Statistical Inference on Random Dot Product Graphs: a Survey

Avanti Athreya   DATHREY1@JHU.EDU
Donniell E. Fishkind   DEF@JHU.EDU
Minh Tang   MTANG10@JHU.EDU
Carey E. Priebe   CEP@JHU.EDU
Department of Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD, 2128, USA

Youngser Park   YOUNGSER@JHU.EDU
Center for Imaging Science
Johns Hopkins University
Baltimore, MD, 21218, USA

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

Keith Levin   KLEVIN@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI, 48109, USA

Vince Lyzinski   VLYZINSKI@UMASS.EDU
Department of Mathematics and Statistics
University of Massachusetts
Amherst, MA, 01003-9305, USA

Yichen Qin   QINYN@UCMAIL.UC.EDU
Department of Operations, Business Analytics, and Information Systems, College of Business
University of Cincinnati
Cincinnati, OH, 45221-0211, USA

Daniel L Sussman   SUSSMAN@BU.EDU
Department of Mathematics and Statistics
Boston University
Boston, MA, 02215, USA

Abstract

The random dot product graph (RDPG) is an independent-edge random graph that is analytically tractable and, simultaneously, either encompasses or can successfully approximate

a wide range of random graphs, from relatively simple stochastic block models to complex latent position graphs. In this survey paper, we describe a comprehensive paradigm for statistical inference on random dot product graphs, a paradigm centered on spectral embeddings of adjacency and Laplacian matrices. We examine the graph-inferential analogues of several canonical tenets of classical Euclidean inference. In particular, we summarize a body of existing results on the consistency and asymptotic normality of the adjacency and Laplacian spectral embeddings, and the role these spectral embeddings can play in the construction of single- and multi-sample hypothesis tests for graph data. We investigate several real-world applications, including community detection and classification in large social networks and the determination of functional and biologically relevant network properties from an exploratory data analysis of the *Drosophila* connectome. We outline requisite background and current open problems in spectral graph inference.

Keywords: Random dot product graph, adjacency spectral embedding, Laplacian spectral embedding, multi-sample graph hypothesis testing, semiparametric modeling

1. Introduction

Random graph inference is an active, interdisciplinary area of current research, bridging combinatorics, probability, statistical theory, and machine learning, as well as a wide spectrum of application domains from neuroscience to sociology. Statistical inference on random graphs and networks, in particular, has witnessed extraordinary growth over the last decade: see, for example, Goldenberg et al. (2010) and Kolaczyk (2009) for a discussion of the considerable applications in recent network science of several canonical random graph models.

Of course, combinatorial graph theory itself is centuries old—indeed, in his resolution to the problem of the bridges of Königsberg, Leonard Euler first formalized graphs as mathematical objects consisting of vertices and edges. The notion of a random graph, however, and the modern theory of inference on such graphs, is comparatively new, and owes much to the pioneering work of Erdős, Rényi, and others in the late 1950s. E.N. Gilbert's short 1959 paper (Gilbert, 1959) considered a random graph for which the existence of edges between vertices are independent Bernoulli random variables with common probability p; roughly concurrently, Erdős and Rényi provided the first detailed analysis of the probabilities of the emergence of certain types of subgraphs within such graphs (Erdős and Rényi, 1960), and today, graphs in which the edges arise independently and with common probability p are known as *Erdős-Rényi* (or ER) graphs.

The Erdős-Rényi (ER) model is one of the simplest generative models for random graphs, but this simplicity belies astonishingly rich behavior (see Alon and Spencer, 2008; Bollobás et al., 2007). Nevertheless, in many applications, the requirement of a common connection probability is too stringent: graph vertices often represent heterogeneous entities, such as different people in a social network or cities in a transportation graph, and the connection probability p_{ij} between vertex i and j may well change with i and j or depend on underlying attributes of the vertices. Moreover, these heterogeneous vertex attributes may not be observable; for example, given the adjacency matrix of a Facebook community, the specific interests of the individuals may remain hidden. To more effectively model such real-world

networks, we consider *latent position* random graphs (Hoff et al., 2002). In a latent position graph, to each vertex i in the graph there is associated an element x_i of the so-called *latent space* $\mathcal{X}$, and the probability of connection p_{ij} between any two edges i and j is given by a *link* or *kernel* function $\kappa : \mathcal{X} \times \mathcal{X} \to [0, 1]$. That is, the edges are generated independently (so the graph is an *independent-edge* graph) and $p_{ij} = \kappa(x_i, x_j)$.

The *random dot product graph* (RDPG) of Young and Scheinerman (Young and Scheinerman, 2007) is an especially tractable latent position graph; here, the latent space is an appropriately constrained subspace of Euclidean space $\mathbb{R}^d$, and the link function is simply the dot or inner product of the pair of d-dimensional latent positions. Thus, in a d-dimensional random dot product graph with n vertices, the latent positions associated to the vertices can be represented by an $n \times d$ matrix $\mathbf{X}$ whose rows are the latent positions, and the matrix of connection probabilities $\mathbf{P} = (\mathbf{P}_{ij})$ is given by $\mathbf{P} = \mathbf{X}\mathbf{X}^\top$. Conditional on this matrix $\mathbf{P}$, the RDPG has an adjacency matrix $\mathbf{A} = (\mathbf{A}_{ij})$ whose entries are Bernoulli random variables with probability $\mathbf{P}_{ij}$. For simplicity, we will typically consider symmetric, *hollow* RDPG graphs; that is, undirected, unweighted graphs in which $\mathbf{A}_{ii} = 0$, so there are no self-edges. In our real data analysis of a neural connectome in Section 6.3, however, we describe how to adapt our results to weighted and directed graphs.

In any latent position graph, the latent positions associated to graph vertices can themselves be random; for instance, the latent positions may be independent, identically distributed random variables with some distribution F on $\mathbb{R}^d$. The well-known *stochastic blockmodel* (SBM), in which each vertex belongs to one of K subsets known as *blocks*, with connection probabilities determined solely by block membership (Holland et al., 1983), can be represented as a random dot product graph in which all the vertices in a given block have the same latent positions (or, in the case of random latent positions, an RDPG for which the distribution F is supported on a finite set). Despite their structural simplicity, stochastic block models are the building blocks for all independent-edge random graphs; in Wolfe and Olhede (2013), the authors demonstrate that any independent-edge random graph can be well-approximated by a stochastic block model with a sufficiently large number of blocks. Since stochastic block models can themselves be viewed as random dot product graphs, we see that suitably high-dimensional random dot product graphs can provide accurate approximations of latent position graphs (Tang et al., 2013), and, in turn, independent-edge graphs. Thus, the architectural simplicity of the random dot product graph makes it particularly amenable to analysis, and its near-universality in graph approximation renders it expansively applicable. In addition, the cornerstone of our analysis of random dot product graphs is a set of classical probabilistic and linear algebraic techniques that are useful in much broader settings, such as random matrix theory. As such, the random dot product graph is both a rich and interesting object of study in its own right and a natural point of departure for wider graph inference.

A classical inference task for Euclidean data is to estimate, from sample data, certain underlying distributional parameters. Similarly, for a latent position graph, a classical graph inference task is to infer the graph parameters from an observation of the adjacency matrix $\mathbf{A}$. Indeed, our overall paradigm for random graph inference is inspired by the fundamental tenets of classical statistical inference for Euclidean data. Namely, our goal

is to construct methods and estimators of graph parameters or graph distributions; and, for these estimators, to analyze their (1) consistency; (2) asymptotic distributions; (3) asymptotic relative efficiency; (4) robustness to model misspecification; and (5) implications for subsequent inference including one- and multi-sample hypothesis testing. In this paper, we summarize and synthesize a considerable body of work on spectral methods for inference in random dot product graphs, all of which not only advance fundamental tenets of this paradigm, but do so within a unified and parsimonious framework. The random graph estimators and test statistics we discuss all exploit the *adjacency spectral embedding* (ASE) or the *Laplacian spectral embedding* (LSE), which are eigendecompositions of the adjacency matrix $\mathbf{A}$ and *normalized* Laplacian matrix $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is the diagonal degree matrix $\mathbf{D}_{ii} = \sum_{j \neq i} \mathbf{A}_{ij}$.

The ambition and scope of our approach to graph inference means that mere upper bounds on discrepancies between parameters and their estimates will not suffice. Such bounds are legion. In our proofs of consistency, we improve several bounds of this type, and in some cases improve them so drastically that concentration inequalities and asymptotic limit distributions emerge in their wake. We stress that aside from specific cases (see Füredi and Komlós, 1981; Tao and Vu, 2012; Lei, 2016), limiting distributions for eigenvalues and eigenvectors of random graphs are notably elusive. For the adjacency and Laplacian spectral embedding, we discuss not only consistency, but also asymptotic normality, robustness, and the use of the adjacency spectral embedding in the nascent field of multi-graph hypothesis testing. We illustrate how our techniques can be meaningfully applied to thorny and very sizable real data, improving on previously state-of-the-art methods for inference tasks such as community detection and classification in networks. What is more, as we now show, spectral graph embeddings are relevant to many complex and seemingly disparate aspects of graph inference.

A bird's-eye view of our methodology might well start with the stochastic blockmodel. For an SBM with a finite number of blocks of stochastically equivalent vertices, in Sussman et al. (2012) and Fishkind et al. (2013), we establish that k-means clustering of the rows of the adjacency spectral embedding accurately partitions the vertices into the correct blocks, even when the embedding dimension is misspecified or the number of blocks is unknown. Furthermore, in Lyzinski et al. (2014) and Lyzinski et al. (2017) we give a significant improvement in the misclassification rate, by exhibiting an almost-surely perfect clustering in which, in the limit, no vertices whatsoever are misclassified. For random dot product graphs more generally, we show in Sussman et al. (2014) that the latent positions are consistently estimated by the embedding, which then allows for accurate learning in a supervised vertex classification framework. In Tang et al. (2013), these results are extended to more general latent position models, establishing a powerful universal consistency result for vertex classification in general latent position graphs, and also exhibiting an efficient embedding of vertices which were not observed in the original graph. In Athreya et al. (2016) and Tang and Priebe (2018), we supply distributional results, akin to a central limit theorem, for both the adjacency and Laplacian spectral embedding, respectively; the former leads to a nontrivially superior algorithm for the estimation of block memberships in a stochastic block model (Suwan et al., 2016), and the latter resolves, through an elegant

comparison of Chernoff information, a long-standing open question of the relative merits of the adjacency and Laplacian graph representations.

Morever, graph embedding plays a central role in the foundational work on hypothesis testing of Tang et al. (2017a) and Tang et al. (2017b) for two-sample graph comparison: these papers provide theoretically justified, valid and consistent hypothesis tests for the semiparamatric problem of determining whether two random dot product graphs have the same latent positions and the nonparametric problem of determining whether two random dot product graphs have the same underlying distributions. This, then, yields a systematic framework for determining statistical similarity across graphs, which in turn underpins yet another provably consistent algorithm for the decomposition of random graphs with a hierarchical structure Lyzinski et al. (2017). In Levin et al. (2017), distributional results are given for an omnibus embedding of multiple random dot product graphs on the same vertex set, and this embedding performs well both for latent position estimation and for multi-sample graph testing. For the critical inference task of vertex nomination, in which the inference goal is to produce an ordering of vertices of interest (see, for instance Coppersmith, 2014), we find in Fishkind et al. (2015a) an array of principled vertex nomination algorithms —-the canonical, maximum likelihood and spectral vertex nomination schemes—and a demonstration of the algorithms' effectiveness on both synthetic and real data. In Lyzinski et al. (2016b) the consistency of the maximum likelihood vertex nomination scheme is established, a scalable restricted version of the algorithm is introduced, and the algorithms are adapted to incorporate general vertex features.

Overall, we stress that these principled techniques for random dot product graphs exploit the Euclidean nature of graph embeddings but are general enough to yield meaningful results for a wide variety of random graphs. Because our focus is, in part, on spectral methods, and because the adjacency matrix $\mathbf{A}$ of an independent-edge graph can be regarded as a noisy version of the matrix of probabilities $\mathbf{P}$ (Oliveira, 2009), we rely on several classical results on matrix perturbations, most prominently the Davis-Kahan Theorem (see Bhatia (1997) for the theorem itself, Rohe et al. (2011) for an illustration of its role in graph inference, and Yu et al. (2015) for a very useful variant). We also depend on the aforementioned spectral bounds in Oliveira (2009) and a more recent sharpening due to Lu and Peng (Lu and Peng, 2013). We leverage probabilistic concentration inequalities, such as those of Hoeffding and Bernstein (Tropp, 2015). Finally, several of our results do require suitable eigengaps for $\mathbf{P}$ and lower bounds on graph density, as measured by the maximum degree and the size of the smallest eigenvalue of $\mathbf{P}$. It is important to point out that in our analysis, we assume that the embedding dimension d of our graphs is known and fixed. In real data applications, such an embedding dimension is not known, and in Section 6.3, we discuss approaches (see Chatterjee, 2015; Zhu and Ghodsi, 2006) to estimating the embedding dimension. Robustness of our procedures to errors in embedding dimension is a problem of current investigation.

In the study of stochastic blockmodels, there has been a recent push to understand the fundamental information-theoretic limits for community detection and graph partitioning (see Mossel et al., 2018, 2014; Abbe and Sandon, 2015; Abbe et al., 2016). These bounds are typically algorithm-free and focus on stochastic blockmodels with constant or logarithmic

average degree, in which differences between vertices in different blocks are assumed to be at the boundary of detectability. Our efforts have a somewhat different flavor, in that we seek to understand the precise behavior of a widely applicable procedure in a more general model. Additionally, we treat sparsity as a secondary concern, and typically do not broach the question of the exact limits of our procedures. Our spectral methods may not be optimal for all random graph models, of course (Krzakala et al., 2013; Kawamoto and Kabashima, 2015), but they are very useful, in that they rely on well-optimized computational methods, can be implemented quickly in many standard languages, extend readily to other models, and serve as a foundation for more complex analyses.

Finally, we would be remiss not to point out that while spectral decompositions and clusterings of the adjacency matrix are appropriate for graph inference, they are also of considerable import in combinatorial graph theory: readers may recall, for instance, the combinatorial *ratio-cut* problem, whose objective is to partition the vertex set of a graph into two disjoint sets in a way that minimizes the number of edges between vertices in the two sets. The minimizer of a relaxation to the ratio-cut problem (Fiedler, 1973) is the eigenvector associated to the second smallest eigenvalue of the graph Laplacian $\mathbf{L}$. While we do not pursue more specific combinatorial applications of spectral methods here, we note that Chung (1997) provides a comprehensive overview, and von Luxburg (2007) gives an accessible tutorial on spectral methods.

We organize the paper as follows. In Section 2, we define random dot product graphs and the adjacency spectral embedding, and we recall important linear algebraic background. In Section 4, we discuss consistency, asymptotic normality, and hypothesis testing, as well as inference for hierarchical models. In Section 6, we discuss applications of these results to real data. Finally, in Section 7 we discuss current theoretical and computational difficulties and open questions, including issues of optimal embedding dimension, model limitations, robustness to errorful observations, and joint graph inference. Finally, in A, we provide further details in the proofs of our main results.

2. Definitions, notation, and background

2.1 Preliminaries and notation

We begin by establishing notation. For a positive integer n, we let $[n] = \{1, 2, \cdots, n\}$. For a vector $\mathbf{v} \in \mathbb{R}^n$, we let $\|\mathbf{v}\|$ denote the Euclidean norm of $\mathbf{v}$. We denote the identity matrix, zero matrix, and the square matrix of all ones by, $\mathbf{I}$, $\mathbf{0}$, and $\mathbf{J}$, respectively. We use $\otimes$ to denote the Kronecker product. For an $n_1 \times n_2$ matrix $\mathbf{H}$, we let $\mathbf{H}_{ij}$ denote its i, jth entry; we denote by $\mathbf{H}_{\cdot j}$ the column vector formed by the j-th column of $\mathbf{H}$; and we denote by $\mathbf{H}_{i\cdot}$ the row vector formed by the i-th row of $\mathbf{H}$. For a slight abuse of notation, we also let $\mathbf{H}_i \in \mathbb{R}^{n_2}$ denote the *column* vector formed by transposing the i-th row of $\mathbf{H}$. That is, $\mathbf{H}_i = (\mathbf{H}_{i\cdot})^\top$. Given any suitably specified ordering on eigenvalues of a square matrix $\mathbf{H}$, we let $\lambda_i(\mathbf{H})$ denote the i-th eigenvalue (under such an ordering) of $\mathbf{H}$ and $\sigma_i(\mathbf{H}) = \sqrt{\lambda_i(\mathbf{H}^\top \mathbf{H})}$ the i-th singular value of $\mathbf{H}$. We let $\|\mathbf{H}\|$ denote the spectral norm of $\mathbf{H}$ and $\|\mathbf{H}\|_F$ denote the Frobenius norm of $\mathbf{H}$. We let $\|\mathbf{H}\|_{2\to\infty}$ denote the maximum of the Euclidean norms of

the rows of $\mathbf{H}$, so that $\|\mathbf{H}\|_{2\to\infty} = \max_i \|\mathbf{H}_i\|$. We denote the trace of a matrix $\mathbf{H}$ by $\mathrm{tr}(\mathbf{H})$. For an $n \times n$ symmetric matrix $\mathbf{H}$ whose entries are all non-negative, we will frequently have to account for terms related to matrix sparsity, and we define $\delta(\mathbf{H})$ and $\gamma(\mathbf{H})$ as follows:

$$\delta(\mathbf{H}) = \max_{1 \le i \le n} \sum_{j=1}^n \mathbf{H}_{ij} \qquad \gamma(\mathbf{H}) = \frac{\sigma_d(\mathbf{H}) - \sigma_{d+1}(\mathbf{H})}{\delta(\mathbf{H})} \tag{1}$$

(We remark that when H is a matrix of connection probabilities for a random graph, δ can be interpreted as the maximum expected degree, and γ relates this quantity to an eigengap that is especially useful in the case of probability matrices of suitably low rank.) In a number of cases, we need to consider a sequence of matrices. We will denote such a sequence by $\mathbf{H}_n$, where n is typically used to denote the index of the sequence. The distinction between a particular element $\mathbf{H}_n$ in a sequence of matrices and a particular row $\mathbf{H}_i$ of a matrix will be clear from context, and our convention is typically to use n to denote the index of a sequence and i or h to denote a particular row of a matrix. In the case where we need to consider the ith row of a matrix that is itself the nth element of a sequence, we will use the notation $(\mathbf{H}_n)_i$.

We define a *graph* G to be an ordered pair of (V, E) where V is the so-called *vertex* or *node* set, and E, the set of *edges*, is a subset of the Cartesian product of $V \times V$. In a graph whose vertex set has cardinality n, we will usually represent V as $V = \{1, 2, \cdots, n\}$, and we say there *is an edge between* i and j if $(i, j) \in E$. The *adjacency* matrix $\mathbf{A}$ provides a compact representation of such a graph:

$$\mathbf{A}_{ij} = 1 \text{ if } (i, j) \in E, \text{ and } \mathbf{A}_{ij} = 0 \text{ otherwise.}$$

Where there is no danger of confusion, we will often refer to a graph G and its adjacency matrix $\mathbf{A}$ interchangeably.

Our focus is random graphs, and thus we will let Ω denote our sample space, $\mathcal{F}$ the σ-algebra of subsets of Ω and $\mathbb{P}$ our probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$. We will denote the expectation of a (potentially multi-dimensional) random variable X with respect to this measure by $\mathbb{E}$. Given an event $F \in \mathcal{F}$, we denote its complement by F^c, and we let $\mathrm{Pr}(F)$ denote the probability of F. As we will see, in many cases we can choose Ω to be subset of Euclidean space. Because we are interested in large-graph inference, we will frequently need to demonstrate that probabilities of certain events decay at specified rates. This motivates the following definition.

Definition 1 (Convergence a.a.s. and w.h.p.) *Given a sequence of events* $\{F_n\} \in \mathcal{F}$, *where* $n = 1, 2, \cdots$, *we say that* F_n *occurs* asymptotically almost surely *(a.a.s.) if* $\mathrm{Pr}(F_n) \to 1$ *as* $n \to \infty$. *We say that* F_n *occurs with high probability (w.h.p.), and write* F_n *w.h.p. , if for any* $c_0 > 1$, *there exists finite positive constant* C_0 *depending on* c_0 *such that* $\mathrm{Pr}[F_n^c] \le C_0 n^{-c_0}$ *for all* n. *We note that* F_n *occurring w.h.p. is stronger than* F_n *occurring asymptotically almost surely. Morever,* F_n *occurring with high probability implies, by the Borel-Cantelli Lemma Chung (2001), that with probability* 1 *there exists an* n_0 *such that* F_n *holds for all* $n \ge n_0$.

Moreover, since our goal is often to understand large-graph inference, we need to consider asymptotics as a function of graph size n. As such, we recall familiar asymptotic notation:

Definition 2 (Asymptotic notation) *If $w(n)$ is a quantity depending on n, we will say that w is of order $\alpha(n)$ and use the notation $w(n) \sim \Theta(\alpha(n))$ to denote that there exist positive constants c, C such that for n sufficiently large,*

$$c\alpha(n) \leq w(n) \leq C\alpha(n).$$

When the quantity $w(n)$ is clear and $w(n) \sim \Theta(\alpha(n))$, we sometimes simply write "w is of order $\alpha(n)$". We write $w(n) \sim O(n)$ if there exists a constant C such that for n sufficiently large, $w(n) \leq Cn$. We write $w(n) \sim o(n)$ if $w(n)/n \to 0$ as $n \to \infty$, and $w(n) \sim o(1)$ if $w(n) \to 0$ as $n \to \infty$. We write $w(n) \sim \Omega(n)$ if there exists a constant C such that for all n sufficiently large, $w(n) \geq Cn$.

Throughout, we will use $C > 0$ to denote a constant, not depending on n, which may vary from one line to another.

2.2 Models

Since our focus is on d-dimensional random dot product graphs, we first define an *an inner product distribution* as a probability distribution over a suitable subset of $\mathbb{R}^d$, as follows:

Definition 3 (d-dimensional inner product distribution) *Let F be a probability distribution whose support is given by $\operatorname{supp} F = \mathcal{X}_d \subset \mathbb{R}^d$. We say that F is a d-dimensional inner product distribution on $\mathbb{R}^d$ if for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}_d = \operatorname{supp} F$, we have $\mathbf{x}^\top \mathbf{y} \in [0, 1]$.*

Next, we define a random dot product graph as an independent-edge random graph for which the edge probabilities are given by the dot products of the latent positions associated to the vertices. We restrict our attention here to graphs that are undirected and in which no vertex has an edge to itself.

Definition 4 (RDPG with distribution F) *Let F be a d-dimensional inner product distribution with $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \overset{i.i.d.}{\sim} F$, collected in the rows of the matrix*

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n]^\top \in \mathbb{R}^{n \times d}.$$

Suppose $\mathbf{A}$ is a random adjacency matrix given by

$$\Pr[\mathbf{A}|\mathbf{X}] = \prod_{i<j} (\mathbf{X}_i^\top \mathbf{X}_j)^{\mathbf{A}_{ij}} (1 - \mathbf{X}_i^\top \mathbf{X}_j)^{1-\mathbf{A}_{ij}} \tag{2}$$

We then write $(\mathbf{A}, \mathbf{X}) \sim \operatorname{RDPG}(F, n)$ and say that $\mathbf{A}$ is the adjacency matrix of a random dot product graph (RDPG) of dimension or rank at most d and with latent positions given by the rows of $\mathbf{X}$. If $\mathbf{X}\mathbf{X}^\top$ is, in fact, a rank d matrix, we say $\mathbf{A}$ is the adjacency matrix of a rank d random dot product graph.

Next, for such random dot product graphs, we frequently need to consider the $d \times d$ *second moment matrix* Δ, defined as follows:

Definition 5 (Second moment matrix) *Suppose A is an adjacency matrix from the random dot product graph model with distribution F. Let $X_1 \sim F$ be one of the i.i.d vectors of latent positions. Define the* second moment matrix Δ *via the expectation*

$$\Delta = E[\mathbf{X}_1 \mathbf{X}_1^\top]$$

While our notation for a random dot product graph with distribution F is $(\mathbf{A}, \mathbf{X}) \sim$ RDPG(F), we emphasize that in this paper the latent positions $\mathbf{X}$ are always assumed to be unobserved. An almost identical definition holds for random dot product graphs with fixed but unobserved latent positions:

Definition 6 (RDPG with fixed latent positions) *In the definition 4 given above, the latent positions are themselves random. If, instead, the latent positions are given by a fixed matrix $\mathbf{X}$ and, given this matrix, the graph is generated according to Eq.(2), we say that $\mathbf{A}$ is a realization of a random dot product graph with latent positions $\mathbf{X}$, and we write $\mathbf{A} \sim$ RDPG($\mathbf{X}$).*

Remark 7 (Nonidentifiability) *Given a graph distributed as an RDPG, the natural task is to recover the latent positions $\mathbf{X}$ that gave rise to the observed graph. However, the RDPG model has an inherent nonidentifiability: let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a matrix of latent positions and let $\mathbf{W} \in \mathbb{R}^{d \times d}$ be a unitary matrix. Since $\mathbf{X}\mathbf{X}^\top = (\mathbf{X}\mathbf{W})(\mathbf{X}\mathbf{W})^\top$, it is clear that the latent positions $\mathbf{X}$ and $\mathbf{X}\mathbf{W}$ give rise to the same distribution over graphs in Equation (2). Note that most latent position models, as defined below, also suffer from similar types of non-identifiability as edge-probabilities may be invariant to various transformations.*

As we mentioned, the random dot product graph is a specific instance of the more general *latent position random graph* with *link* or *kernel* function κ. Indeed, the latent positions themselves need not belong to Euclidean space per se, and the link function need not be an inner product.

