}^r$ as the vector of dimension $r = \frac{d(d+1)}{2}$ that contains those elements of $\mathbf{R}$, so that for $k, l \in [d], k \leq l$,

$$[\mathrm{vec}(\mathbf{R})]_{\frac{2k+l(l-1)}{2}} := \mathbf{R}_{kl}.$$

Given the matrices $\mathbf{V}$ and $\mathbf{P}^{(i)} = \mathbf{V}\mathbf{R}^{(i)}\mathbf{V}^\top$ we define $\mathbf{\Sigma}^{(i)} \in \mathbb{R}^{r\times r}$ as

$$\mathbf{\Sigma}^{(i)}_{\frac{2k+l(l-1)}{2},\frac{2k'+l'(l'-1)}{2}} := \sum_{s=1}^{n-1}\sum_{t=s+1}^{n} \mathbf{P}^{(i)}_{st}(1-\mathbf{P}^{(i)}_{st})\left[(\mathbf{V}_{sk}\mathbf{V}_{tl} + \mathbf{V}_{tk}\mathbf{V}_{sl})(\mathbf{V}_{sk'}\mathbf{V}_{tl'} + \mathbf{V}_{sk'}\mathbf{V}_{tl'})\right]. \quad (8)$$

The above matrix can change with n since it is a function of $\mathbf{V}$ and $\mathbf{R}^{(i)}$. As it will be formalized in Theorem 11, when n is large enough the covariance of the upper triangular and diagonal entries of $\widehat{\mathbf{R}}^{(i)}$ (after some proper alignment) is approximated by $\mathbf{\Sigma}^{(i)}$. The following assumption, which is slightly stronger than Assumption 3, provides a sufficient condition to derive the asymptotic joint distribution of such entries of $\widehat{\mathbf{R}}^{(i)}$.

Assumption 4 (Score matrix covariance)**.** *The magnitude of the smallest eigenvalue of $\mathbf{\Sigma}^{(i)}$ satisfies $|\lambda_{\min}(\mathbf{\Sigma}^{(i)})| = \omega(n^{-2})$.*

The next theorem presents the asymptotic distribution of the estimated score matrices as the size of the graphs increases. The proof is given on the Appendix.

Theorem 11. *Suppose* $(\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(m)}) \sim \text{COSIE}\left(\mathbf{V}; \mathbf{R}^{(1)}, \dots, \mathbf{R}^{(m)}\right)$ *are a sample of adjacency matrices from the COSIE model such that the parameters satisfy* $\min_{i\in[m]} \delta(\mathbf{P}^{(i)}) = \omega(\log n)$, $\varepsilon = O\left(\frac{1}{\max \sqrt{\delta(\mathbf{P}^{(i)})}}\right)$, *as well as the delocalization requirements given in Assumption 2.*

a) Suppose that Assumption 3 holds. Let $\mathbf{W} \in \mathcal{O}_d$ *be an orthogonal matrix such that* $\mathbf{W} = \operatorname{argmin}_{\mathbf{W}\in\mathcal{O}_d} \|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F$. *Then, the estimates of* $\{\mathbf{R}^{(i)}\}_{i=1}^m$ *provided by MASE (Algorithm 1),* $\{\widehat{\mathbf{R}}^{(i)}\}_{i=1}^m$, *satisfy*

$$\left(\mathbf{\Sigma}^{(i)}_{\frac{2k+l(l-1)}{2}, \frac{2k+l(l-1)}{2}}\right)^{-1/2} \left(\mathbf{W}\widehat{\mathbf{R}}^{(i)}\mathbf{W}^\top - \mathbf{R}^{(i)} + \mathbf{H}_m^{(i)}\right)_{kl} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$, *where* $\mathcal{N}(0,1)$ *is a univariate standard normal distribution,* $\mathbf{H}_m^{(i)}$ *is a random matrix that satisfies* $\mathbb{E}[\|\mathbf{H}_m^{(i)}\|_F] = O(\frac{d}{\sqrt{m}})$, *and and* $\mathbf{\Sigma}^{(i)}_{\frac{2k+l(l-1)}{2}, \frac{2k+l(l-1)}{2}} \sim \frac{s^2(\mathbf{P}^{(i)})}{n^2}$.

b) If in addition to the requirements in a), Assumption 4 holds, then

$$\left(\mathbf{\Sigma}^{(i)}\right)^{-1/2} \operatorname{vec}\left(\mathbf{W}\widehat{\mathbf{R}}^{(i)}\mathbf{W}^\top - \mathbf{R}^{(i)} + \mathbf{H}_m^{(i)}\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}_r, \mathbf{I}_r),$$

where $\mathcal{N}(\mathbf{0}_r, \mathbf{I}_r)$ *is an* r*-dimensional standard normal distribution.*

In addition to Assumptions 2, 3 and 4, Theorem 11 imposes slightly stronger assumptions on ε than Theorem 7 to obtain an asymptotically normal distribution for the entries of $\mathbf{R}^{(i)}$.

Due to the non-identifiability of the specific matrix $\mathbf{V}$, the results in Theorem 11 about each estimated score matrix $\widehat{\mathbf{R}}^{(i)}$ are stated in terms of an orthogonal matrix $\mathbf{W}$. The next corollary shows that this unknown matrix $\mathbf{W}$ can be removed and each individual entry of the matrix $\widehat{\mathbf{R}}^{(i)}$ also has an asymptotic normal distribution with a unspecified but bounded variance.

Corollary 12. *Suppose that the assumptions of Theorem 11 part b) hold. Then, there exists a sequence of orthogonal matrices* $\mathbf{W} \in \mathcal{O}_d$ *such that*

$$\frac{1}{\sigma_{i,k,l}}(\widehat{\mathbf{R}}^{(i)} - \mathbf{W}^\top\mathbf{R}^{(i)}\mathbf{W} + \mathbf{H}_m^{(i)})_{kl} \xrightarrow{d} \mathcal{N}(0,1),$$

where $\mathbb{E}[\|\mathbf{H}_m^{(i)}\|_F] = O\left(\frac{d}{\sqrt{m}}\right)$ *and* $\sigma^2_{i,k,l} = O\left(\frac{d^2 s^2(\mathbf{P}^{(i)})}{n^2}\right) = O(1)$.

Using Corollary 12 in combination with a tail bounds of a normal distribution, for sufficiently large n, the mean square error between the estimated score matrix and its expectation satisfies

$$\|\widehat{\mathbf{R}}^{(i)} - \mathbf{W}^\top\mathbf{R}^{(i)}\mathbf{W} + \mathbf{H}_m\|_F = O_P\left(\frac{d^2\sqrt{s^2(\mathbf{P}^{(i)})}}{n}\right). \tag{9}$$

Because $s^2(\mathbf{P}^{(i)}) = O(n^2)$, the right hand side of Equation (9) is at most of an order $O_P(d^2)$. On the other hand, a lower bound for the norm of the score matrices can be obtained by observing that

$$\|\mathbf{R}^{(i)}\|_F^2 = \|\mathbf{P}^{(i)}\|_F^2 = \sum_{k=1}^d \left(\lambda_k(\mathbf{P}^{(i)})\right)^2.$$

The eigenvalues of $\mathbf{P}^{(i)}$ are of order $\delta(\mathbf{P}^{(i)})$ (see Remark 24 in Athreya et al. (2017)), and hence, under the assumptions of Theorem 11, $\|\mathbf{R}^{(i)}\|_F = \omega(\log n)$. Therefore, the relative mean square

error of the estimated score matrices with respect to its expectation goes to zero as n increases, and if the sample size m is also growing, the estimation bias $\mathbf{H}_m$ vanishes.

To illustrate the result, we consider the problem of estimating the eigenvalues of a random graph under the COSIE model. For that goal, we generate a sample of m adjacency matrices from the COSIE model, and use the eigenvalues of $\widehat{\mathbf{R}}^{(1)}$ as an estimate of the eigenvalues of $\mathbf{P}^{(1)}$. According to Theorem 11, the entries of $\widehat{\mathbf{R}}^{(1)}$ are asymptotically normally distributed and centered around a bias term $\mathbf{H}$ that vanishes as the sample size grows. We conjecture that there is a similar phenomenon in the eigenvalues of $\mathbf{R}^{(1)}$; namely, that the eigenvalues of $\widehat{\mathbf{R}}^{(i)}$ are asymptotically normally distributed about the eigenvalues of $\mathbf{P}$, albeit with some bias. More formally, we conjecture that

$$\lambda_k(\widehat{\mathbf{R}}^{(i)}) - \lambda_k(\mathbf{P}^{(i)}) - \eta_m \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

where $\eta_m = O_P(1/\sqrt{m})$. These results dovetail nicely the theory of Tang (2018), which presents a normality result, with bias, for the eigenvalues of stochastic blockmodel graphs. To illustrate the bias-reduction impact of multiple graphs in our case, consider the problem of eigenvalue estimation in both the single and multiple-graph setting. Figure 2 shows that as the sample size increases, the distribution of the two leading eigenvalues of a SBM graph estimated with MASE approaches the true eigenvalues. When the number of vertices increases but the sample size remains constant, this is not the case, as observed in Figure 3.

5. Simulations

In this section, we study the empirical performance of MASE for estimating the parameters of the COSIE model, as well the performance of this method in subsequent inference tasks, including community detection, graph classification and hypothesis testing. In all cases, we compare with state-of-the-art models and methods for multiple-graph data as baselines. The results show that the COSIE model is able to effectively handle heterogeneous distributions of graphs and leverage the information across all of them, and demonstrates a good empirical performance in different settings.

5.1 Subspace Estimation Error

We first study the performance in estimating the common subspace $\mathbf{V}$ by using the estimator $\widehat{\mathbf{V}}$ obtained by MASE in a setting where the number of graphs m increases. Given a pair of matrices with orthonormal columns $\mathbf{V}$ and $\widehat{\mathbf{V}}$, we measure the distance between their invariant subspaces via the spectral norm of the difference between the projections, given by $\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top\|$. This distance is zero only when there exist an orthogonal matrix $W \in \mathcal{O}_d$ such that $\widehat{\mathbf{V}} = \mathbf{V}\mathbf{W}$.

Given a sample of graphs $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}$, the ASE of the average of the graphs, given by

$$\bar{\mathbf{A}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{A}^{(i)},$$

provides a good estimator of the invariant subspace when all the graphs have the same expectation matrix (Tang et al., 2019). However, when each graph has a different expected value, $\bar{\mathbf{A}}$ approaches to the population average, which might not contain useful information about the structure of the invariant subspace. MASE, on the other hand, is able to handle heterogeneous structure, and does not require such strict assumptions. We compare this method with the performance of MASE these two settings, i.e., equal and different distribution of the graphs.

We simulate graphs from a three-block multilayer stochastic blockmodel with membership matrix $\mathbf{Z} \in \{0,1\}^{n\times 3}$ and connectivity matrices $\mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(m)}$. The number of vertices is fixed to $3^6 = 729$, and equal-sized communities. We simulate two scenarios. In the first scenario, all graphs have the

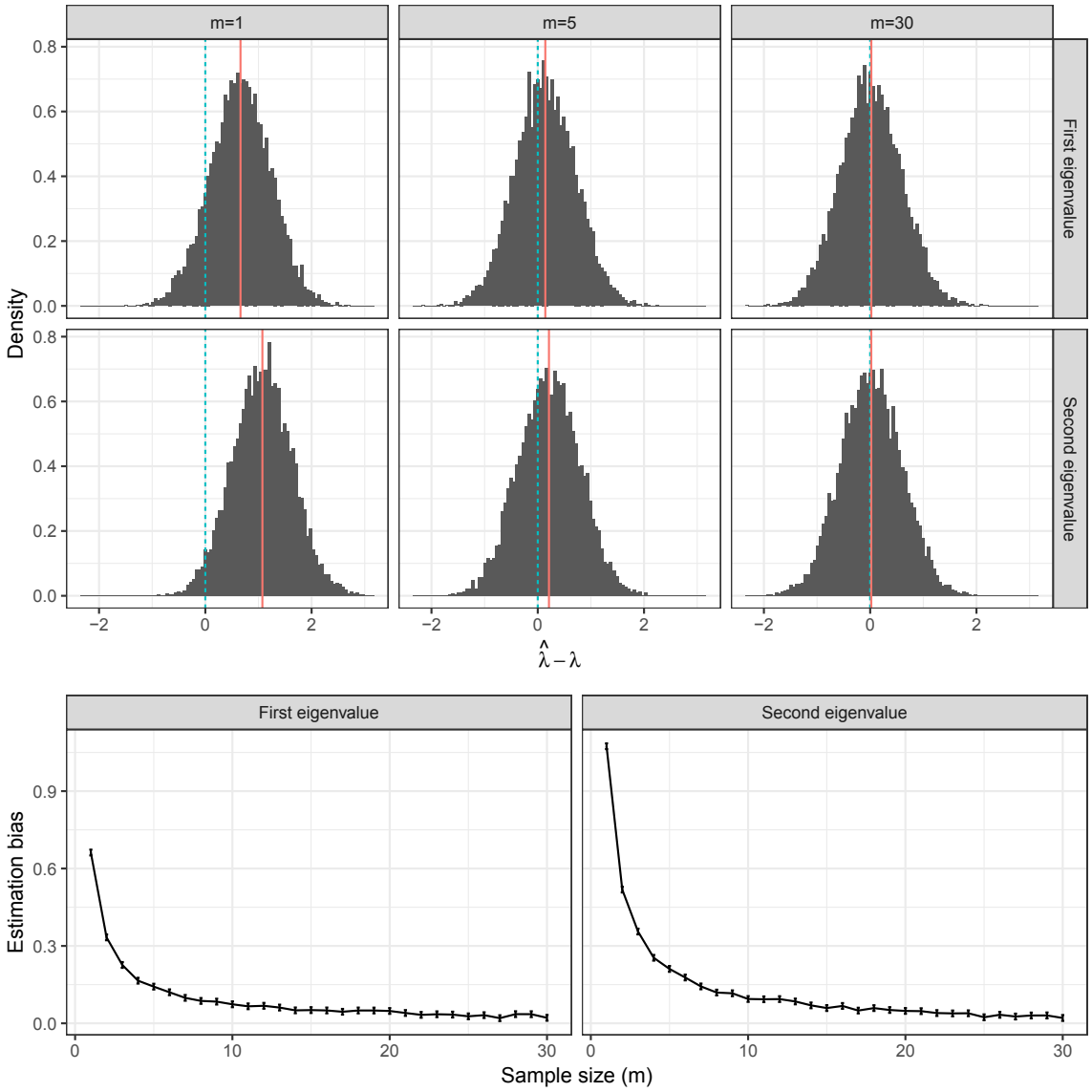


Figure 2: The top panel shows the distribution of the difference between the MASE estimates and true eigenvalues of a single graph. The graphs in the sample are distributed according to a two-block multilayer SBM with $n = 300$ vertices, block connection probability matrix $\mathbf{B}^{(i)} = 0.3\mathbf{I} + 0.1\mathbf{1}\mathbf{1}^\top$. The distributions appear to be Gaussian, and as the sample size m increases, the mean of the distribution (red solid line) approaches to zero (blue dashed line). This phenomenon is also observed in the bottom panel, which shows that the estimation bias decreases with the sample size.

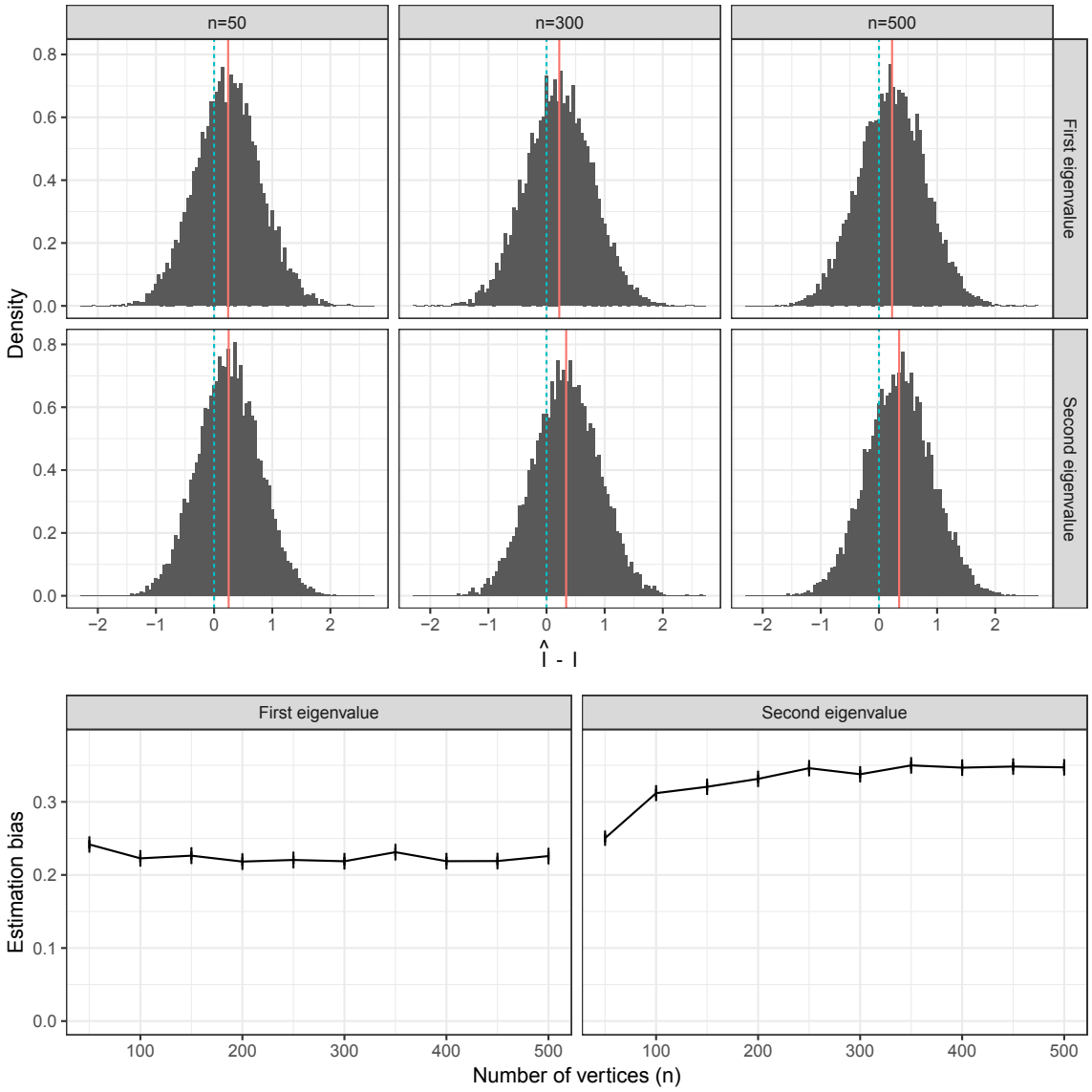


Figure 3: The top panel shows the distribution of the difference between the MASE estimates and true eigenvalues of a single graph. The graphs in the sample are distributed according to a two-block multilayer SBM, block connection probability matrix $\mathbf{B}^{(i)} = 0.3\mathbf{I} + 0.1\mathbf{1}\mathbf{1}^\top$, and a fixed sample size $m = 3$ but different number of vertices n. The estimation bias remains positive even with a large graph size (bottom panel), and for any fixed n the mean of the distributions (red solid line) is away from zero (blue dashed line), but the distributions appear to be Gaussian.

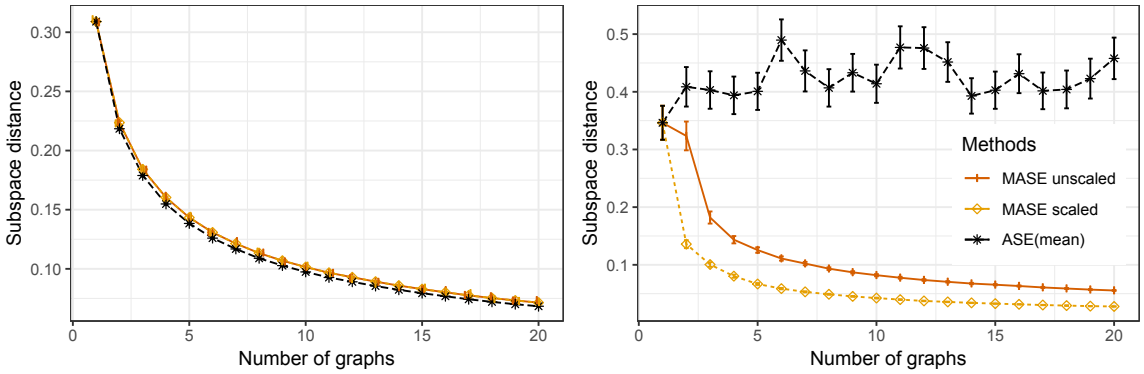


Figure 4: Distance between the true and estimated invariant subspaces computed with MASE (scaled and unscaled ASE) and ASE of the mean of the graphs for a sample of graphs distributed according to a multilayer SBM with 3 communities. On the left panel, all graphs have the same expected matrix. The three methods perform almost the same, reducing the error as the sample size increases. On the right panel, the connection probabilities of the SBM are chosen uniformly at random for each graph. ASE on the average adjacency matrix performs poorly, while MASE still improves with a larger sample size.

same connectivity, given by

$$\mathbf{B}^{(i)} = \begin{pmatrix} .4 & .1 & .1 \\ .1 & .4 & .2 \\ .1 & .2 & .3 \end{pmatrix}, \qquad \forall i \in [m].$$

In the second scenario, each matrix is generated randomly by independently sampling the entries as $\mathbf{B}^{(i)}_{uv} \sim U(0,1)$, for each $1 \leq u \leq v \leq 3$ and $i \in [m]$.

We then compare the performance of the different methods as the number of graphs increases. We consider both the unscaled and scaled versions of MASE, and ASE on the mean of the adjacency matrices (ASE(mean)). Figure 4 shows the average subspace estimation error of several Monte Carlo experiments (25 simulations in the first scenario, and 100 in the second). In the first scenario, the three methods exhibit a similar performance, decreasing the error as a function of the sample size, and although ASE(mean) has a slightly better performance, the performance of MASE is statistically not worse than ASE(mean). In the second scenario, when all the graphs have different expected values, ASE(mean) performs poorly, even with a large number of graphs, while MASE mantains a small error and is still able to improve the performance with larger sample size. These results corroborate our theory that MASE is effective in leveraging the information across multiple heterogeneous graphs, and it performs similarly to the ASE on the average of the graphs when the sample has a homogeneous distribution.

The results of Figure 4 also show that in some circumstances the scaled MASE can perform significantly better than its unscaled counterpart in estimating the common invariant subspace. When some eigenvalues of a matrix $\mathbf{B}^{(i)}$ are close to zero, the estimation of the eigenvectors corresponding to the smallest non-zero eigenvalues is difficult. The eigenvalue scaling in ASE can help MASE in discerning the dimensions of $\widehat{\mathbf{V}}^{(i)}$ that are more accurately estimated, which might explain the improved performance of the scaled MASE in this simulation. Our current theoretical results only address the unscaled MASE, but the simulations encourage extensions of the analysis to the scaled MASE.

5.2 Modeling Heterogeneous Graphs

Now, we study the performance of MASE in modeling the structure of graphs with different distributions. First, simulated graphs are distributed according to the multilayer SBM with two blocks, but now each graph i has an associated class label $y_i \in \{1, 2, 3, 4\}$. The connection matrices $\mathbf{B}^{(i)}$ have four different classes depending on the label, so if $y_i = k$ then

$$\mathbf{B}^{(i)} = 0.25(\mathbf{1}\mathbf{1}^\top) + \alpha \mathbf{C}^{(k)},$$

where $\alpha \geq 0$ is a constant that controls the separation between the classes, and the matrices $\mathbf{C}^{(k)}$ are defined as

$$\mathbf{C}^{(1)} = \begin{pmatrix} .1 & 0 \\ 0 & .1 \end{pmatrix} \qquad \mathbf{C}^{(2)} = -\begin{pmatrix} .1 & 0 \\ 0 & .1 \end{pmatrix}$$
$$\mathbf{C}^{(3)} = \begin{pmatrix} .1 & 0 \\ 0 & 0 \end{pmatrix} \qquad \mathbf{C}^{(4)} = \begin{pmatrix} 0 & 0 \\ 0 & .1 \end{pmatrix}.$$

This choice of graphs allows to simultaneously study the effects on the difference between the classes and the smallest eigenvalue of the score matrices. When $\alpha = 0$, all score matrices are the same, but also the smallest eigenvalue is zero, so both subspace estimation and graph classification are hard problems, and as α increases, the signal for these problems improves.

We simulate 40 graphs, with 10 on each class, all with $n = 256$ vertices and equal-sized communities. We compare the performance of MASE with respect to other latent space approaches for modeling multiple graphs: joint embedding of graphs (JE) (Wang et al., 2019c), multiple RDPG (MRDPG) (Nielsen and Witten, 2018), and the omnibus embedding (OMNI) (Levin et al., 2017). The first two methods, JE and MRDPG, are based on a model related to COSIE. In particular, the expected value of each adjacency matrix is described as

$$\mathbf{P}^{(i)} = \mathbf{H}\mathbf{\Lambda}^{(i)}\mathbf{H}^\top,$$

where $\mathbf{H}$ is a $n \times d$ matrix, and $\mathbf{\Lambda}$ is a $d \times d$ diagonal matrix. In JE, the matrix $\mathbf{H}$ is only restricted to have columns with norm 1, which makes the model non-identifiable, while MRDPG imposes further restrictions on $\mathbf{H}$ to be orthogonal and $\mathbf{\Lambda}$ to have nonnegative entries. Both models are fitted by optimizing a least squares loss function to obtain estimates $\widehat{\mathbf{H}}$ and $\{\widehat{\mathbf{\Lambda}}^{(i)}\}_{i=1}^m$. The omnibus embedding obtains estimates for individual latent positions for each graph by jointly embedding the graphs into the same space. Given a sample of m adjacency matrices, the omnibus embedding obtains an omnibus matrix $\mathbf{M} \in \mathbb{R}^{mn \times mn}$, in which each off-diagonal submatrix is the average of each pair of graphs in the sample, so that

$$\mathbf{M}_{u+(i-1)m, v+(j-1)m} = \frac{\left(\mathbf{A}^{(j)}_{uv} + \mathbf{A}^{(i)}_{uv}\right)}{2}.$$

The omnibus embedding is generated by the ASE of $\mathbf{M}$, and each of the ith blocks of n rows of the embedding corresponds to the estimated latent positions of the ith graph, such that

$$\left((\widehat{\mathbf{X}}^{(1,\text{OMNI})})^\top, \ldots, (\widehat{\mathbf{X}}^{(m,\text{OMNI})})^\top\right)^\top = \text{ASE}(\mathbf{M}).$$

For MASE and OMNI, we use an embedding dimension $d = 2$; JRDPG and MRDPG represent the scores of each graph with a diagonal matrix, so in order to capture all the variability in the data we use $d = 3$ for the embedding dimension. In all comparisons, the two versions of MASE (scaled and unscaled ASE) performed very similarly, and therefore we only report the results of unscaled MASE.

5.2.1 Semi-supervised Graph Classification

First, we measure the accuracy of these methods in graph classification. We use the simulated graphs as training data; for test data, we generate a second set of graphs with the same characteristics as the first. In order to determine classification accuracy, we first use each method to obtain an embedding of all the graphs, including training and test data. We then use the Frobenius distance between the individual score matrices of each graph to build a 1-nearest neighbor classifier for the test data, and use the labels of the training data to classify the test data. The individual score matrices correspond to $\{\widehat{\mathbf{R}}^{(i)}\}_{i=1}^{m}$ in MASE, $\{\widehat{\mathbf{H}}^{(i)}\}_{i=1}^{m}$ in JE and MRDPG, and $\{\widehat{\mathbf{X}}^{(i,\text{OMNI})}\}_{i=1}^{m}$ in OMNI. Note that both training and test data are used to generate the estimates of the score matrices; in other words, to produce an *unsupervised* joint embedding for all the graphs. Such a joint embedding allows to avoid cumbersome Procrustes alignments that are required by the OMNI and the MRDPG procedures when embedding test and training data separately; neither OMNI and MRDPG have established out-of-sample extensions for their respective embeddings. MASE can generate the embedding without using the test data by constructing the invariant subspace matrix $\widehat{\mathbf{V}}$ only using the training data, but we use the semi-supervised procedure described before for consistency between the different methods.

Figure 5 shows the average classification accuracy as a function of α for 50 Monte Carlo simulations. For all methods, the accuracy increases as the separation between each class model increases with α, and MASE performs best among all the methods. This result is not surprising, considering the fact that MASE has better flexibility to model heterogeneous graphs. The performance of OMNI is also excellent, but it is worth noting that OMNI employs a much larger number of parameters $(n \times d)$ to represent each graph. JE is based on a model that can also represent each expected adjacency correctly, but the non-convexity and over-parametertization of the objective function can make the optimization problem challenging, and the performance depends on the initial value, which is random, resulting in inferior performance on average. The identifiability constraints of MRDPG limit the type of graphs that this model can represent, and this method is never able to separate two of the classes correctly, resulting in a limited performance in practice.

5.2.2 Model Estimation

We also compare the performance of different methods in terms of approximating the probability matrices of the model. For each method and each graph, we obtain an estimate $\widehat{\mathbf{P}}^{(i)}$ of the generative probability matrix of the model $\mathbf{P}^{(i)}$, and measure the model estimation error with the normalized mean squared error as

$$L(\widehat{\mathbf{P}}^{(i)}, \mathbf{P}^{(i)}) = \frac{\|\widehat{\mathbf{P}}^{(i)} - \mathbf{P}^{(i)}\|_F}{\|\mathbf{P}^{(i)}\|_F},$$

and report the average estimation error over all the graphs. Figure 6 reports the average Monte Carlo estimation error as a function of α. Again, we observe that MASE has the best performance among all the methods. This is not surprising, considering that MASE is designed for this model, but it shows the limitations of the other methods. JE also shows an improvement in estimation as the separation between the classes increases, since this is the only other method that can represent correctly the four classes, but as in the classification task, the variance is larger. While OMNI can discriminate the graphs correctly, as previously observed, the error in estimating the generative probability matrix is large. MRDPG is again not able to succeed due to the model limitations.

5.2.3 Community Detection

We use the joint latent positions obtained by each method to perform community detection, by clustering the rows of each embedding using a Gaussian mixture model (GMM). For MASE, we use the matrix $\widehat{\mathbf{V}}$ as the estimated joint latent positions, while for JE and MRDPG we use the matrix $\widehat{\mathbf{H}}$. For OMNI, we take the average of the individual latent positions of the graphs $\sum_{i=1}^{m} \widehat{\mathbf{X}}^{(m,\text{OMNI})}$.

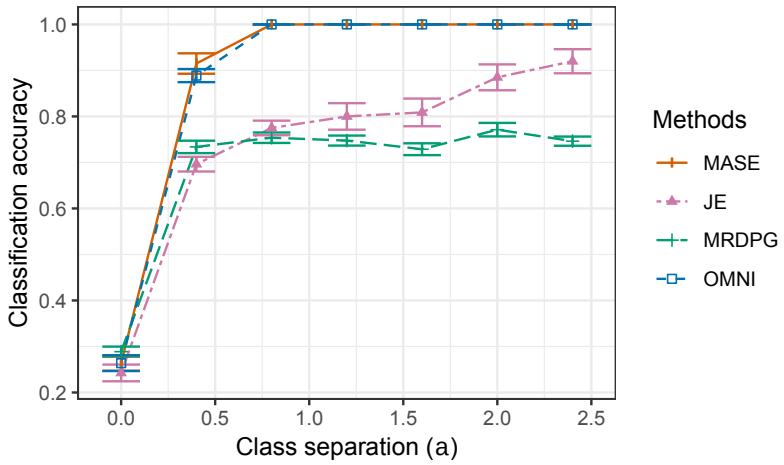


Figure 5: Out-of-sample classification accuracy as a function of the difference between classes for different graph embedding methods. Graphs in the sample are distributed as a multilayer SBM with $n = 256$ vertices, two communities, four different connectivity matrices that correspond to the class labels, and a training and test samples with $m = 40$ graphs. As the separation between the classes increase, all methods improve performance, but MASE and OMNI show the most gains.

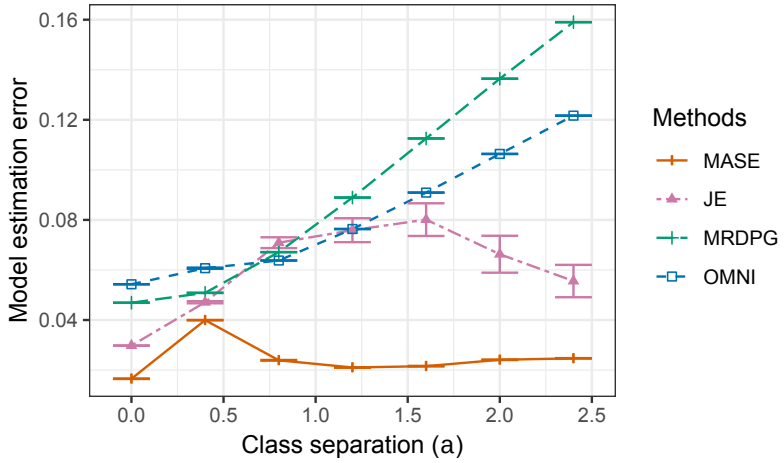


Figure 6: Average normalized mean squared error of estimating the expected adjacency matrix of a sample of graphs using different embedding methods. Graphs in the sample are distributed as a multilayer SBM with $n = 256$ vertices, two communities, four different classes of connectivity matrices, and a training and test samples with $m = 40$ graphs. MASE and JE are the only methods that are flexible enough to capture the heterogeneous structure of the graphs, but MASE shows superior performance among these two. As the graphs become more different, the error of MRDPG and OMNI increases.

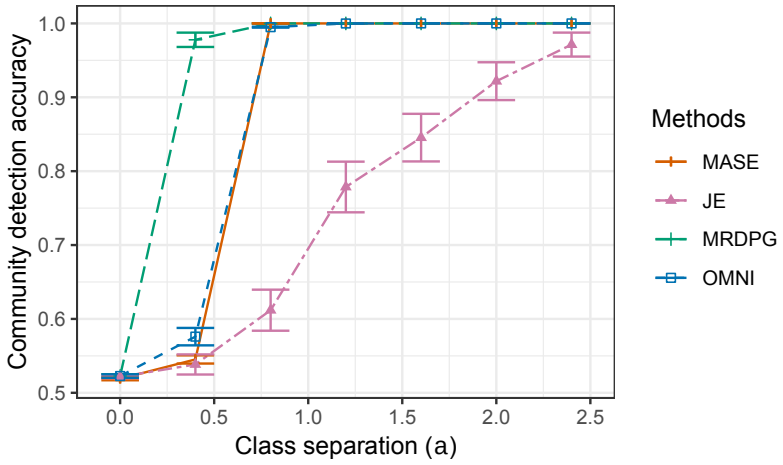


Figure 7: Performance in community detection of a Gaussian mixture clustering applied to a common set of vertex latent positions estimated with different methods, for a sample of $m = 40$ two-block multilayer SBM graphs, with $n = 256$ vertices, and four classes of connectivity matrices. The class separation controls the magnitude of the smallest eigenvalue of the graphs, and as this increases, all methods show better accuracy. MRDPG, which is based on a non-convex optimization problem, performance the best for small α, but as long as α is large enough, MASE and OMNI also show a great performance.