Definition 8 (Latent position random graph with kernel κ) *Let $\mathcal{X}$ be a set and $\kappa : \mathcal{X} \times \mathcal{X} \to [0,1]$ a symmetric function. Suppose to each $i \in [n]$ there is associated a point $\mathbf{X}_i \in \mathcal{X}$. Given $\mathbf{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ consider the graph with adjacency matrix $\mathbf{A}$ defined by*

$$\Pr[\mathbf{A}|\mathbf{X}] = \prod_{i<j} \kappa(\mathbf{X}_i, \mathbf{X}_j)^{\mathbf{A}_{ij}} (1 - \kappa(\mathbf{X}_i, \mathbf{X}_j))^{1-\mathbf{A}_{ij}} \tag{3}$$

Then $\mathbf{A}$ is the adjacency matrix of a latent position random graph with latent position $\mathbf{X}$ and link function κ.

Similarly, we can define independent edge graphs for which latent positions need not play a role.

Definition 9 (Independent-edge graphs) *For a matrix symmetric matrix $\mathbf{P}$ of probabilities, we say that $\mathbf{A}$ is distributed as an independent edge graph with probabilities $\mathbf{P}$ if*

$$\Pr[\mathbf{A}|\mathbf{X}] = \prod_{i<j} \mathbf{P}_{ij}^{\mathbf{A}_{ij}} (1 - \mathbf{P}_{ij})^{1-\mathbf{A}_{ij}} \tag{4}$$

By their very structure, latent position random graphs, for fixed latent positions, are independent-edge random graphs. In general, for any latent position graph the matrix of edge probabilities $\mathbf{P}$ is given by $\mathbf{P}_{ij} = \kappa(\mathbf{X}_i, \mathbf{X}_j)$ Of course, in the case of an random dot product graph with latent position matrix $\mathbf{X}$, the probability $\mathbf{P}_{ij}$ of observing an edge between vertex i and vertex j is simply $\mathbf{X}_i^\top \mathbf{X}_j$. Thus, for an RDPG with latent positions $\mathbf{X}$, the matrix $\mathbf{P} = [p_{ij}]$ is given by $\mathbf{P} = \mathbf{X}\mathbf{X}^\top$.

In order to more carefully relate latent position models and RDPGs, we can consider the set of positive semidefinite latent position graphs. Namely, we will say that a latent position random graph is positive semidefinite if the matrix $\mathbf{P}$ is positive semidefinite. In this case, we note that an RDPG can be used to approximate the latent position random graph distribution. (In fact, as we describe further below, new work of Rubin-Delanchy, Tang, and Priebe (Rubin-Delanchy et al., 2017) presents a generalization of the RDPG which allows us to drop the positive semidefiniteness requirement.) The best rank-d approximation of $\mathbf{P}$, in terms of the Frobenius norm (Eckart and Young, 1936), will correspond to a RDPG with d-dimensional latent positions. In this sense, by allowing d to be as large as necessary, any positive semi-definite latent position random graph distribution can be approximated by a RDPG distribution to arbitrary precision (Tang et al., 2013).

While latent position models generalize the random dot product graph, RDPGs can be easily related to the more limited *stochastic blockmodel* (SBM) graph (Holland et al., 1983). The stochastic block model is also an independent-edge random graph whose vertex set is partitioned into K groups, called *blocks*, and the stochastic blockmodel is typically parameterized by (1) a $K \times K$ matrix of probabilities $\mathbf{B}$ of adjacencies between vertices in each of the blocks, and (2) a *block-assignment vector* $\tau : [n] \to [K]$ which assigns each vertex to its block. That is, for any two vertices i, j, the probability of their connection is

$$\mathbf{P}_{ij} = \mathbf{B}_{\tau(i), \tau(j)},$$

and we typically write $\mathbf{A} \sim \text{SBM}(\mathbf{B}, \tau)$. Here we present an alternative definition in terms of the RDPG model.

Definition 10 (Positive semidefinite k-block SBM) *We say an RDPG with latent positions $\mathbf{X}$ is an SBM with K blocks if the number of distinct rows in $\mathbf{X}$ is K, denoted $\mathbf{X}_{(1)}, \cdots, \mathbf{X}_{(K)}$ In this case, we define the block membership function $\tau : [n] \mapsto [K]$ to be a function such that $\tau(i) = \tau(j)$ if and only if $\mathbf{X}_i = \mathbf{X}_j$. We then write*

$$\mathbf{A} \sim \text{SBM}(\tau, \{\mathbf{X}_{(i)}\}_{i=1}^{K})$$

In addition, we also consider the case of a stochastic block model in which the block memberships of each vertex is randomly assigned. More precisely, let $\pi \in (0,1)^K$ with $\sum_{k=1}^{n} \pi_k = 1$ and suppose that $\tau(1), \tau(2), \ldots, \tau(n)$ are now i.i.d. random variables with distribution Categorical(π); that is, $\Pr(\tau(i) = k) = \pi_k$ for all k. Then we say $\mathbf{A}$ is an SBM with i.i.d block memberships, and we write

$$\mathbf{A} \sim \text{SBM}(\pi, \{X_{(i)}\}).$$

We also consider the *degree-corrected* stochastic block model (Karrer and Newman, 2011):

Definition 11 (Degree Corrected SBM) *We say an RDPG is a degree-corrected stochastic block model (DCSBM) with K blocks if there exist K unit vectors $y_1, \ldots, y_K \in \mathbb{R}^d$ such that for each $i \in [n]$, there exists $k \in [K]$ and $c_i \in (0,1)$ such that $X_i = c_i y_k$.*

Remark 12 *The degree-corrected stochastic blockmodel model is inherently more flexible than the standard SBM because it allows for vertices within each block/community to have different expected degrees. This flexibility has made it a popular choice for modeling network data (again, see Karrer and Newman, 2011).*

For an important generalization that offers more flexibility in block memberships, we consider the *mixed-membership stochastic block model* (MMSBM) of Airoldi et al. (2008):

Definition 13 (Mixed Membership SBM) *We say an RDPG is a mixed membership stochastic block model (MMSBM) with K blocks if there exists K unit vectors $y_1, \ldots, y_K \in \mathbb{R}^d$ such that for each $i \in [n]$, there exists $\alpha_1, \ldots, \alpha_K > 0$ such that $\sum_{k=1}^K \alpha_k = 1$ and $X_i = \sum_{k=1}^K \alpha_k y_k$.*

Remark 14 *The mixed membership SBM allows for each vertex to be in a mixture of different blocks. Additionally, note that every RDPG is a MMSBM for some choice of K.*

Our next theorem summarizes the relationship between these models.

Theorem 15 *Considered as statistical models for graphs—to wit, sets of probability distributions on graphs—the positive-semidefinite K-block SBM is a subset of the K-block DCSBM and the K-block MMSBM. Both the positive semidefinite K-block DCSBM and K-block MMSBM are subsets of the RDPG model with K-dimensional latent positions. Finally, the union of all possible RDPG models, without restriction of latent position dimension, is dense in the set of positive semidefinite latent position models.*

We emphasize that, as mentioned previously, very recent work of Rubin-Delanchy, Priebe, and Tang (Rubin-Delanchy et al., 2017) on the *generalized random dot product graph* (gRDPG) extends the essential construction of the random dot product graph to encompass non-positive-semidefinite stochastic block models and mixed membership block models, and the gRDPG possesses an elegant uniqueness property that permits the representation of mixed membership as a convex combination of extreme points in a simplex. So, while we impose some restrictions on positive-definiteness here, these are primarily for ease of exposition. Our methodology is general enough to encompass networks that exhibit both homophily and heterophily (see, for example, Hoff, 2007). Moreoever, when considering normalized latent positions for random dot product graphs, our requirement of positive definiteness is, in effect, a degree-corrected "affinity." Also, as we show in Section 6.3, our techniques have proven empirically appropriate in the analysis of certain directed graphs, wherein we consider right- and left-singular vectors in a singular value decomposition of the adjacency matrix.

2.3 Embeddings

Since we rely on spectral decompositions, we begin with describing the notations for the spectral decomposition of the rank d positive semidefinite matrix $\mathbf{P} = \mathbf{X}\mathbf{X}^\top$.

Definition 16 (Spectral Decomp. of P) *Since $\mathbf{P}$ is symmetric and positive semidefinite, let $\mathbf{P} = \mathbf{U_P}\mathbf{S_P}\mathbf{U_P}^\top$ denote its spectral decomposition, with $\mathbf{U_P} \in \mathbb{R}^{n \times d}$ having orthonormal columns and $\mathbf{S_P} \in \mathbb{R}^{d \times d}$ a diagonal matrix with nonincreasing entries*

$$(\mathbf{S_P})_{1,1} \geq (\mathbf{S_P})_{2,2} \geq \cdots \geq (\mathbf{S_P})_{d,d} > 0.$$

As with the spectral decomposition of the matrix $\mathbf{P}$, given an adjacency matrix $\mathbf{A}$, we define its adjacency spectral embedding as follows;

Definition 17 (Adjacency spectral embedding (ASE)) *Given a positive integer $d \geq 1$, the* adjacency spectral embedding *(ASE) of* $\mathbf{A}$ *into* $\mathbb{R}^d$ *is given by* $\hat{\mathbf{X}} = \mathbf{U}_{\mathbf{A}}\mathbf{S}_{\mathbf{A}}^{1/2}$ *where*

$$|\mathbf{A}| = [\mathbf{U}_{\mathbf{A}}|\mathbf{U}_{\mathbf{A}}^{\perp}][\mathbf{S}_{\mathbf{A}}\bigoplus\mathbf{S}_{\mathbf{A}}^{\perp}][\mathbf{U}_{\mathbf{A}}|\mathbf{U}_{\mathbf{A}}^{\perp}]^{\top}$$

is the spectral decomposition of $|\mathbf{A}| = (\mathbf{A}^{\mathbf{T}}\mathbf{A})^{1/2}$ *and* $\mathbf{S}_{\mathbf{A}}$ *is the diagonal matrix of the d largest eigenvalues of* $|\mathbf{A}|$ *and* $\mathbf{U}_{\mathbf{A}}$ *is the* $n \times d$ *matrix whose columns are the corresponding eigenvectors.*

The intuition behind the notion of adjacency spectral embedding is as follows. Our goal is to estimate the latent position matrix $\mathbf{X}$. Now, if the matrix $\mathbf{P}$ were actually observable, then the spectral embedding of $\mathbf{P}$, given by $\mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{1/2}$, is simply some orthogonal transformation of $\mathbf{X}$. Of course, $\mathbf{P}$ is typically not observable; instead we observe $\mathbf{A}$, a noisy version of $\mathbf{P}$. The ASE will be a good estimate of $\mathbf{X}$ provided that the noise does not greatly impact the embedding—that is, if $\mathbf{A}$ and $\mathbf{P}$ are suitably close. As we will see shortly, one can show that $\|\mathbf{A}-\mathbf{X}\mathbf{X}^{\top}\| = O(\|\mathbf{X}\|) = o(\|\mathbf{X}\mathbf{X}^{\top}\|)$ with high probability (Oliveira, 2009; Lu and Peng, 2013; Tropp, 2015; Lei and Rinaldo, 2015). That is to say, $\mathbf{A}$ can be viewed as a comparatively small perturbation of $\mathbf{X}\mathbf{X}^{\top}$. Weyl's inequality or the Kato-Temple inequality (Cape et al., 2017a; Kato, 1950) guarantee that the eigenvalues of $\mathbf{A}$ are "close" to the eigenvalues of $\mathbf{X}\mathbf{X}^{\top}$. In addition, by the Davis-Kahan theorem (Davis and Kahan, 1970), the subspace spanned by the top d eigenvectors of $\mathbf{X}\mathbf{X}^{\top}$ is well-approximated by the subspace spanned by the top d eigenvectors of $\mathbf{A}$.

Having defined the adjacency spectral embedding, we next define the analogous Laplacian spectral embedding, which uses the spectral decomposition of the normalized Laplacian matrix.

Definition 18 (Laplacian Spectral Embedding (LSE)) *Let* $\mathcal{L}(\mathbf{A}) = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ *denote the normalized Laplacian of* $\mathbf{A}$ *where* $\mathbf{D}$ *is the diagonal matrix whose diagonal entries* $\mathbf{D}_{ii} = \sum_{j\neq i}\mathbf{A}_{ij}$. *Given a positive integer* $d \geq 1$, *the* Laplacian spectral embedding *(LSE) of* $\mathbf{A}$ *into* $\mathbb{R}^d$ *is given by* $\check{\mathbf{X}} = \mathbf{U}_{\mathcal{L}(\mathbf{A})}\tilde{\mathbf{S}}_{\mathbf{A}}^{1/2}$ *where*

$$|\mathcal{L}(\mathbf{A})| = \left[\mathbf{U}_{\mathcal{L}(\mathbf{A})}|\mathbf{U}_{\mathcal{L}(\mathbf{A})}^{\perp}\right]\left[\mathbf{S}_{\mathcal{L}(\mathbf{A})}\bigoplus\mathbf{S}_{\mathcal{L}(\mathbf{A})}^{\perp}\right]\left[\mathbf{U}_{\mathcal{L}(\mathbf{A})}|\mathbf{U}_{\mathcal{L}(\mathbf{A})}^{\perp}\right]^{\top}$$

is the spectral decomposition of $|\mathcal{L}(\mathbf{A})| = (\mathcal{L}(\mathbf{A})^{\top}\mathcal{L}(\mathbf{A}))^{1/2}$ *and* $\mathbf{S}_{\mathcal{L}(\mathbf{A})}$ *is the diagonal matrix containg the d largest eigenvalues of* $|\mathcal{L}(\mathbf{A})|$ *on the diagonal and* $\mathbf{U}_{\mathcal{L}(\mathbf{A})}$ *is the* $n \times d$ *matrix whose columns are the corresponding eigenvectors.*

Finally, there are a variety other matrices for which spectral decompositions may be applied to yield an embedding of the graph (Le et al., 2017). These are often dubbed as "regularized" embeddings, and many such embeddings are constructed with the aim of improving robustness of spectral decompositions to graph sparsity. While we do not analyze these embeddings directly, a number of our approaches can be adapted to these other embeddings.

3. Core proof techniques: probabilistic and linear algebraic bounds

In this section, we give an overview of the core background results used in our proofs. While we defer to the appendix the proofs of our primary results on consistency and normality for spectral graph estimates, we provide key background here to allow the reader a consolidated, accessible guide to the essential linear algebraic foundations of our approach. The key tools to several of our results on consistency and normality of the adjacency spectral embedding depend on a triumvirate of matrix concentration inequalities, the Davis-Kahan Theorem, and detailed bounds via the power method.

3.1 Concentration inequalities

Concentration inequalities for real- and matrix-valued data are a critical component to our proofs of consistency for spectral estimates. We make use of classical inequalities, such as Hoeffding's inequality, for real-valued random variables, and we also exploit more recent work on the concentration of sums of random matrices and matrix martingales around their expectation. For a careful study of several important matrix concentration inequalities, see Tropp (2015).

We begin by recalling Hoeffding's inequality, which bounds the deviations between a sample mean of independent random variables and the expected value of that sample mean.

Theorem 19 *Let X_i, $1 \leq i \leq n$, be independent, bounded random variables defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose a_i, b_i are real numbers such that $a_i \leq X_i \leq b_i$. Let $\bar{X}$ be their sample mean:*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then

$$\Pr\left(\bar{X} - \mathbb{E}(\bar{X}) \geq t \right) \leq \exp\left(\frac{[-2n^2 t^2]}{\sum_{i=1}^{n} (b_i - a_i)^2} \right) \tag{5}$$

and

$$\Pr\left(|\bar{X} - \mathbb{E}(\bar{X})| \geq t \right) \leq 2 \exp\left(\frac{[-2n^2 t^2]}{\sum_{i=1}^{n} (b_i - a_i)^2} \right) \tag{6}$$

For an undirected, hollow RDPG with probability matrix $\mathbf{P}$, $\mathbb{E}(\mathbf{A}_{ij}) = \mathbf{P}_{ij}$ for all $i \neq j$. As such, one can regard $\mathbf{A}$ as a "noisy" version of $\mathbf{P}$. It is tempting to believe that $\mathbf{A}$ and $\mathbf{P}$ are close in terms of the Frobenius norm, but this is sadly not true; indeed, it is easy to see that

$$\|\mathbf{A} - \mathbf{P}\|_F^2 = \Theta(\|\mathbf{P}\|_F^2)$$

To overcome this using only Hoeffding's inequality, we can instead consider the difference $(\mathbf{A}^2 - \mathbf{P}^2)_{ij}$, which is a sum of independent random variables. Hence, Hoeffding's inequality implies that

$$|(\mathbf{A}^2 - \mathbf{P}^2)_{ij}|^2 = o(|\mathbf{P}^2{}_{ij}|^2)$$

Since the eigenvectors of $\mathbf{A}$ and $\mathbf{A}^2$ coincide, this is itself sufficient to show concentration of the adjacency spectral embedding (Sussman et al., 2012; Rohe et al., 2011). However,

somewhat stronger and more elegant results can be shown by considering the spectral norm instead. In particular, a nontrivial body of recent work on matrix concentration implies that, under certain assumptions on the sparsity of $\mathbf{P}$, the spectral norm of $\mathbf{A} - \mathbf{P}$ can be well-controlled. We focus on the following important results of Oliveira (2009) and Tropp (2015) and further improvements of Lu and Peng (2013) and Lei and Rinaldo (2015), all of which establish that the $\mathbf{A}$ and $\mathbf{P}$ are close in spectral norm.

Theorem 20 (Spectral norm control of $\mathbf{A} - \mathbf{P}$) *Suppose Let $\mathbf{A}$ be the adjacency matrix of an independent-edge random graph on $[n]$ with matrix of edge probabilities $\mathbf{P}$. For any constant c, there exists another constant C, independent of n and $\mathbf{P}$, such that if $\delta(\mathbf{P}) > C \ln n$, then for any $n^{-c} < \eta < 1/2$,*

$$\Pr\left(\|\mathbf{A} - \mathbf{P}\| \le 4\sqrt{\delta(P)\ln(n/\eta)} \right) \ge 1 - \eta. \tag{7}$$

In Lu and Peng (2013), the authors give an improvement under slightly stronger density assumptions[1]:

Theorem 21 (Refined spectral norm control of $\mathbf{A} - \mathbf{P}$) *With notation as above, suppose there exist positive constants such that for n sufficiently large, $\delta(\mathbf{P}) > (\log n)^{4+a}$. Then for any $c > 0$ there exists a constant C depending on c such that*

$$\mathbb{P}\left(\|\mathbf{A} - \mathbf{P}\| \le 2\sqrt{\delta(\mathbf{P})} + C\delta^{1/4}(\mathbf{P})\ln n \right) \ge 1 - n^{-c}. \tag{8}$$

3.2 Matrix perturbations and spectral decompositions

The above results formalize our intuition that $\mathbf{A}$ provides a "reasonable" estimate for $\mathbf{P}$. Moreover, in the RDPG case, where $\mathbf{P}$ is of low rank and is necessarily positive semidefinite, these results have implications about the nonnegativity of the eigenvalues of $\mathbf{A}$. Specifically, we use Weyl's Theorem to infer bounds on the differences between the eigenvalues of $\mathbf{A}$ and $\mathbf{P}$ from the spectral norm of their difference, and the Gerschgorin Disks Theorem to infer lower bounds on the maximum row sums of $\mathbf{P}$ from assumptions on the eigengap of $\mathbf{P}$ (since both $\mathbf{P}$ and $\mathbf{A}$ are nonnegative matrices, one could also obtain the same lower bounds by invoking the Perron-Frobenius Theorem). For completeness, we recall the Gerschgorin Disks Theorem and Weyl's Theorem (these can be found, for example, in Horn and Johnson, 1985). The former relates the eigenvalues of a matrix to the sums of the absolute values of the entries in each row, and the latter establishes bounds on the differences in eigenvalues between a matrix and a perturbation.

Theorem 22 (Gerschgorin Disks) *Let $\mathbf{H}$ be a complex $n \times n$ matrix, with entries $\mathbf{H}_{ij}$. For $i \in \{1, \cdots, n\}$ let $R_i = \sum_{j \ne i} |\mathbf{H}_{ij}|$. Let the ith Gerschgorin disk $D(\mathbf{H}_{ii}, R_i)$ be the closed disk centered at $\mathbf{H}_{ii}$ with radius R_i. Then every eigenvalue of $\mathbf{H}$ lies within at least one of the Gershgorin discs $D(\mathbf{H}_{ii}, R_i)$.*

---

1. A similar bound is provided in Lei and Rinaldo (2015), but with $\delta(\mathbf{P})$ defined as $\delta(\mathbf{P}) = n \max_{ij} \mathbf{P}_{ij}$ and a density assumption of the form $(n \max_{ij} \mathbf{P}_{ij}) > (\log n)^{1+a}$.

Theorem 23 (Weyl) *Let* $\mathbf{M}, \mathbf{H}$, *and* $\mathbf{R}$ *be* $n \times n$ *Hermitian matrices, and suppose* $\mathbf{M} = \mathbf{H} + \mathbf{R}$. *Suppose* $\mathbf{H}$ *and* $\mathbf{R}$ *have eigenvalues* $\nu_1 \geq \cdots \geq \nu_n$ *and* $r_1 \geq \cdots \geq r_n$, *respectively. Suppose the eigenvalues of* $\mathbf{M}$ *are given by* $\mu_1 \geq \cdots \geq \mu_n$. *Then*

$$\nu_i + r_n \leq \mu_i \leq \nu_i + r_1$$

From our random graph model assumptions and our graph density assumptions, we can conclude that with for sufficiently large n, the top d eigenvalues of $\mathbf{A}$ will be nonnegative:

Remark 24 (Nonnegativity of the top d eigenvalues of A) *Suppose* $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X})$. *Since* $\mathbf{P} = \mathbf{X}\mathbf{X}^\top$, *it is necessarily positive semidefinite, and thus has nonnegative eigenvalues. If we now assume that* $\gamma(\mathbf{P}) > c_0$ *for some constant* c_0, *then along with the Gershgorin Disks Theorem, guarantee that the top d eigenvalues of* $\mathbf{P}$ *are all of order* $\delta(\mathbf{P})$, *and our rank assumption on* $\mathbf{P}$ *mandates that the remaining eigenvalues be zero. If* $\delta(\mathbf{P}) > \log^{4+a'} n$, *the spectral norm bound in* (8) *applies, ensuring that for n sufficiently large,* $\|A - P\|_2 \sim O(\sqrt{\delta(\mathbf{P})})$ *with high probability. Thus, by Weyl's inequality, we see that the top d eigenvalues of* $\mathbf{A}$ *are, with high probability, of order δ, and the remaining are, with high probability, within $\sqrt{\delta}$ of zero.*

Since $\mathbf{P} = \mathbf{X}\mathbf{X}^\top = \mathbf{U}_\mathbf{P}\mathbf{S}_\mathbf{P}^{1/2}(\mathbf{U}_\mathbf{P}\mathbf{S}_\mathbf{P}^{1/2})^\top$ and $\mathbf{A}$ is close to $\mathbf{P}$, it is intuitively appealing to conjecture that, in fact, $\hat{\mathbf{X}} = \mathbf{U}_\mathbf{A}\mathbf{S}_\mathbf{A}^{1/2}$ should be close to some rotation of $\mathbf{U}_\mathbf{P}\mathbf{S}_\mathbf{P}^{1/2}$. That is, if $\mathbf{X}$ is the matrix of true latent positions—so $\mathbf{X}\mathbf{W} = \mathbf{U}_\mathbf{P}\mathbf{S}_\mathbf{P}^{1/2}$ for some orthogonal matrix $\mathbf{W}$—then it is plausible that $\|\hat{\mathbf{X}} - \mathbf{X}\mathbf{W}\|_F$ ought to be comparatively small. To make this precise, however, we need to understand how both eigenvalues and eigenvectors of a matrix behave when the matrix is perturbed. Weyl's inequality addresses the former. The impact of matrix perturbations on associated eigenspaces is significantly more complicated, and the Davis-Kahan Theorem (Davis and Kahan, 1970; Bhatia, 1997) provides one approach to the latter. The Davis-Kahan has a significant role in several approaches to spectral estimation for graphs: for example, Rohe, Chatterjee, and Yu leverage it in Rohe et al. (2011) to prove the accuracy of spectral estimates (specifically, the graph Laplacian) in high-dimensional stochastic blockmodels. The version we give below is from Yu et al. (2015), which is a user-friendly guide to the the Davis-Kahan Theorem and its statistical implications.

The Davis-Kahan Theorem is often stated as a result on canonical angles between subspaces. To that end, we recall that if $\mathbf{U}$ and $\mathbf{V}$ are two $n \times d$ matrices with orthonormal columns, then we define the vector of d *canonical* or *principal* angles between their column spaces to be the vector Θ such that

$$\Theta = \left(\theta_1 = \cos^{-1}\sigma_1, \cdots, \theta_d = \cos^{-1}\sigma_d\right)^\top$$

where $\sigma_1, \cdots, \sigma_d$ are the singular values of $\mathbf{U}^\top\mathbf{V}$. We define the matrix $\sin(\Theta)$ to be the $d \times d$ diagonal matrix for which $\sin(\theta)_{ii} = \sin\theta_i$.