In all cases, because the number of communities is known, we fit two clusters using GMM, and compare with the true communities of the multilayer SBM.

The average accuracy in clustering the communities is reported in Figure 7. Spectral clustering methods require to have enough separation on the smallest non-zero eigenvalue. This is the case for MASE and OMNI, which show poor performance with small α, but once the signal is strong enough, both methods perform excellent. MRDPG shows a good performance, even for small α, which could be due to the optimization procedure employed by the method in fitting $\mathbf{V}$. As in all the other tasks, JE shows a high variance in performance.

5.3 Graph Hypothesis Testing

Our goal is to test the hypothesis that for a given pair of random adjacency matrices $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$, the underlying probability matrices $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$ are the same. Using the COSIE model, for any pair of matrices $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$, there exists an embedding dimension d and a matrix with orthonormal columns $\mathbf{V} \in \mathbb{R}^{n \times d}$ such that $\mathbf{P}^{(i)} = \mathbf{V}\mathbf{R}^{(i)}\mathbf{V}^\top$. Therefore, our framework reduces the problem to testing the hypothesis $H_0 : \mathbf{R}^{(1)} = \mathbf{R}^{(2)}$. We evaluate the performance of our method and compare it with the omnibus embedding (OMNI) method of Levin et al. (2017), which is one of the few principled methods for this problem in latent space models.

We proceed by generating $\mathbf{A}^{(k)} \sim \text{Ber}(\mathbf{P}^{(k)})$ with $\mathbf{P}^{(k)} = \mathbf{Z}^{(k)}\mathbf{B}^{(k)}\mathbf{Z}^{(k)^\top}$ a mixed membership SBM (MMSBM) with three communities (Airoldi et al., 2007), such that for each $i = 1, \ldots, n$, the row that corresponds to vertex i is distributed as $\mathbf{Z}^{(k)}_{i\cdot} \sim \text{Dir}(0.1, 3)$. The matrix $\mathbf{B}^{(1)} \in \mathbb{R}^{3\times 3}$ of the first model is fixed as $\mathbf{B}^{(1)} = 0.3\mathbf{I} + 0.1(\mathbf{1}\mathbf{1}^\top)$, where $\mathbf{1}$ is the 3-dimensional vector with all entries equal to one. We simulate two scenarios.

1. Same community assignments, but different connectivity matrices: fix the block memberships in both graphs to be the same, $\mathbf{Z}^{(1)} = \mathbf{Z}^{(2)}$, and vary $\|\mathbf{B}^{(1)} - \mathbf{B}^{(2)}\|_F$ by increasing $\mathbf{B}^{(2)}_{11}$ while keeping all the entries of $\mathbf{B}^{(2)}$ equal to $\mathbf{B}^{(1)}$.

2. Different community assignments for some of the vertices, same connectivity matrices: fix $\mathbf{B}^{(2)} = \mathbf{B}^{(1)}$, and sample the memberships of the first t vertices independently in the two graphs, while keeping the other $n - t$ memberships equal: that is, $\{\mathbf{Z}^{(1)}_{i\cdot}, \mathbf{Z}^{(2)}_{j\cdot} : i \in [n], j \in [t]\}$ are i.i.d Dir$((0.1, 3))$, and $\mathbf{Z}^{(2)}_{k\cdot} = \mathbf{Z}^{(1)}_{k\cdot}$ for $k > t$.

The first scenario can be represented as a COSIE model with dimension $d = d_1 = d_2 = 3$, where d_i is the rank of the matrix $\mathbf{P}^{(i)}$, but for the second one, to exactly represent these graphs in the COSIE model, we use $d_1 = d_2 = 3$ and $d = 5$.

To test the null hypothesis, we use the square Frobenius norm of the difference between the score matrices $\|\widehat{\mathbf{R}}^{(1)} - \widehat{\mathbf{R}}^{(2)}\|^2_F$. A similar test statistic is used for OMNI by calculating the distance between the estimated latent positions $\|\widehat{\mathbf{X}}^{(1,\text{OMNI})} - \widehat{\mathbf{X}}^{(2,\text{OMNI})}\|^2_F$. For both MASE and OMNI, we estimate the exact distribution of the test statistic via Monte Carlo simulations using the correct expected value of each graph $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$. For MASE, we also evaluate the performance of empirical p-values calculated using a parametric bootstrap approach, or the asymptotic null distribution of the score matrices, which are described below.

a) A semiparametric bootstrap method introduced in Tang et al. (2017a) uses estimates $\widetilde{\mathbf{P}}^{(1)}$ and $\widetilde{\mathbf{P}}^{(2)}$ (constructed by the ASE of $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$) in place of the true matrices $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$—the latter of which may not, of course, even been known in practice. A total of 1000 independent pairs of adjacency matrices $\widetilde{\mathbf{A}}^{(1)}$ and $\widetilde{\mathbf{A}}^{(2)}$ are generated by fixing the distribution of the two graphs to $\widetilde{\mathbf{P}}^{(1)}$ for the first 500 pairs, and to $\widetilde{\mathbf{P}}^{(2)}$ for the remaining pairs, and these pairs are used to approximate the empirical null distribution of the test statistic, and calculate the p-value as the proportion of bootstrapped test statistic values that are smaller than the observed one.

b) When $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ are both sampled from $\mathbf{P}^{(1)}$, Theorem 11 suggests that the distribution of the difference between the entries of the estimated score matrices (after a proper orthogonal alignment) is approximated by a multivariate normal, with a mean proportional to the bias term $\mathbf{H}^{(1)}$ and covariance depending on $\mathbf{\Sigma}^{(1)}$. Letting $\widehat{\mathbf{P}}^{(1)} = \widehat{\mathbf{V}}\widehat{\mathbf{R}}^{(1)}\widehat{\mathbf{V}}^\top$ be the estimate for $\mathbf{P}^{(1)}$ obtained by MASE, an estimate $\widehat{\mathbf{\Sigma}}^{(1)}$ of the asymptotic covariance in Theorem 11 can be derived from Equation (8) by plugin in $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{P}}^{(1)}$. Neglecting the effect of the bias term $\mathbf{H}^{(1)}$ (which vanishes as the number of graphs increases), we use a random variable $\mathbf{y} \sim \mathcal{N}(\mathbf{0}_r, 2\widehat{\mathbf{\Sigma}}^{(1)})$ to approximate the null distribution of $\text{vec}(\mathbf{W}(\widehat{\mathbf{R}}^{(1)} - \widehat{\mathbf{R}}^{(2)})\mathbf{W}^\top)$ (with $\mathbf{W}$ as defined in Theorem 11). Therefore, the null distribution of the test statistic $\|\widehat{\mathbf{R}}^{(1)} - \widehat{\mathbf{R}}^{(2)}\|^2_F = \|\mathbf{W}(\widehat{\mathbf{R}}^{(1)} - \widehat{\mathbf{R}}^{(2)})\mathbf{W}^\top\|^2_F$ is approximately distributed as a generalized chi-square distribution (since it is a function of the squared entries of $\mathbf{y}$), or it can be estimated using Monte Carlo simulations of $\mathbf{y}$. The same process is repeated by using $\widehat{\mathbf{P}}^{(2)}$, after which the empirical p-value can be estimated using a mixture of the empirical distributions obtained from $\widehat{\mathbf{P}}^{(1)}$ and $\widehat{\mathbf{P}}^{(2)}$.

Figure 8 shows the result in terms of power for rejecting the null hypothesis that the two graphs have equal distribution (the confidence level was chosen as 0.05). As the separation between null and alternative increases, the result from both methods approach full power when the p-value is calculated using the exact null distribution (MASE and OMNI). The first scenario shows an advantage for MASE, which exploits both the fact that the common subspace is well represented on each graph and that there are a smaller number of parameters to estimate than in the omnibus model. The second scenario is more challenging for MASE since the number of parameters is increased, but it still performs competitively. In both cases, the p-values obtained with the bootstrap method (MASE-bootstrap) perform similarly to the empirical power when using the true $\mathbf{P}$ matrices.

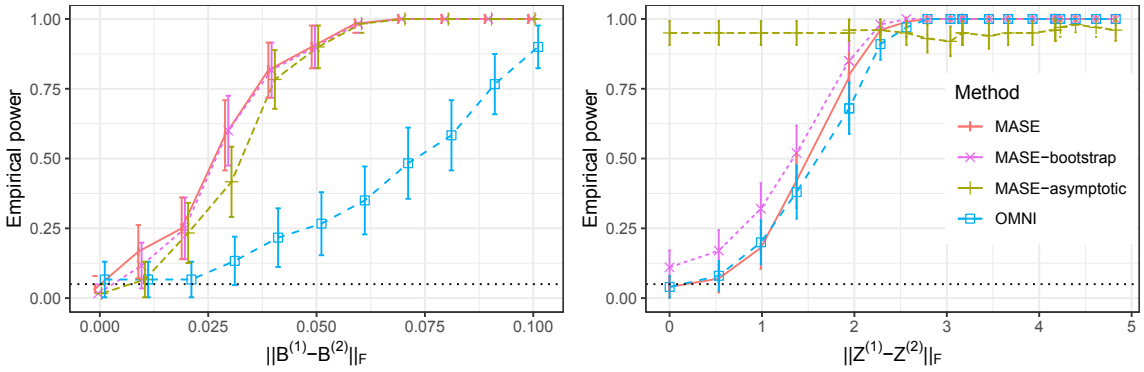


Figure 8: Empirical power for rejecting the null hypothesis that two graphs have the same distribution as a function of the difference in their parameters at a 0.05 significance level (black dotted line). Both graphs are sample from a mixed-membership SBM $\mathbf{A}^{(i)} \sim \text{MMSBM}(\mathbf{Z}^{(i)}, \mathbf{B}^{(i)})$. In the left panel, the community assignments $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ are the same, while the connection probability changes on the x-axis. In the right panel, $\mathbf{B}^{(1)} = \mathbf{B}^{(2)}$, while the community assignments of a few nodes change. Both methods improve their power as the difference between the graphs get larger, but MASE can effectively use the common subspace structure to outperform OMNI when the difference between the $\mathbf{B}$ matrices is small. The p-values of MASE are estimated accurately usign a parametric bootstrap, while the asymptotic distribution is only valid in the first scenario when the common invariant subspace is represented in both graphs.

The asymptotic distribution of the score matrices provides a computationally efficient way to estimate the p-values, but this approach relies on the validity of the assumptions of Theorem 11 and how small the magnitude of the bias term is, which depends on the estimation error of $\mathbf{V}$. Having only two graphs to estimate the parameters of the model, there is no guarantee that this bias term will be sufficiently small. Nevertheless, in the first scenario (Figure 8, left panel), the estimated power using the asymptotic distribution (MASE-asymptotic) provides an accurate result; here, $\mathbf{V}$ is well represented in both graphs, which helps in reducing its estimation error. In the second scenario (Figure 8, left panel), the dimension of $\mathbf{V}$ ($d = 5$) is higher than the rank of the expectation of each graph ($d_1 = d_2 = 3$), hence, the score matrices are not full rank, affecting the magnitude of the bias term and invalidating some conditions of Theorem 11. As a result, the estimated p-values with this method appear to be invalid and significantly inflated on this scenario.

6. Modeling the Connectivity of Brain Networks

We evaluate the ability of our method to characterize differences in brain connectivity between subjects using a set of graphs constructed from diffusion magnetic resonance imaging (dMRI). The data corresponds to the HNU1 study (Zuo et al., 2014) which consists of dMRI records of 30 different healthy subjects that were scanned ten times each over a period of one month. Based on these recordings, we constructed 300 graphs (one per subject and scan) using the NeuroData's MRI to Graphs (NDMG) pipeline (Kiar et al., 2018). The vertices were registered to the CC200 atlas (Craddock et al., 2012), which identified 200 vertices in the brain.

We first apply our multiple adjacency spectral embedding method to the HNU1 data. To choose the dimension of the embedding for each individual graph, we use the automatic scree plot selection method of Zhu and Ghodsi (2006), which chooses between 5 and 18 dimensions for each graph (see

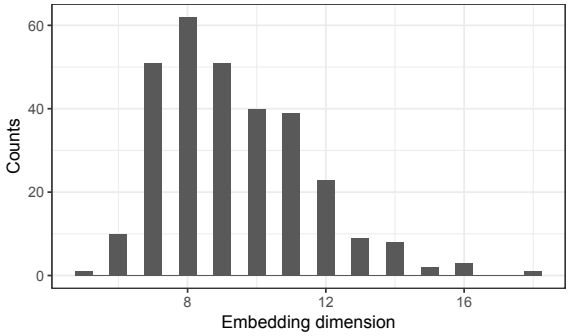


Figure 9: Histogram of the embedding dimension selected by the method of Zhu and Ghodsi (2006) for each of the 300 graphs in the HNU1 data. The graphs are composed by 200 vertices, and the values of the embedding dimensions ranges between 5 and 18.

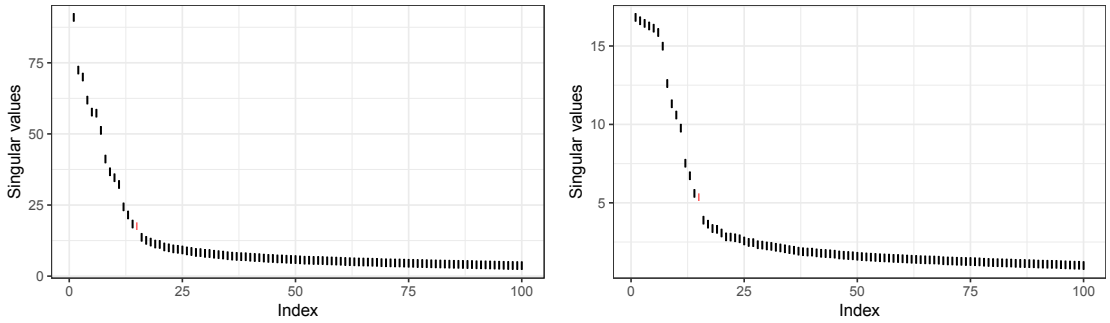


Figure 10: Scree plot of the top singular values of the concatenated scaled (left) and unscaled (right) adjacency spectral embeddings of the HNU1 data graphs. The plots show the 100 largest singular values ordered by magnitude. In both scaled and unscaled methods, we identified an elbow at $d = 15$.

Figure 9). The joint model dimension for MASE is chosen based on the scree plot of the singular values of the matrix of concatenated adjacency spectral embeddings. We perform the method using both the scaled and unscaled versions of ASE. In both cases, we observe an elbow on the scree plot at $d = 15$ (see Figure 10), and thus we selected this value for our model.

When we apply the MASE method to the HNU1 data, we obtain a matrix of joint latent positions $\widehat{\mathbf{V}} \in \mathbf{R}^{200\times 15}$ and a set of symmetric matrices $\{\widehat{\mathbf{R}}^{(i)}\}_{i=1}^{300}$ of size 15×15 for each individual graph, yielding 120 parameters to represent each graph. Figure 11 shows a three-dimensional plot of the latent positions of the vertices for $\widehat{\mathbf{V}}$. Most of the vertices of the CC200 atlas are labeled according to their spatial location as left or right hemisphere, and this structure is reflected in Figure 11—in fact, the plane $v_3 = 0$ appears to be a useful discriminant boundary between left and right hemisphere. This is additional fortifying evidence that the embedding obtained by MASE is anatomically meaningful.

The individual graph parameters $\{\widehat{\mathbf{R}}^{(i)}\}_{i=1}^{300}$ of dimension 15×15 are difficult to interpret because their values are identifiable only with respect to $\widehat{\mathbf{V}}$, but also because of the large dimensionality. Nevertheless, the matrix of Frobenius distances $\mathbf{D} \in \mathbf{R}^{300\times 300}$ such that $\mathbf{D}_{ij} = \|\widehat{\mathbf{R}}^{(i)} - \widehat{\mathbf{R}}^{(j)}\|_F$, is invariant to any rotations on $\widehat{\mathbf{V}}$ according to Proposition 5. This matrix is shown in Figure 12, and

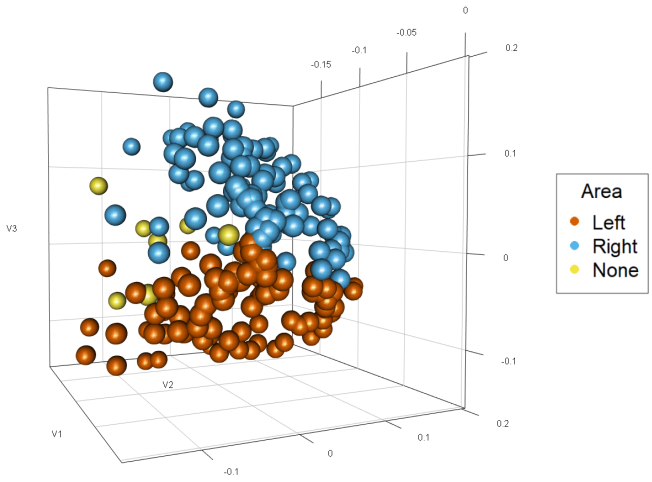


Figure 11: Estimated latent positions obtained by MASE on the HNU1 data graphs with a 3-dimensional embedding.

reflects that graphs coming from the same individual tend to be more similar to each other than graphs of different individuals. We perform classical multidimensional scaling (CMDS) (Borg and Groenen, 2003) on the matrix $\mathbf{D}$ with 5 dimensions for the scaling. Figure 13 shows scatter plots of the positions discovered by MDS for a subset of the graphs corresponding to 5 subjects; these subjects were chosen because the distances between them represent a useful snapshot of the variability in the data. We observe that the points representing graphs of the same subject usually cluster together. The location of the points corresponding to subjects 12 and 14 suggests some similarity between their connectomes, while subject 13 shows a larger spread, and hence more variability in the brain network representations.

The ability of our method to distinguish between graphs of different subjects is further evaluated via classification and hypothesis testing. In the classification analysis, we performed a comparison to other methods that are based on low-rank embeddings following a similar procedure to Section 5.2, namely, the joint embedding of graphs of Wang et al. (2019c) (JE), the multiple random dot product graphs of Nielsen and Witten (2018) (MRDPG), and the omnibus embedding of Levin et al. (2017) (OMNI). Additionally, we compare with the common and individual structure explained (CISE) algorithm and its variant 2 (CISEv2), introduced in Wang et al. (2019b), which fit a graph embedding based on a logistic regression for the edges. In a real data setting, low-rank models are only an approximation to the true generation mechanism, and a more parsimonious model that requires fewer parameters to approximate the data accurately is preferable. Hence, for model selection, we measure the accuracy as a function of the number of embedding dimensions, which controls the description length, defined as the total number of parameters used by the model. Figure 14 shows the 10-fold cross-validated error of a 1-nearest neighbor classifier of the individual graph parameters obtained by the methods. Note that both MASE and OMNI are able to obtain an almost perfect classification, with only one graph being misclassified, in both cases corresponding to subject 20. This graph can be observed to be different to all the other graphs in Figure 12. Although OMNI can achieve good accuracy with only one embedding dimensions, the description length is much larger because OMNI requires dn parameters for each graph, while MASE only uses $d \times d$ symmetric matrices. MASE achieves an optimal classification error with only $d = 9$ dimensions; this classification analysis also offers a supervised way to choose the number of embedding dimensions for MASE suggesting that 9 might be enough, but we keep the more conservative choice of $d = 15$.

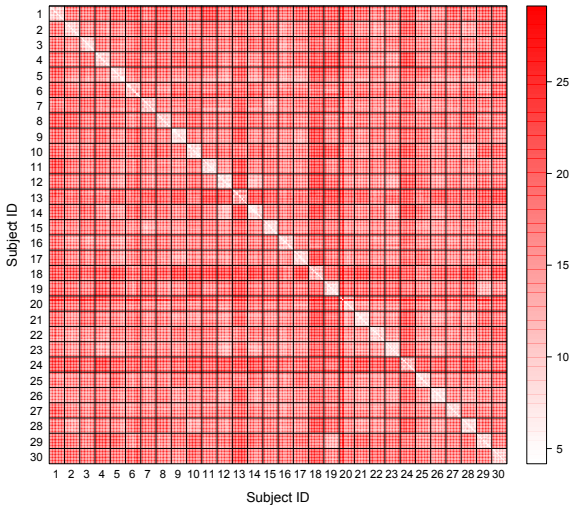


Figure 12: Matrix of distances between the estimated score parameters. Each entry of the matrix corresponds to the distance between the MASE estimated score matrices of a pair of graphs from the HNU1 data, composed by 30 subjects with 10 graphs each. Distances between graphs of the same subject are usually smaller than distances between different subjects.

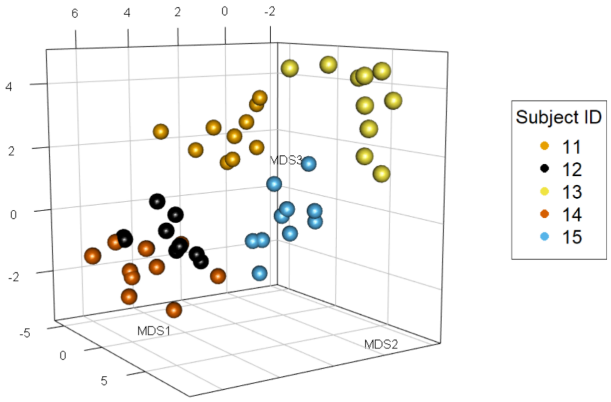


Figure 13: Multidimensional scaling of the individual graph representations obtained by MASE for the HNU1 data. The plot shows the positions discovered by MDS for the graphs corresponding to 5 subjects of the data, showing that graphs corresponding to the same subject are more similar between each other.

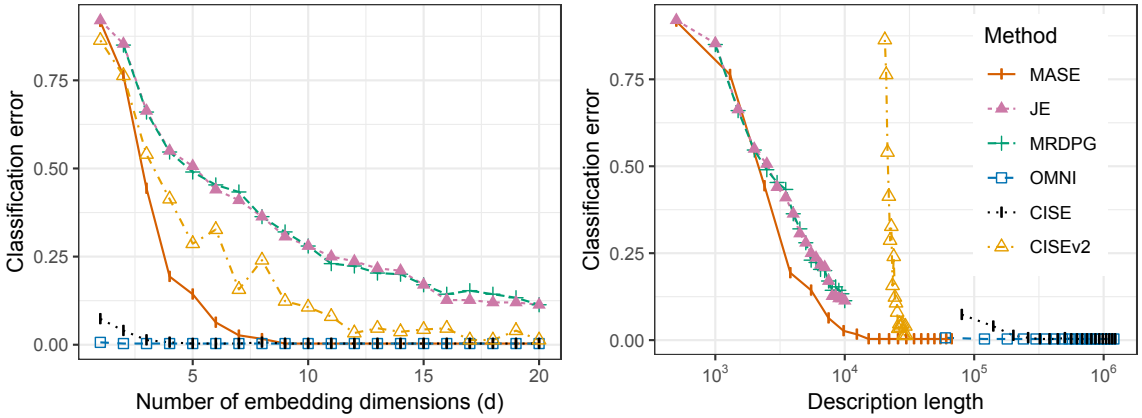
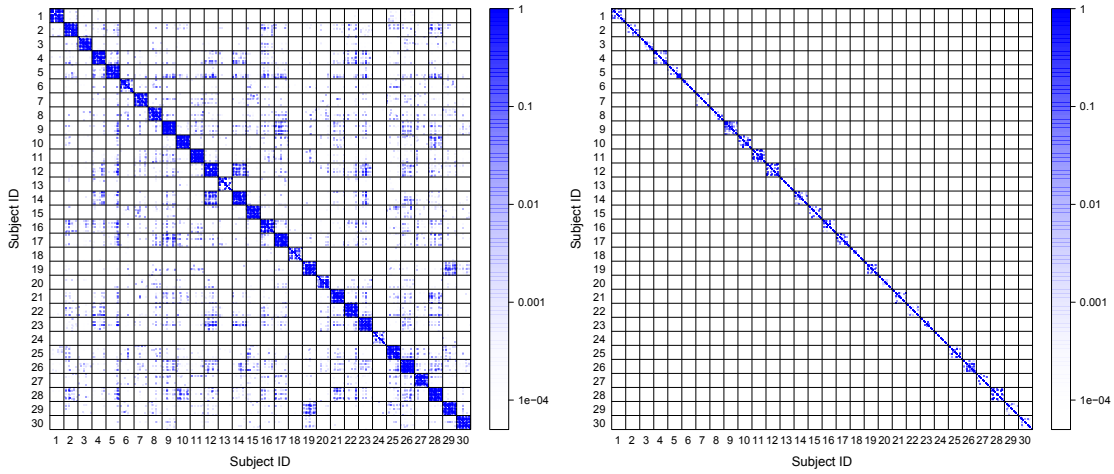


Figure 14: Cross-validation error in classifying the subject labels of the HNU1 data. For each embedding estimated with a different method, a 1-NN classifier was trained using the distance between the individual graph embeddings, varying the number of embedding dimensions (left plot) which control the description length of the different embeddings (right plot). MASE, OMNI and CISE are the only methods that achieve perfect accuracy, but MASE requires much fewer parameters.

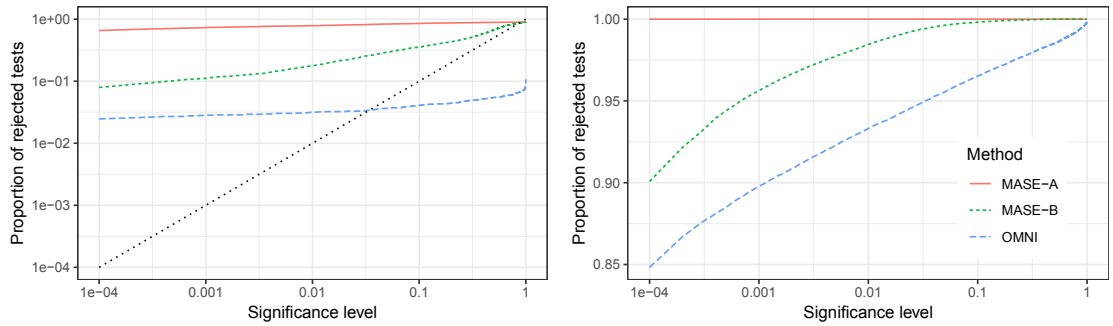
Neither JE or MRDPG achieve an optimal classification error even with $d = 20$ dimensions, and they are outperformed by MASE even when fixing the description length in all methods. CISE and CISEv2 are able to classify the graphs accurately, but require a larger number of parameters, and are computationally more demanding since are based on a non-convex optimization problem. This result suggests that MASE is not only accurate, but more parsimonious than other similar low-rank embeddings.

The matrix of distance between subject parameters obtained by MASE can be used in combination with other distance-based methods to find differences and similarities between subjects, as the previous analyses show. However, a more principled approach to the identification of graph similarity is to measure, for instance, the extent to which differences between graphs can be attributed to random noise. In this spirit, we use the COSIE model to test the hypothesis that the expected edge connectivity of each pair of adjacency matrices is the same. That is, for each pair of graphs $\mathbf{A}^{(i)}$ and $\mathbf{A}^{(j)}$ we test the hypothesis $H_0 : \mathbf{R}^{(i)} = \mathbf{R}^{(j)}$. We follow the same procedure described in Section 5.3 and used the parametric bootstrap method of Tang et al. (2017a) and the asymptotic distribution of the score matrices to estimate the null distribution of the test statistic. Figure 15 shows the p-values of the test performed for every pair of graphs. The diagonal blocks of the matrix correspond to graphs representing the same subject, and in most cases these p-values are large, especially when using the parametric bootstrap to estimate the null distribution. This suggests that in most cases we cannot reject the hypothesis that these graphs have the same distribution, which is reasonable given that these paired graphs represent brain scans of the same individual. Off-diagonal elements of the matrices in Figure 15 represent the result of the equality test for a pair of graphs corresponding to different subjects, and exhibit small p-values in general. For some pairs of subjects, such as 12 and 14, the p-values are also high, suggesting some possible similarities in the brain connectivity of these two subjects, which is also observed on Figure 13. The left panel in Figure 16 shows the percentage of rejections of the test for the p-values in the diagonal of the matrix, and the black dotted line represents the expected number of rejections under the null hypothesis. Note that both OMNI and MASE have slightly more rejections than what would be expected by chance under their corresponding models, especially for the p-values calculated using the asymptotic distribution



(a) Parametric bootstrap (b) Asymptotic distribution

Figure 15: Matrices of p-values for the hypothesis test that a pair of graphs has the same score matrix in the COSIE model. Each entry corresponds to the aforementioned test for a pair of the graphs in the HNU1 data. The test generally assigns small p-values to pairs of graphs corresponding to different subjects, and large p-values to same-subject pairs.



(a) Same subject graph pairs (b) Different subject graph pairs

Figure 16: Proportion of rejected test as a function of the significance level for pairwise equal distribution tests between pairs of graphs representing scans of the same subject (left) and different subjects (right). The black dotted line on the left figure corresponds to the identity, indicating the expected proportion of rejections assuming that the distribution of same-subject graphs is the same. The test constructed with MASE estimates either using the asymptotic distribution (MASE-A) or parametric bootstrap (MASE-B) to approximate the null is rejected more often than the OMNI test.

(MASE-A). The right panel in Figure 16 shows the percentage of rejected tests for pairs of graphs corresponding to different subjects. Both MASE and OMNI reject a large portion of those tests.

7. Discussion

The COSIE model presents a flexible, adaptable model for multiple graphs, one that encompasses a rich collection of independent-edge random graphs and yet retains identifiability and model parsimony. As described in the text, several single-graph models extend to the multiple-graph setting (Airoldi et al., 2007; Zhang et al., 2020; Lyzinski et al., 2016) within the COSIE framework. The presence of a common invariant subspace and the allowance for different score matrices guarantee that COSIE can be used to approximate real-world graph heterogeneity. The MASE procedure builds on the long history of spectral graph inference; it is intuitive, scalable, and consistent for the recovery of the common subspace in COSIE. The accurate inference of the common subspace, in turn, enables to further determine the score parameters, which can then be deployed to discern distinctions between graphs and even subgraphs. Moreover, the MASE procedure enables graph eigenvalue estimation, community detection, and testing whether graphs arise from a common distribution.

Much work remains and includes the following: allowing the score matrices to grow in rank or capture important population-level properties such as variation in time or multilayer structure; to consider estimation and limit results with more relaxed sparsity and delocalization assumptions; and to investigate other spectral embeddings, such as the normalized or regularized Laplacian, which has been shown to uncover different network structure (Priebe et al., 2019) and provide consistent estimators of communities in sparse networks (Le et al., 2017). Indeed, a rich array of open problems can begin to be addressed within the scope of the COSIE model. Open `R` source code for MASE is available at `https://github.com/jesusdaniel/mase`, and in the Python `GraSPy` package at `https://neurodata.io/graspy` (Chung et al., 2019).

Acknowledgments

This research has been supported by the Lifelong Learning Machines (L2M) program of the Defence Advanced Research Projects Agency (DARPA) via contract number HR0011-18-2-0025. This work is also supported in part by the D3M program of DARPA, the NSF DMS award number 1902755, and funding from Microsoft Research. We would like to thank Keith Levin and Elizaveta Levina for helpful discussions.

Appendix A. Proofs

A.1 Stochastic Blockmodels are COSIE

Proof [Proof of Proposition 4] Define $\mathbf{V} = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1/2}$. Because $\mathbf{Z}^\top\mathbf{Z}$ is diagonal and full rank, it is easy to observe that the columns of $\mathbf{V}$ are orthogonal. By writing $\mathbf{R}^{(i)} = (\mathbf{Z}^\top\mathbf{Z})^{1/2}\mathbf{B}^{(i)}(\mathbf{Z}^\top\mathbf{Z})^{1/2}$ we obtain the result. ■

A.2 Identifiability

Proof [Proof of Proposition 5] Suppose that there exists a matrix of orthonormal columns $\mathbf{U} \in \mathbb{R}^{n\times d}$ and symmetric matrices $\mathbf{S}^{(1)}, \ldots, \mathbf{S}^{(m)} \in \mathbb{R}^{d\times d}$ such that

$$\mathbf{V}\mathbf{R}^{(i)}\mathbf{V}^\top = \mathbf{U}\mathbf{S}^{(i)}\mathbf{U}^\top \tag{10}$$

for all $i \in [m]$.

a) Observe that

$$\begin{aligned}\|\mathbf{R}^{(i)} - \mathbf{R}^{(j)}\| &= \|\mathbf{V}(\mathbf{R}^{(i)} - \mathbf{R}^{(j)})\mathbf{V}^\top\| \\ &= \|\mathbf{U}(\mathbf{S}^{(i)} - \mathbf{S}^{(j)})\mathbf{U}^\top\| \\ &= \|\mathbf{S}^{(i)} - \mathbf{S}^{(j)}\|.\end{aligned}$$

The claim follows from the fact that the same equalities hold with Frobenius norm in place of the spectral norm.

b) Squaring the matrices in Equation (10) yields

$$\begin{aligned}\mathbf{U}\left(\sum_{i=1}^{m}\left(\mathbf{S}^{(i)}\right)^2\right)\mathbf{U}^\top &= \mathbf{V}\left(\sum_{i=1}^{m}\left(\mathbf{R}^{(i)}\right)^2\right)\mathbf{V}^\top \\ &= \mathbf{V}(\widetilde{\mathbf{R}}\widetilde{\mathbf{R}}^\top)\mathbf{V}^\top.\end{aligned}$$

Note that the full-rank condition of $\widetilde{\mathbf{R}}$ implies that $\widetilde{\mathbf{R}}\widetilde{\mathbf{R}}^\top$ is also full-rank, so the inverse of $\widetilde{\mathbf{R}}\widetilde{\mathbf{R}}^\top$ exists. Set

$$\mathbf{W} = \left(\sum_{i=1}^{m}\left(\mathbf{S}^{(i)}\right)^2\right)\mathbf{U}^\top\mathbf{V}(\widetilde{\mathbf{R}}\widetilde{\mathbf{R}}^\top)^{-1},$$

so that $\mathbf{UW} = \mathbf{V}$. To check that $\mathbf{W}$ is orthogonal, note that $\mathbf{U}^\top\mathbf{UW} = \mathbf{W} = \mathbf{U}^\top\mathbf{V}$ and $\mathbf{V}^\top\mathbf{UW} = \mathbf{I} \in \mathbb{R}^{d\times d}$, which imply $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$. Since $\mathbf{W}$ is square ($d \times d$), this implies that $\mathbf{W}$ is of full rank. The uniqueness of the matrix inverse guarantees that $\mathbf{W}^{-1} = \mathbf{W}^\top$, and, in turn, that $\mathbf{WW}^\top = \mathbf{I}$; this implies that $\mathbf{W}$ is orthonormal.

c) Suppose that $\mathbf{U} = \mathbf{V}$. Then, multiplying Equation (10) by $\mathbf{V}^\top$ and $\mathbf{V}$ on the left and right establishes that $\mathbf{S}^{(i)} = \mathbf{R}^{(i)}$.