Theorem 25 (A variant of Davis-Kahan) *Suppose* $\mathbf{H}$ *and* $\mathbf{H}'$ *are two symmetric* $n \times n$ *matrices with real entries with spectrum given by* $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_n$ *and* $\lambda'_1 \geq \lambda'_2 \cdots \geq \lambda'_n$, *respectively; and let* $\mathbf{v}_1, \cdots, \mathbf{v}_n$ *and* $\mathbf{v}'_1, \cdots, \mathbf{v}'_n$ *denote their corresponding orthonormal eigenvectors.*

Let $0 < d \le n$ be fixed, and let $\mathbf{V}$ be the matrix of whose columns are the eigenvectors $v_1, \cdots, v_d$, and similarly $\mathbf{V}'$ the matrix whose columns are the eigenvectors $\mathbf{v}'_1, \cdots \mathbf{v}'_n$. Then

$$\| \sin(\Theta) \|_F \le \frac{2\sqrt{d} \| \mathbf{V} - \mathbf{V}' \|}{\lambda_d(\mathbf{H}) - \lambda_{d+1}(\mathbf{H})}.$$

Observe that if we assume that $\mathbf{P}$ is of rank d and has a sufficient eigengap, the Davis-Kahan Theorem gives us an immediate bound on the spectral norm of the difference between $\mathbf{U_A U_A^\top}$ and $\mathbf{U_P U_P^\top}$ in terms of this eigengap and the spectral norm difference of $\mathbf{A} - \mathbf{P}$, namely:

$$\| \mathbf{U_A U_A^\top} - \mathbf{U_P U_P^\top} \| = \max_i \| \sin(\theta_i) \| \le \frac{C \| \mathbf{A} - \mathbf{P} \|}{\lambda_d(\mathbf{P})}.$$

Recall that the Frobenius norm of a matrix $\mathbf{H}$ satisfies

$$\left(\| \mathbf{H} \|_F \right)^2 = \sum_{i,j} \mathbf{H}_{ij}^2 = \operatorname{tr}(\mathbf{H^\top H}) \ge \| \mathbf{H} \|^2$$

and further that if $\mathbf{H}$ is of rank d, then

$$\left(\| \mathbf{H} \|_F \right)^2 \le d \| \mathbf{H} \|^2$$

and hence for rank d matrices, spectral norm bounds are easily translated into bounds on the Frobenius norm. It is worth noting that Rohe et al. (2011) guarantees that a difference in projections can be transformed into a difference between eigenvectors themselves: namely, given the above bound for $\| \mathbf{U_A U_A^\top} - \mathbf{U_P U_P^\top} \|_F$, there is a constant C and an orthonormal matrix $W \in \mathbb{R}^{d \times d}$ such that

$$\| \mathbf{U_P} W - \mathbf{U_A} \|_F \le C \sqrt{d} \frac{\| \mathbf{A} - \mathbf{P} \|}{\lambda_d(\mathbf{P})}. \tag{9}$$

While these important results provide the backbone for much of our theory, the detailed bounds and distributional results described in the next section rely on a decomposition of $\hat{\mathbf{X}}$ in terms of $(\mathbf{A} - \mathbf{P}) \mathbf{U_P S_P}^{-1/2}$ and a remainder. This first term can be viewed as an application of the power method for finding eigenvectors. Additionally, standard univariate and multivariate concentration inequalities and distributional results can be readily applied to this term. On the other hand, the remainder term can be shown to be of smaller order than the first, and much of the technical challenges of this work rely on carefully bounding the remainder term.

4. Spectral embeddings and estimation for RDPGs

There is a wealth of literature on spectral methods for estimating model parameters in random graphs, dating back more than half a century to estimation in simple Erdős-Rényi models. More specifically, for Erdős-Rényi graphs, we would be remiss not to point to Furedi and Komlos's classic work (Füredi and Komlós, 1981) on the eigenvalues and eigenvectors

of the adjacency matrix of an ER graph. Again, despite their model simplicity, Erdős-Rényi graphs veritably teem with open questions; to cite but one example, in a very recent manuscript, Arias-Castro and Verzhelen (Arias-Castro and Verzelen, 2014) address, in the framework of hypothesis testing, the question of subgraph detection within an ER graph.

Moving up to the slightly more heterogeneous stochastic block model, we again find a rich literature on consistent block estimation in stochastic block models. Fortunato (Fortunato, 2010) provides an overview of partitioning techniques for graphs in general, and consistent partitioning of stochastic block models for two blocks was accomplished by Snijders and Nowicki (Snijders and K.Nowicki, 1997) and for equal-sized blocks by Condon and Karp in 2001. For the more general case, Bickel and Chen (Bickel and Chen, 2009) demonstrate a stronger version of consistency via maximizing Newman-Girvan modularity (Newman, 2006) and other modularities. For a growing number of blocks, Choi et al. (Bickel et al., 2013) prove consistency of likelihood based methods, and Bickel et al. (Bickel et al., 2011) provide a method to consistently estimate the stochastic block model parameters using subgraph counts and degree distributions. This work and the work of Bickel and Chen (Bickel and Chen, 2009) both consider the case of very sparse graphs. In Airoldi et al. (2008), Airoldi et al. define the important generalization of a *mixed-membership* stochastic blockmodel, in which block membership may change depending on vertex-to-vertex interactions, and the authors demonstrate methods of inference for the mixed membership and block probabilities.

Rohe, Chatterjee and Yu show in Rohe et al. (2011) that spectral embeddings of the Laplacian give consistent estimates of block memberships in a stochastic block model, and they also demonstrate how to deploy spectral approaches to estimate block memberships even in strongly heterophilic block models. One of the earliest corresponding results on the consistency of the adjacency spectral embedding the estimation of block memberships is given by Sussman and coauthors in Sussman et al. (2012), where it is proved that for a stochastic block model with K blocks and a rank d block probability matrix B, clustering the rows of the adjacency spectral embedding via k-means clustering (see Pollard, 1981) results in at most $\log n$ vertices being misclassified. An improvement to this can be found in Fishkind et al. (2013), where consistency recovery is possible even if the rank of the embedding dimension is unknown.

In Lyzinski et al. (2014), under the assumption of distinct eigenvalues for the second moment matrix Δ (recall we defined this in 5) of a random dot product graph, it is shown that clustering the rows of the adjacency spectral embedding results in asymptotically almost surely perfect recovery of the block memberships in a stochastic blockmodel—that is, for sufficiently large n, the probability of all vertices being correctly assigned is close to 1. An especially strong recovery is exhibited here: it is shown that in the $2 \to \infty$ norm, $\hat{\mathbf{X}}$ is sufficiently close to a rotation of the the true latent positions. In fact, each *row* in $\hat{\mathbf{X}}$ is within $C \log n / \sqrt{n}$ of the corresponding row in $\mathbf{X}$. Unlike a Frobenius norm bound, in which it is possible that some rows of $\hat{\mathbf{X}}$ may be close to the true positions but others may be significantly farther away, this $2 \to \infty$ bound implies that the adjacency spectral embedding has a *uniform* consistency.

Furthermore, in Tang et al. (2017a), one finds a nontrivial tightening of the Frobenius norm bound for the difference between the (rotated) true and estimated latent positions: in fact the Frobenius norm is not merely bounded from above by a term of order $\log n$, but rather concentrates around a *constant*. This constant-order Frobenius bound forms the basis of a principled two-sample hypothesis test for determining whether two random dot product graphs have the same generating latent positions (see Section 5.2).

In Lyzinski et al. (2017), the $2 \to \infty$-norm bound is extended even to the case when the second moment matrix Δ does not have distinct eigenvalues. This turns out to be critical in guaranteeing that the adjacency spectral embedding can be effectively deployed for community detection in hierarchical block models. We present this bound for the $2 \to \infty$ norm in some detail here; it illustrates the confluence of our key techniques and provides an effective roadmap for several subsequent results on asymptotic normality and two-sample testing.

4.1 Consistency of latent position estimates

We state here one of our lynchpin results on consistency, in the $2 \to \infty$ norm, of the adjacency spectral embedding for latent position estimation of a random dot product graph. We given an outline of the proof here, and refer the reader to the Appendix A for the details, which essentially follow the proof given in Lyzinski et al. (2017). We emphasize our setting is a sequence of random dot product graphs $\mathbf{A}_n \sim \mathrm{RDPG}(\mathbf{X}_n)$ for increasing n and thus we make the following density assumption on $\mathbf{P}_n$ as n increases:

Assumption 1 *Let $\mathbf{A}_n \sim \mathrm{RDPG}(\mathbf{X}_n)$ for $n \geq 1$ be a sequence of random dot product graphs with $\mathbf{A}_n$ being a $n \times n$ adjacency matrix. Suppose that $\mathbf{X}_n$ is of rank d for all n sufficiently large. Suppose also that there exists constants $a > 0$ and $c_0 > 0$ such that for all n sufficiently large,*

$$\delta(\mathbf{P}_n) = \max_i \sum_{j=1}^{n} (\mathbf{P}_n)_{ij} \geq \log^{4+a}(n); \quad \gamma(\mathbf{P}_n) = \frac{\lambda_d(\mathbf{P}_n)}{\delta(\mathbf{P}_n)} \geq c_0.$$

We remark that this assumption is, in essence, a restriction on graph sparsity, and not a particular onerous one; d-dimensional random dot product graphs with i.i.d latent positions that are not "too" degenerate will meet these requirements, and bounds on δ of the polylogarithmic order given above are quite common (see Oliveira, 2009; Lu and Peng, 2013).

Our consistency result for the $2 \to \infty$ norm is Theorem 26 below. In the proof of this particular result, we consider a particular random dot product graph with non-random (that is, fixed) latent positions, but our results apply also to the case of random latent positions. In Section 4.2, where we provide a central limit theorem, we focus on the case in which the latent positions are themselves random. Similarly, in Section 5.2, in our analysis of the semiparametric two-sample hypothesis test for the equality of latent positions in a pair of random dot product graphs, we return to the setting in which the latent positions are fixed, and in the nonparametric hypothesis test of equality of distributions, we analyze again the case when the latent positions are random. It is convenient to be able to move

18

fluidly between the two versions of a random dot product graph, adapting our results as appropriate in each case.

In the Appendix, we give a detailed proof of Theorem 26 and we point out the argument used therein also sets the stage for the central limit theorem for the rows of the adjacency spectral embedding given in Subsection 4.2.

Theorem 26 *Let* $\mathbf{A}_n \sim \mathrm{RDPG}(\mathbf{X}_n)$ *for* $n \geq 1$ *be a sequence of random dot product graphs where the* $\mathbf{X}_n$ *is assumed to be of rank* d *for all* n *sufficiently large. Denote by* $\hat{\mathbf{X}}_n$ *the adjacency spectral embedding of* $\mathbf{A}_n$ *and let* $(\hat{\mathbf{X}}_n)_i$ *and* $(\mathbf{X}_n)_i$ *be the* i-*th row of* $\hat{\mathbf{X}}_n$ *and* $\mathbf{X}_n$, *respectively. Let* E_n *be the event that there exists an orthogonal transformation* $\mathbf{W}_n \in \mathbb{R}^{d \times d}$ *such that*

$$\max_i \|(\hat{\mathbf{X}}_n)_i - \mathbf{W}_n(\mathbf{X}_n)_i\| \leq \frac{Cd^{1/2}\log^2 n}{\delta^{1/2}(\mathbf{P}_n)}$$

where $C > 0$ *is some fixed constant and* $\mathbf{P}_n = \mathbf{X}_n\mathbf{X}_n^\top$. *Then* E_n *occurs asymptotically almost surely; that is,* $\Pr(E_n) \to 1$ *as* $n \to \infty$.

Under the stochastic blockmodel, previous bounds on $\|\mathbf{X} - \hat{\mathbf{X}}\|_F$ implied that k-means applied to the rows of $\hat{\mathbf{X}}$ would approximately correctly partition the vertices into their the true blocks with up to $O(\log n)$ errors. However, this Frobenius norm bound does not preclude the possibility of large outliers in the individual rows of $\mathbf{X} - \hat{\mathbf{X}}$. The improvements provided by Theorem 26 overcome this hurdle and, as shown in Lyzinski et al. (2014), under suitable sparsity and eigengap assumptions, k-means applied to $\hat{\mathbf{X}}$ will *exactly correctly* partition the vertices into their true blocks. This implication demonstrates the importance of improving the overall bounds and in focusing on the correct metrics for a given task—in this case, for instance, block identification.

While we defer the proof of this result to the Appendix, in A.1, we provide a brief outline and note several key ingredients here. First is a lemma guaranteeing the existence of an orthogonal matrix $\mathbf{W}^*$ such that

$$\|\mathbf{W}^*\mathbf{S_A} - \mathbf{S_P}\mathbf{W}^*\|_F = O((\log n)\delta^{1/2}(P))$$

That is, there is an approximate commutativity between right and left multiplication of the corresponding matrices of eigenvalues by this orthogonal transformation. The second essential component is, at heart, a bound inspired by the power method. Specifically, we show that there exists an orthogonal matrix

$$\|\hat{\mathbf{X}} - \mathbf{X}\mathbf{W}\| = \|(\mathbf{A} - \mathbf{P})\mathbf{U_P}\mathbf{S_P}^{-1/2}\|_F + O((\log n)\delta^{-1/2})$$

Finally, from this point, establishing the bound on the $2 \to \infty$ norm is a consequence of Hoeffding's inequality applied to sums of the form

$$\sum_k (\mathbf{A}_{ik} - \mathbf{P}_{ik}\mathbf{U}_{kj})$$

.

The $2 \to \infty$ bound in Theorem 26 has several important implications. As we mentioned, the results in Lyzinski et al. (2014) establish an earlier form of this bound, with more

restrictive assumptions on the the second moment matrix, and illustrate how this can be used to cluster vertices in an SBM perfectly; that is, with no vertices misclassified. In addition, in Lyzinski et al. (2014), we show that clustering the rows of the ASE can be useful for inference in a degree-corrected stochastic block model as well. In Section 6, we see that because of Theorem 26, the adjacency spectral embedding and a novel angle-based clustering procedure can be used for accurately identifying subcommunities in an affinity-structured, hierarchical stochastic blockmodel (Lyzinski et al., 2017). In the next section, we see how our proof technique here can be used to obtain distributional results for the rows of the adjacency spectral embedding.

4.2 Distributional results for the ASE

In the classical statistical task of parametric estimation, one observes a collection of i.i.d observations $X_1, \cdots, X_n$ from some family of distributions $F_\theta : \theta \in \Theta$, where Θ is some subset of Euclidean space, and one seeks to find a consistent estimator $T(X_1, \cdots, X_n)$ for θ. As we mentioned in Section 1, often a next task is to determine the asymptotic distribution, as $n \to \infty$, of a suitable normalization of this estimator T. Such distributional results, in turn, can be useful for generating confidence intervals and testing hypotheses.

We adopt a similar framework for random graph inference. In the previous section, we demonstrate the consistency of the adjacency spectral embedding for the true latent position of a random dot product graph. In this section, we establish the asymptotic normality of the rows of this embedding and the Laplacian spectral embedding. In the subsequent section, we examine how our methodology can be deployed for multisample graph hypothesis testing.

We emphasize that distributional results for spectral decompositions of random graphs are comparatively few. The classic results of Füredi and Komlós (Füredi and Komlós, 1981) describe the eigenvalues of the Erdős-Rényi random graph and the work of Tao and Vu (Tao and Vu, 2012) is focused on distributions of eigenvectors of more general random matrices under moment restrictions, but Athreya et al. (2016) and Tang and Priebe (2018) are among the only references for central limit theorems for spectral decompositions of adjacency and Laplacian matrices for a wider class of independent-edge random graphs than merely the Erdős-Rényi model. Apart from their inherent interest, these limit theorems point us to current open questions on efficient estimation and the relative merits of different estimators and embeddings, in part by rendering possible a comparison of asymptotic variances and allowing us to quantify relative efficiency (see Tang et al. (2016) and to precisely conjecture a decomposition of the sources of variance in different spectral embeddings for multiple graphs (see Levin et al., 2017).

Specifically, we show that for a d-dimensional random dot product graph with i.i.d latent positions, there exists a sequence of $d \times d$ orthogonal matrices $\mathbf{W}_n$ such that for any row index i, $\sqrt{n}(\mathbf{W}_n(\hat{\mathbf{X}}_n)_i - (\mathbf{X}_n)_i)$ converges as $n \to \infty$ to a mixture of multivariate normals.

Theorem 27 (Central Limit Theorem for rows of ASE) *Let* $(\mathbf{A}_n, \mathbf{X}_n) \sim \mathrm{RDPG}(F)$ *be a sequence of adjacency matrices and associated latent positions of a d-dimensional random dot product graph according to an inner product distribution F. Let* $\Phi(\mathbf{x}, \mathbf{\Sigma})$ *denote*

the cdf of a (multivariate) Gaussian with mean zero and covariance matrix $\boldsymbol{\Sigma}$, evaluated at $\mathbf{x} \in \mathbb{R}^d$. Then there exists a sequence of orthogonal d-by-d matrices $(\mathbf{W}_n)_{n=1}^{\infty}$ such that for all $\boldsymbol{z} \in \mathbb{R}^d$ and for any fixed index i,

$$\lim_{n \to \infty} \Pr\left[n^{1/2} \left(\hat{\mathbf{X}}_n \mathbf{W}_n - \mathbf{X}_n \right)_i \le \boldsymbol{z} \right] = \int_{\operatorname{supp} F} \Phi\left(\boldsymbol{z}, \boldsymbol{\Sigma}(\mathbf{x}) \right) dF(\mathbf{x}),$$

where

$$\boldsymbol{\Sigma}(\mathbf{x}) = \Delta^{-1} \mathbb{E}\left[(\mathbf{x}^\top \mathbf{X}_1 - (\mathbf{x}^\top \mathbf{X}_1)^2) \mathbf{X}_1 \mathbf{X}_1^\top \right] \Delta^{-1}; \quad and \ \Delta = \mathbb{E}[\mathbf{X}_1 \mathbf{X}_1^\top]. \tag{10}$$

We also note the following important corollary of Theorem 27 for when F is a mixture of K point masses; that is, $(\mathbf{X}, \mathbf{A}) \sim \mathrm{RDPG}(F)$ is a K-block stochastic blockmodel graph. Then for any fixed index i, the event that $\mathbf{X}_i$ is assigned to block $k \in \{1, 2, \ldots, K\}$ has non-zero probability and hence one can condition on the block assignment of $\mathbf{X}_i$ to show that the conditional distribution of $\sqrt{n}(\mathbf{W}_n(\hat{\mathbf{X}}_n)_i - (\mathbf{X}_n)_i)$ converges to a multivariate normal. This is in contrast to the unconditional distribution being a mixture of multivariate normals as given in Theorem 27.

Corollary 28 (SBM) *Assume the setting and notations of Theorem 27 and let*

$$F = \sum_{k=1}^{K} \pi_k \delta_{\nu_k}, \quad \pi_1, \cdots, \pi_K > 0, \sum_k \pi_k = 1$$

be a mixture of K point masses in $\mathbb{R}^d$ where δ_{ν_k} is the Dirac delta measure at ν_k. Then there exists a sequence of orthogonal matrices $\mathbf{W}_n$ such that for all $\boldsymbol{z} \in \mathbb{R}^d$ and for any fixed index i,

$$\mathbb{P}\left\{ \sqrt{n}(\mathbf{W}_n \hat{\mathbf{X}}_n - \mathbf{X}_n)_i \le \boldsymbol{z} \mid \mathbf{X}_i = \nu_k \right\} \longrightarrow \Phi(\boldsymbol{z}, \Sigma_k) \tag{11}$$

where $\Sigma_k = \Sigma(\nu_k)$ is as defined in Eq. (52).

We relegate the full details of the proof of this central limit theorem to the Appendix, in Section A.2, but a few points bear noting here. First, both Theorem 27 and Corollary 28 are very similar to results proved in Athreya et al. (2016), but with the crucial difference being that we no longer require that the second moment matrix $\Delta = \mathbb{E}[\mathbf{X}_1 \mathbf{X}_1^\top]$ of $\boldsymbol{X}_1 \sim F$ have distinct eigenvalues (for more details, see Tang and Priebe (2018)). As in Athreya et al. (2016), our proof here depends on writing the difference between a row of the adjacency spectral embedding and its corresponding latent position as a pair of summands: the first, to which a classical Central Limit Theorem can be applied, and the second, essentially a combination of residual terms, which we show, using techniques similar to those in the proof of Theorem 26, converges to zero. The weakening of the assumption of distinct eigenvalues necessitates significant changes from Athreya et al. (2016) in how to bound the residual terms, because the arguments in Athreya et al. (2016) require an adaptation of a result of Bickel and Sarkar (2015)—the latter of which depends on the assumption of distinct eigenvalues—to control these terms. Here, we resort to somewhat different methodology: we prove instead that analogous bounds to those in Lyzinski et al. (2017) and in Tang and Priebe (2018) hold for the estimated latent positions. This enables us to establish that here, too, the rows of the adjacency spectral embedding are also approximately normally distributed.

We stress that this central limit theorem depends on a delicate bounding of a sequence of so-called residual terms, but its essence is straightforward. Specifically, there exists an orthogonal transformation $\mathbf{W}^*$ such that we can write the ith row of the matrix

$$\sqrt{n}(\mathbf{U_A S_A^{1/2}} - \mathbf{U_P S_P^{1/2} W^*})$$

as

$$\sqrt{n}(\mathbf{U_A S_A^{1/2}} - \mathbf{U_P S_P^{1/2} W^*})_i = \sqrt{n}(\mathbf{A} - \mathbf{P})\mathbf{U_P S_P^{-1/2} W^*})_i + \text{Residual terms} \qquad (12)$$

where the residual terms are all of order $O(n^{-1/2}\log n)$ in probability. Now, to handle the first term in Eq.(12), we can condition on a fixed latent position $\mathbf{X}_i = \mathbf{x}_i$, and when this is fixed, the classical Lindeberg-Feller Central Limit Theorem establishes the asymptotic normality of this term. The order of the residual terms then guarantees, by Slutsky's Theorem, the desired asymptotic normality of the gap between estimated and true latent positions, and finally we need only integrate over the possible latent positions to obtain a mixture of normals.

4.3 An example under the stochastic block model

To illustrate Theorem 27, we refer the reader to Athreya et al. (2016), from which we reproduce the following simulation. Consider random graphs generated according to a stochastic block model with parameters

$$B = \begin{bmatrix} 0.42 & 0.42 \\ 0.42 & 0.5 \end{bmatrix} \quad \text{and} \quad \pi = (0.6, 0.4). \qquad (13)$$

In this model, each node is either in block 1 (with probability 0.6) or block 2 (with probability 0.4). Adjacency probabilities are determined by the entries in B based on the block memberships of the incident vertices. The above stochastic blockmodel corresponds to a random dot product graph model in $\mathbb{R}^2$ where the distribution F of the latent positions is a mixture of point masses located at $x_1 \approx (0.63, -0.14)$ (with probability 0.6) and $x_2 \approx (0.69, 0.13)$ (with probability 0.4).

We sample an adjacency matrix $\mathbf{A}$ for graphs on n vertices from the above model for various choices of n. For each graph G, let $\hat{\mathbf{X}} \in \mathbb{R}^{n \times 2}$ denote the embedding of A and let $\hat{\mathbf{X}}_{i\cdot}$ denote the ith row of $\hat{\mathbf{X}}$. In Figure 1, we plot the n rows of $\hat{\mathbf{X}}$ for the various choices of n. The points are denoted by symbols according to the block membership of the corresponding vertex in the stochastic blockmodel. The ellipses show the 95% level curves for the distribution of $\hat{X}_i$ for each block as specified by the limiting distribution.

We estimate the covariance matrices for the residuals. The theoretical covariance matrices are given in the last column of Table 1, where Σ_1 and Σ_2 are the covariance matrices for the residual $\sqrt{n}(\hat{\mathbf{X}}_i - \mathbf{X}_i)$ when X_i is from the first block and second block, respectively. The empirical covariance matrices, denoted $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$, are computed by evaluating the sample covariance of the rows of $\sqrt{n}\hat{\mathbf{X}}_i$ corresponding to vertices in block 1 and 2 respectively. The estimates of the covariance matrices are given in Table 1. We see that as n increases, the sample covariances tend toward the specified limiting covariance matrix given in the last column.

(a) $n = 1000$
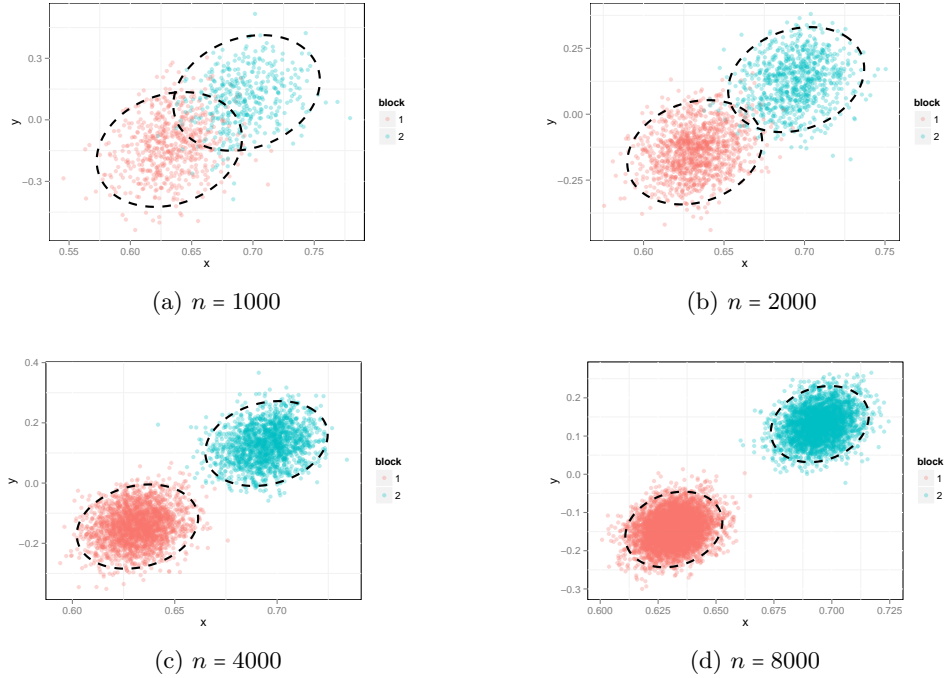
(b) $n = 2000$

(c) $n = 4000$

(d) $n = 8000$

Figure 1: Plot of the estimated latent positions in a two-block stochastic blockmodel graph on n vertices. The points are denoted by symbols according to the block membership of the corresponding vertices. Dashed ellipses give the 95% level curves for the distributions as specified in Theorem 27. Figure duplicated from Athreya et al. (2016).

n	2000	4000	8000	16000	∞
$\hat{\Sigma}_1$	$\begin{bmatrix} .58 & .54 \\ .54 & 16.56 \end{bmatrix}$	$\begin{bmatrix} .58 & .63 \\ .63 & 14.87 \end{bmatrix}$	$\begin{bmatrix} .60 & .61 \\ .61 & 14.20 \end{bmatrix}$	$\begin{bmatrix} .59 & .58 \\ .58 & 13.96 \end{bmatrix}$	$\begin{bmatrix} .59 & .55 \\ .55 & 13.07 \end{bmatrix}$
$\hat{\Sigma}_2$	$\begin{bmatrix} .58 & .75 \\ .75 & 16.28 \end{bmatrix}$	$\begin{bmatrix} .59 & .71 \\ .71 & 15.79 \end{bmatrix}$	$\begin{bmatrix} .58 & .54 \\ .54 & 14.23 \end{bmatrix}$	$\begin{bmatrix} .61 & .69 \\ .69 & 13.92 \end{bmatrix}$	$\begin{bmatrix} .60 & .59 \\ .59 & 13.26 \end{bmatrix}$

Table 1: The sample covariance matrices for $\sqrt{n}(\hat{X}_i - X_i)$ for each block in a stochastic blockmodel with two blocks. Here $n \in \{2000, 4000, 8000, 16000\}$. In the last column are the theoretical covariance matrices for the limiting distribution. Table reproduced from Athreya et al. (2016).