■

A.3 Estimation of the Common Invariant Subspace

Before presenting the proof of Theorem 7, we first establish the following lemma.

Lemma 13. *Let $(\mathbf{A}_n)_{n=1}^{\infty}$ be a sequence of the adjacency matrices of independent-edge Bernoulli random graphs such that $\mathbf{A}_n \in \{0,1\}^{n\times n}$ and $\mathbb{E}[\mathbf{A}_n|\mathbf{P}_n] = \mathbf{P}_n$ for some $\mathbf{P}_n \in \mathbb{R}^{n\times n}$. Suppose that $\delta(\mathbf{P}_n) = \omega(\log n)$. Then, there exists some $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, the random variable $\|\mathbf{A}_n - \mathbf{P}_n\|$ is sub-Gaussian and*

$$\mathbb{E}\left[\|\mathbf{A}_n - \mathbf{P}_n\|^p\right] \leq 2^{2p-1}p^{p/2+1}\left(\delta(\mathbf{P}_n)\right)^{p/2}.$$

Proof By Corollary 3.12 in Bandeira and van Handel (2016), there exists $C > 0$ such that for $\tilde{\sigma}(\mathbf{P}^{(i)}) = \max_{u\in[n]}(\sum_{v=1}^{n}\mathbf{P}^{(i)}_{uv}(1-\mathbf{P}^{(i)}_{uv}))^{1/2}$ and any $\eta \geq 0$,

$$\mathbb{P}\left(\|\mathbf{A}_n - \mathbf{P}_n\| \leq 3\sqrt{2}\tilde{\sigma} + C\eta\right) \geq 1 - ne^{-\eta^2}.$$

Taking $\eta = \beta\sqrt{\delta(\mathbf{P}_n)}/C$ and n_0 such that $\delta(\mathbf{P}_n) > C\log n + 1$ holds for all $n \geq n_0$ and $\beta \geq 1$,

$$\mathbb{P}\left(\|\mathbf{A}_n - \mathbf{P}_n\| \geq 4\beta\sqrt{2\delta(\mathbf{P}_n)}\right) \leq \exp(-\beta^2).$$

Hence, for all $t \geq 4\sqrt{2\delta(\mathbf{P}_n)}$,

$$\mathbb{P}\left(\|\mathbf{A}_n - \mathbf{P}_n\| > t\right) \leq \exp\left(-\frac{t^2}{32\delta(\mathbf{P}_n)}\right).$$

Therefore, by Lemma 5.5 in Vershynin (2010), $\|\mathbf{A}_n - \mathbf{P}_n\|$ is sub-Gaussian for all $n \geq n_0$, and

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{A}_n - \mathbf{P}_n\|^p\right] &= \int_0^\infty \mathbb{P}(\|\mathbf{A}_n - \mathbf{P}_n\| > t) p t^{p-1} dt \\ &\leq \left(\frac{p}{2}\right)^{1+p/2} (32\delta(\mathbf{P}_n))^{p/2}.\end{aligned}$$

■

Proof [Proof of Theorem 7] We introduce some notation. Define the following matrices as

$$\begin{aligned}\widehat{\mathbf{\Pi}}^{(i)} &= \widehat{\mathbf{V}}^{(i)}(\widehat{\mathbf{V}}^{(i)})^\top, & \widehat{\mathbf{\Pi}} &= \frac{1}{m}\sum_{i=1}^m \widehat{\mathbf{\Pi}}^{(i)}, \\ \widetilde{\mathbf{\Pi}}^{(i)} &= \mathbb{E}\left[\widehat{\mathbf{V}}^{(i)}(\widehat{\mathbf{V}}^{(i)})^\top\right] & \widetilde{\mathbf{\Pi}} &= \frac{1}{m}\sum_{i=1}^m \widetilde{\mathbf{\Pi}}^{(i)}, \\ \mathbf{\Pi} &= \mathbf{V}\mathbf{V}^\top.\end{aligned}$$

Note that $\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top = \widehat{\mathbf{\Pi}}$, and therefore $\widehat{\mathbf{V}}$ corresponds to the d leading eigenvectors of $\widehat{\mathbf{\Pi}}$. Recall that $\widetilde{\mathbf{V}}$ is the matrix containing the d leading eigenvectors of $\widetilde{\mathbf{\Pi}} = \frac{1}{m}\sum_{i=1}^m \mathbb{E}\big[\widehat{\mathbf{V}}^{(i)}(\widehat{\mathbf{V}}^{(i)})^\top\big]$.

To prove Theorem 7, we follow the main arguments of Fan et al. (2019). We analyze the terms of the right hand side of Equation (4) separately. For the first term, corresponding to $\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top\|_F$, note that by the Davis-Kahan theorem (see Yu et al. (2015)),

$$\begin{aligned}\mathbb{E}\left[\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top\|_F\right] &\leq \frac{2^{3/2}\mathbb{E}\|\widehat{\mathbf{\Pi}} - \widetilde{\mathbf{\Pi}}\|_F}{\lambda_d(\widetilde{\mathbf{\Pi}}) - \lambda_{d+1}(\widetilde{\mathbf{\Pi}})} \\ &= \frac{2^{3/2}\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^m(\widehat{\mathbf{\Pi}}^{(i)} - \widetilde{\mathbf{\Pi}}^{(i)})\right\|_F}{\lambda_d(\widetilde{\mathbf{\Pi}}) - \lambda_{d+1}(\widetilde{\mathbf{\Pi}})}. \qquad (11)\end{aligned}$$

Using again the Davis-Kahan Theorem, the spectral norm of each term inside the norm of the numerator can be bounded as

$$\|\widehat{\mathbf{\Pi}}^{(i)} - \mathbf{\Pi}\|_F \leq \frac{2^{3/2}\sqrt{d}\|\mathbf{A}^{(i)} - \mathbf{P}^{(i)}\|}{|\lambda_{\min}(\mathbf{R}^{(i)})|}. \qquad (12)$$

Now, to control the error of each term in the numerator of Equation (11), we use the triangle and Jensen inequalities, and the previous bound in Equation (12) as follows

$$\begin{aligned}\mathbb{E}\left[\|\widehat{\mathbf{\Pi}}^{(i)} - \widetilde{\mathbf{\Pi}}^{(i)}\|_F^p\right]^{1/p} &\leq \mathbb{E}\left[\|\widehat{\mathbf{\Pi}}^{(i)} - \mathbf{\Pi}\|_F^p\right]^{1/p} + \mathbb{E}\left[\|\mathbb{E}[\mathbf{\Pi} - \widehat{\mathbf{\Pi}}^{(i)}]\|_F^p\right]^{1/p} \\ &\leq 2\mathbb{E}\left[\|\widehat{\mathbf{\Pi}}^{(i)} - \mathbf{\Pi}\|_F^p\right]^{1/p} \\ &\leq \frac{2^{5/2}\sqrt{d}\mathbb{E}\left[\|\mathbf{A}^{(i)} - \mathbf{P}^{(i)}\|^p\right]^{1/p}}{|\lambda_{\min}(\mathbf{R}^{(i)})|} \\ &\leq \frac{2^{9/2}p\left(d\ \delta(\mathbf{P}^{(i)})\right)^{1/2}}{|\lambda_{\min}(\mathbf{R}^{(i)})|},\end{aligned}$$

where the last inequality comes from Lemma 13, assuming that n is sufficiently large. Using Lemma 4 of Fan et al. (2019),

$$\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}(\widehat{\mathbf{\Pi}}^{(i)}-\widetilde{\mathbf{\Pi}}^{(i)})\right\|_F\right] \lesssim \frac{1}{\sqrt{m}}\sqrt{\frac{d}{m}\sum_{i=1}^{m}\frac{\delta(\mathbf{P}^{(i)})}{\lambda_{\min}^2(\mathbf{R}^{(i)})}}.$$

Combining the previous equation and Equation (11),

$$\mathbb{E}\left[\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top-\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top\|_F\right] \lesssim \frac{1}{\sqrt{m}(\lambda_d(\widetilde{\mathbf{\Pi}})-\lambda_{d+1}(\widetilde{\mathbf{\Pi}}))}\sqrt{\frac{d}{m}\sum_{i=1}^{m}\frac{\delta(\mathbf{P}^{(i)})}{\lambda_{\min}^2(\mathbf{R}^{(i)})}}. \tag{13}$$

To control the error of the second term in Equation (4), we use the triangle inequality and Lemma 2 of Fan et al. (2019). For each $i \in [m]$, suppose that $U_1^{(i)},\dots,U_n^{(i)} \in \mathbb{R}^n$ is an orthonormal basis of $\mathbb{R}^n$ such that $\text{span}\{U_1^{(i)},\dots,U_d^{(i)}\} = \text{span}(\mathbf{V})$ and $\mathbf{P}^{(i)}U_j^{(i)} = \lambda_j(\mathbf{P}^{(i)})U_j^{(i)}$ for each $i \in [d]$. Thus, the vectors $U_1^{(i)},\dots,U_d^{(i)}$ correspond to the d leading eigenvectors of $\mathbf{P}^{(i)}$ when $\lambda_j(\mathbf{P}^{(i)}) \neq 0$, and they span the same invariant subspace as $\mathbf{V}$. Define $\mathbf{U}^{(i)} \in \mathbb{R}^{n\times d}$ as

$$\mathbf{U}^{(i)} = \begin{pmatrix} U_1^{(i)} & \cdots & U_d^{(i)} \end{pmatrix}.$$

Note that $\mathbf{U}^{(i)}\mathbf{U}^{(i)^\top} = \mathbf{V}\mathbf{V}^\top$. Also, define $\mathbf{\Lambda}^{(i)} \in \mathbb{R}^{d\times d}$ be a diagonal matrix containing the d leading eigenvalues in magnitude of $\mathbf{P}^{(i)}$, such that $\mathbf{\Lambda}_{jj}^{(i)} = \lambda_j(\mathbf{P}^{(i)})$. Set $\mathbf{G}^{(i)} = \sum_{k=d+1}^{n} U_k^{(i)}(U_k^{(i)})^\top$. Define $\mathbf{H}^{(i)}$ as

$$\mathbf{H}^{(i)} = \mathbf{G}^{(i)}(\mathbf{A}^{(i)}-\mathbf{P}^{(i)})\mathbf{U}^{(i)}(\mathbf{\Lambda}^{(i)})^{-1}\mathbf{U}^{(i)^\top}.$$

Note that $\mathbb{E}[\mathbf{H}^{(i)}] = 0$, and

$$\begin{aligned}\|\mathbf{H}^{(i)}+\mathbf{H}^{(i)^\top}\|_F &\le 2\|\mathbf{G}^{(i)}(\mathbf{A}^{(i)}-\mathbf{P}^{(i)})\mathbf{U}^{(i)}(\mathbf{\Lambda}^{(i)})^{-1}\mathbf{U}^{(i)^\top}\|_F \\ &\le 2\|(\mathbf{\Lambda}^{(i)})^{-1}\|\|\mathbf{G}^{(i)}\|\|\mathbf{A}^{(i)}-\mathbf{P}^{(i)}\|\|\mathbf{U}^{(i)}\|_F \\ &\le 2\sqrt{d}\frac{\|\mathbf{A}^{(i)}-\mathbf{P}^{(i)}\|}{|\lambda_{\min}(\mathbf{R}^{(i)})|}. \end{aligned} \tag{14}$$

By Lemma 2 of Fan et al. (2019), if $\|\mathbf{A}^{(i)}-\mathbf{P}^{(i)}\| \le \lambda_{\min}(\mathbf{R}^{(i)})/10$ then

$$\|\widehat{\mathbf{\Pi}}^{(i)}-\mathbf{\Pi}-\mathbf{H}^{(i)}-\mathbf{H}^{(i)^\top}\|_F \le 24\sqrt{d}\frac{\|\mathbf{A}^{(i)}-\mathbf{P}^{(i)}\|^2}{\lambda_{\min}(\mathbf{R}^{(i)})^2}.$$

On the other hand, if $\|\mathbf{A}^{(i)}-\mathbf{P}^{(i)}\| > \lambda_{\min}(\mathbf{R}^{(i)})/10$, by using the Davis-Kahan theorem and Equation (14),

$$\begin{aligned}\|\widehat{\mathbf{\Pi}}^{(i)}-\mathbf{\Pi}-\mathbf{H}^{(i)}-\mathbf{H}^{(i)^\top}\|_F &\le \|\widehat{\mathbf{\Pi}}^{(i)}-\mathbf{\Pi}\|_F + \|\mathbf{H}^{(i)}+\mathbf{H}^{(i)^\top}\|_F \\ &\le 4\sqrt{d}\frac{\|\mathbf{A}^{(i)}-\mathbf{P}^{(i)}\|}{|\lambda_{\min}(\mathbf{R}^{(i)})|}.\end{aligned}$$

Define $\mathcal{D}$ as the event in which $\|\mathbf{A}^{(i)} - \mathbf{P}^{(i)}\| > \lambda_{\min}(\mathbf{R}^{(i)})/10$. Observe that

$$\begin{aligned}
\|\widetilde{\mathbf{\Pi}}^{(i)} - \mathbf{\Pi}\|_F &= \|\mathbb{E}[\widehat{\mathbf{\Pi}}^{(i)} - \mathbf{\Pi} - \mathbf{H}^{(i)} - \mathbf{H}^{(i)^\top}]\|_F \\
&\leq \mathbb{E}[\|\widehat{\mathbf{\Pi}}^{(i)} - \mathbf{\Pi} - \mathbf{H}^{(i)} - \mathbf{H}^{(i)^\top}\|_F(\mathbb{1}_{\mathcal{D}} + \mathbb{1}_{\mathcal{D}^c})] \\
&\leq 4\sqrt{d}\frac{\mathbb{E}[\mathbb{1}_{\mathcal{D}}\|\mathbf{A}^{(i)} - \mathbf{P}^{(i)}\|]}{|\lambda_{\min}(\mathbf{R}^{(i)})|} + 24\sqrt{d}\frac{\mathbb{E}[\|\mathbf{A}^{(i)} - \mathbf{P}^{(i)}\|^2]}{\lambda_{\min}(\mathbf{R}^{(i)})^2} \\
&\leq 40\sqrt{d}\frac{\mathbb{E}[\|\mathbf{A}^{(i)} - \mathbf{P}^{(i)}\|^2]}{\lambda_{\min}(\mathbf{R}^{(i)})^2} + 24\sqrt{d}\frac{\mathbb{E}[\|\mathbf{A}^{(i)} - \mathbf{P}^{(i)}\|^2]}{\lambda_{\min}(\mathbf{R}^{(i)})^2} \\
&\leq 64\sqrt{d}\frac{\mathbb{E}[\|\mathbf{A}^{(i)} - \mathbf{P}^{(i)}\|^2]}{\lambda_{\min}(\mathbf{R}^{(i)})^2} \\
&\lesssim \sqrt{d}\frac{\delta(\mathbf{P}^{(i)})}{\lambda_{\min}(\mathbf{R}^{(i)})^2}.
\end{aligned}$$

By the triangle inequality,

$$\|\widetilde{\mathbf{\Pi}} - \mathbf{\Pi}\|_F \lesssim \frac{\sqrt{d}}{m}\sum_{i=1}^{m}\frac{\delta(\mathbf{P}^{(i)})}{\lambda_{\min}^2(\mathbf{R}^{(i)})}. \tag{15}$$

Combining Equations (13) and (15), we obtain that

$$\mathbb{E}[\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top\|_F] \lesssim \frac{1}{\sqrt{m}(\lambda_d(\widetilde{\mathbf{\Pi}}) - \lambda_{d+1}(\widetilde{\mathbf{\Pi}}))}\sqrt{\frac{d}{m}}\varepsilon + \sqrt{d}\varepsilon^2.$$

Finally, the denominator of the first term can be bounded using Weyl's inequality, by observing that

$$\begin{aligned}
\lambda_d(\widetilde{\mathbf{\Pi}}) &\geq 1 - \|\widetilde{\mathbf{\Pi}} - \mathbf{\Pi}\|, \\
\lambda_{d+1}(\widetilde{\mathbf{\Pi}}) &\leq \|\widetilde{\mathbf{\Pi}} - \mathbf{\Pi}\| \\
\lambda_d(\widetilde{\mathbf{\Pi}}) - \lambda_{d+1}(\widetilde{\mathbf{\Pi}}) &\geq 1 - 2\|\widetilde{\mathbf{\Pi}} - \mathbf{\Pi}\|_F.
\end{aligned}$$

Finally, Equation (15) and the fact that $\varepsilon = o(1)$ ensures that this denominator is bounded away from zero, which completes the proof. ■

A.4 Community Detection Results

Proof [Proof of Corollary 8] Define the parameters of the COSIE model

$$\begin{aligned}
\mathbf{V} &= \mathbf{Z}\mathbf{\Xi}^{-1/2}, \\
\mathbf{R}^{(i)} &= \mathbf{\Xi}^{1/2}\mathbf{B}^{(i)}\mathbf{\Xi}^{1/2},
\end{aligned}$$

so that $\mathbf{V}\mathbf{R}^{(i)}\mathbf{V}^\top = \mathbf{Z}\mathbf{B}^{(i)}\mathbf{V}^\top$. To obtain a bound on ε defined according to Theorem 7, observe that the largest expected degree can be bounded above as

$$\begin{aligned}
\delta(\mathbf{P}^{(i)}) &= \max_{j\in[K]}\sum_{k=1}^{K}\mathbf{B}^{(i)}_{jk}n_k \\
&\leq \left(\sum_{j=1}^{K}n_j^2\right)^{1/2}\|\mathbf{B}^{(i)}\|_1 \\
&\leq \sqrt{K}\left(\sum_{j=1}^{K}n_j^2\right)^{1/2}\lambda_1(\mathbf{B}^{(i)}).
\end{aligned}$$

Also, the smallest eigenvalue of the corresponding score matrix $\mathbf{R}^{(i)}$ is bounded below as

$$\begin{aligned}|\lambda_{\min}(\mathbf{R}^{(i)})| &= |\lambda_{\min}(\mathbf{\Xi}^{1/2}\mathbf{B}^{(i)}\mathbf{\Xi}^{1/2})| \\ &\geq \lambda_{\min}(\mathbf{\Xi})|\lambda_{\min}(\mathbf{B}^{(i)})| = n_{\min}|\lambda_{\min}(\mathbf{B}^{(i)})|.\end{aligned}$$

Therefore, using the above inequalities and Assumption 1,

$$\begin{aligned}\varepsilon^2 = \frac{1}{m}\sum_{i=1}^{m}\frac{\delta(\mathbf{P}^{(i)})}{\lambda^2_{\min}(\mathbf{R}^{(i)})} &\leq \frac{1}{m}\sum_{i=1}^{m}\frac{\sqrt{K}\left(\sum_{j=1}^{K} n_j^2\right)^{1/2}\lambda_1(\mathbf{B}^{(i)})}{n^2_{\min}\lambda^2_{\min}(\mathbf{B}^{(i)})} \\ &\leq \frac{\gamma}{n^{\kappa}_{\min}}.\end{aligned}$$

Based on the above calculations, Theorem 7 implies the result. ■

Proof [Proof of Theorem 9] Consider the permutation matrix $\widehat{\mathbf{Q}}$ and the orthogonal matrix $\widehat{\mathbf{W}}$ such that

$$\widehat{\mathbf{Q}} = \underset{\mathbf{Q}\in\mathcal{P}_K}{\operatorname{argmin}} \|\widehat{\mathbf{Z}}\mathbf{Q} - \mathbf{Z}\|_F,$$

$$\widehat{\mathbf{W}} = \underset{\mathbf{W}\in\mathcal{O}_K}{\operatorname{argmin}} \|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F.$$

Using Lemma 3.2 of Rohe et al. (2011), observe that if $\|\widehat{Z}_u\widehat{\mathbf{C}}\widehat{\mathbf{Q}} - V_u\widehat{\mathbf{W}}\| < 1/\sqrt{2n_{\max}}$ then $\widehat{Z}_u\mathbf{Q} = Z_u$. Hence,

$$\begin{aligned}\|\widehat{\mathbf{Z}}\widehat{\mathbf{Q}} - \mathbf{Z}\|_F &\leq \sum_{u=1}^{n}\mathbb{1}\{\widehat{Z}_u\widehat{\mathbf{Q}} \neq Z_u\} \\ &\leq \sqrt{2n_{\max}}\|\widehat{\mathbf{Z}}\widehat{\mathbf{C}}\widehat{\mathbf{Q}} - \mathbf{V}\widehat{\mathbf{W}}\|_F \\ &\leq \sqrt{2n_{\max}}\left(\|\widehat{\mathbf{Z}}\widehat{\mathbf{C}}\widehat{\mathbf{Q}} - \widehat{\mathbf{V}}\|_F + \|\widehat{\mathbf{V}} - \mathbf{V}\widehat{\mathbf{W}}\|_F\right) \\ &\leq 2^{3/2}\sqrt{n_{\max}}\|\widehat{\mathbf{V}} - \mathbf{V}\widehat{\mathbf{W}}\|_F.\end{aligned}$$

Corollary 8 immediately implies the result. ■

A.5 Asymptotic Normality Results

Proof [Proof of Proposition 10]

a) Define $\mathbf{\Xi} = \operatorname{diag}(n_1, \ldots, n_K)$ as before. Note that $\mathbf{Z}\mathbf{\Xi}^{-1/2}$ is the common subspace of a multilayer SBM with community membership $\mathbf{Z}$. Therefore, it is enough to show that there exist some $\mathbf{W} \in \mathcal{O}_K$ such that Assumption 2 holds for $\mathbf{Z}\mathbf{\Xi}^{-1/2}\mathbf{W}$. Define an orthogonal matrix $\mathbf{W}$ such that

$$\frac{b_1}{\sqrt{K}} \leq |\mathbf{W}_{kl}| \leq \frac{b_2}{\sqrt{K}}$$

for all $k, l \in [K]$. Such a matrix exists for any $K \geq 1$ (see for example Jaming and Matolcsi (2015)). If the node u belongs to community k, then

$$\frac{b_1}{\sqrt{Kn_{\max}}} \leq |(\mathbf{Z}\mathbf{W})_{uv}| = \frac{|\mathbf{W}_{uv}|}{\sqrt{n_k}} \leq \frac{b_2}{\sqrt{Kn_{\min}}},$$

which completes the proof.

b) Under the SBM, $\mathbf{P}^{(i)} = \mathbf{Z}\mathbf{B}^{(i)}\mathbf{Z}^\top$. Therefore

$$\sum_{u=1}^{n}\sum_{v=1}^{n}\mathbf{P}^{(i)}_{uv}(1-\mathbf{P}^{(i)}_{uv}) = \sum_{s=1}^{K}\sum_{t=1}^{K} n_s n_t \mathbf{B}^{(i)}_{st}(1-\mathbf{B}^{(i)}_{st}) = \omega(1).$$

Also,

$$\sum_{s=1}^{n}\sum_{t=1}^{n}\mathbf{P}^{(i)}_{st}(1-\mathbf{P}^{(i)}_{st})(n\mathbf{V}_{sk}\mathbf{V}_{tl} + n\mathbf{V}_{tk}\mathbf{V}_{sl})^2 \geq \omega(1)\sum_{s=1}^{n}\sum_{t=1}^{n}(\mathbf{V}_{sk}\mathbf{V}_{tl}+\mathbf{V}_{tk}\mathbf{V}_{sl})^2 = \omega(1).$$

■

We now turn to our proof of the asymptotic distribution of $\widehat{\mathbf{R}}^{(i)}$. Recall that the score matrices $\mathbf{R}^{(i)} = \mathbf{R}^{(i,n)}$ and the basis of the invariant subspace $\mathbf{V} = \mathbf{V}^{(n)}$ depend on n, and so does the distribution of the random matrices $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}$. In this section, we write

$$\mathbf{E}^{(i)} := \mathbf{A}^{(i)} - \mathbf{V}\mathbf{R}^{(i)}\mathbf{V}^\top.$$

We emphasize that $\mathbf{E}^{(i)}$ also depends on n. The asymptotic normality ultimately depends on sums of the entries for a sequence of matrices $\mathbf{E}^{(i)}$ that increase their dimension and change the distribution as n increases.

We recall an important elementary result relating the Frobenius norm and the $\sin(\Theta)$ distance (see Bhatia (1997)) for any pair of matrices: if two matrices $\widehat{\mathbf{U}}, \mathbf{U} \in \mathbf{R}^{n\times d}$ each have orthonormal columns, then

$$\|\sin\Theta(\widehat{\mathbf{U}},\mathbf{U})\|_F \leq \inf_{\mathbf{W}\in\mathcal{O}_d}\|\widehat{\mathbf{U}}-\mathbf{U}\mathbf{W}\|_F \leq \sqrt{2}\|\sin\Theta(\widehat{\mathbf{U}},\mathbf{U})\|_F. \tag{16}$$

(Observe that the expectation bound stated in Theorem 7 immediately applies to $\sin\Theta$ distances up to an additional absolute constant factor.) Now for $\mathbf{W} = \arg\inf_{\mathbf{W}\in\mathcal{O}_d}\|\widehat{\mathbf{U}}-\mathbf{U}\mathbf{W}\|_F$, then

$$\|\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{W}\|_F \leq \|\sin\Theta(\widehat{\mathbf{U}},\mathbf{U})\|_F^2. \tag{17}$$

To observe that the previous inequality holds, write $\widetilde{\mathbf{Q}}_1\widetilde{\mathbf{D}}\widetilde{\mathbf{Q}}_2^\top = \mathbf{U}\mathbf{U}^\top$ as the singular value decomposition of $\mathbf{U}\mathbf{U}^\top$, with $\widetilde{\mathbf{Q}}_1, \widetilde{\mathbf{Q}}_2 \in \mathbb{R}^{d\times d}$ orthogonal matrices, and $\widetilde{\mathbf{D}} \in \mathbb{R}^{d\times d}$ a diagonal matrix with $0 \leq \widetilde{\mathbf{D}}_{kk} \leq 1$. Then,

$$\|\sin\Theta(\widehat{\mathbf{U}},\mathbf{U})\|_F^2 = \sum_{k=1}^{d}(1-\widetilde{\mathbf{D}}_{kk}^2).$$

Now, by taking $\mathbf{W} = \widetilde{\mathbf{Q}}_1\widetilde{\mathbf{Q}}_2^\top$ (which is an orthogonal matrix), we have that $\mathbf{U}^\top\widehat{\mathbf{U}} - \widetilde{\mathbf{Q}}_1\widetilde{\mathbf{Q}}_2^\top = \widetilde{\mathbf{Q}}_1(\widetilde{\mathbf{D}}-\mathbf{I})\widetilde{\mathbf{Q}}_2^\top$, and hence

$$\|\mathbf{U}^\top\widehat{\mathbf{U}} - \widetilde{\mathbf{Q}}_1\widetilde{\mathbf{Q}}_2^\top\|_F = \left(\sum_{k=1}^{d}(1-\widetilde{\mathbf{D}}_{kk})^2\right)^{1/2} \leq \sum_{k=1}^{d}(1-\widetilde{\mathbf{D}}_{kk}^2).$$

Finally, notice that $\mathbf{W} = \arg\inf_{\mathbf{W}\in\mathcal{O}_d}\|[\mathbf{U},\mathbf{U}_\perp]^\top\left(\widehat{\mathbf{U}}-\mathbf{U}\mathbf{W}\right)\|_F = \arg\inf_{\mathbf{W}\in\mathcal{O}_d}\|\mathbf{U}^\top\widehat{\mathbf{U}}-\mathbf{W}\|_F$.

We now proceed to further quantify the estimation of $\mathbf{R}^{(i)}$ via $\widehat{\mathbf{R}}^{(i)} := \widehat{\mathbf{V}}^\top\mathbf{A}^{(i)}\widehat{\mathbf{V}}$. In particular, we will prove the following lemma.

Lemma 14. *Under the general assumptions of Theorem 11, there exist sequences of matrices* $\mathbf{B}^{(i)}, \mathbf{N}^{(i)}, \mathbf{W} \in \mathbb{R}^{d\times d}$ *depending on* d, m, *and* n, *such that*

$$\mathbf{W}\widehat{\mathbf{R}}^{(i)}\mathbf{W}^\top - \mathbf{R}^{(i)} - \mathbf{W}\mathbf{B}^{(i)}\mathbf{W}^\top - \mathbf{W}\mathbf{N}^{(i)}\mathbf{W}^\top = \mathbf{V}^\top\mathbf{E}^{(i)}\mathbf{V}, \tag{18}$$

with $\mathbf{H}_m^{(i)} = -(\mathbf{W}\mathbf{B}^{(i)}\mathbf{W}^\top + \mathbf{W}\mathbf{N}^{(i)}\mathbf{W}^\top)$ *satisfying* $\mathbb{E}[\|\mathbf{H}_m^{(i)}\|_F] = O_P(d/\sqrt{m})$, *and* $\mathbf{W} = \arg\inf_{\mathbf{W}\in\mathcal{O}_d} \|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F$ *is an orthogonal matrix*

Proof Letting $\mathbf{E}^{(i)} := \mathbf{A}^{(i)} - \mathbf{P}^{(i)}$, we obtain the expansion

$$\begin{aligned}\widehat{\mathbf{R}}^{(i)} &= \widehat{\mathbf{V}}^\top\mathbf{A}^{(i)}\widehat{\mathbf{V}} \\ &= \widehat{\mathbf{V}}^\top\mathbf{P}^{(i)}\widehat{\mathbf{V}} + \widehat{\mathbf{V}}^\top(\mathbf{A}^{(i)} - \mathbf{P}^{(i)})\widehat{\mathbf{V}} \\ &= (\widehat{\mathbf{V}}^\top\mathbf{V})\mathbf{R}^{(i)}(\mathbf{V}^\top\widehat{\mathbf{V}}) + \widehat{\mathbf{V}}^\top\mathbf{E}^{(i)}\widehat{\mathbf{V}}. \end{aligned}\tag{19}$$

For any orthogonal matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$, the first term can be further expanded as

$$(\widehat{\mathbf{V}}^\top\mathbf{V})\mathbf{R}^{(i)}(\mathbf{V}^\top\widehat{\mathbf{V}}) = \mathbf{W}^\top\mathbf{R}^{(i)}\mathbf{W} + (\widehat{\mathbf{V}}^\top\mathbf{V} - \mathbf{W}^\top)\mathbf{R}^{(i)}(\mathbf{V}^\top\widehat{\mathbf{V}}) + \mathbf{W}^\top\mathbf{R}^{(i)}(\mathbf{V}^\top\widehat{\mathbf{V}} - \mathbf{W}).$$

Furthermore, let $\mathbf{W} = \arg\inf_{\mathbf{W}\in\mathcal{O}_d} \|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F$. The transpose invariance of the Frobenius and spectral norms and Equations (16) and (17) guarantee that terms two and three on the right hand side above can be bounded above as

$$\begin{aligned}\|\mathbf{W}^\top\mathbf{R}^{(i)}(\mathbf{V}^\top\widehat{\mathbf{V}} - \mathbf{W})\|_F &\le \|\mathbf{W}^\top\|\|\mathbf{R}^{(i)}\|\|\mathbf{V}^\top\widehat{\mathbf{V}} - \mathbf{W}\|_F &&\le 2\|\mathbf{R}^{(i)}\|\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F^2, \\ \|(\widehat{\mathbf{V}}^\top\mathbf{V} - \mathbf{W}^\top)\mathbf{R}^{(i)}(\mathbf{V}^\top\widehat{\mathbf{V}})\|_F &\le \|\widehat{\mathbf{V}}^\top\mathbf{V} - \mathbf{W}^\top\|_F\|\mathbf{R}^{(i)}\|\|\mathbf{V}^\top\widehat{\mathbf{V}}\| &&\le 2\|\mathbf{R}^{(i)}\|\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F^2.\end{aligned}$$

On the other hand, the second term in Equation (19) can be further expanded as the sum of four matrices, namely

$$\begin{aligned}\widehat{\mathbf{V}}^\top\mathbf{E}^{(i)}\widehat{\mathbf{V}} = \mathbf{W}^\top\mathbf{V}^\top\mathbf{E}^{(i)}\mathbf{V}\mathbf{W} &\\ + \mathbf{W}^\top\mathbf{V}^\top\mathbf{E}^{(i)}(\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}) &\\ - (\mathbf{W}^\top\mathbf{V}^\top - \widehat{\mathbf{V}}^\top)\mathbf{E}^{(i)}\mathbf{V}\mathbf{W} &\\ - (\mathbf{W}^\top\mathbf{V}^\top - \widehat{\mathbf{V}}^\top)\mathbf{E}^{(i)}(\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}).&\end{aligned}$$

The final term in the above expansion can be bounded via

$$\|(\mathbf{W}^\top\mathbf{V}^\top - \widehat{\mathbf{V}}^\top)\mathbf{E}^{(i)}(\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W})\|_F \le \|\mathbf{E}^{(i)}\|\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F^2.$$

Observe that terms two and three above are simply the transpose of one another, hence it suffices to analyze the former, which satisfies the bound

$$\|\mathbf{W}^\top\mathbf{V}^\top\mathbf{E}^{(i)}(\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W})\|_F \le \|\mathbf{E}^{(i)}\|\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F.$$

Thus, there exist (sequences of) matrices $\mathbf{B}, \mathbf{N} \in \mathbb{R}^{d\times d}$ depending on d, m, and n, such that

$$\mathbf{W}\widehat{\mathbf{R}}^{(i)}\mathbf{W}^\top - \mathbf{R}^{(i)} = \mathbf{V}^\top\mathbf{E}^{(i)}\mathbf{V} + \mathbf{W}\mathbf{B}\mathbf{W}^\top + \mathbf{W}\mathbf{N}\mathbf{W}^\top, \tag{20}$$

for which

$$\begin{aligned}\|\mathbf{B}\|_F &\le 2\|\mathbf{E}^{(i)}\|\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F, \\ \|\mathbf{N}\|_F &\le 4\left(\|\mathbf{E}^{(i)}\| + \|\mathbf{R}^{(i)}\|\right)\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F^2.\end{aligned}$$

Put $\mathbf{H}_m = \mathbf{W}\mathbf{B}\mathbf{W}^\top + \mathbf{W}\mathbf{N}\mathbf{W}^\top$; we call $\mathbf{H}_m$ a bias term.