We also investigate the effects of the multivariate normal distribution as specified in Theorem 27 on inference procedures. It is shown in Sussman et al. (2012, 2014) that the approach of embedding a graph into some Euclidean space, followed by inference (for example, clustering or classification) in that space can be consistent. However, these consistency results are, in a sense, only first-order results. In particular, they demonstrate only that the error of the inference procedure converges to 0 as the number of vertices in the graph increases. We now illustrate how Theorem 27 may lead to a more refined error analysis.

We construct a sequence of random graphs on n vertices, where n ranges from 1000 through 4000 in increments of 250, following the stochastic blockmodel with parameters as given above in Eq. (13). For each graph G_n on n vertices, we embed G_n and cluster the embedded vertices of G_n via Gaussian mixture model and K-means clustering. Gaussian mixture model-based clustering is done using the MCLUST implementation of (Fraley and Raftery, 1999). We then measure the classification error of the clustering solution. We repeat this procedure 100 times to obtain an estimate of the misclassification rate. The results are plotted in Figure 2. For comparison, we plot the Bayes optimal classification error rate under the assumption that the embedded points do indeed follow a multivariate normal mixture with covariance matrices Σ_1 and Σ_2 as given in the last column of Table 1. We also plot the misclassification rate of $(C \log n)/n$ as given in Sussman et al. (2012) where the constant C was chosen to match the misclassification rate of K-means clustering for $n = 1000$. For the number of vertices considered here, the upper bound for the constant C from Sussman et al. (2012) will give a vacuous upper bound of the order of 10^6 for the misclassification rate in this example. Finally, we recall that the $2 \to \infty$ norm bound of Theorem 26 implies that, for large enough n, even the k-means algorithm will exactly recover the true block memberships with high probability (Lyzinski et al., 2016a).

For yet another application of the central limit theorem, we refer the reader to Suwan et al. (2016), where the authors discuss the assumption of multivariate normality for estimated latent positions and how this can lead to a significantly improved empirical-Bayes framework for the estimation of block memberships in a stochastic blockmodel.

4.4 Distributional results for Laplacian spectral embedding

We now present the analogous central limit theorem results of Section 4.2 for the normalized *Laplacian spectral embedding* (see Definition 18). We first recall the definition of the Laplacian spectral embedding.

Theorem 29 (Central Limit Theorem for rows of LSE) *Let $(\mathbf{A}_n, \mathbf{X}_n) \sim \mathrm{RDPG}(F)$ for $n \geq 1$ be a sequence of d-dimensional random dot product graphs distributed according to some inner product distribution F. Let $\boldsymbol{\mu}$ and $\tilde{\Delta}$ denote the quantities*

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}_1] \in \mathbb{R}^d; \quad \tilde{\Delta} = \mathbb{E}\Big[\frac{\mathbf{X}_1\mathbf{X}_1^\top}{\mathbf{X}_1^\top\boldsymbol{\mu}}\Big] \in \mathbb{R}^{d\times d}. \tag{14}$$

Also denote by $\tilde{\Sigma}(\mathbf{x})$ the $d \times d$ matrix

$$\mathbb{E}\Big[\Big(\frac{\tilde{\Delta}^{-1}\mathbf{X}_1}{\mathbf{X}_1^\top\boldsymbol{\mu}} - \frac{\mathbf{x}}{2\mathbf{x}^\top\boldsymbol{\mu}}\Big)\Big(\frac{\mathbf{X}_1^\top\tilde{\Delta}^{-1}}{\mathbf{X}_1^\top\boldsymbol{\mu}} - \frac{\mathbf{x}^\top}{2\mathbf{x}^\top\boldsymbol{\mu}}\Big)\frac{(\mathbf{x}^\top\mathbf{X}_1 - \mathbf{x}^\top\mathbf{X}_1\mathbf{X}_1^\top\mathbf{x})}{\mathbf{x}^\top\boldsymbol{\mu}}\Big]. \tag{15}$$
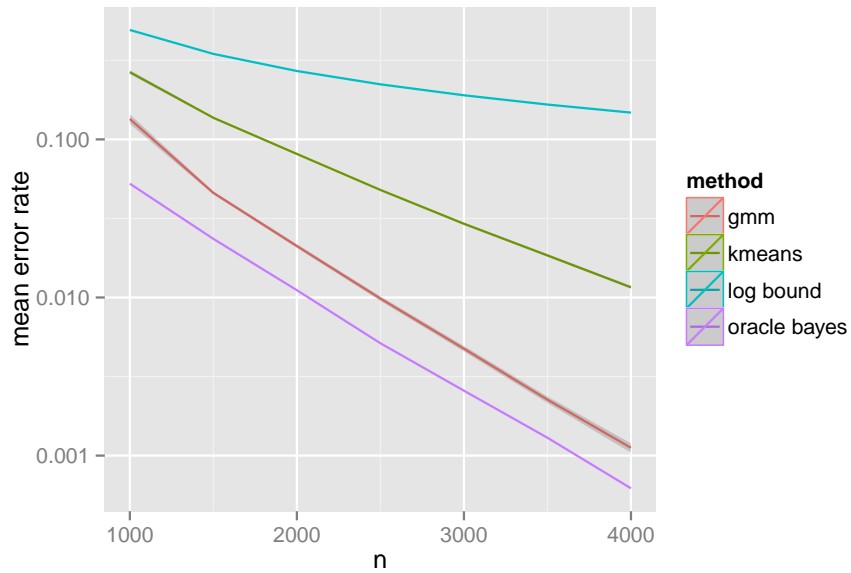
Figure 2: Comparison of classification error for Gaussian mixture model (red curve), K-Means (green curve), and Bayes classifier (cyan curve). The classification errors for each $n \in \{1000, 1250, 1500, \ldots, 4000\}$ were obtained by averaging 100 Monte Carlo iterations and are plotted on a $\log_{10}$ scale. The plot indicates that the assumption of a mixture of multivariate normals can yield non-negligible improvement in the accuracy of the inference procedure. The log-bound curve (purple) shows an upper bound on the error rate as derived in Sussman et al. (2012). Figure duplicated from Athreya et al. (2016).

Then there exists a sequence of $d \times d$ orthogonal matrices $(\mathbf{W}_n)$ such that for each fixed index i and any $\boldsymbol{x} \in \mathbb{R}^d$,

$$\Pr\left\{n\big(\mathbf{W}_n(\breve{\mathbf{X}}_n)_i - \frac{(\mathbf{X}_n)_i}{\sqrt{\Sigma_j (\mathbf{X}_n)_i^\top (\mathbf{X}_n)_j}}\big) \le \boldsymbol{x}\right\} \longrightarrow \int \Phi(\boldsymbol{x}, \tilde{\Sigma}(\boldsymbol{y})) dF(\boldsymbol{y}) \tag{16}$$

When F is a mixture of point masses—specifically, when $\mathbf{A} \sim \mathrm{RDPG}(F)$ is a stochastic blockmodel graph—we also have, below, the following limiting conditional distribution for

$$n\left(\mathbf{W}_n(\breve{\mathbf{X}}_n)_i - \frac{(\mathbf{X}_n)_i}{\sqrt{\Sigma_j (\mathbf{X}_n)_i^\top (\mathbf{X}_n)_j}}\right).$$

Theorem 30 *Assume the setting and notations of Theorem 29 and let*

$$F = \sum_{k=1}^{K} \pi_k \delta_{\nu_k}, \quad \pi_1, \cdots, \pi_K > 0, \sum_k \pi_k = 1$$

be a mixture of K point masses in $\mathbb{R}^d$. Then there exists a sequence of $d \times d$ orthogonal matrices $\mathbf{W}_n$ such that for any fixed index i,

$$\mathbb{P}\left\{n\big(\mathbf{W}_n(\breve{\mathbf{X}}_n)_i - \frac{\nu_k}{\sqrt{\Sigma_l n_l \nu_k^\top \nu_l}}\big) \le \boldsymbol{z} \mid (\mathbf{X}_n)_i = \nu_k\right\} \longrightarrow \boldsymbol{\Phi}(\boldsymbol{z}, \tilde{\Sigma}_k) \tag{17}$$

where $\tilde{\Sigma}_k = \tilde{\Sigma}(\nu_k)$ is as defined in Eq. (15) and n_k for $k \in \{1, 2, \ldots, K\}$ denote the number of vertices in $\mathbf{A}$ that are assigned to block k.

Remark 31 *As a special case of Theorem 30, we note that if $\mathbf{A}$ is an Erdős-Rényi graph on n vertices with edge probability p^2 – which corresponds to a random dot product graph where the latent positions are identically p – then for each fixed index i, the normalized Laplacian embedding satisfies*

$$n\big(\breve{\mathbf{X}}_i - \frac{1}{\sqrt{n}}\big) \overset{\mathrm{d}}{\longrightarrow} \mathcal{N}\big(0, \frac{1-p^2}{4p^2}\big).$$

Recall that $\breve{\mathbf{X}}_i$ is proportional to $1/\sqrt{d_i}$ where d_i is the degree of the i-th vertex. On the other hand, the adjacency spectral embedding satisfies

$$\sqrt{n}(\hat{\mathbf{X}}_i - p) \overset{\mathrm{d}}{\longrightarrow} \mathcal{N}(0, 1 - p^2).$$

As another example, let $\mathbf{A}$ be sampled from a stochastic blockmodel with block probability matrix $\mathbf{B} = \begin{bmatrix} p^2 & pq \\ pq & q^2 \end{bmatrix}$ and block assignment probabilities $(\pi, 1 - \pi)$. Since $\mathbf{B}$ has rank 1, this model corresponds to a random dot product graph where the latent positions are either p with probability π or q with probability $1 - \pi$. Then for each fixed index i, the normalized Laplacian embedding satisfies

$$n\big(\breve{\mathbf{X}}_i - \frac{p}{\sqrt{n_1 p^2 + n_2 pq}}\big) \overset{\mathrm{d}}{\longrightarrow} \mathcal{N}\big(0, \frac{\pi p(1-p^2) + (1-\pi)q(1-pq)}{4(\pi p + (1-\pi)q)^3}\big) \text{ if } \mathbf{X}_i = p, \tag{18}$$

$$n\big(\breve{\mathbf{X}}_i - \frac{q}{\sqrt{n_1 pq + n_2 q^2}}\big) \overset{\mathrm{d}}{\longrightarrow} \mathcal{N}\big(0, \frac{\pi p(1-pq) + (1-\pi)q(1-q^2)}{4(\pi p + (1-\pi)q)^3}\big) \text{ if } \mathbf{X}_i = q. \tag{19}$$

where n_1 and $n_2 = n - n_1$ are the number of vertices of $\mathbf{A}$ with latent positions p and q. The adjacency spectral embedding, meanwhile, satisfies

$$\sqrt{n}(\hat{\mathbf{X}}_i - p) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \frac{\pi p^4(1-p^2) + (1-\pi)pq^3(1-pq)}{(\pi p^2 + (1-\pi)q^2)^2}\right) \text{ if } \mathbf{X}_i = p, \tag{20}$$

$$\sqrt{n}(\hat{\mathbf{X}}_i - q) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \frac{\pi p^3 q(1-pq) + (1-\pi)q^4(1-q^2)}{(\pi p^2 + (1-\pi)q^2)^2}\right) \text{ if } \mathbf{X}_i = q. \tag{21}$$

We present a sketch of the proof of Theorem 29 in the Appendix, in Section A.3. Due to the intricacy of the proof, however, even in the Appendix we do not provide full details; we instead refer the reader to Tang and Priebe (2018) for the complete proof. Moving forward, we focus on the important implications of these distributional results for subsequent inference, including a mechanism by which to assess the relative desirability of ASE and LSE, which *vary depending on inference task.*

5. Implications for subsequent inference

The previous sections are devoted to establishing the consistency and asymptotic normality of the adjacency and Laplacian spectral embeddings for the estimation of latent positions in an RDPG. In this section, we describe several subsequent graph inference tasks, all of which depend on this consistency: specifically, nonparametric clustering, semiparametric and nonparametric two-sample graph hypothesis testing, and multi-sample graph inference.

5.1 Nonparametric clustering: a comparison of ASE and LSE via Chernoff information

We now discuss how the limit results of Section 4.2 and Section 4.4 can provide insight into subsequent inference. In a recent pioneering work, the authors of Bickel and Sarkar (2015) analyze, in the context of stochastic blockmodel graphs, how the choice of spectral embedding by either the adjacency matrix or the normalized Laplacian matrix impacts subsequent recovery of the block assignments. In particular, they show that a metric constructed from the average distance between the vertices of a block and its cluster centroid for the spectral embedding can be used as a surrogate measure for the performance of the subsequent inference task: namely, the metric is a surrogate measure for the error rate in recovering the vertices to block assignments using the spectral embedding. It is shown in Bickel and Sarkar (2015) that for a large regime of parameters in a two-block stochastic block model, the normalized Laplacian spectral embedding reduces the within-block variance (occasionally by a factor of four) while preserving the between-block variance, as compared to that of the adjacency spectral embedding. This suggests that for a large region of the parameter space for two-block stochastic blockmodels, the spectral embedding of the Laplacian is preferable to the spectral embedding of the adjacency matrix for subsequent inference. However, we observe that the metric in Bickel and Sarkar (2015) is intrinsically tied to the use of K-means as the clustering procedure: specifically, a smaller value of the metric for the Laplacian spectral embedding as compared to that for the adjacency spectral embedding

only implies that clustering the Laplacian spectral embedding using K-means is possibly better than clustering the adjacency spectral embedding using K-means.

Motivated by the above observation, in Tang and Priebe (2018), we propose a metric that is *independent* of any specific clustering procedure, a metric that characterizes the minimum error achievable by *any* clustering procedure that uses only the spectral embedding for the recovery of block assignments in stochastic blockmodel graphs. For a given embedding method, the metric used in Tang and Priebe (2018) is based on the notion of statistical information between the limiting distributions of the blocks. Roughly speaking, smaller statistical information implies less information to discriminate between the blocks of the stochastic blockmodel. More specifically, the limit result in Section 4.2 and Section 4.4 state that, for stochastic blockmodel graphs, conditional on the block assignments the entries of the scaled eigenvectors corresponding to the few largest eigenvalues of the adjacency matrix and the normalized Laplacian matrix converge to a multivariate normal as the number of vertices increases. Furthermore, the associated covariance matrices are not spherical, so K-means clustering for the adjacency spectral embedding or Laplacian spectral embedding does not yield minimum error for recovering the block assignment. Nevertheless, these limiting results also facilitate comparison between the two embedding methods via the classical notion of Chernoff information (Chernoff, 1952).

We now recall the notion of the Chernoff α-divergences (for $\alpha \in (0, 1)$) and Chernoff information. Let F_0 and F_1 be two absolutely continuous multivariate distributions in $\Omega = \mathbb{R}^d$ with density functions f_0 and f_1, respectively. Suppose that $Y_1, Y_2, \ldots, Y_m$ are independent and identically distributed random variables, with Y_i distributed either F_0 or F_1. We are interested in testing the simple null hypothesis $\mathbb{H}_0: F = F_0$ against the simple alternative hypothesis $\mathbb{H}_1: F = F_1$. A test T can be viewed as a sequence of mappings $T_m: \Omega^m \mapsto \{0, 1\}$ such that given $Y_1 = y_1, Y_2 = y_2, \ldots, Y_m = y_m$, the test rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $T_m(y_1, y_2, \ldots, y_m) = 1$; similarly, the test favors $\mathbb{H}_0$ if $T_m(y_1, y_2, \ldots, y_m) = 0$.

The Neyman-Pearson lemma states that, given $Y_1 = y_1, Y_2 = y_2, \ldots, Y_m = y_m$ and a threshold $\eta_m \in \mathbb{R}$, the likelihood ratio test which rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ whenever

$$\left(\sum_{i=1}^m \log f_0(y_i) - \sum_{i=1}^m \log f_1(y_i) \right) \le \eta_m$$

is the most powerful test at significance level $\alpha_m = \alpha(\eta_m)$, so that the likelihood ratio test minimizes the Type II error β_m subject to the constraint that the Type I error is at most α_m. Assuming that $\pi \in (0, 1)$ is a prior probability that $\mathbb{H}_0$ is true. Then, for a given $\alpha_m^* \in (0, 1)$, let $\beta_m^* = \beta_m^*(\alpha_m^*)$ be the Type II error associated with the likelihood ratio test when the Type I error is at most α_m^*. The quantity

$$\inf_{\alpha_m^* \in (0,1)} \pi \alpha_m^* + (1 - \pi) \beta_m^*$$

is the Bayes risk in deciding between $\mathbb{H}_0$ and $\mathbb{H}_1$ given the m independent random variables $Y_1, Y_2, \ldots, Y_m$. A classical result of Chernoff (Chernoff, 1952, 1956) states that the Bayes risk is intrinsically linked to a quantity known as the *Chernoff information*. More specifically,

let $C(F_0, F_1)$ be the quantity

$$C(F_0, F_1) = -\log\Big[\inf_{t\in(0,1)}\int_{\mathbb{R}^d} f_0^t(\boldsymbol{x})f_1^{1-t}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\Big]$$
$$= \sup_{t\in(0,1)}\Big[-\log\int_{\mathbb{R}^d} f_0^t(\boldsymbol{x})f_1^{1-t}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\Big]. \tag{22}$$

Then we have

$$\lim_{m\to\infty}\frac{1}{m}\inf_{\alpha_m^*\in(0,1)}\log(\pi\alpha_m^* + (1-\pi)\beta_m^*) = -C(F_0, F_1). \tag{23}$$

Thus $C(F_0, F_1)$, the Chernoff information between F_0 and F_1, is the *exponential* rate at which the Bayes error $\inf_{\alpha_m^*\in(0,1)}\pi\alpha_m^* + (1-\pi)\beta_m^*$ decreases as $m \to \infty$; we note that the Chernoff information is independent of π. We also define, for a given $t \in (0,1)$ the Chernoff divergence $C_t(F_0, F_1)$ between F_0 and F_1 by

$$C_t(F_0, F_1) = -\log\int_{\mathbb{R}^d} f_0^t(\boldsymbol{x})f_1^{1-t}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

The Chernoff divergence is an example of a f-divergence as defined in, for example, the work of Csizár (1967) and Ali and Shelvey (1966). When $t = 1/2$, $C_t(F_0, F_1)$ is the Bhattacharyya distance between F_0 and F_1. Recall that any f-divergence satisfies the Information Processing Lemma and is invariant with respect to invertible transformations (Liese and Vadja, 2006). Therefore, any f-divergence such as the Kullback-Liebler divergence can also be used to compare the two embedding methods; the Chernoff information is particularly attractive because of its explicit relationship with the Bayes risk.

The characterization of Chernoff information in Eq. (23) can be extended to $K + 1 \geq 2$ hypotheses. Let $F_0, F_1, \ldots, F_K$ be distributions on $\mathbb{R}^d$ and suppose that $Y_1, Y_2, \ldots, Y_m$ are independent and identically distributed random variables with Y_i distributed $F \in \{F_0, F_1, \ldots, F_K\}$. We are thus interested in determining the distribution of the Y_i among the $K+1$ hypothesis $\mathbb{H}_0: F = F_0, \ldots, \mathbb{H}_K: F = F_K$. Suppose also that hypothesis $\mathbb{H}_k$ has *a priori* probability π_k. Then for any decision rule δ, the risk of δ is $r(\delta) = \sum_k \pi_k \sum_{l\neq k} \alpha_{lk}(\delta)$ where $\alpha_{lk}(\delta)$ is the probability of accepting hypothesis $\mathbb{H}_l$ when hypothesis $\mathbb{H}_k$ is true. Then we have, from Leang and Johnson (1997),

$$\inf_{\delta}\lim_{m\to\infty}\frac{r(\delta)}{m} = -\min_{k\neq l} C(F_k, F_l). \tag{24}$$

where the infimum is over all decision rules δ. That is, for any δ, $r(\delta)$ decreases to 0 as $m \to \infty$ at a rate no faster than $\exp(-m\min_{k\neq l} C(F_k, F_l))$. It is also shown in Leang and Johnson (1997) that the *Maximum A Posterior* decision rule achieves this rate.

Finally, if $F_0 = \mathcal{N}(\mu_0, \Sigma_0)$ and $F_1 = \mathcal{N}(\mu_1, \Sigma_1)$, then denoting by $\Sigma_t = t\Sigma_0 + (1-t)\Sigma_1$, the Chernoff information $C(F_0, F_1)$ between F_0 and F_1 is given by

$$C(F_0, F_1) = \sup_{t\in(0,1)}\Big(\frac{t(1-t)}{2}(\mu_1-\mu_2)^\top\Sigma_t^{-1}(\mu_1-\mu_2) + \frac{1}{2}\log\frac{|\Sigma_t|}{|\Sigma_0|^t|\Sigma_1|^{1-t}}\Big).$$

Comparison of the performance of the Laplacian spectral embedding and the adjacency spectral embedding for recovering the block assignments now proceeds as follows. Let $\mathbf{B} \in [0,1]^{K \times K}$ and $\boldsymbol{\pi} \in \mathbb{R}^K$ be the matrix of block probabilities and the vector of block assignment probabilities for a K-block stochastic blockmodel. We shall assume that $\mathbf{B}$ is positive semidefinite. Then given an n vertex instantiation of the SBM graph with parameters $(\boldsymbol{\pi}, \mathbf{B})$, for sufficiently large n, the large-sample optimal error rate for recovering the block assignments when adjacency spectral embedding is used as the initial embedding step can be characterized by the quantity $\rho_{\mathrm{A}} = \rho_{\mathrm{A}}(n)$ defined by

$$\rho_{\mathrm{A}} = \min_{k \neq l} \sup_{t \in (0,1)} \frac{1}{2} \log \frac{|\Sigma_{kl}(t)|}{|\Sigma_k|^t |\Sigma_l|^{1-t}} + \frac{nt(1-t)}{2}(\nu_k - \nu_l)^\top \Sigma_{kl}^{-1}(t)(\nu_k - \nu_l) \qquad (25)$$

where $\Sigma_{kl}(t) = t\Sigma_k + (1-t)\Sigma_l$; $\Sigma_k = \Sigma(\nu_k)$ and $\Sigma_l = \Sigma(\nu_l)$ are as defined in Theorem 27. We recall Eq. (24), in particular the fact that as ρ_{A} increases, the large-sample optimal error rate decreases. Similarly, the large-sample optimal error rate when Laplacian spectral embedding is used as the pre-processing step can be characterized by the quantity $\rho_{\mathrm{L}} = \rho_{\mathrm{L}}(n)$ defined by

$$\rho_{\mathrm{L}} = \min_{k \neq l} \sup_{t \in (0,1)} \frac{1}{2} \log \frac{|\tilde{\Sigma}_{kl}(t)|}{|\tilde{\Sigma}_k|^t |\tilde{\Sigma}_l|^{1-t}} + \frac{nt(1-t)}{2}(\tilde{\nu}_k - \tilde{\nu}_l)^\top \tilde{\Sigma}_{kl}^{-1}(t)(\tilde{\nu}_k - \tilde{\nu}_l) \qquad (26)$$

where $\tilde{\Sigma}_{kl}(t) = t\tilde{\Sigma}_k + (1-t)\tilde{\Sigma}_l$ with $\tilde{\Sigma}_k = \tilde{\Sigma}(\nu_k)$ and $\tilde{\Sigma}_l = \tilde{\Sigma}(\nu_l)$ as defined in Theorem 30, and $\tilde{\nu}_k = \nu_k / (\sum_{k'} \pi_{k'} \nu_{k'}^\top \nu_{k'})^{1/2}$. Note that ρ_{A} and ρ_{L} are similar to a weighted Mahalanobis distance. We emphasize that we have made the simplifying assumption that $n_k = n\pi_k$ in our expression for $\tilde{\nu}_k$ in Eq. (26). This is for ease of comparison between ρ_{A} and ρ_{L} in our subsequent discussion.

The ratio $\rho_{\mathrm{A}}/\rho_{\mathrm{L}}$ is a surrogate measure of the relative large-sample performance of the adjacency spectral embedding as compared to the Laplacian spectral embedding for subsequent inference, at least in the context of stochastic blockmodel graphs. That is to say, for given parameters $\boldsymbol{\pi}$ and $\mathbf{B}$, if $\rho_{\mathrm{A}}/\rho_{\mathrm{L}} > 1$, then adjacency spectral embedding is preferred over Laplacian spectral embedding when n, the number of vertices in the graph, is sufficiently large; similarly, if $\rho_{\mathrm{A}}/\rho_{\mathrm{L}} < 1$, then Laplacian spectral embedding is preferred over adjacency spectral embedding.