Using Lemma 4 of Fan et al. (2019) as in Equation (13), we have

$$\mathbb{E}\left[\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top\|_F^2\right] \lesssim \frac{d}{m}\varepsilon^2,$$

and hence, by the same arguments as in the proof of Theorem 7,

$$\mathbb{E}\left[\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F^2\right] \leq \mathbb{E}\left[\left(\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top\|_F + \|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top\|_F\right)^2\right] \lesssim \frac{d}{m}\varepsilon^2.$$

Recall that we have a high probability bound on $\mathbf{E}^{(i)}$ in spectral norm (see the proof of Lemma 13), namely for n sufficiently large,

$$\mathbb{P}\left(\|\mathbf{E}^{(i)}\| > 8\sqrt{\delta(\mathbf{P}^{(i)})}\right) \leq ne^{-[\delta(\mathbf{P}^{(i)})]^2/C} \leq n^{-2}.$$

Also, recall that

$$\delta(\mathbf{P}^{(i)}) = \max_{v=1,\dots,n}\sum_{u=1}^{n}(\mathbf{P}^{(i)})_{uv},$$

and as such

$$\|\mathbf{R}^{(i)}\| = \|\mathbf{V}\mathbf{R}^{(i)}\mathbf{V}^\top\| = \|\mathbf{P}^{(i)}\| \leq \delta(\mathbf{P}^{(i)}).$$

Observe that $\|\mathbf{E}^{(i)}\| \leq n$ and $\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F \leq 2\sqrt{d}$. Hence, for any $t > 0$,

$$\begin{aligned}
\mathbb{E}[\|\mathbf{E}^{(i)}\|\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F] \leq& t\mathbb{E}[\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F] + 2n\sqrt{d}\mathbb{P}(\|\mathbf{E}^{(i)}\| > t),\\
\mathbb{E}[\|\mathbf{E}^{(i)}\|\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F^2] \leq& t\mathbb{E}[\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F^2] + 2nd\mathbb{P}(\|\mathbf{E}^{(i)}\| > t).
\end{aligned}$$

Setting $t = 8\sqrt{\delta(\mathbf{P}^{(i)})}$ guarantees that the Frobenius norm of the bias term $\mathbf{H}_m$ vanishes in probability as $n, m \to \infty$. In particular, note that under the assumption that $(\delta(\mathbf{P}^{(i)}))^{1/2}\varepsilon = O(1)$,

$$\begin{aligned}
\mathbb{E}[\|\mathbf{H}_m\|_F] &\lesssim \mathbb{E}[\|\mathbf{E}^{(i)}\|\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F] + \mathbb{E}[(\|\mathbf{E}^{(i)}\| + \delta(\mathbf{P}^{(i)}))\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_F^2]\\
&\lesssim \sqrt{\delta(\mathbf{P}^{(i)})}\left(\sqrt{\frac{d}{m}}\varepsilon + \sqrt{d}\varepsilon^2\right)\left(1 + \sqrt{\delta(\mathbf{P}^{(i)})}\left(\sqrt{\frac{d}{m}}\varepsilon + \sqrt{d}\varepsilon^2\right)\right)\\
&\lesssim \frac{d}{\sqrt{m}}.
\end{aligned}$$

■

The central limit theorem will be complete once we show that the entries of $\mathbf{V}\mathbf{E}^{(i)}\mathbf{V}^\top$ converge to a normal distribution as $n \to \infty$. We address this in our remaining lemma.

Lemma 15. *Let $\mathbf{\Sigma}^{(i)} \in \mathbb{R}^{r\times r}$ be the matrix defined on Equation (8).*

a) Under the assumptions of Theorem 11 part a), for any $k, l \in [d]$, $k \leq l$, it holds that

$$\left(\mathbf{\Sigma}^{(i)}_{\frac{2k+l(l-1)}{2},\frac{2k+l(l-1)}{2}}\right)^{-1/2}(\mathbf{V}^\top\mathbf{E}^{(i)}\mathbf{V})_{kl} \xrightarrow{d} \mathcal{N}(0,1).$$

where

$$\left(\frac{c_1^4}{n^2} + \frac{c_1^8}{n^4}\right)s^2(\mathbf{P}^{(i)}) \leq \mathbf{\Sigma}^{(i)}_{\frac{2k+l(l-1)}{2},\frac{2k+l(l-1)}{2}} \leq \left(\frac{c_2^4}{n^2} + \frac{c_2^8}{n^4}\right)s^2(\mathbf{P}^{(i)}) \tag{21}$$

b) Under the assumptions of Theorem 11 part b), it holds that

$$(\mathbf{\Sigma}^{(i)})^{-1/2}\operatorname{vec}(\mathbf{V}^\top\mathbf{E}^{(i)}\mathbf{V}) \xrightarrow{d} \mathcal{N}(\mathbf{0}_r, \mathbf{I}_r).$$

Proof To obtain the asymptotic distribution of the entries of $\mathbf{V}^\top\mathbf{E}^{(i)}\mathbf{V}$, observe that, first, the entire matrix $\mathbf{V}^\top\mathbf{E}^{(i)}\mathbf{V}$ depends on n. Second, note that

$$(\mathbf{V}^\top\mathbf{E}^{(i)}\mathbf{V})_{kl} = \sum_{s=1}^{n}\sum_{t=1}^{n}\mathbf{E}^{(i)}_{st}\mathbf{V}_{sk}\mathbf{V}_{tl} \tag{22}$$

$$= \sum_{s=1}^{n-1}\sum_{t=s+1}^{n}\mathbf{E}^{(i)}_{st}\left(\mathbf{V}_{sk}\mathbf{V}_{tl} + \mathbf{V}_{tk}\mathbf{V}_{sl}\right), \tag{23}$$

so each of the entries of the matrix $\mathbf{V}^\top\mathbf{E}^{(i)}\mathbf{V}$ is a sum of the independent random variables. The expected value of the above sum is

$$\mathbb{E}[(\mathbf{V}^\top\mathbf{E}^{(i)}\mathbf{V})_{kl}] = \left(\mathbf{V}^\top\mathbb{E}[\mathbf{E}^{(i)}]\mathbf{V}\right)_{kl} = 0.$$

To ease notation, we write $\mathbf{\Sigma} = \mathbf{\Sigma}^{(i)}$. For each pair (k,l) and (k',l') with $k \le l$ and $k' \le l'$, the covariance between those corresponding entries of $\mathbf{V}^\top\mathbf{E}^{(i)}\mathbf{V}$ is given by

$$\begin{aligned}
&\operatorname{Cov}\left((\mathbf{V}^\top\mathbf{E}^{(i)}\mathbf{V})_{kl},\ (\mathbf{V}^\top\mathbf{E}^{(i)}\mathbf{V})_{k'l'}\right)\\
&\quad= \sum_{s=1}^{n-1}\sum_{t=s+1}^{n}\sum_{s'=1}^{n-1}\sum_{t'=s'+1}^{n}\Big\{\operatorname{Cov}\left(\mathbf{E}^{(i)}_{st},\mathbf{E}^{(i)}_{s't'}\right)\times\\
&\qquad\qquad(\mathbf{V}_{sk}\mathbf{V}_{tl}+\mathbf{V}_{tk}\mathbf{V}_{sl})(\mathbf{V}_{s'k'}\mathbf{V}_{t'l'}+\mathbf{V}_{s'l'}\mathbf{V}_{t'k'})\Big\}\\
&\quad= \sum_{s=1}^{n-1}\sum_{t=s+1}^{n}\operatorname{Cov}\left(\mathbf{E}^{(i)}_{st},\mathbf{E}^{(i)}_{st}\right)(\mathbf{V}_{sk}\mathbf{V}_{tl}+\mathbf{V}_{tk}\mathbf{V}_{sl})(\mathbf{V}_{sk'}\mathbf{V}_{tl'}+\mathbf{V}_{sl'}\mathbf{V}_{tk'})\\
&\quad= \sum_{s=1}^{n-1}\sum_{t=s+1}^{n}\mathbf{P}^{(i)}_{st}(1-\mathbf{P}^{(i)}_{st})\left[(\mathbf{V}_{sk}\mathbf{V}_{tl}+\mathbf{V}_{tk}\mathbf{V}_{sl})(\mathbf{V}_{sk'}\mathbf{V}_{tl'}+\mathbf{V}_{sl'}\mathbf{V}_{tk'})\right]\\
&\quad= \mathbf{\Sigma}_{\frac{2k+l(l-1)}{2},\frac{2k'+l'(l'-1)}{2}}.
\end{aligned} \tag{24}$$

Recall that $s^2(\mathbf{P}^{(i)})$ is defined as

$$s^2(\mathbf{P}^{(i)}) = \sum_{s=1}^{n}\sum_{t=1}^{n}\mathbf{P}^{(i)}_{st}(1-\mathbf{P}^{(i)}_{st}).$$

Using Assumption 2 and Equation (22), the variance of $(\mathbf{V}^\top\mathbf{E}^{(i)}\mathbf{V})_{kl}$ can be bounded from below by

$$\begin{aligned}
\mathbf{\Sigma}_{\frac{2k+l(l-1)}{2},\frac{2k+l(l-1)}{2}} &= \operatorname{Var}\left((\mathbf{V}^\top\mathbf{E}^{(i)}\mathbf{V})_{kl}\right)\\
&= \sum_{s=1}^{n}\sum_{t=1}^{n}\operatorname{Var}\left(\mathbf{E}^{(i)}_{st}\right)(\mathbf{V}_{sk}\mathbf{V}_{tl})^2 + 2\sum_{s=1}^{n-1}\sum_{t=s+1}^{n}\operatorname{Cov}\left(\mathbf{E}^{(i)}_{st},\mathbf{E}^{(i)}_{ts}\right)(\mathbf{V}_{sk}\mathbf{V}_{tl})(\mathbf{V}_{sl}\mathbf{V}_{tk})\\
&= \sum_{s=1}^{n}\sum_{t=1}^{n}\mathbf{P}^{(i)}_{st}(1-\mathbf{P}^{(i)}_{st})(\mathbf{V}_{sK}\mathbf{V}_{tl})^2 + 2\sum_{s=1}^{n-1}\sum_{t=s+1}^{n}\mathbf{P}^{(i)}_{st}(1-\mathbf{P}^{(i)}_{st})(\mathbf{V}_{sk}\mathbf{V}_{tl})(\mathbf{V}_{sl}\mathbf{V}_{tk})\\
&= \omega\left(\frac{1}{n^2}\right),
\end{aligned} \tag{25}$$

where the last equality follows from Assumption 3. The same calculations allow to obtain a similar upper bound:

$$\mathbf{\Sigma}_{\frac{2k+l(l-1)}{2},\frac{2k+l(l-1)}{2}} \leq \left(\frac{c_2^4}{n^2}+\frac{c_2^4}{n^2}\right) s^2(\mathbf{P}^{(i)}).$$

This shows Equation (21).

To prove part a), define $\mathbf{F}^{(i,n)}$ as

$$\mathbf{F}_{st}^{(i,n)} := \mathbf{E}_{st}^{(i)} \left(\mathbf{V}_{sk}\mathbf{V}_{tl}+\mathbf{V}_{tk}\mathbf{V}_{sl}\right),$$

which is bounded above, because of Assumption 2, by

$$\left|\mathbf{F}_{st}^{(i,n)}\right| \leq \frac{2c_2^2}{n}. \tag{26}$$

Equations (26) and (25) imply that for any $\epsilon > 0$ and sufficiently large n,

$$\left|\mathbf{F}_{st}^{(i,n)}\right| < \epsilon \left[\mathrm{Var}\left((\mathbf{V}^\top \mathbf{E}^{(i)}\mathbf{V})_{kl}\right)\right]^{1/2},$$

for all s,t with $1 \leq s \leq n, 1 \leq t \leq n$. Hence, for all n sufficiently large,

$$\frac{1}{\mathrm{Var}\left((\mathbf{V}^\top \mathbf{E}^{(i)}\mathbf{V})_{kl}\right)} \sum_{s=1}^{n-1}\sum_{t=s+1}^{n} \mathbb{E}\left[|\mathbf{F}_{st}^{(i,n)}|^2 \mathbb{1}\left\{|\mathbf{F}_{st}^{(i,n)}| > \epsilon \left[\mathrm{Var}\left((\mathbf{V}^\top \mathbf{E}^{(i)}\mathbf{V})_{kl}\right)\right]^{1/2}\right\}\right] = 0.$$

This shows that the Lindeberg's condition is satisfied, and by the Lindeberg-Feller Central Limit Theorem applied to Equation (23), we have that

$$\left(\mathbf{\Sigma}_{\frac{2k+l(l-1)}{2},\frac{2k+l(l-1)}{2}}\right)^{-1/2} (\mathbf{V}^\top \mathbf{E}^{(i)}\mathbf{V})_{kl} \xrightarrow{d} \mathcal{N}(0,1).$$

To prove part b), first observe that Assumption 3 holds whenever Assumption 4 holds. This can be observed from Equation (25) and the fact that

$$\lambda_{\min}(\mathbf{\Sigma}^{(i)}) = \min_{x\in\mathbb{R}^r} x^\top \mathbf{\Sigma}^{(i)} x \leq \min_{j\in[r]} \mathbf{\Sigma}_{jj}^{(i)}.$$

Let $\mathbf{s} \in \mathbb{R}^r$ be an arbitrary vector. We show that for any $\mathbf{s}$,

$$\mathbf{s}^\top \mathbf{\Sigma}^{-1/2} \mathrm{vec}(\mathbf{V}^\top \mathbf{E}^{(i)}\mathbf{V}) \xrightarrow{d} \mathbf{s}^\top \mathcal{N}(0,\mathbf{I}_r) = \mathcal{N}(0,\mathbf{s}^\top\mathbf{s}). \tag{27}$$

The above quantity can be expressed as a sum of independent random variables

$$\begin{aligned}
&\mathbf{s}^\top \mathbf{\Sigma}^{-1/2} \mathrm{vec}(\mathbf{V}^\top \mathbf{E}^{(i)}\mathbf{V}) \\
&\quad= \sum_{k=1}^{d}\sum_{l=k}^{d} \mathbf{s}_{\frac{2k+l(l-1)}{2}} \left\{\sum_{k'=1}^{d}\sum_{l'=k'}^{d} \mathbf{\Sigma}^{-1/2}_{\frac{2k+l(l-1)}{2},\frac{2k'+l'(l'-1)}{2}} \left[(\mathbf{V}^\top \mathbf{E}^{(i)}\mathbf{V})_{k'l'}\right]\right\} \\
&\quad= \sum_{s=1}^{n-1}\sum_{t=s+1}^{n} \mathbf{E}_{st}^{(i)} \sum_{k=1}^{d}\sum_{l=k}^{d} \mathbf{s}_{\frac{2k+l(l-1)}{2}} \sum_{k'=1}^{d}\sum_{l'=k'}^{d} \mathbf{\Sigma}^{-1/2}_{\frac{2k+l(l-1)}{2},\frac{2k'+l'(l'-1)}{2}} \left(\mathbf{V}_{sk'}\mathbf{V}_{tl'}+\mathbf{V}_{tk'}\mathbf{V}_{sl'}\right).
\end{aligned}$$

Define

$$\mathbf{G}_{st}^{(i,n)} := \mathbf{E}_{st}^{(i)} \sum_{k=1}^{d}\sum_{l=k}^{d} \mathbf{s}_{\frac{2k+l(l-1)}{2}} \sum_{k'=1}^{d}\sum_{l'=k'}^{d} \mathbf{\Sigma}^{-\frac{1}{2}}_{\frac{2k+l(l-1)}{2},\frac{2k'+l'(l'-1)}{2}} \left(\mathbf{V}_{sk'}\mathbf{V}_{tl'}+\mathbf{V}_{tk'}\mathbf{V}_{sl'}\right).$$