As an illustration of the ratio $\rho_{\mathrm{A}}/\rho_{\mathrm{L}}$, we first consider the collection of 2-block stochastic blockmodels where $\mathbf{B} = \begin{bmatrix} p^2 & pq \\ pq & q^2 \end{bmatrix}$ for $p, q \in (0,1)$ and $\boldsymbol{\pi} = (\pi_1, \pi_2)$ with $\pi_1 + \pi_2 = 1$. We note that these $\mathbf{B}$ also have rank 1 and thus the Chernoff information can be computed explicitly. Then for sufficiently large n, ρ_{A} is approximately

$$\rho_{\mathrm{A}} \approx \sup_{t \in (0,1)} \frac{nt(1-t)}{2}(p-q)^2 (t\sigma_1^2 + (1-t)\sigma_2^2)^{-1}$$

where σ_1 and σ_2 are as specified in Eq. (20) and Eq. (21), respectively. Simple calculations yield

$$\rho_{\mathrm{A}} \approx \frac{n(p-q)^2(\pi_1 p^2 + \pi_2 q^2)^2}{2\big(\sqrt{\pi_1 p^4(1-p^2) + \pi_2 pq^3(1-pq)} + \sqrt{\pi_1 p^3 q(1-pq) + \pi_2 q^4(1-q^2)}\big)^2}$$
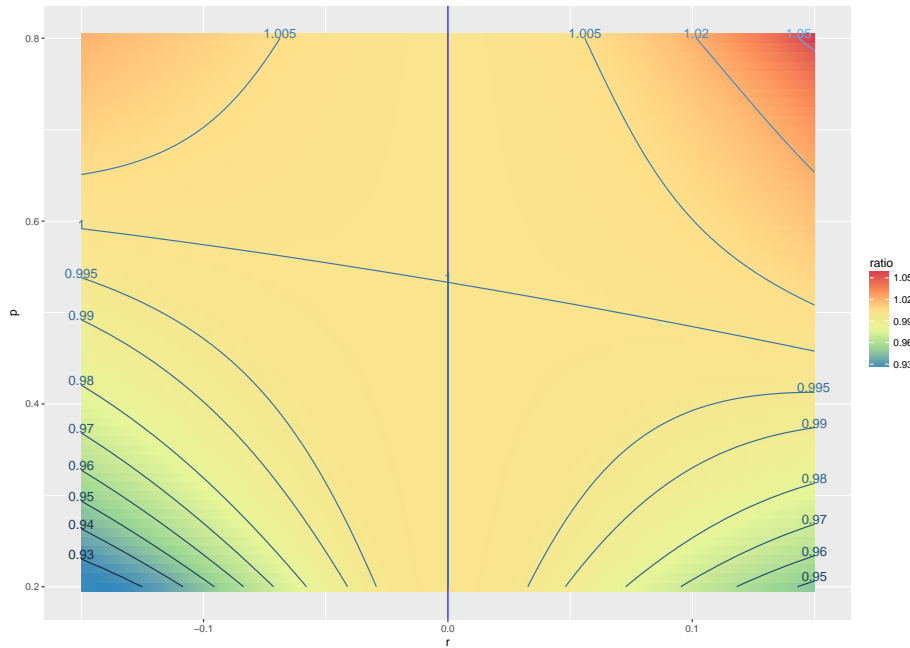
Figure 3: The ratio $\rho_\mathrm{A}/\rho_\mathrm{L}$ displayed for various values of $p \in [0.2, 0.8]$ and $r = q - p \in [-0.15, 0.15]$. The labeled lines are the contour lines for $\rho_\mathrm{A}/\rho_\mathrm{L}$. Figure duplicated from Tang and Priebe (2018).

for sufficiently large n. Similarly, denoting by $\tilde{\sigma}_1^2$ and $\tilde{\sigma}_2^2$ the variances specified in Eq. (18) and Eq. (19), we have

$$
\begin{aligned}
\rho_\mathrm{L} &\approx \sup_{t \in (0,1)} \frac{nt(1-t)}{2} \left(\frac{p}{\sqrt{\pi_1 p^2 + \pi_2 pq}} - \frac{q}{\sqrt{\pi_1 pq + \pi_2 q^2}} \right)^2 (t\tilde{\sigma}_1^2 + (1-t)\tilde{\sigma}_2^2)^{-1} \\
&\approx \frac{2n(\sqrt{p} - \sqrt{q})^2 (\pi_1 p + \pi_2 q)^2}{\left(\sqrt{\pi_1 p(1-p^2) + \pi_2 q(1-pq)} + \sqrt{\pi_1 p(1-pq) + \pi_2 q(1-q^2)} \right)^2} \\
&\approx \frac{2n(p-q)^2 (\pi_1 p + \pi_2 q)^2}{(\sqrt{p} + \sqrt{q})^2 \left(\sqrt{\pi_1 p(1-p^2) + \pi_2 q(1-pq)} + \sqrt{\pi_1 p(1-pq) + \pi_2 q(1-q^2)} \right)^2}
\end{aligned}
$$

for sufficiently large n. Fixing $\boldsymbol{\pi} = (0.6, 0.4)$, we compute the ratio $\rho_\mathrm{A}/\rho_\mathrm{L}$ for a range of p and q values, with $p \in [0.2, 0.8]$ and $q = p + r$ where $r \in [-0.15, 0.15]$. The results are plotted in Figure 3. The y-axis of Figure 3 denotes the values of p and the x axis are the values of r. We see from the above figure that, in general, neither of the methods, namely adjacency spectral clustering or normalized Laplacian spectral embedding followed by clustering via Gaussian mixture models, dominates over the whole (p, r) parameter space. However, in general, one can easily show that LSE is preferable over ASE whenever the block probability matrix $\mathbf{B}$ becomes sufficiently sparse. Determination of similarly intuitive conditions for which ASE dominates over LSE is considerably more subtle and is the topic of current research. But in general, we observe that ASE dominates over LSE whenever the entries of $\mathbf{B}$ are relatively large.

Finally we consider the collection of stochastic blockmodels with parameters $\boldsymbol{\pi}$ and $\mathbf{B}$ where

$$\mathbf{B} = \begin{bmatrix} p & q & q \\ q & p & q \\ q & q & p \end{bmatrix}, \quad p, q \in (0, 1), \text{ and } \boldsymbol{\pi} = (0.8, 0.1, 0.1). \tag{27}$$

First we compute the ratio $\rho_{\mathrm{A}}/\rho_{\mathrm{L}}$ for $p \in [0.3, 0.9]$ and $r = q - p$ with $r \in [-0.2, -0.01]$. The results are plotted in Figure 4, with the y-axis of Figure 4 being the values of p and the x-axis being the values of r. Once again we see that for the purpose of subsequent inference, neither embedding methods dominates over the whole parameter space and that LSE is still preferable to ASE for smaller values of p and q and that ASE is preferable to LSE for larger values of p and q.
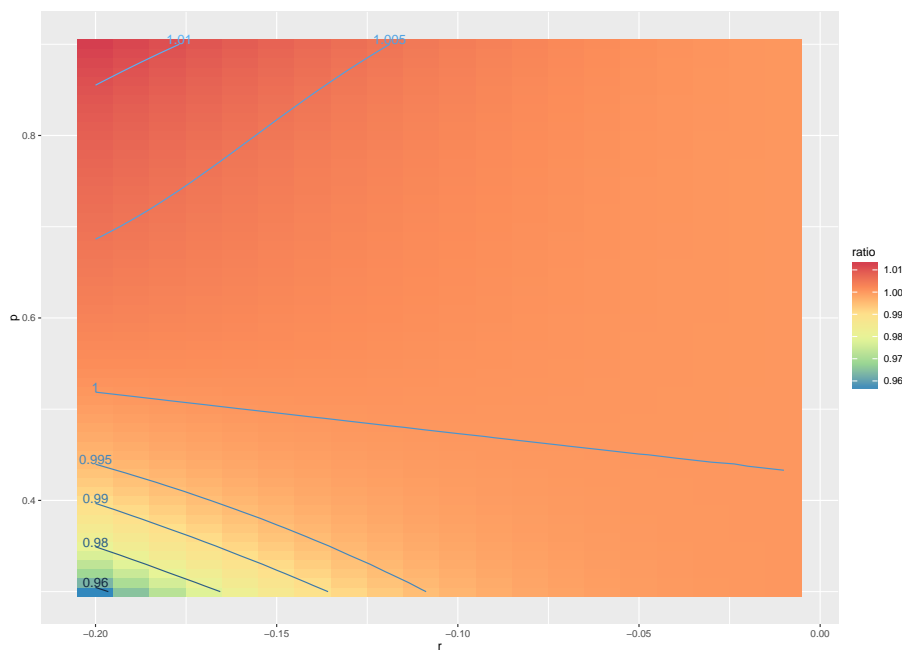
Figure 4: The ratio ρ_A/ρ_L displayed for various values of $p \in [0.2, 0.8]$ and $r = q - p \in [-0.2, -0.01]$ for the 3-block stochastic blockmodel of Eq. (27). The labeled lines are the contour lines for $\rho_{\mathrm{A}}/\rho_{\mathrm{L}}$. Figure duplicated from Tang and Priebe (2018).

5.2 Hypothesis testing

The field of multi-sample graph inference is comparatively new, and the development of a comprehensive machinery for two-sample hypothesis testing for random graphs is of both theoretical and practical importance. The test procedures in Tang et al. (2017a) and Tang et al. (2017b), both of which leverage the adjacency spectral embedding to test hypotheses of equality or equality in distribution for random dot product graphs, are among the only principled methodologies currently available. In both cases, the accuracy of the adjacency spectral embedding as an estimate of the latent positions is the key to constructing a

test statistic. Specifically, in Tang et al. (2017a), we give a new and improved bound, Theorem 32 below, for the Frobenius norm of the difference between the original latent positions and the estimated latent positions obtained from the embedding. This bound is then used to establish a valid and consistent test for the semiparameteric hypothesis test of equality for latent positions in a pair of vertex-aligned random dot product graphs. In the nonparametric case, in Tang et al. (2017b), we demonstrate how the adjacency spectral embedding can be integrated with a kernel density estimator to accurately estimate the underlying distribution F for in a random dot product graph with i.i.d latent positions.

We reproduce here several important points from Tang et al. (2017a,b). To begin our summary of two-sample hypothesis testing for graphs, we first consider the problem of developing a test for the hypothesis that two random dot product graphs on the same vertex set, with known vertex correspondence, have the same generating latent position or have generating latent positions that are scaled or diagonal transformations of one another, modulo a possible (non-identifiable) orthogonal transformation. This framework includes, as a special case, a test for whether two stochastic blockmodels have the same or related block probability matrices. In this two-sample testing problem, though, the parameter dimension grows as the sample size grows. Therefore, the problem is not precisely analogous to classical two-sample tests for, say, the difference of two parameters belonging to some fixed Euclidean space, in which an increase in data has no effect on the dimension of the parameter. The problem is also not nonparametric, since we view our latent positions as fixed and impose specific distributional requirements on the data—that is, on the adjacency matrices. Indeed, we regard the problem as semiparametric, and Tang et al. (2017a) adapts the traditional definition of consistency to this setting. In particular, for the test procedure we describe, power will increase to one for alternatives in which the difference between the two latent positions grows with the sample size.

Our test procedure is, at first glance, deceptively simple: given a pair of adjacency matrices $\mathbf{A}$ and $\mathbf{B}$ for two d-dimensional random dot product graphs, we generate their adjacency spectral embeddings, denoted $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$, respectively, and compute an appropriately normalized version of the so-called *Procrustes fit* or *Procrustes distance* between the two embeddings:

$$\min_{\mathbf{W} \in \mathcal{O}^{d \times d}} \|\hat{\mathbf{X}} - \hat{\mathbf{Y}}\mathbf{W}\|_F$$

(Recall that such a fit is necessary because of the inherent nonidentifiability of the random dot product model.)

Understanding the limiting distribution of this test statistic is more complicated, however, and appropriately framing the set of null and alternative hypothesis for which the test is valid and consistent (that is, a level α-test with power converging to 1 as $n \to \infty$) is delicate. To that end, we first state the key concentration inequality for $\min_{\mathbf{W} \in \mathcal{O}^{d \times d}} \|\hat{\mathbf{X}} - \hat{\mathbf{Y}}\mathbf{W}\|_F$.

Theorem 32 *Suppose* $\mathbf{P} = \mathbf{X}\mathbf{X}^\top$ *is an* $n \times n$ *probability matrix of rank* d. *Suppose also that there exists* $\epsilon > 0$ *such that* $\delta(\mathbf{P}) > (\log n)^{2+\epsilon}$. *Let* $c > 0$ *be arbitrary but fixed. Then there exists a* $n_0(c)$ *and a universal constant* $C \geq 0$ *such that if* $n \geq n_0$ *and* $n^{-c} < \eta < 1/2$, *then*

33

there exists a deterministic $\mathbf{W} \in \mathcal{O}(d)$ *such that, with probability at least* $1 - 3\eta$,

$$\left| \|\hat{\mathbf{X}} - \mathbf{X}\mathbf{W}\|_F - C(\mathbf{X}) \right| \leq \frac{Cd \log (n/\eta)}{C(\mathbf{X})\sqrt{\gamma^5(\mathbf{P})\delta(\mathbf{P})}} \tag{28}$$

where $C(\mathbf{X})$ *is a function of* $\mathbf{X}$ *given by*

$$C(\mathbf{X}) = \sqrt{\operatorname{tr} \mathbf{S}_{\mathbf{P}}^{-1/2}\mathbf{U}_{\mathbf{P}}^{\top}\mathbb{E}[(\mathbf{A} - \mathbf{P})^2]\mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{-1/2}} \tag{29}$$

where $\mathbb{E}[(\mathbf{A} - \mathbf{P})^2]$ *is taken with respect to* $\mathbf{A}$ *and conditional on* $\mathbf{X}$.

We note that the proof of this theorem consists of two pieces: it is straightforward to show that the Frobenius norm bound of Lemma 50 implies that

$$\|\hat{\mathbf{X}} - \mathbf{X}\mathbf{W}\|_F = \|(\mathbf{A} - \mathbf{P})\mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{-1/2}\|_F + O(d\log(n)\delta^{-1/2}(\mathbf{P})\gamma^{-5/2}(\mathbf{P}))$$

To complete the theorem, then, in Tang et al. (2017a) we provide a concentration inequality for $\|(\mathbf{A} - \mathbf{P})\mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{-1/2}\|_F^2$, showing that

$$\left| \|(\mathbf{A} - \mathbf{P})\mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{-1/2}\|_F^2 - C^2(\mathbf{X}) \right| \leq \frac{14\sqrt{2d}\log (n/\eta)}{\gamma(\mathbf{P})\sqrt{\delta(\mathbf{P})}}. \tag{30}$$

where $C(\mathbf{X})$ is as defined in (29). We do not go into the details of this concentration inequality here, but rather point the reader to Tang et al. (2017a). We observe, however, that this inequality has immediate consequences for two-sample testing for random dot product graphs. For two random dot product graphs with probability matrices $\mathbf{P} = \mathbf{X}\mathbf{X}^{\top}$ and $\mathbf{Q} = \mathbf{Y}\mathbf{Y}^{\top}$, consider the null hypothesis $\mathbf{X} = \mathbf{Y}\mathbf{W}$ for some orthogonal $\mathbf{W}$. It can be shown that $\min_{\mathbf{W} \in \mathcal{O}^{d \times d}} \|\hat{\mathbf{X}} - \hat{\mathbf{Y}}\mathbf{W}\|_F$ is the basis for a valid and consistent test. We emphasize, though, that as the graph size n increases, the $n \times d$ matrix of latent positions also increases in size. As a consequence of this, we consider the following notion of consistency in this semiparametric setting. As an aside on the notation, in this section, we consider a sequence of graphs with latent positions, all indexed by n; thus, as we noted in our preliminary remarks on notation, $\mathbf{X}_n$ and $\hat{\mathbf{X}}_n$ refer to the *matrices* of true and estimated latent positions in this sequence. We let

$$\mathbf{X}_n =_W \mathbf{Y}_n$$

denote that $\mathbf{X}_n$ and $\mathbf{Y}_n$ are equivalent up to orthogonal transformation; that is

$$\mathbf{X}_n = \mathbf{Y}_n\mathbf{W}$$

for some orthogonal $\mathbf{W} \in \mathbb{R}^{d \times d}$.

Definition 33 *Let* $(\mathbf{X}_n, \mathbf{Y}_n)_{n \in \mathbb{N}}$, *be a given sequence of latent positions, where* $\mathbf{X}_n$ *and* $\mathbf{Y}_n$ *are both in* $\mathbb{R}^{n \times d}$. *A test statistic* T_n *and associated rejection region* R_n *to test the null hypothesis*

$$H_0^n : \mathbf{X}_n =_W \mathbf{Y}_n \quad against \quad H_a^n : \mathbf{X}_n \neq_W \mathbf{Y}_n$$

is a consistent, asymptotically level α test *if for any* $\eta > 0$, *there exists* $n_0 = n_0(\eta)$ *such that*

34

(i) If $n > n_0$ and H_a^n is true, then $P(T_n \in R_n) > 1 - \eta$

(ii) If $n > n_0$ and H_0^n is true, then $P(T_n \in R_n) \leq \alpha + \eta$

With this definition of consistency, we obtain the following theorem on two-sample testing for random dot products on the same vertex set and with known vertex correspondence.

Theorem 34 *For each fixed n, consider the hypothesis test*

$$H_0^n : \mathbf{X}_n =_W \mathbf{Y}_n \quad versus \quad H_a^n : \mathbf{X}_n \neq_W \mathbf{Y}_n$$

where $\mathbf{X}_n$ and $\mathbf{Y}_n \in \mathbb{R}^{n \times d}$ are matrices of latent positions for two random dot product graphs. Let $\hat{\mathbf{X}}_n$ and $\hat{\mathbf{Y}}_n$ be the adjacency spectral embeddings of $\mathbf{A}_n \sim \mathrm{Bernoulli}(\mathbf{X}_n \mathbf{X}_n^\top)$ and $\mathbf{B}_n \sim \mathrm{Bernoulli}(\mathbf{Y}_n \mathbf{Y}_n^\top)$, respectively. Define the test statistic T_n as follows:

$$T_n = \frac{\min\limits_{\mathbf{W} \in \mathcal{O}(d)} \|\hat{\mathbf{X}}_n \mathbf{W} - \hat{\mathbf{Y}}_n\|_F}{\sqrt{d\gamma^{-1}(\mathbf{A}_n)} + \sqrt{d\gamma^{-1}(\mathbf{B}_n)}}. \tag{31}$$

Let $\alpha \in (0,1)$ be given. Then for all $C > 1$, if the rejection region is $R := \{t \in \mathbb{R} : t \geq C\}$, then there exists an $n_1 = n_1(\alpha, C) \in \mathbb{N}$ such that for all $n \geq n_1$, the test procedure with T_n and rejection region R is an at most level α test, that is, for all $n \geq n_1$, if $\mathbf{X}_n =_W \mathbf{Y}_n$, then $\mathbb{P}(T_n \in R) \leq \alpha$. Furthermore, suppose the sequence of latent positions $\{\mathbf{X}_n\}$ and $\{\mathbf{Y}_n\}$, $n \in \mathbb{N}$, adhere to the following requirements:

(i) *The latent positions satisfy the eigengap assumptions in Assumption 1;*

(ii) *With d_n denoting the quantity $d_n := \min\limits_{\mathbf{W} \in \mathcal{O}(d)} \|\mathbf{X}_n \mathbf{W} - \mathbf{Y}_n\|$, suppose $d_n \neq 0$ for infinitely many n. Let $t_1 = \min\{k > 0 : d_k > 0\}$ and sequentially define $t_n = \min\{k > t_{n-1} : d_k > 0\}$. Let $b_n = d_{t_n}$. Suppose $\liminf b_n = \infty$.*

Then this test procedure is consistent in the sense of Definition 33 over this sequence of latent positions.

Remark 35 *This result does not require that $\mathbf{A}_n$ and $\mathbf{B}_n$ be independent for any fixed n, nor that the sequence of pairs $(\mathbf{A}_n, \mathbf{B}_n)$, $n \in \mathbb{N}$, be independent. Further, with regard to requirement (ii) above, we note that Theorem 34 is written to emphasize consistency in the sense of Definition 33, even in a case when, for example, the latent position sequence is such that $\mathbf{X}_n =_W \mathbf{Y}_n$ for all even n, but $\mathbf{X}_n$ and $\mathbf{Y}_n$ are sufficiently far apart for odd n. In addition, the insistence that $\liminf b_k = \infty$ can also be relaxed somewhat. Specifically, consistency is achieved as long as*

$$\liminf_{n \to \infty} \left(\|\mathbf{X}_n \mathbf{W} - \mathbf{Y}_n\|_F - C(\mathbf{X}_n) - C(\mathbf{Y}_n) \right) > 0.$$

It also possible to construct analogous tests for latent positions related by scaling factors, or, in the case of the degree-corrected stochastic block model, by projection. We summarize these below, beginning with the case of scaling.

For the scaling case, let $\mathcal{C} = \mathcal{C}(\mathbf{Y}_n)$ denote the class of all positive constants c for which all the entries of $c^2 \mathbf{Y}_n \mathbf{Y}_n^\top$ belong to the unit interval. We wish to test the null hypothesis

H_0:$\mathbf{X}_n =_W c_n \mathbf{Y}_n$ for some $c_n \in \mathcal{C}$ against the alternative H_a:$\mathbf{X}_n \neq_W c_n \mathbf{Y}_n$ for any $c_n \in \mathcal{C}$. Recall, again, that our notation denotes equivalence up to $d \times d$ orthogonal transformation $\mathbf{W}$. In what follows below, we will only write $c_n > 0$, but will always assume that $c_n \in \mathcal{C}$, since the problem is ill-posed otherwise. The test statistic T_n is now a simple modification of the one used in Theorem 34: for this test, we compute a Procrustes distance between scaled adjacency spectral embeddings for the two graphs.

Theorem 36 *For each fixed n, consider the hypothesis test*

$$H_0^n\text{:}\mathbf{X}_n =_W c_n \mathbf{Y}_n \quad \text{for some } c_n > 0 \text{ versus } \quad H_a^n\text{:}\mathbf{X}_n \neq_W c_n \mathbf{Y}_n \quad \text{for all } c_n > 0$$

where $\mathbf{X}_n$ and $\mathbf{Y}_n \in \mathbb{R}^{n \times d}$ are latent positions for two random dot product graphs with adjacency matrices $\mathbf{A}_n$ and $\mathbf{B}_n$, respectively. Define the test statistic T_n as follows:

$$T_n = \frac{\min\limits_{\mathbf{W} \in \mathcal{O}(d)} \|\hat{\mathbf{X}}_n \mathbf{W}/\|\hat{\mathbf{X}}_n\|_F - \hat{\mathbf{Y}}_n/\|\hat{\mathbf{Y}}_n\|_F\|_F}{2\sqrt{d\gamma^{-1}(\mathbf{A}_n)}/\|\hat{\mathbf{X}}_n\|_F + 2\sqrt{d\gamma^{-1}(\mathbf{B}_n)}/\|\hat{\mathbf{Y}}_n\|_F}. \tag{32}$$

Let $\alpha \in (0,1)$ be given. Then for all $C > 1$, if the rejection region is $R := \{t \in \mathbb{R} : t \geq C\}$, then there exists an $n_1 = n_1(\alpha, C) \in \mathbb{N}$ such that for all $n \geq n_1$, the test procedure with T_n and rejection region R is an at most level α test. Furthermore, consider the sequence of latent position $\{\mathbf{X}_n\}$ and $\{\mathbf{Y}_n\}$, $n \in \mathbb{N}$, satisfying Assumption 1 and denote by d_n the quantity

$$d_n := \frac{\min\limits_{\mathbf{W} \in \mathcal{O}(d)} \|\mathbf{X}_n \mathbf{W}/\|\mathbf{X}_n\|_F - \mathbf{Y}_n/\|\mathbf{Y}_n\|_F\|_F}{1/\|\mathbf{X}_n\|_F + 1/\|\mathbf{Y}_n\|_F} = \frac{\min\limits_{\mathbf{W} \in \mathcal{O}(d)} \|\mathbf{X}_n\|\mathbf{Y}_n\|_F \mathbf{W} - \mathbf{Y}_n\|\mathbf{X}_n\|_F\|_F}{\|\mathbf{X}_n\|_F + \|\mathbf{Y}_n\|_F} \tag{33}$$

Suppose $d_n \neq 0$ for infinitely many n. Let $t_1 = \min\{k > 0 : d_k > 0\}$ and sequentially define $t_n = \min\{k > t_{n-1} : d_k > 0\}$. Let $b_n = d_{t_n}$. If $\liminf b_n = \infty$, then this test procedure is consistent in the sense of Definition 33 over this sequence of latent positions.

We next consider the case of testing whether the latent positions are related by a diagonal transformation: that is, whether H_0:$\mathbf{X}_n =_W \mathbf{D}_n \mathbf{Y}_n$ for some diagonal matrix $\mathbf{D}_n$. We proceed analogously to the scaling case, above, by defining the class $\mathcal{E} = \mathcal{E}(\mathbf{Y}_n)$ to be all positive diagonal matrices $\mathbf{D}_n \in \mathbb{R}^{n \times n}$ such that $\mathbf{D}_n \mathbf{Y}_n \mathbf{Y}_n^\top \mathbf{D}_n$ has all entries in the unit interval.

As before, we will always assume that $\mathbf{D}_n$ belongs to $\mathcal{E}$, even if this assumption is not explicitly stated. The test statistic T_n in this case is again a simple modification of the one used in Theorem 34. However, for technical reasons, our proof of consistency requires an additional condition on the minimum Euclidean norm of each row of the matrices $\mathbf{X}_n$ and $\mathbf{Y}_n$. To avoid certain technical issues, we impose a slightly stronger density assumption on our graphs for this test. These assumptions can be weakened, but at the cost of interpretability. The assumptions we make on the latent positions, which we summarize here, are moderate restrictions on the sparsity of the graphs.

Assumption 2 *We assume that there exists $d \in \mathbb{N}$ such that for all n, $\mathbf{P}_n$ is of rank d. Further, we assume that there exist constants $\epsilon_1 > 0$, $\epsilon_2 \in (0,1)$, $c_0 > 0$ and $n_0(\epsilon_1, \epsilon_2, c) \in \mathbb{N}$*

such that for all $n \geq n_0$:

$$\gamma(\mathbf{P}_n) \geq c_0; \qquad \delta(\mathbf{P}_n) \geq (\log n)^{2+\epsilon_1}; \qquad \min_i \|X_i\| > \left(\frac{\log n}{\sqrt{\delta(\mathbf{P}_n)}}\right)^{1-\epsilon_2} \tag{34}$$

We then have the following result.

Theorem 37 *For each fixed n, consider the hypothesis test*

$$H_0^n : \mathbf{X}_n =_W \mathbf{D}_n \mathbf{Y}_n \quad \text{for some } \mathbf{D}_n \in \mathcal{E} \quad \text{versus} \quad H_a^n : \mathbf{X}_n \neq_W \mathbf{D}_n \mathbf{Y}_n \quad \text{for any } \mathbf{D}_n \in \mathcal{E}$$

where $\mathbf{X}_n$ and $\mathbf{Y}_n \in \mathbb{R}^{n \times d}$ are matrices of latent positions for two random dot product graphs. For any matrix $\mathbf{Z} \in \mathbb{R}^{n \times d}$, let $\mathcal{D}(\mathbf{Z})$ be the diagonal matrix whose diagonal entries are the Euclidean norm of the rows of $\mathbf{Z}$ and let $\mathcal{P}(\mathbf{Z})$ be the matrix whose rows are the projection of the rows of $\mathbf{Z}$ onto the unit sphere. We define the test statistic as follows:

$$T_n = \frac{\min_{\mathbf{W} \in \mathcal{O}(d)} \|\mathcal{P}(\hat{\mathbf{X}}_n)\mathbf{W} - \mathcal{P}(\hat{\mathbf{Y}}_n)\|_F}{2\sqrt{d\gamma^{-1}(\mathbf{A})}\|\mathcal{D}^{-1}(\hat{\mathbf{X}}_n)\| + 2\sqrt{d\gamma^{-1}(\mathbf{B}_n)}\|\mathcal{D}^{-1}(\hat{\mathbf{Y}}_n)\|}. \tag{35}$$

where we write $\mathcal{D}^{-1}(\mathbf{Z})$ for $(\mathcal{D}(\mathbf{Z}))^{-1}$. Note that $\|\mathcal{D}^{-1}(\mathbf{Z})\| = 1/(\min_i \|Z_i\|)$.

Let $\alpha \in (0,1)$ be given. Then for all $C > 1$, if the rejection region is $R := \{t \in \mathbb{R} : t \geq C\}$, then there exists an $n_1 = n_1(\alpha, C) \in \mathbb{N}$ such that for all $n \geq n_1$, the test procedure with T_n and rejection region R is an at most level-α test. Furthermore, consider the sequence of latent position $\{\mathbf{X}_n\}$ and $\{\mathbf{Y}_n\}$, $n \in \mathbb{N}$, satisfying Assumption 2 and denote by d_n the quantity

$$d_n := \frac{\min_{\mathbf{W} \in \mathcal{O}(d)} \|\mathcal{P}(\mathbf{X}_n)\mathbf{W} - \mathcal{P}(\mathbf{Y}_n)\|_F}{\|\mathcal{D}^{-1}(\mathbf{X})\| + \|\mathcal{D}^{-1}(\mathbf{Y})\|} = D_\mathcal{P}(\mathbf{X}_n, \mathbf{Y}_n) \tag{36}$$

Suppose $d_n \neq 0$ for infinitely many n. Let $t_1 = \min\{k > 0 : d_k > 0\}$ and sequentially define $t_n = \min\{k > t_{n-1} : d_k > 0\}$. Let $b_n = d_{t_n}$. If $\liminf b_n = \infty$, then this test procedure is consistent in the sense of Definition 33 over this sequence of latent positions.

This collection of semiparametric tests has numerous applications in graph comparison; in Section 6 , we describe its use in connectomics and brain scan data. We stress, though, that the Procrustes transformations are rather cumbersome, and they limit our ability to generalize these procedures to graph comparisons involving more than two graphs. As a consequence, it can be useful to consider *joint* or *omnibus* embeddings, in which adjacency matrices for multiple graphs on the same vertex set are jontly embedded into a single (larger-dimensional) space, but *with a distinct representation for each graph.* For an illuminating joint graph inference study on the *C. elegans* connectome that addresses somewhat different questions from semiparametric testing, see Chen et al. (2016). Simultaneously embedding multiple graphs into a shared space allows comparison of graphs without the need to perform pairwise alignments of graph embeddings. Further, a distinct representation of each graph renders the omnibus embedding especially useful for subsequent comparative graph inference.

5.3 Omnibus embedding

In Levin et al. (2017), we show that an omnibus embedding—that is, an embedding of multiple graphs into a single shared space—can yield consistent estimates of underlying latent positions. Moreover, like the adjacency spectral embedding for a single graph, the rows of this omnibus embedding, suitably-scaled, are asymptotically normally distributed. As might be anticipated, the use of multiple independent graphs generated from the same latent positions, as opposed just a single graph, yields a reduction in variance for the estimated latent positions, and since the omnibus embedding provides a distinct representation for each graph, subsequently averaging these estimates reduces the variance further still. Further, the omnibus embedding allows us to compare graphs without cumbersome Procrustes alignments, but with nevertheless very similar empirical power, at least in certain simple cases, as that of the Procrustes-centered semiparametric test decribed earlier. In sum, the omnibus embedding can deliver nearly-optimal estimation of latent positions, and improvements in testing. Finally, when embedding m d-dimensional random dot product graphs, each with n vertices, the omnibus embedding provides a matrix in $\mathbb{R}^{nm \times d}$, with a separate representation in $\mathbb{R}^d$ for each vertex *in each graph*. We surmise that this distinct representation may be especially helpful for inference tasks such as graph regression, estimation of additional vertex covariates, or changepoint detection in a time series of graphs, and the impact of the omnibus embedding on such problems is a topic of curren investigation.

The discussion we give here is condensed and reproduced in significant parts from Levin et al. (2017). To construct the omnibus embedding, we consider a collection of m random dot product graphs, all with the same generating latent positions. This motivates the following definition:

Definition 38 (Conditionally independent, identically distributed joint RDPG)
Let F be a d-dimensional inner product distribution on $\mathbb{R}^d$. We say that random graphs $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(m)}$ are distributed as a conditionally independent, identically distributed joint random dot product graph (ciJRDPG) *and write*

$$(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(m)}, \mathbf{X}) \sim \text{ciJRDPG}(F, n, m)$$

if $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n]^\top \in \mathbb{R}^{n \times d}$ has its (transposed) rows distributed i.i.d. as $\mathbf{X}_i \sim F$, and we have marginal distributions $(\mathbf{A}^{(k)}, \mathbf{X}) \sim \text{RDPG}(F, n)$ for each $k = 1, 2, \ldots, m$. That is, the $\mathbf{A}^{(k)}$ are conditionally independent given $\mathbf{X}$, with edges independently distributed as $\mathbf{A}_{i,j}^{(k)} \sim \text{Bernoulli}((\mathbf{X}\mathbf{X}^\top)_{ij})$ for all $1 \le i < j \le n$ and all $k \in [m]$.

Remark 39 *We observe that the notion of a* joint *random dot product graph can be much more general than this definition alone, and indeed can allow for different mechanisms of dependence across graphs. Here, however, we are interested in the simple case in which the graphs are all on the same vertex set, and latent positions for the graphs have the same distribution; given the latent position matrix, the m graphs are generated independently of one another from the same latent positions.*

Given a set of m adjacency matrices distributed as

$$(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(m)}, \mathbf{X}) \sim \text{ciJRDPG}(F, n, m)$$

for distribution F on $\mathbb{R}^d$, a natural inference task is to recover the n latent positions $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \in \mathbb{R}^d$ shared by the vertices of the m graphs. To estimate the underlying latent postions from these m graphs, Tang et al. (2016) provides justification for the estimate $\bar{\mathbf{X}} = \mathrm{ASE}(\bar{\mathbf{A}}, d)$, where $\bar{\mathbf{A}}$ is the sample mean of the adjacency matrices $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(m)}$. However, $\bar{\mathbf{X}}$ is ill-suited to any task that requires comparing latent positions across the m graphs, since the $\bar{\mathbf{X}}$ estimate collapses the m graphs into a single set of n latent positions. This motivates the *omnibus embedding*, which still yields a single spectral decomposition, but with a separate d-dimensional representation for each of the m graphs. This makes the omnibus embedding useful for *simultaneous* inference across all m observed graphs.

Definition 40 (Omnibus embedding) *Let* $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(m)} \in \mathbb{R}^{n \times n}$ *be adjacency matrices of a collection of m undirected graphs. We define the mn-by-mn omnibus matrix of* $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(m)}$ *by*

$$\mathbf{M} = \begin{bmatrix} \mathbf{A}^{(1)} & \frac{1}{2}(\mathbf{A}^{(1)} + \mathbf{A}^{(2)}) & \ldots & \frac{1}{2}(\mathbf{A}^{(1)} + \mathbf{A}^{(m)}) \\ \frac{1}{2}(\mathbf{A}^{(2)} + \mathbf{A}_2^{(1)}) & \mathbf{A}^{(2)} & \ldots & \frac{1}{2}(\mathbf{A}^{(2)} + \mathbf{A}^{(m)}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}(\mathbf{A}^{(m)} + \mathbf{A}^{(1)}) & \frac{1}{2}(\mathbf{A}^{(m)} + \mathbf{A}^{(2)}) & \ldots & \mathbf{A}^{(m)} \end{bmatrix}, \tag{37}$$

and the d-dimensional omnibus embedding *of* $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(m)}$ *is the adjacency spectral embedding of* $\mathbf{M}$:

$$\mathrm{OMNI}(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(m)}, d) = \mathrm{ASE}(\mathbf{M}, d).$$

where ASE is the d-dimensional adjacency spectral embedding. Under the ciJRDPG model, given the latent positions $\mathbf{X}$, the omnibus matrix has expected value

$$\mathbb{E}\mathbf{M} = \tilde{\mathbf{P}} = \mathbf{J}_m \otimes \mathbf{P} = \mathbf{U}_{\tilde{\mathbf{P}}} \mathbf{S}_{\tilde{\mathbf{P}}} \mathbf{U}_{\tilde{\mathbf{P}}}^\top$$

for $\mathbf{U}_{\tilde{\mathbf{P}}} \in \mathbb{R}^{mn \times d}$ having d orthonormal columns and $\mathbf{S}_{\tilde{\mathbf{P}}} \in \mathbb{R}^{d \times d}$ diagonal. Since $\mathbf{M}$ is a reasonable estimate for $\tilde{\mathbf{P}} = \mathbb{E}\mathbf{M}$, the matrix $\hat{\mathbf{Z}} = \mathrm{OMNI}(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(m)}, d)$ is a natural estimate of the mn latent positions collected in the matrix $\mathbf{Z} = [\mathbf{X}^\top \mathbf{X}^\top \ldots \mathbf{X}^\top]^\top \in \mathbb{R}^{mn \times d}$. Here again, as in Remark 7, $\hat{\mathbf{Z}}$ only recovers the true latent positions $\mathbf{Z}$ up to an orthogonal rotation. The matrix

$$\mathbf{Z}^* = \begin{bmatrix} \mathbf{X}^* \\ \mathbf{X}^* \\ \vdots \\ \mathbf{X}^* \end{bmatrix} = \mathbf{U}_{\tilde{\mathbf{P}}} \mathbf{S}_{\tilde{\mathbf{P}}}^{1/2} \in \mathbb{R}^{mn \times d}, \tag{38}$$

provides a reasonable canonical choice of latent positions, so that $\mathbf{Z} = \mathbf{Z}^* \mathbf{W}$ for some suitably-chosen orthogonal matrix $\mathbf{W} \in \mathbb{R}^{d \times d}$; again, just as for a single random dot product graph, spectral embedding of the omnibus matrix is a consistent estimator for the latent positions (up to rotation).

Below, we state precise results on consistency and asymptotic normality of the embedding of the omnibus matrix $\mathbf{M}$. The proofs are similar to, but somewhat more involved than, the aforementioned analogues for the adjacency spectral embedding for one graph. We also demonstrate from simulations that the omnibus embedding can be successfully leveraged for subsequent inference, specifically two-sample testing.

First, Lemma 41 shows that the omnibus embedding provides uniformly consistent estimates of the true latent positions, up to an orthogonal transformation, roughly analogous to Lemma 5 in Lyzinski et al. (2014). Lemma 41 shows consistency of the omnibus embedding under the $2 \to \infty$ norm, implying that all mn of the estimated latent positions are near (a rotation of) their corresponding true positions.

Lemma 41 *With $\tilde{\mathbf{P}}$, $\mathbf{M}$, $\mathbf{U_M}$, and $\mathbf{U}_{\tilde{\mathbf{P}}}$ defined as above, there exists an orthogonal matrix $\tilde{\mathbf{W}} \in \mathbb{R}^{d \times d}$ such that with high probability,*

$$\|\mathbf{U_M S_M}^{1/2} - \mathbf{U}_{\tilde{\mathbf{P}}} \mathbf{S}_{\tilde{\mathbf{P}}}^{1/2} \tilde{\mathbf{W}}\|_{2 \to \infty} \leq \frac{C m^{1/2} \log mn}{\sqrt{n}}. \tag{39}$$

As with the adjacency spectral embedding, we once again can show the asymptotic normality of the individual rows of the omnibus embedding. Note that the covariance matrix does change with m, and for m large, this results in a nontrivial variance reduction.

Theorem 42 *Let $(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \dots, \mathbf{A}^{(m)}, \mathbf{X}) \sim \text{ciJRDPG}(F, n, m)$ for some d-dimensional inner product distribution F and let $\mathbf{M}$ denote the omnibus matrix as in (37). Let $\mathbf{Z} = \mathbf{Z}^* \mathbf{W}$ with $\mathbf{Z}^*$ as defined in Equation (38), with estimate $\hat{\mathbf{Z}} = \text{OMNI}(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \dots, \mathbf{A}^{(m)}, d)$. Let $h = m(s - 1) + i$ for $i \in [n], s \in [m]$, so that $\hat{\mathbf{Z}}_h$ denotes the estimated latent position of the i-th vertex in the s-th graph $\mathbf{A}^{(s)}$. That is, $\hat{\mathbf{Z}}_h$ is the column vector formed by transposing the h-th row of the matrix $\hat{\mathbf{Z}} = \mathbf{U_M S_M}^{1/2} = \text{OMNI}(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \dots, \mathbf{A}^{(m)}, d)$. Let $\Phi(\mathbf{x}, \boldsymbol{\Sigma})$ denote the cumulative distribution function of a (multivariate) Gaussian with mean zero and covariance matrix $\boldsymbol{\Sigma}$, evaluated at $\mathbf{x} \in \mathbb{R}^d$. There exists a sequence of orthogonal d-by-d matrices $(\tilde{\mathbf{W}}_n)_{n=1}^{\infty}$ such that for all $\mathbf{x} \in \mathbb{R}^d$,*

$$\lim_{n \to \infty} \Pr\left[n^{1/2} \left(\hat{\mathbf{Z}} \tilde{\mathbf{W}}_n - \mathbf{Z} \right)_h \leq \mathbf{x} \right] = \int_{\text{supp} F} \Phi\left(\mathbf{x}, \boldsymbol{\Sigma}(\mathbf{y}) \right) dF(\mathbf{y}),$$

where $\boldsymbol{\Sigma}(\mathbf{y}) = (m + 3)\boldsymbol{\Delta}^{-1}\tilde{\boldsymbol{\Sigma}}(\mathbf{y})\boldsymbol{\Delta}^{-1}/(4m)$, $\boldsymbol{\Delta} = \mathbb{E}[\mathbf{X}_1 \mathbf{X}_1^\top]$ and

$$\tilde{\boldsymbol{\Sigma}}(\mathbf{y}) = \mathbb{E}\left[(\mathbf{y}^\top \mathbf{X}_1 - (\mathbf{y}^\top \mathbf{X}_1)^2)\mathbf{X}_1 \mathbf{X}_1^\top \right].$$

Next, we summarize from Levin et al. (2017) experiments on synthetic data exploring the efficacy of the omnibus embedding described above. If we merely wish to estimate the latent positions $\mathbf{X}$ of a set of m graphs $(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \dots, \mathbf{A}^{(m)}, \mathbf{X}) \sim \text{ciJRDPG}(F, n, m)$, the estimate $\bar{\mathbf{X}} = \text{ASE}(\sum_{i=1}^{m} \mathbf{A}^{(i)}/m, d)$, the embedding of the sample mean of the adjacency matrices performs well asymptotically (see Tang et al., 2016). Indeed, all else equal, the embedding $\bar{\mathbf{X}}$ is preferable to the omnibus embedding if only because it requires an eigendecomposition of an n-by-n matrix rather than the much larger mn-by-mn omnibus matrix.

Of course, the omnibus embedding can still be used to to estimate the latent positions, potentially at the cost of increased variance. Figure 5 compares the mean-squared error of various techniques for estimating the latent positions for a random dot product graph. The figure plots the (empirical) mean squared error in recovering the latent positions of a 3-dimensional ciJRDPG as a function of the number of vertices n. Each point in the plot is the empirical mean of 50 independent trials; in each trial, the latent positions are
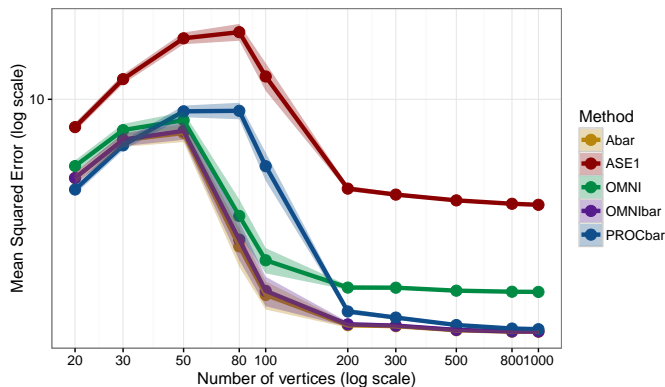
Figure 5: Mean squared error (MSE) in recovery of latent positions (up to rotation) in a 2-graph joint RDPG model as a function of the number of vertices. Performance of ASE applied to a single graph (red), ASE embedding of the mean graph (gold), the Procrustes-based pairwise embedding (blue), the omnibus embedding (green) and the mean omnibus embedding (purple). Each point is the mean of 50 trials; error bars indicate $\pm 2(SE)$. Mean omnibus embedding (OMNIbar) is competitive with $\text{ASE}(\bar{\mathbf{A}}, d)$; Procrustes alignment estimation is notably inferior to the other two-graph techniques for graphs of size between 80 and 200 vertices (note that the gap appears to persist at larger graph sizes, though it shrinks). Figure duplicated from Levin et al. (2017).

drawn i.i.d. from a Dirichlet with parameter $[1, 1, 1]^\top \in \mathbb{R}^3$. Once the latent positions are so obtained, we independently generate two random dot product graphs, $\mathbf{A}^{(1)}, \mathbf{A}^{(2)} \in \mathbb{R}^{n \times n}$ with these latent positions. The figure is interpreted as follows:

1. **ASE1 (red)**: we embed only one of the two observed graphs, and use only the ASE of that graph to estimate the latent positions in $\mathbf{X}$, ignoring entirely the information present in $\mathbf{A}^{(2)}$. This condition serves as a baseline for how much additional information is provided by the second graph $\mathbf{A}^{(2)}$.

2. **Abar (gold)**: we embed the average of the two graphs, $\bar{\mathbf{A}} = (\mathbf{A}^{(1)} + \mathbf{A}^{(2)})/2$ as $\hat{\mathbf{X}} = \text{ASE}(\bar{\mathbf{A}}, 3)$.

3. **OMNI (green)**: We apply the omnibus embedding to obtain $\hat{\mathbf{Z}} = \text{ASE}(\mathbf{M}, 3)$, where $\mathbf{M}$ is as in Equation (37). We then use only the first n rows of $\hat{\mathbf{Z}} \in \mathbb{R}^{2n \times d}$ as our estimate of $\mathbf{X}$. This embedding incorporates information available in both graphs $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$, but does not weight them equally, since the first rows of $\hat{\mathbf{Z}}$ are based primarily on the information contained in $\mathbf{A}^{(1)}$.

4. **OMNIbar (purple)**: We again apply the omnibus embedding to obtain estimated latent positions $\hat{\mathbf{Z}} = \text{ASE}(\mathbf{M}, 3)$, but this time we use all available information by averaging the first n rows and the second n rows of $\hat{\mathbf{Z}}$.

5. **PROCbar (blue)**: We separately embed the graphs $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$, perform a Procrustes alignment between the two embeddings, and average the aligned embeddings to obtain our final estimate of the latent positions.

41

First, let us note that ASE applied to a single graph (red) lags all other methods, as expected, since this discards all information from the second graph. For very small graphs, the dearth of signal is such that no method will recover the latent positions accurately.

Crucially, however, we see that the OMNIbar estimate (purple) performs nearly identically to the Abar estimate (gold), the natural choice among spectral methods for the estimation latent positions The Procrustes estimate (in blue) provides a two-graph analogue of ASE (red): it combines two ASE estimates via Procrustes alignment, but does not enforce an *a priori* alignment of the estimated latent positions. As predicted by the results in Lyzinski et al. (2014) and Tang et al. (2017a), the Procrustes estimate is competitive with the Abar (gold) estimate for suitably large graphs. The OMNI estimate (in green) serves, in a sense, as an intermediate, since it uses information available from both graphs, but in contrast to Procrustes (blue), OMNIbar (purple) and Abar (gold), it does not make complete use of the information available in the second graph. For this reason, it is noteworthy that the OMNI estimate outperforms the Procrustes estimate for graphs of 80-100 vertices. That is, for certain graph sizes, the omnibus estimate appears to more optimally leverage the information in both graphs than the Procrustes estimate does, despite the fact that the information in the second graph has comparatively little influence on the OMNI embedding.

The omnibus embedding can also be applied to testing the semiparametric hypothesis that two observed graphs are drawn from the same underlying latent positions. Consider a collection of latent positions $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n, \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n \in \mathbb{R}^d$. Let the graph G_1 with adjacency matrix $\mathbf{A}^{(1)}$ have edges distributed independently as $\mathbf{A}_{ij}^{(1)} \sim \text{Bernoulli}(\mathbf{X}_i^\top \mathbf{X}_j)$. Similarly, let G_2 have adjacency matrix $\mathbf{A}^{(2)}$ with edges distributed independently as $\mathbf{A}_{ij}^{(2)} \sim \text{Bernoulli}(\mathbf{Y}_i^\top \mathbf{Y}_j)$. Consider the following null hypothesis:

$$H_0 : \mathbf{X} =_W \mathbf{Y} \tag{40}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are the matrices whose rows are the latent positions $\mathbf{X}_i$, $\mathbf{Y}_i$, respectively. The omnibus embedding provides a natural test of this null hypothesis (40) by comparing the first n and last n rows of a lower-dimensional embedding of the omnibus matrix

$$\mathbf{M} = \begin{bmatrix} \mathbf{A}^{(1)} & (\mathbf{A}^{(1)} + \mathbf{A}^{(2)})/2 \\ (\mathbf{A}^{(1)} + \mathbf{A}^{(2)})/2 & \mathbf{A}^{(2)} \end{bmatrix}.$$

Intuitively, when H_0 holds, the distributional result in Theorem 42 holds, and the i-th and $(n+i)$-th rows of $\text{OMNI}(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, d)$ are equidistributed (though they are not independent). On the other hand, when H_0 fails to hold, there exists at least one $i \in [n]$ for which the i-th and $(n+i)$-th rows of $\mathbf{M}$ are *not* identically distributed, and thus the corresponding embeddings are also distributionally distinct. This suggests a test that compares the first n rows of $\text{OMNI}(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, d)$ against the last n rows. That is, our omnibus test statistic simply considers the Frobenius norm of the difference between the top $n \times d$ submatrix of the omnibus embedding and the bottom $n \times d$ submatrix of the omnibus embedding, *with no Procrustes alignment.* Here, we empirically explore the power this test against its Procrustes-based alternative from Tang et al. (2017a).

We draw $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \in \mathbb{R}^3$ i.i.d. according to a Dirichlet distribution F with parameter $\vec{\alpha} = [1, 1, 1]^\top$. With $\mathbf{X}$ defined as the matrix $\mathbf{X} = [\mathbf{X}_1 \mathbf{X}_2 \ldots \mathbf{X}_n]^\top \in \mathbb{R}^{n \times 3}$, let graph G_1 have

adjacency matrix $\mathbf{A}^{(1)}$, where $\mathbf{A}_{ij}^{(1)} \sim \text{Bernoulli}((\mathbf{XX}^\top)_{ij})$. We generate a second graph G_2 by first drawing random points $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n \overset{\text{i.i.d.}}{\sim} F$. Selecting a set of indices $I \subset [n]$ of size $k < n$ uniformly at random from among all such $\binom{n}{k}$ sets, we let G_2 have latent positions

$$\mathbf{Y}_i = \begin{cases} \mathbf{Z}_i & \text{if } i \in I \\ \mathbf{X}_i & \text{otherwise.} \end{cases}$$

With $\mathbf{Y}$ the matrix $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n]^\top \in \mathbb{R}^{n \times 3}$, we generate graph G_2 with adjacency matrix $\mathbf{A}^{(2)}$, where $\mathbf{A}_{ij}^{(2)} \sim \text{Bernoulli}((\mathbf{YY}^\top)_{ij})$. We wish to test

$$H_0 : \mathbf{X} =_W \mathbf{Y}. \tag{41}$$

Consider two different tests, one based on a Procrustes alignment of the adjacency spectral embeddings of G_1 and G_2, as in Tang et al. (2017a), and the other based on the omnibus embedding. Both approaches are based on estimates of the latent positions of the two graphs. In both cases we use a test statistic that is some variant of the form $T = \sum_{i=1}^n \|\hat{\mathbf{X}}_i - \hat{\mathbf{Y}}_i\|_F^2$, and accept or reject based on a Monte Carlo estimate of the critical value of T under the null hypothesis, in which $\mathbf{X}_i = \mathbf{Y}_i$ for all $i \in [n]$. In each trial, we use 500 Monte Carlo iterates to estimate the distribution of T.