The magnitude of each of these summands can be bounded from above as

$$\begin{aligned}
|\mathbf{G}_{st}^{(i,n)}| &\leq |\mathbf{E}_{st}^{(i)}| \sum_{k=1}^{d}\sum_{l=k}^{d} \left|\mathbf{s}_{\frac{2k+l(l-1)}{2}}\right| \sum_{k'=1}^{d}\sum_{l'=k'}^{d} \left|\mathbf{\Sigma}^{-\frac{1}{2}}_{\frac{2k+l(l-1)}{2},\frac{2k'+l'(l'-1)}{2}}\right| |\mathbf{V}_{sk'}\mathbf{V}_{tl'} + \mathbf{V}_{tk'}\mathbf{V}_{sl'}| \\
&\leq \frac{2c_2^2}{n} \sum_{k=1}^{d}\sum_{l=k}^{d} \left|\mathbf{s}_{\frac{2k+l(l-1)}{2}}\right| \sum_{k'=1}^{d}\sum_{l'=k'}^{d} \left|\mathbf{\Sigma}^{-\frac{1}{2}}_{\frac{2k+l(l-1)}{2},\frac{2k'+l'(l'-1)}{2}}\right| \\
&\leq \frac{2c_2^2}{n} \|\mathbf{s}\|_2 \left\{\sum_{k=1}^{d}\sum_{l=k}^{d} \left(\sum_{k'=1}^{d}\sum_{l'=k'}^{d} \left|\mathbf{\Sigma}^{-\frac{1}{2}}_{\frac{2k+l(l-1)}{2},\frac{2k'+l'(l'-1)}{2}}\right|\right)^2\right\}^{\frac{1}{2}} \\
&\leq \frac{2c_2^2}{n} \|\mathbf{s}\|_2 \left\{\sum_{k=1}^{d}\sum_{l=k}^{d}\sum_{k'=1}^{d}\sum_{l'=k'}^{d} r \left|\mathbf{\Sigma}^{-\frac{1}{2}}_{\frac{2k+l(l-1)}{2},\frac{2k'+l'(l'-1)}{2}}\right|^2\right\}^{\frac{1}{2}} \\
&= \frac{2\sqrt{r}c_2^2}{n} \|\mathbf{s}\|_2 \left\|\mathbf{\Sigma}^{-1/2}\right\|_F \\
&\leq \frac{2\sqrt{r}c_2^2}{n} \|\mathbf{s}\|_2 \left(\sqrt{r}|\lambda_{\max}(\mathbf{\Sigma}^{-1/2})|\right) \\
&\leq \frac{2rc_2^2 \|\mathbf{s}\|_2}{n|\lambda_{\min}(\mathbf{\Sigma})|^{1/2}} = o(\|\mathbf{s}\|_2), \qquad (28)
\end{aligned}$$

where the last equation follows by Assumption 4. The bilinearity of the covariance and Equation (24) gives that

$$\mathrm{Var}\left(\mathbf{s}^\top \mathbf{\Sigma}^{-1/2} \mathrm{vec}(\mathbf{V}^\top \mathbf{E}^{(i)} \mathbf{V})\right) = \mathbf{s}^\top \mathbf{s}. \qquad (29)$$

Equations (29) and (28) imply that for any $\epsilon > 0$ and sufficiently large n,

$$\left|\mathbf{G}_{st}^{(i,n)}\right| < \epsilon \,\mathrm{Var}\left(\mathbf{s}^\top \mathbf{\Sigma}^{-1/2} \mathrm{vec}(\mathbf{V}^\top \mathbf{E}^{(i)} \mathbf{V})\right)^{1/2} = \epsilon \|\mathbf{s}\|_2,$$

uniformly for all s, t with $1 \leq s \leq n, 1 \leq t \leq n$. Hence, for all n sufficiently large,

$$\frac{1}{\|\mathbf{s}\|_2^2} \sum_{s=1}^{n-1}\sum_{t=s+1}^{n} \mathbb{E}\left[|\mathbf{G}_{st}^{(i,n)}|^2 \mathbb{1}\left\{|\mathbf{G}_{st}^{(i,n)}| > \epsilon \|\mathbf{s}\|_2\right\}\right] = 0.$$

This shows that Lindeberg's condition is satisfied for all n sufficiently large, and by the Lindeberg-Feller Central Limit Theorem, we have that the convergence in distribution given by Equation (27) holds. Since this convergence is valid for any $\mathbf{s} \in \mathbb{R}^r$, the Cramér-Wold theorem implies that

$$\mathbf{\Sigma}^{-1/2} \,\mathrm{vec}(\mathbf{V}^\top \mathbf{E}^{(i)} \mathbf{V}) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}_r, \mathbf{I}_r\right).$$

■

Proof [Proof of Theorem 11] The results follow by combining the decomposition given in the left hand side of Equation 18 in Lemma 14 and the asymptotic distribution for the right hand side obtained in Lemma 15. ■

Proof [Proof of Corollary 12] By Lemma 14, there exist a sequence of orthogonal matrices $\mathbf{W}$ such that

$$\widehat{\mathbf{R}}^{(i)} - \mathbf{W}^\top \mathbf{R}^{(i)} \mathbf{W} + \mathbf{H}_m^{(i)} = \mathbf{W}\mathbf{V}^\top \mathbf{E}^{(i)} \mathbf{V}\mathbf{W}.$$

The term of the right hand side converges to a multivariate normal distribution by Lemma 15, and using the same arguments as in Equation (25), the variance of its entry $k, l \in [d], k \leq l$, can be bounded from above as

$$\begin{aligned}\mathrm{Var}([(\mathbf{VW})^\top \mathbf{E}^{(i)}(\mathbf{VW})]_{kl}) =& \sum_{s=1}^{n}\sum_{t=1}^{n} \mathbf{P}^{(i)}_{st}(1-\mathbf{P}^{(i)}_{st})\left((\mathbf{VW})_{sk}(\mathbf{VW})_{tl}\right)^2 + \\ & 2\sum_{s=1}^{n-1}\sum_{t=s+1}^{n} \mathbf{P}^{(i)}_{st}(1-\mathbf{P}^{(i)}_{st})\left((\mathbf{VW})_{sk}(\mathbf{VW})_{tl}\right)\left((\mathbf{VW})_{sl}(\mathbf{VW})_{tk}\right) \\ & \leq \max_{s,k}\left((\mathbf{VW})^4_{sk} + 2(\mathbf{VW})^4_{sk}\right) s^2(\mathbf{P}^{(i)}).\end{aligned}$$

Finally, observe that $|(\mathbf{VW})_{sk}| \leq \|\mathbf{V}_{s\cdot}\|_2 \|\mathbf{W}_{\cdot k}\|_2 \leq \sqrt{\frac{d}{n}}$, which completes the proof. ■

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18:1–86, 2017.

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2007.

Jesus Arroyo and Elizabeth Hou. Efficient distributed estimation of inverse covariance matrices. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pages 1–5. IEEE, jun 2016.

Jesús D. Arroyo Relión, Daniel Kessler, Elizaveta Levina, and Stephan F. Taylor. Network classification with applications to brain connectomics. *Ann. Appl. Stat.*, 13(3):1648–1677, 09 2019.

Avanti Athreya, Carey E Priebe, Minh Tang, Vince Lyzinski, David J Marchette, and Daniel L. Sussman. A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A*, 78: 1–18, 2016.

Avanti Athreya, Donniell E. Fishkind, Keith Levin, Vince Lyzinski, Youngser Park, Yichen Qin, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18:1–92, 2017.

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

Rajendra Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer New York, New York, NY, 1997.

Sharmodeep Bhattacharyya and Shirshendu Chatterjee. Spectral clustering for multiple disassociative sparse networks. *Working Paper*, 2017.

Sharmodeep Bhattacharyya and Shirshendu Chatterjee. Spectral Clustering for Multiple Sparse Networks: I. *arXiv preprint arXiv:1805.10594*, 2018.

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

S. Boccaletti, G. Bianconi, R. Criado, C. I. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, nov 2014.

I. Borg and P. Groenen. Modern Multidimensional Scaling: Theory and Applications. *Journal of Educational Measurement*, 40(3):277–280, sep 2003.

Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews. Neuroscience*, 10(3):186–98, mar 2009.

Jaewon Chung, Benjamin D Pedigo, Eric W Bridgeford, Bijan K Varjavand, Hayden S Helm, and Joshua T Vogelstein. Graspy: Graph statistics in python. *Journal of Machine Learning Research*, 20(158):1–7, 2019.

R. Cameron Craddock, G.Andrew James, Paul E. Holtzheimer, Xiaoping P. Hu, and Helen S. Mayberg. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, 33(8):1914–1928, aug 2012.

Ciprian M Crainiceanu, Brian S Caffo, Sheng Luo, Vadim M Zipunnikov, and Naresh M Punjabi. Population value decomposition, a framework for the analysis of image populations. *Journal of the American Statistical Association*, 106(495):775–790, 2011.

Daniele Durante and David B Dunson. Bayesian inference and testing of group differences in brain networks. *Bayesian Analysis*, 13(1):29–58, 2018.

Daniele Durante, David B. Dunson, and Joshua T. Vogelstein. Nonparametric Bayes Modeling of Populations of Networks. *Journal of the American Statistical Association*, 112(520):1516–1530, oct 2017.

Paul Erdős and Alfred Rényi. On random graphs. *Publ. Math. Debrecen*, 6:290–297, 1959.

Jianqing Fan, Dong Wang, Kaizheng Wang, and Ziwei Zhu. Distributed estimation of principal eigenspaces. *Ann. Statist.*, 47(6):3009–3031, 12 2019.

E. N. Gilbert. Random graphs. *Ann. Math. Statist.*, 30(4):1141–1144, 1959.

Cedric E. Ginestet, Jun Li, Prakash Balachandran, Steven Rosenberg, and Eric D. zyk. Hypothesis testing for network data in functional neuroimaging. *Ann. Appl. Stat.*, 11(2):725–750, 06 2017.

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

Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2009.

Jing Dong J Han, Nicolas Berlin, Tong Hao, Debra S. Goldberg, Gabriel F. Berriz, Lan V. Zhang, Denis Dupuy, Albertha J. M. Walhout, Michael E. Cusick, Frederick P. Roth, and Marc Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93, jul 2004.

Qiuyi Han, Kevin Xu, and Edoardo Airoldi. Consistent estimation of dynamic and multi-layer block models. In *International Conference on Machine Learning*, pages 1511–1520, 2015.

Peter D. Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

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

Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, oct 2012.

Philippe Jaming and M Matolcsi. On the existence of flat orthogonal matrices. In *Acta Mathematica Hungarica*, volume 147, pages 179–188. Springer Verlag, 2015.

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

Gregory Kiar, Eric Bridgeford, Will Gray Roncal, Consortium for Reliability (CoRR), Reproducibility, Vikram Chandrashekhar, Disa Mhembere, Sephira Ryman, Xi-Nian Zuo, Daniel S Marguiles, R Cameron Craddock, Carey E Priebe, Rex Jung, Vince Calhoun, Brian Caffo, Randal Burns, Michael P Milham, and Joshua Vogelstein. A high-throughput pipeline identifies robust connectomes but troublesome variability. *bioRxiv*, page 188706, 2018.

Yura Kim and Elizaveta Levina. Graph-aware linear mixed effects models for brain connectivity networks. *arXiv preprint arXiv:1903.02129*, 2019.

Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, Mason A Porter, and Ernesto Estrada. Multilayer networks. *Journal of Complex Networks*, 2:203–271, 2014.

Eric D. Kolaczyk. *Topics at the Frontier of Statistics and Network Analysis*. Cambridge University Press, Cambridge, 2017.

Pierre Latouche, Etienne Birmelé, and Christophe Ambroise. Overlapping stochastic block models with application to the french political blogosphere. *Ann. Appl. Stat.*, 5(1):309–336, 03 2011.

Can M Le, Elizaveta Levina, and Roman Vershynin. Concentration and regularization of random graphs. *Random Structures and Algorithms*, 51(3):538–561, 2017.

Jason D. Lee, Qiang Liu, Yuekai Sun, and Jonathan E. Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(1):115—-144, 2017.

Keith Levin, Avanti Athreya, Minh Tang, Vince Lyzinski, and Carey E. Priebe. A central limit theorem for an omnibus embedding of multiple random dot product graphs. In *arXiv preprint arXiv:1705.09355*, 2017.

Wenyuan Li, Chun Chi Liu, Tong Zhang, Haifeng Li, Michael S. Waterman, and Xianghong Jasmine Zhou. Integrative analysis of many weighted co-expression networks using tensor computation. *PLoS Computational Biology*, 7(6):e1001106, 2011.

Vince Lyzinski, Daniel L. Sussman, Minh Tang, Avanti Athreya, and Carey E. Priebe. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic Journal of Statistics*, 8:2905–2922, 2014.

Vince Lyzinski, Minh Tang, Avanti Athreya, Youngser Park, and Carey E Priebe. Community detection and classification in hierarchical stochastic blockmodels. *IEEE Transactions on Network Science and Engineering*, 4(1):13–26, 2016.

Catherine Matias and Vincent Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 79(4):1119–1141, 2017.

Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980): 876–878, 2010.

Agnes Martine Nielsen and Daniela Witten. The Multiple Random Dot Product Graph Model. *arXiv preprint arXiv:1811.12172*, 2018.

Subhadeep Paul and Yuguo Chen. Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *Ann. Statist.*, 48(1):230–250, 02 2020.

Dragana M. Pavlović, Bryan R.L. Guillaume, Emma K. Towlson, Nicole M.Y. Kuek, Soroosh Afyouni, Petra E. Vértes, B.T. Thomas Yeo, Edward T. Bullmore, and Thomas E. Nichols. Multi-subject stochastic blockmodels for adaptive analysis of individual differences in human brain network cluster structure. *NeuroImage*, 220:116611, 2020.

Tiago P Peixoto. Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Physical Review E*, 92(4):042807, 2015.

Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, and Bradley L Schlaggar. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.

Carey E. Priebe, Youngser Park, Joshua T. Vogelstein, John M. Conroy, Vince Lyzinski, Minh Tang, Avanti Athreya, Joshua Cape, and Eric Bridgeford. On a two-truths phenomenon in spectral graph clustering. *Proceedings of the National Academy of Sciences*, 116(13):5995–6000, 2019.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.*, 39(4):1878–1915, 08 2011.

Patrick Rubin-Delanchy, Carey E. Priebe, Minh Tang, and Joshua Cape. A statistical interpretation of spectral embedding: the generalised random dot product graph. *arXiv preprint arXiv:1709.05506*, 2017.

Daniel L. Sussman, Minh Tang, Donniell E. Fishkind, and Carey E. Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

Daniel L. Sussman, Minh Tang, Carey E. Priebe, Minh Tang, and Carey E. Priebe. Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):48–57, jan 2014.

Michael Szell, Renaud Lambiotte, and Stefan Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31): 13636—-13641, 2010.

Minh Tang. The eigenvalues of stochastic blockmodel graphs. *arXiv preprint arXiv:1803.11551*, 2018.

Minh Tang, Avanti Athreya, Daniel L. Sussman, Vince Lyzinski, and Carey E. Priebe. A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics*, 26(2):344–354, 2017a.

Minh Tang, Avanti Athreya, Daniel L Sussman, Vince Lyzinski, and Carey E Priebe. A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli*, 23(3):1599–1630, 2017b.

Runze Tang, Michael Ketcha, Alexandra Badea, Evan D. Calabrese, Daniel S. Margulies, Joshua T. Vogelstein, Carey E. Priebe, and Daniel L. Sussman. Connectome smoothing via low-rank approximations. *IEEE Transactions on Medical Imaging*, 38(6):1446–1456, 2019.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Lu Wang, Zhengwu Zhang, and David Dunson. Symmetric Bilinear Regression for Signal Subgraph Estimation. *IEEE Transactions on Signal Processing*, 67(7):1929–1940, 2019a.

Lu Wang, Zhengwu Zhang, and David Dunson. Common and individual structure of brain networks. *Ann. Appl. Stat.*, 13(1):85–112, 03 2019b.

Shangsi Wang, Jesús Arroyo, Joshua T Vogelstein, and Carey E Priebe. Joint Embedding of Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, (in press)*, 2019c.

Stephen J. Young and Edward R. Scheinerman. Random Dot Product Graph Models for Social Networks. In *Algorithms and Models for the Web-Graph*, pages 138–149. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Anru Zhang and Dong Xia. Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338, 2018.

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Detecting overlapping communities in networks using spectral methods. *SIAM Journal on Mathematics of Data Science*, 2(2):265–283, 2020.

Zhengwu Zhang, Genevera I. Allen, Hongtu Zhu, and David Dunson. Tensor network factorizations: Relationships between brain structural connectomes and traits. *NeuroImage*, 197:330–343, 2019.

Mu Zhu and Ali Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930, 2006.

Xi-Nian Zuo, Jeffrey S Anderson, Pierre Bellec, Rasmus M Birn, Bharat B Biswal, Janusch Blautzik, John C.S Breitner, Randy L Buckner, Vince D Calhoun, F. Xavier Castellanos, Antao Chen, Bing Chen, Jiangtao Chen, Xu Chen, Stanley J Colcombe, William Courtney, R Cameron Craddock, Adriana Di Martino, Hao-Ming Dong, Xiaolan Fu, Qiyong Gong, Krzysztof J Gorgolewski, Ying Han, Ye He, Yong He, Erica Ho, Avram Holmes, Xiao-Hui Hou, Jeremy Huckins, Tianzi Jiang, Yi Jiang, William Kelley, Clare Kelly, Margaret King, Stephen M LaConte, Janet E Lainhart, Xu Lei, Hui-Jie Li, Kaiming Li, Kuncheng Li, Qixiang Lin, Dongqiang Liu, Jia Liu, Xun Liu, Yijun Liu, Guangming Lu, Jie Lu, Beatriz Luna, Jing Luo, Daniel Lurie, Ying Mao, Daniel S Margulies, Andrew R Mayer, Thomas Meindl, Mary E Meyerand, Weizhi Nan, Jared A Nielsen, David O'Connor, David Paulsen, Vivek Prabhakaran, Zhigang Qi, Jiang Qiu, Chunhong Shao, Zarrar Shehzad, Weijun Tang, Arno Villringer, Huiling Wang, Kai Wang, Dongtao Wei, Gao-Xia Wei, Xu-Chu Weng, Xuehai Wu, Ting Xu, Ning Yang, Zhi Yang, Yu-Feng Zang, Lei Zhang, Qinglin Zhang, Zhe Zhang, Zhiqiang Zhang, Ke Zhao, Zonglei Zhen, Yuan Zhou, Xing-Ting Zhu, and Michael P Milham. An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific Data*, 1:140049, 2014.