We note that in the experiments presented here, we assume that the latent positions $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ of graph G_1 are known for sampling purposes, so that the matrix $\mathbf{P} = \mathbb{E}\mathbf{A}^{(1)}$ is known exactly, rather than estimated from the observed adjacency matrix $\mathbf{A}^{(1)}$. This allows us to sample from the true null distribution. As proved in Lyzinski et al. (2014), the estimated latent positions $\hat{\mathbf{X}}_1 = \text{ASE}(\mathbf{A}^{(1)})$ and $\hat{\mathbf{X}}_2 = \text{ASE}(\mathbf{A}^{(2)})$ recover the true latent positions $\mathbf{X}_1$ and $\mathbf{X}_2$ (up to rotation) to arbitrary accuracy in $(2, \infty)$-norm for suitably large n (Lyzinski et al., 2014). Without using this known matrix $\mathbf{P}$, we would require that our matrices have tens of thousands of vertices before the variance associated with estimating the latent positions would no longer overwhelm the signal present in the few altered latent positions.

Three major factors influence the complexity of testing the null hypothesis in Equation (41): the number of vertices n, the number of changed latent positions $k = |I|$, and the distances $\|\mathbf{X}_i - \mathbf{Y}_i\|_F$ between the latent positions. The three plots in Figure 6 illustrate the first two of these three factors. These three plots show the power of two different approaches to testing the null hypothesis (41) for different sized graphs and for different values of k, the number of altered latent positions. In all three conditions, both methods improve as the number of vertices increases, as expected, especially since we do not require estimation of the underlying expected matrix $\mathbf{P}$ for Monte Carlo estimation of the null distribution of the test statistic. We see that when only one vertex is changed, neither method has power much above 0.25. However, in the case of $k = 5$ and $k = 10$, is it clear that the omnibus-based test achieves higher power than the Procrustes-based test, especially in the range of 30 to 250 vertices. A more detailed examination of the relative impact of these factors in testing is given in Levin et al. (2017). We note that our present power analysis is limited to an empirical study in a few special cases, and a more detailed analysis of theoretical power guarantees for this embedding is ongoing.
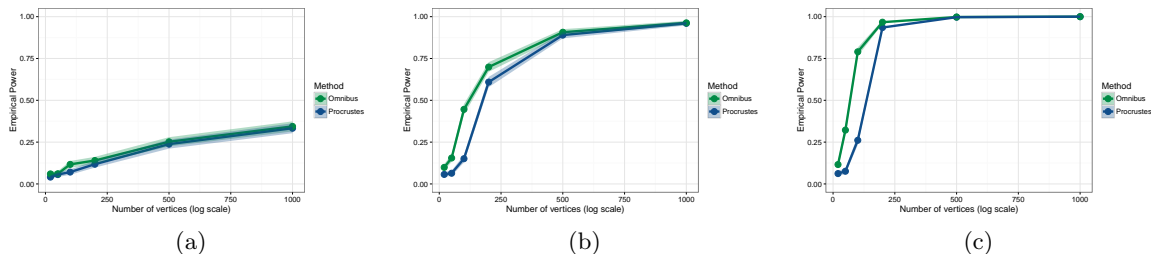
(a) (b) (c)

Figure 6: Power of the ASE-based (blue) and omnibus-based (green) tests to detect when the two graphs being testing differ in (a) one, (b) five, and (c) ten of their latent positions. Each point is the proportion of 1000 trials for which the given technique correctly rejected the null hypothesis, and error bars denote two standard errors of this empirical mean in either direction. Figure duplicated from Levin et al. (2017).

In sum, our omnibus embedding provides a natural mechanism for the simultaneous embedding of multiple graphs into a single vector space. This eliminates the need for multiple Procrustes alignments, which were required in previously-explored approaches to multiple-graph testing (see Tang et al., 2017a). In the Procrustes-based approach, each graph is embedded separately, yielding estimates $\mathbf{X}_1$ and $\mathbf{X}_2$, and the test statistic is

$$\min_{\mathbf{W} \in \mathcal{O}_d} \|\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_2 \mathbf{W}\|_F, \tag{42}$$

Under the null hypothesis, a suitable rescaling of this converges as $n \to \infty$. The effect of this Procrustes alignment on subsequent inference is ill-understood; it has the potential to introduce variance, and our simulations results suggest that it negatively impacts performance in both estimation and testing settings. Furthermore, when the matrix $\mathbf{P} = \mathbf{X}\mathbf{X}^\top$ does not have distinct eigenvalues (and is thus not uniquely diagonalizable), this Procrustes step is unavoidable, since the difference $\|\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_2\|_F$ need not converge at all.

In contrast, our omnibus embedding builds an alignment of the graphs into its very structure. To see this, consider, for simplicity, the $m = 2$ case. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the matrix whose rows are the latent positions of both graphs G_1 and G_2, and let $\mathbf{M} \in \mathbb{R}^{2n \times 2n}$ be their omnibus matrix. Then

$$\mathbb{E}\mathbf{M} = \tilde{\mathbf{P}} = \begin{bmatrix} \mathbf{P} & \mathbf{P} \\ \mathbf{P} & \mathbf{P} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{X} \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ \mathbf{X} \end{bmatrix}^\top.$$

Suppose now that we wish to factorize $\tilde{\mathbf{P}}$ as

$$\tilde{\mathbf{P}} = \begin{bmatrix} \mathbf{X} \\ \mathbf{X}\mathbf{W}^* \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ \mathbf{X}\mathbf{W}^* \end{bmatrix}^\top = \begin{bmatrix} \mathbf{P} & \mathbf{X}(\mathbf{W}^*)^\top \mathbf{X}^\top \\ \mathbf{X}\mathbf{W}^*\mathbf{X}^\top & \mathbf{P} \end{bmatrix}.$$

That is, we want to consider graphs G_1 and G_2 as being generated from the same latent positions, but in one case, say, under a *different* rotation. This possibility necessitates the Procrustes alignment in the case of separately-embedded graphs. In the case of the omnibus matrix, the structure of the $\tilde{\mathbf{P}}$ matrix implies that $\mathbf{W}^* = \mathbf{I}_d$. In contrast to the Procrustes alignment, the omnnibus matrix incorporates an alignment *a priori*. Simulations show

44

that the omnibus embedding outperforms the Procrustes-based test for equality of latent positions, especially in the case of moderately-sized graphs.

To further illustrate the utility of this omnibus embedding, consider the case of testing whether three different random dot product graphs have the same generating latent positions. The omnibus embedding gives us a *single* canonical representation of all three graphs: Let $\hat{\mathbf{X}}_1^O$, $\hat{\mathbf{X}}_2^O$, and $\hat{\mathbf{X}}_3^O$ be the estimates for the three latent position matrices generated from the omnibus embedding. To test whether any two of these random graphs have the same generating latent positions, we merely have to compare the Frobenius norms of their differences, as opposed to computing three separate Procrustes alignments. In the latter case, in effect, we do not have a canonical choice of coordinates in which to compare our graphs simultaneously.

5.4 Nonparametric graph estimation and testing

The semiparametric and omnibus testing procedures we describe both focus on the estimation of the latent positions themselves. But a very natural concern, for a random dot product graph with distribution F, is the estimation of the *distribution F*. We next address how the adjacency spectral embedding, judiciously integrated with kernel density estimation, can be used for nonparametric estimation and testing in random graphs.

Throughout our discussion on nonparametric estimation, we shall always assume that the distributions of the latent positions satisfy the following distinct eigenvalues assumption. The assumption implies that the estimates of the latent position obtained by the adjacency spectral embedding will, in the limit, be uniquely determined.

Assumption 3 *The distribution F for the latent positions $X_1, X_2, \ldots, \sim F$ is such that the second moment matrix $\mathbb{E}[X_1 X_1^\top]$ has d distinct eigenvalues and d is known.*

We realize that Assumption 3 is restrictive – in particular, it is not satisfied by the stochastic block model with $K > 2$ blocks of equal size and edge probabilities p within communities and q between communities – it is a necessary technical condition for us to obtain the limiting results of Theorem 3. The motivation behind this assumption is as follows: the matrix $\mathbb{E}[X_1 X_1^\top]$ is of rank d with d known so that given a graph $\mathbf{A} \sim \text{RDPG}(F)$, one can construct the adjacency spectral embedding of $\mathbf{A}$ into the "right" Euclidean space. The requirement that $\mathbb{E}[X_1 X_1^\top]$ has d distinct eigenvalues is—once again—due to the intrinsic property of non-identifiability of random dot product graphs. As always, for any random dot product graph $\mathbf{A}$, the latent position $\mathbf{X}$ associated with $\mathbf{A}$ can only be estimated up to some true but unknown orthogonal transformation. Because we are concerned with two-sample hypothesis testing, we must guard against the scenario in which we have two graphs $\mathbf{A}$ and $\mathbf{B}$ with latent positions $\mathbf{X} = \{X_i\}_{i=1}^n \overset{\text{i.i.d}}{\sim} F$ and $\mathbf{Y} = \{Y_k\}_{k=1}^m \overset{\text{i.i.d}}{\sim} F$ but whose estimates $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ lie in different, incommensurate subspaces of $\mathbb{R}^d$. That is to say, the estimates $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ satisfy $\hat{\mathbf{X}} \approx \mathbf{X}\mathbf{W}_1$ and $\hat{\mathbf{Y}} \approx \mathbf{Y}\mathbf{W}_2$, but $\|\mathbf{W}_1 - \mathbf{W}_2\|_F$ does not converge to 0 as $n, m \to \infty$. See also Fishkind et al. (2015b) for exposition of a related so-called "incommensurability phenomenon."

Our main point of departure for this subsection compared to Section 5.2 is the assumption that, given a sequence of pairs of random dot product graphs with adjacency matrices $\mathbf{A}_n$ and $\mathbf{B}_n$, the rows of the latent positions $\mathbf{X}_n$ and $\mathbf{Y}_n$ are independent samples from some fixed distributions F and G, respectively. The corresponding tests are therefore tests of equality between F and G. More formally, we consider the following two-sample nonparametric testing problems for random dot product graphs. Let F and G be two inner product distributions. Given $\mathbf{A} \sim \mathrm{RDPG}(F)$ and $\mathbf{B} \sim \mathrm{RDPG}(G)$, we consider the tests:

1. *(Equality, up to orthogonal transformation)*

$$H_0 \colon F \perp\!\!\!\perp G \quad \text{against} \quad H_A \colon F \not\!\perp\!\!\!\perp G,$$

   where $F \perp\!\!\!\perp G$ denotes that there exists a unitary operator U on $\mathbb{R}^d$ such that $F = G \circ U$ and $F \not\!\perp\!\!\!\perp G$ denotes that $F \neq G \circ U$ for any unitary operator U on $\mathbb{R}^d$.

2. *(Equality, up to scaling)*

$$H_0 \colon F \perp\!\!\!\perp G \circ c \quad \text{for some } c > 0 \quad \text{against} \quad H_A \colon F \not\!\perp\!\!\!\perp G \circ c \quad \text{for any } c > 0,$$

   where $Y \sim F \circ c$ if $cY \sim F$.

3. *(Equality, up to projection)*

$$H_0 \colon F \circ \pi^{-1} \perp\!\!\!\perp G \circ \pi^{-1} \quad \text{against} \quad H_A \colon F \circ \pi^{-1} \not\!\perp\!\!\!\perp G \circ \pi^{-1},$$

   where π is the projection $x \mapsto x/\|x\|$; hence $Y \sim F \circ \pi^{-1}$ if $\pi^{-1}(Y) \sim F$.

We note that the above null hypotheses are nested; $F \perp\!\!\!\perp G$ implies $F \perp\!\!\!\perp G \circ c$ for $c = 1$ while $F \perp\!\!\!\perp G \circ c$ for some $c > 0$ implies $F \circ \pi^{-1} \perp\!\!\!\perp G \circ \pi^{-1}$.

We shall address the above hypothesis testing problem by combining the framework of adjacency spectral embedding and the kernel-based hypothesis testing framework of Gretton et al. (2012). The testing procedure in Gretton et al. (2012) is based on the following notion of the maximum mean discrepancy between distributions. Let Ω be a compact metric space and $\kappa \colon \Omega \times \Omega \mapsto \mathbb{R}$ a continuous, symmetric, and positive definite kernel on Ω. Denote by $\mathcal{H}$ the reproducing kernel Hilbert space associated with κ. Now let F be a probability distribution on Ω. Under mild conditions on κ, the map $\mu[F]$ defined by

$$\mu[F] \coloneqq \int_\Omega \kappa(\omega, \cdot)\, \mathrm{d}F(\omega)$$

belongs to $\mathcal{H}$. Now, for given probability distributions F and G on Ω, the *maximum mean discrepancy* between F and G with respect to $\mathcal{H}$ is the measure

$$\mathrm{MMD}(F, G; \mathcal{H}) \coloneqq \|\mu[F] - \mu[G]\|_{\mathcal{H}}.$$

We now summarize some important properties of the maximum mean discrepancy from Gretton et al. (2012).

Theorem 43 *Let $\kappa: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a positive definite kernel and denote by $\mathcal{H}$ the reproducing kernel Hilbert space associated with κ. Let F and G be probability distributions on Ω; X and X' independent random variables with distribution F, Y and Y' independent random variables with distribution G, and X is independent of Y. Then*

$$\|\mu[F] - \mu[G]\|_{\mathcal{H}}^2 = \sup_{h \in \mathcal{H}: \|h\|_{\mathcal{H}} \leq 1} |\mathbb{E}_F[h] - \mathbb{E}_G[h]|^2$$
$$= \mathbb{E}[\kappa(X, X')] - 2\mathbb{E}[\kappa(X, Y)] + \mathbb{E}[\kappa(Y, Y')]. \tag{43}$$

Given $\mathbf{X} = \{X_i\}_{i=1}^n$ and $\mathbf{Y} = \{Y_k\}_{k=1}^m$ with $\{X_i\} \stackrel{i.i.d}{\sim} F$ and $\{Y_i\} \stackrel{i.i.d}{\sim} G$, the quantity $U_{n,m}(\mathbf{X}, \mathbf{Y})$ defined by

$$U_{n,m}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n(n-1)} \sum_{j \neq i} \kappa(X_i, X_j) - \frac{2}{mn} \sum_{i=1}^n \sum_{k=1}^m \kappa(X_i, Y_k)$$
$$+ \frac{1}{m(m-1)} \sum_{l \neq k} \kappa(Y_k, Y_l) \tag{44}$$

is an unbiased consistent estimate of $\|\mu[F] - \mu[G]\|_{\mathcal{H}}^2$. Denote by $\tilde{\kappa}$ the kernel

$$\tilde{\kappa}(x, y) = \kappa(x, y) - \mathbb{E}_z \kappa(x, z) - \mathbb{E}_{z'} \kappa(z', y) + \mathbb{E}_{z,z'} \kappa(z, z')$$

where the expectation is taken with respect to $z, z' \sim F$. Suppose that $\frac{m}{m+n} \to \rho \in (0,1)$ as $m, n \to \infty$. Then under the null hypothesis of $F = G$,

$$(m+n) U_{n,m}(\mathbf{X}, \mathbf{Y}) \xrightarrow{d} \frac{1}{\rho(1-\rho)} \sum_{l=1}^{\infty} \lambda_l (\chi_{1l}^2 - 1) \tag{45}$$

where $\{\chi_{1l}^2\}_{l=1}^{\infty}$ is a sequence of independent χ^2 random variables with one degree of freedom, and $\{\lambda_l\}$ are the eigenvalues of the integral operator $\mathcal{I}_{F,\tilde{\kappa}}: \mathcal{H} \mapsto \mathcal{H}$ defined as

$$I_{F,\tilde{\kappa}}(\phi)(x) = \int_{\Omega} \phi(y) \tilde{\kappa}(x, y) dF(y).$$

Finally, if κ is a universal or characteristic kernel (Sriperumbudur et al., 2011; Steinwart, 2001), then μ is an injective map: that is, $\mu[F] = \mu[G]$ if and only if $F = G$.

Remark 44 *A kernel $\kappa: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is universal if κ is a continuous function of both its arguments and if the reproducing kernel Hilbert space $\mathcal{H}$ induced by κ is dense in the space of continuous functions on $\mathcal{X}$ with respect to the supremum norm. Let $\mathcal{M}$ be a family of Borel probability measures on $\mathcal{X}$. A kernel κ is characteristic for $\mathcal{M}$ if the map $\mu \in \mathcal{M} \mapsto \int \kappa(\cdot, z) \mu(dz)$ is injective. If κ is universal, then κ is characteristic for any $\mathcal{M}$ (Sriperumbudur et al., 2011). As an example, let $\mathcal{X}$ be a finite dimensional Euclidean space and define, for any $q \in (0,2)$, $k_q(x, y) = \frac{1}{2}(\|x\|^q + \|y\|^q - \|x - y\|^q)$. The kernels k_q are then characteristic for the collection of probability distributions with finite second moments (Lyons, 2013; Sejdinovic et al., 2013). In addition, by Eq. (43), the maximum mean discrepancy with reproducing kernel k_q can be written as*

$$\mathrm{MMD}^2(F, Q; k_q) = 2\mathbb{E}\|X - Y\|^q - \mathbb{E}\|X - X'\|^q - \mathbb{E}\|Y - Y'\|^q.$$

where X, X' are independent with distribution F, Y, Y' are independent with distribution G, and X, Y are independent. This coincides with the notion of the energy distances of Székely and Rizzo (2013), or, when $q = 1$, a special case of the one-dimensional interpoint comparisons of Maa et al. (1996). Finally, we note that $(m + n)U_{n,m}(\mathbf{X}, \mathbf{Y})$ under the null hypothesis of $F = G$ in Theorem 43 depends on the $\{\lambda_l\}$ which, in turn, depend on the distribution F; thus the limiting distribution is not distribution-free. Moreover the eigenvalues $\{\lambda_l\}$ can, at best, be estimated; for finite n, they cannot be explicitly determined when F is unknown. In practice, generally the critical values are estimated through a bootstrap resampling or permutation test.

We focus on the nonparametric two-sample hypothesis test of $\mathbb{H}_0 \colon F \perp\!\!\!\perp G$ against $\mathbb{H}_A \colon F \not\perp\!\!\!\perp G$. For our purposes, we shall assume henceforth that κ is a twice continuously-differentiable radial kernel and that κ is also universal. To justify this assumption on our kernel, we point out that in Theorem 45 below, we show that the test statistic $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ based on the estimated latent positions converges to the corresponding statistic $U_{n,m}(\mathbf{X}, \mathbf{Y})$ for the true but unknown latent positions. Due to the non-identifiability of the random dot product graph under unitary transformation, *any* estimate of the latent positions is close, only up to an appropriate orthogonal transformation, to $\mathbf{X}$ and $\mathbf{Y}$. For a radial kernel κ, this implies the approximations $\kappa(\hat{X}_i, \hat{X}_j) \approx \kappa(X_i, X_j)$, $\kappa(\hat{Y}_k, \hat{Y}_l) \approx \kappa(Y_k, Y_l)$ and the convergence of $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ to $U_{n,m}(\mathbf{X}, \mathbf{Y})$. If κ is not a radial kernel, the above approximations might not hold and $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ need not converge to $U_{n,m}(\mathbf{X}, \mathbf{Y})$. The assumption that κ is twice continuously-differentiable is technical. Finally, the assumption that κ is universal allows the test procedure to be consistent against a large class of alternatives.

Theorem 45 *Let* $(\mathbf{X}, \mathbf{A}) \sim \mathrm{RDPG}(F)$ *and* $(\mathbf{Y}, \mathbf{B}) \sim \mathrm{RDPG}(G)$ *be independent random dot product graphs with latent position distributions F and G. Furthermore, suppose that both F and G satisfies the distinct eigenvalues condition in Assumption 3. Consider the hypothesis test*

$$H_0 \colon F \perp\!\!\!\perp G \quad against \quad H_A \colon F \not\perp\!\!\!\perp G.$$

Denote by $\hat{\mathbf{X}} = \{\hat{X}_1, \ldots, \hat{X}_n\}$ *and* $\hat{\mathbf{Y}} = \{\hat{Y}_1, \ldots, \hat{Y}_m\}$ *the adjacency spectral embedding of* $\mathbf{A}$ *and* $\mathbf{B}$, *respectively. Let κ be a twice continuously-differentiable radial kernel and* $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ *be defined as*

$$U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) = \frac{1}{n(n-1)} \sum_{j \neq i} \kappa(\hat{X}_i, \hat{X}_j) - \frac{2}{mn} \sum_{i=1}^{n} \sum_{k=1}^{m} \kappa(\hat{X}_i, \hat{Y}_k) + \frac{1}{m(m-1)} \sum_{l \neq k} \kappa(\hat{Y}_k, \hat{Y}_l).$$

Let $\mathbf{W}_1$ *and* $\mathbf{W}_2$ *be $d \times d$ orthogonal matrices in the eigendecomposition* $\mathbf{W}_1 \mathbf{S}_1 \mathbf{W}_1^\top = \mathbf{X}^\top \mathbf{X}$, $\mathbf{W}_2 \mathbf{S}_2 \mathbf{W}_2 = \mathbf{Y}^\top \mathbf{Y}$, *respectively. Suppose that $m, n \to \infty$ and $m/(m+n) \to \rho \in (0,1)$. Then under the null hypothesis of $F \perp\!\!\!\perp G$, the sequence of matrices* $\mathbf{W}_{n,m} = \mathbf{W}_2 \mathbf{W}_1^\top$ *satisfies*

$$(m+n)(U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_{n,m})) \xrightarrow{\text{a.s.}} 0. \tag{46}$$

Under the alternative hypothesis of $F \not\perp\!\!\!\perp G$, the sequence of matrices $\mathbf{W}_{n,m}$ *satisfies*

$$\frac{m+n}{\log^2(m+n)}(U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_{n,m})) \xrightarrow{\text{a.s.}} 0. \tag{47}$$

Eq.(46) and Eq.(47) state that the test statistic $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ using the *estimated* latent positions is almost identical to the statistic $U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_{n,m})$ using the true latent positions, under both the null and alternative hypothesis. If we assume that κ is a universal kernel, then $U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_{n,m})$ converges to 0 under the null and converges to a positive number under the alternative. The test statistic $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ therefore yields a test procedure that is consistent against any alternative, provided that both F and G satisfy Assumption 3, namely that the second moment matrices have d distinct eigenvalues.

We next consider the case of testing the hypothesis that the distributions F and G are equal up to scaling or equal up to projection. For the test of equality up to scaling,

$$H_0\colon F \perp\!\!\!\perp G \circ c \quad \text{for some } c > 0 \quad \text{against} \quad H_A\colon F \not\perp\!\!\!\perp G \circ c \quad \text{for any } c > 0,$$

where $Y \sim F \circ c$ if $cY \sim F$, we modified Theorem 45 by first scaling the adjacency spectral embeddings by the norm of the empirical means before computing the kernel test statistic. In particular, let

$$\hat{s}_X = n^{-1/2}\|\hat{\mathbf{X}}\|_F, \quad \hat{s}_Y = m^{-1/2}\|\hat{\mathbf{Y}}\|_F, \quad s_X = n^{-1/2}\|\mathbf{X}\|_F, \quad s_Y = m^{-1/2}\|\mathbf{Y}\|_F,$$

then the conclusions of Theorem 45 hold when we replace $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ and $U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_{n,m})$ with $U_{n,m}(\hat{\mathbf{X}}/\hat{s}_X, \hat{\mathbf{Y}}/\hat{s}_Y)$ and $U_{n,m}(\mathbf{X}/s_X, \mathbf{Y}\mathbf{W}_{n,m}/s_Y)$, respectively.

For the test of equality up to projection,

$$H_0\colon F \circ \pi^{-1} \perp\!\!\!\perp G \circ \pi^{-1} \quad \text{against} \quad H_A\colon F \circ \pi^{-1} \not\perp\!\!\!\perp G \circ \pi^{-1},$$

where π is the projection $x \mapsto x/\|x\|$ that maps x onto the unit sphere in $\mathbb{R}^d$, the conclusions of Theorem 45 hold when we replace $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ and $U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_{n,m})$ with $U_{n,m}(\pi(\hat{\mathbf{X}}), \pi(\hat{\mathbf{Y}}))$ and $U_{n,m}(\pi(\mathbf{X}), \pi(\mathbf{Y})\mathbf{W}_{n,m})$, respectively, provided that that 0 is not an atom of either F or G (and hence that $F(\{0\}) = G(\{0\}) = 0$). The assumption that 0 is not an atom is necessary, because otherwise the problem is possibly ill-posed: specifically, $\pi(0)$ is undefined. To contextualize the test of equality up to projection, consider the very specific case of the degree-corrected stochastic blockmodel of Karrer and Newman (2011). As mentioned earlier, the degree-corrected stochastic blockmodel can be regarded as a random dot product graph whose latent position X_v for an arbitrary vertex v is of the form $X_v = \theta_v \nu_v$ where ν_v is sampled from a mixture of point masses and θ_v (the degree-correction factor) is sampled from a distribution on $(0, 1]$. Thus, given two degree-corrected stochastic blockmodel graphs, equality up to projection tests whether the underlying mixtures of point masses (that is, the distribution of the ν_v) are the same modulo the distribution of the degree-correction factors θ_v.

6. Applications

We begin this section by first presenting an application of the two-sample semiparametric test procedure in Section 5.2, demonstrating how it can be applied to compare data from a collection of neural images.

Algorithm 1 Bootstrapping procedure for the test $\mathbb{H}_0 \colon \mathbf{X} =_W \mathbf{Y}$.

1: **procedure** BOOTSTRAP($\mathbf{X}, T, bs$)   ▷ Returns the p-value associated with T.
2:     $d \leftarrow \mathrm{ncol}(\mathbf{X}); \quad \mathcal{S}_X \leftarrow \varnothing$ ▷ Set d to be the number of columns of $\mathbf{X}$.
3:     **for** $b \leftarrow 1{:}bs$ **do**
4:        $\mathbf{A}_b \leftarrow \mathrm{RDPG}(\hat{\mathbf{X}}); \quad \mathbf{B}_b \leftarrow \mathrm{RDPG}(\hat{\mathbf{X}})$
5:        $\hat{\mathbf{X}}_b \leftarrow \mathrm{ASE}(\mathbf{A}_b, d); \quad \hat{\mathbf{Y}}_b \leftarrow \mathrm{ASE}(\mathbf{B}_b, d)$
6:        $T_b \leftarrow \min_{\mathbf{W}} \|\hat{\mathbf{X}}_b - \hat{\mathbf{Y}}_b \mathbf{W}\|_F; \qquad \mathcal{S}_X \leftarrow \mathcal{S}_X \cup T_b$
7:     **end for**
8:     **return** $p \leftarrow (|\{s \in \mathcal{S}_X \colon s \geq T\}| + 0.5)/bs$ ▷ Continuity correction.
9: **end procedure**
10:
11: $\hat{\mathbf{X}} \leftarrow \mathrm{ASE}(\mathbf{A}, d); \hat{\mathbf{Y}} \leftarrow \mathrm{ASE}(\mathbf{B}, d)$ ▷ The embedding dimension d is assumed given.
12: $T \leftarrow \min_{\mathbf{W}} \|\hat{\mathbf{X}} - \hat{\mathbf{Y}} \mathbf{W}\|_F$
13: $p_X \leftarrow \mathrm{Bootstrap}(\hat{\mathbf{X}}, T, bs)$ ▷ The number of bootstrap samples bs is assumed given.
14: $p_Y \leftarrow \mathrm{Bootstrap}(\hat{\mathbf{Y}}, T, bs)$
15: $p = \max\{p_X, p_Y\}$ ▷ Returns the maximum of the two p-values.

6.1 Semiparametric testing for brain scan data

We consider neural imaging graphs obtained from the test-retest diffusion MRI and magnetization-prepared rapid acquisition gradient echo (MPRAGE) data of Landman et al. (2011). The raw data consist of 42 images: namely, one pair of neural images from each of 21 subjects. These images are generated, in part, for the purpose of evaluating scan-rescan reproducibility of the MPRAGE image protocol. Table 5 from Landman et al. (2011) indicates that the variability of MPRAGE is quite small; specifically, the cortical gray matter, cortical white matter, ventricular cerebrospinal fluid, thalamus, putamen, caudate, cerebellar gray matter, cerebellar white matter, and brainstem were identified with mean volume-wise reproducibility of 3.5%, with the largest variability being that of the ventricular cerebrospinal fluid at 11%.

We use the MIGRAINE pipeline of Roncal et al. (2013) to convert these scans into spatially-aligned graphs, in which each vertex corrresponds to a particular voxel in a reference coordinate system to which the image is registered. We first consider a collection of small graphs on seventy vertices that are generated from an atlas of seventy brain regions and the fibers connecting them. Given these graphs, we proceed to investigate the similarities and dissimilarities between the scans. We first embed each graph into $\mathbb{R}^4$. We then test the hypothesis of equality up to rotation between the graphs. Since Theorem 34 is a large-sample result, the rejection region specified therein might be excessively conservative for the graphs on $n = 70$ vertices in our current context. We remedy this issue by using the rejection region and p-values reported by the parametric bootstrapping procedure presented in Algorithm 1.

The pairwise comparisons between the 42 graphs are presented in Figure 7. Figure 7 indicates that, in general, the test procedure fails to reject the null hypothesis when the two graphs are for the same subject. This is consistent with the reproducibility finding of Landman et al. (2011). Furthermore, this outcome is also intuitively plausible; in addition

to failing to reject when two scans are from the same subject, we also frequently *do* reject the null hypothesis when the two graphs are from scans of different subjects. Note that our analysis is purely exploratory; as such, we do not grapple with issues of multiple comparisons here.
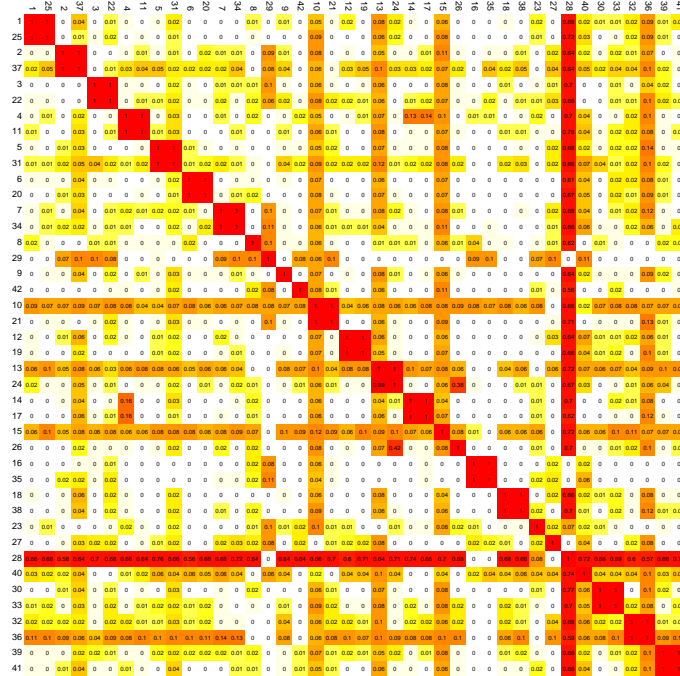

Figure 7: Matrix of p-values (uncorrected) for testing the hypothesis $\mathbb{H}_0 : \mathbf{X} =_W \mathbf{Y}$ for the $42 \times 41/2$ pairs of graphs generated from the KKI test-retest dataset Landman et al. (2011). The labels had been arranged so that the pair $(2i-1, 2i)$ correspond to scans from the same subject. The p-values are color coded to vary in intensity from white (p-value of 0) to dark red (p-value of 1). Figure duplicated from Tang et al. (2017a).

Finally, we note that similar results also hold when we consider the large graphs generated from these test-retest data through the MIGRAINE pipeline. In particular, for each magnetic resonance scan, the MIGRAINE pipeline can generates graphs with upto 10^7 vertices and 10^{10} edges with the vertices of all the graphs aligned. Bootstrapping the test statistics for these large graphs present some practical difficulties; one procedure proposed in Tang et al. (2017a) is based on bootstrapping disjoint induced subgraphs of the original graphs using the bootstrapping procedure in Algorithm 1 and combining the resulting p-values using Fisher's combined probability tests (Mosteller and Fisher, 1948).

6.2 Community detection and classification in hierarchical models

In disciplines as diverse as social network analysis and neuroscience, many large graphs are believed to be composed of loosely connected communities and/or smaller graph primitives, whose structure is more amenable to analysis. We emphasize that while the problem of community detection is very well-studied and there are an abundance of community detection algorithms, these algorithms have focused mostly on uncovering the subgraphs. Recently, however, the characterization and further *classification* of these subgraphs into stochastically similar motifs has emerged as an important area of ongoing research The nonparametric two-sample hypothesis testing procedure in Section 5.4 can be used in conjunction with spectral community detection algorithms to yield a robust, scalable, integrated methodology for *community detection* and *community comparison* in graphs (Lyzinski et al., 2017).

The notion of *hierarchical stochastic block model*—namely, a graph consisting of densely connected subcommunities which are themselves stochastic block models, with this structure iterated repeatedly—is precisely formulated in Lyzinski et al. (2017). In that work, a novel angle-based clustering method is introduced, and this clustering method allows us to isolate appropriate subgraphs. We emphasize that the angle-based clustering procedure in Lyzinski et al. (2017) is designed to identify a particular affinity structure within our hierarchical block model graph. Figure 8 illustrates how an angle-based clustering may differ from a k-means clustering.
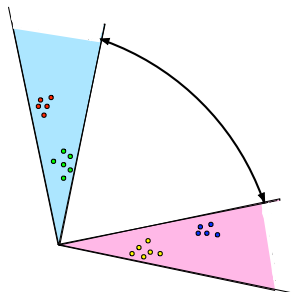
Figure 8: Subgraphs vs. angle-based clustering: Note that if the fraction of points in the pink cone is sufficiently large, K-means clustering (with $K = 2$) will not cluster the vertices in the hierarchical SBM affinity model of Lyzinski et al. (2017) into the appropriate subgraphs. Figure duplicated from Lyzinski et al. (2017).

Our overall community detection algorithm is summarized in Algorithm 2. As an illustrative example of this methodology, we present an analysis of the communities in the Friendster social network. The Friendster social network contains roughly 60 million users and 2 billion connections/edges. In addition, there are roughly 1 million communities at the local scale. Because we expect the social interactions in these communities to inform the function of the different communities, we expect to observe distributional repetition among the graphs associated with these communities.

Algorithm 2 Detecting hierarchical structure for graphs

Input: Adjacency matrix $A \in \{0,1\}^{n \times n}$ for a latent position random graph.

Output: Subgraphs and characterization of their dissimilarity

while Cluster size exceeds threshold **do**

Step 1: Compute the adjacency spectral embedding $\hat{\mathbf{X}}$ of A into $\mathbb{R}^D$;

Step 2: Cluster $\hat{\mathbf{X}}$ to obtain subgraphs $\hat{H}_1, \cdots, \hat{H}_R$ using a novel angle-based clustering procedure given in Lyzinski et al. (2017).

Step 3: For each $i \in [R]$, compute the adjacency spectral embedding for each subgraph $\hat{H}_i$ into $\mathbb{R}^d$, obtaining $\hat{\mathbf{X}}_{\hat{H}_i}$;

Step 4: Compute $\widehat{S} := [U_{\hat{n}_r, \hat{n}_s}(\hat{\mathbf{X}}_{\hat{H}_r}, \hat{\mathbf{X}}_{\hat{H}_s})]$, where U is the test statistic in Theorem 45, producing a pairwise dissimilarity matrix on induced subgraphs;

Step 5: Cluster induced subgraphs into motifs using the dissimilarities given in $\widehat{S}$; for example, use a hierarchical clustering algorithm to cluster the rows of $\widehat{S}$ or the matrix of associated p-values.

Step 6: Recurse on a representative subgraph from each motif (for example, the largest subgraph), embedding into $\mathbb{R}^d$ in Step 1 (not $\mathbb{R}^D$);

end while

Implementing Algorithm 2 on the very large Friendster graph presents several computational challenges and model selection quagmires. To overcome the computational challenge in scalability, we use the specialized SSD-based graph processing engine `FlashGraph` (Zheng et al., 2015), which is designed to analyze graphs with billions of nodes. With `FlashGraph`, we adjacency spectral embed the Friendster adjacency matrix into $\mathbb{R}^{14}$—where $\widehat{D} = 14$ is chosen using universal singular value thresholding on the partial SCREE plot (Chatterjee, 2015). We next cluster the embedded points into $\widehat{R} = 15$ large-scale/coarse-grained clusters ranging in size from 10^6 to 15.6 million vertices (note that to alleviate sparsity concerns, we projected the embedding onto the sphere before clustering); After re-embedding the induced subgraphs associated with these 15 clusters, we use a linear time estimate of the test statistic U to compute $\widehat{S}$, the matrix of estimated pairwise dissimilarities among the subgraphs. See Figure 9 for a heat map depicting $\widehat{S} \in \mathbb{R}^{15 \times 15}$. In the heat map, the similarity of the communities is represented on the spectrum between white and red, with white representing highly similar communities and red representing highly dissimilar communities. From the figure, we can see clear repetition in the subgraph distributions; for example, we see a repeated motif including subgraphs $\{\hat{H}_5, \hat{H}_4, \hat{H}_3, \hat{H}_2\}$ and a clear motif including subgraphs $\{\hat{H}_{10}, \hat{H}_{12}, \hat{H}_9\}$.

Formalizing the motif detection step, we employ hierarchical clustering to cluster $\widehat{S}$ into motifs; see Figure 9 for the corresponding hierarchical clustering dendrogram, which suggests that our algorithm does in fact uncover repeated motif structure at the coarse-grained level in the Friendster graph. While it may be difficult to draw meaningful inference from repeated motifs at the scale of hundreds of thousands to millions of vertices, if these motifs are capturing a common HSBM structure within the subgraphs in the motif, then we can employ our algorithm recursively on each motif to tease out further hierarchical structure.
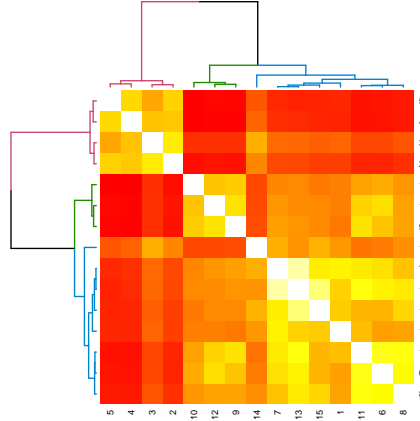
Figure 9: Heat map depiction of the level one Friendster estimated dissimilarity matrix $\widehat{S} \in \mathbb{R}^{15 \times 15}$. In the heat map, the similarity of the communities is represented on the spectrum between white and red, with white representing highly similar communities and red representing highly dissimilar communities. In addition, we cluster $\widehat{S}$ using hierarchical clustering and display the associated hierarchical clustering dendrogram. Figure duplicated from Lyzinski et al. (2017).

Exploring this further, we consider three subgraphs $\{\hat{H}_2, \hat{H}_8, \hat{H}_{15}\}$, two of which are in the same motif (8 and 15) and both differing significantly from subgraph 2 according to $\widehat{S}$. We embed these subgraphs into $\mathbb{R}^{26}$ (26 were chosen once again using the universal singular value thresholding of Chatterjee (2015)) perform a Procrustes alignment of the vertex sets of the three subgraphs, cluster each into 4 clusters (4 chosen to optimize silhouette width in k-means clustering), and estimate both the block connection probability matrices,

$$
\hat{P}_2 = \begin{bmatrix} 0.000045 & 0.00080 & 0.00056 & 0.00047 \\ 0.00080 & 0.025 & 0.0096 & 0.0072 \\ 0.00057 & 0.0096 & 0.012 & 0.0067 \\ 0.00047 & 0.0072 & 0.0067 & 0.023 \end{bmatrix},
$$

$$
\hat{P}_8 = \begin{bmatrix} 0.0000022 & 0.000031 & 0.000071 & 0.000087 \\ 0.000031 & 0.0097 & 0.00046 & 0.00020 \\ 0.000071 & 0.00046 & 0.0072 & 0.0030 \\ 0.000087 & 0.00020 & 0.003 & 0.016 \end{bmatrix},
$$

$$
\hat{P}_{15} = \begin{bmatrix} 0.0000055 & 0.00011 & 0.000081 & 0.000074 \\ 0.00011 & 0.014 & 0.0016 & 0.00031 \\ 0.000081 & 0.0016 & 0.0065 & 0.0022 \\ 0.000074 & 0.00031 & 0.0022 & 0.019 \end{bmatrix},
$$

and the block membership probabilities $\hat{\pi}_2$, $\hat{\pi}_8$, $\hat{\pi}_{15}$, for each of the three graphs. We calculate

$$\|\hat{P}_2 - \hat{P}_8\|_F = 0.033; \quad \|\hat{\pi}_2 - \hat{\pi}_8\| = 0.043;$$
$$\|\hat{P}_8 - \hat{P}_{15}\|_F = 0.0058; \quad \|\hat{\pi}_8 - \hat{\pi}_{15}\| = 0.0010;$$
$$\|\hat{P}_2 - \hat{P}_{15}\|_F = 0.027; \quad \|\hat{\pi}_2 - \hat{\pi}_{15}\| = 0.043;$$

which suggests that the repeated structure our algorithm uncovers is *SBM substructure*, thus ensuring that we can proceed to apply our algorithm recursively to the subsequent motifs.

As a final point, we recursively apply Algorithm 2 to the subgraph $\hat{H}_{11}$. We first embed the graph into $\mathbb{R}^{26}$ (again, with 26 chosen via universal singular value thresholding). We then cluster the vertices into $\widehat{R} = 13$ large-scale clusters of sizes ranging from 500K to 2.7M vertices. We then use a linear time estimate of the test statistic T to compute $\widehat{S}$ (see Figure 10), and note that there appear to be clear repeated motifs (for example, subgraphs 8 and 12) among the $\widehat{H}$'s. We run hierarchical clustering to cluster the 13 subgraphs, and note that the associated dendrogram—as shown in Figure 10—shows that our algorithm again uncovered some repeated level-2 structure in the Friendster network. We can, of course, recursively apply our algorithm still further to tease out the motif structure at increasingly fine-grained scale.

Ideally, when recursively running Algorithm 2, we would like to simultaneously embed and cluster all subgraphs in the motif. In addition to potentially reducing embedding variance, being able to efficiently simultaneously embed all the subgraphs in a motif could greatly increase algorithmic scalability in large networks with a very large number of communities at local-scale. In order to do this, we need to understand the nature of the repeated structure within the motifs. This repeated structure can inform an estimation of a motif average (an averaging of the subgraphs within the motif), which can then be embedded into an appropriate Euclidean space in lieu of embedding all of the subgraphs in the motif separately. However, this averaging presents several novel challenges, as these subgraphs may be of very different orders and may be errorfully obtained, which could lead to compounded errors in the averaging step.

6.3 Structure discovery in the *Drosophila* connectome

In this subsection, we address a cutting-edge application of our techniques to neuroscience: structure discovery in the larval Drosophila connectome, comprehensively described in Priebe et al. (2017), and from which significant portions are reproduced here, with permission. This is a first-of-its-kind exploratory data analysis of a newly-available wiring diagram, and although the connectome graph we analyze is directed, weighted, and also of unknown embedding dimension, our statistical techniques can nonetheless be adapted to this setting.

Specifically, we introduce the *latent structure model* (LSM) for network modeling and inference. The LSM is a generalization of the stochastic block model (SBM) in that the latent
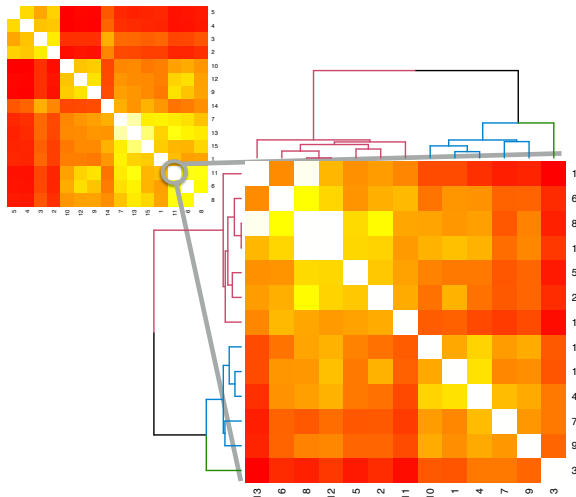
Figure 10: Heat map depiction of the level two Friendster estimated dissimilarity matrix $\widehat{S} \in \mathbb{R}^{13 \times 13}$ of $\widehat{H}_{11}$. In the heat map, the similarity of the communities is represented on the spectrum between white and red, with white representing highly similar communities and red representing highly dissimilar communities. In addition, we cluster $\widehat{S}$ using hierarchical clustering and display the associated hierarchical clustering dendrogram. Figure duplicated from Lyzinski et al. (2017).

positions are allowed to lie on a lower-dimensional curve, and this model is amenable to semiparametric Gaussian mixture modeling (GMM) applied to the adjacency spectral embedding (ASE). The resulting *connectome code* derived via semiparametric GMM composed with ASE, which we denote, in shorthand, by $GMM \circ ASE$, captures latent connectome structure and elucidates biologically relevant neuronal properties.

HHMI Janelia has recently reconstructed the complete wiring diagram of the higher order parallel fiber system for associative learning in the larval *Drosophila* brain, the mushroom body (MB). Memories are thought to be stored as functional and structural changes in connections between neurons, but the complete circuit architecture of a higher-order learning center involved in memory formation or storage has not been known in any organism—until now, that is. Our MB connectome was obtained via serial section transmission electron microscopy of an entire larval *Drosophila* nervous system (Ohyama et al., 2015; Schneider-Mizell et al., 2016). This connectome contains the entirety of MB intrinsic neurons, called Kenyon cells, and all of their pre- and post-synaptic partners (Eichler et al., 2017).

We consider the right hemisphere MB. The connectome consists of four distinct types of neurons – Kenyon Cells (KC), Input Neurons (MBIN), Output Neurons (MBON), Projection Neurons (PN) – with directed connectivity illustrated in Figure 11. There are $n = 213$ neurons[2], with $n_{KC} = 100$, $n_{MBIN} = 21$, $n_{MBON} = 29$, and $n_{PN} = 63$. Figure 12 displays the

---

2. There are 13 isolates, all are KC; removing these isolates makes the (directed) graph one (weakly, but not strongly) connected component with 213 vertices and 7536 directed edges.

observed MB connectome as an adjacency matrix. Note that, in accordance with Figure 11, Figure 12 shows data (edges) in only eight of the 16 blocks.
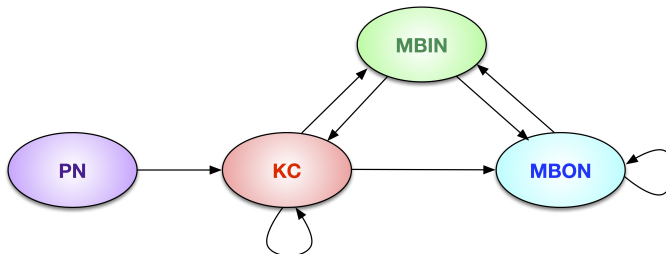

Figure 11: Illustration of the larval *Drosophila* mushroom body connectome as a directed graph on four neuron types. Figure duplicated from Priebe et al. (2017).
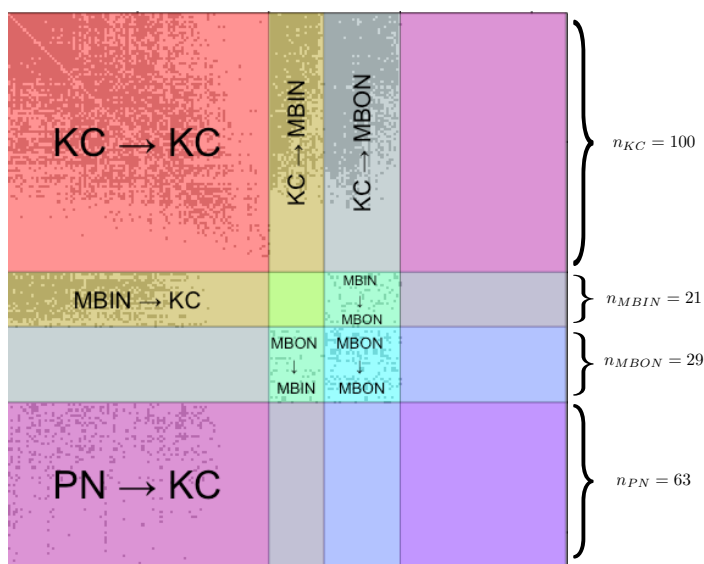

Figure 12: Observed data for our MB connectome as a directed adjacency matrix on four neuron types with 213 vertices ($n_{KC} = 100$, $n_{MBIN} = 21$, $n_{MBON} = 29$, and $n_{PN} = 63$) and 7536 directed edges. (This data matrix is available at <http://www.cis.jhu.edu/~parky/MBstructure.html>.) Figure duplicated from Priebe et al. (2017).

Due to its undeniable four-neuron-type connectivity structure, we might think of our MB connectome, to first order, as an observation from a ($K = 4$)-block directed stochastic block model (SBM) (Wang and Wong, 1987) on n vertices. This model is parameterized by (i) a block membership probability vector $\rho = [\rho_1, \cdots, \rho_K]$ such that $\rho_k \geq 0$ for all k and $\sum_k \rho_k = 1$ and (ii) a $K \times K$ block connectivity probability matrix B with entries $B_{k_1, k_2} \in [0, 1]$ governing the probability of directed edges from vertices in block k_1 to vertices in block k_2. For this

57

model of our MB connectome we have

$$B = \begin{bmatrix} B_{11} & B_{12} & B_{13} & 0 \\ B_{21} & 0 & B_{23} & 0 \\ 0 & B_{32} & B_{33} & 0 \\ B_{41} & 0 & 0 & 0 \end{bmatrix}$$

where the 0 in the B_{31} entry, for example, indicates that there are no directed connections from any MBON neuron to any KC neuron (as seen in Figures 11 and 12).

Theoretical results and methodological advances suggest that Gaussian mixture modeling (see, for example, Fraley and Raftery, 2002) composed with adjacency spectral embedding, denoted $GMM \circ ASE$, can be instructive in analysis of the (directed) SBM.

Since this graph is directed, we adapt our embedding technique just slightly. Given $d \geq 1$, the adjacency spectral embedding (ASE) of a *directed* graph on n vertices employs the singular value decomposition to represent the $n \times n$ adjacency matrix via $\mathbf{A} = \left[\mathbf{U} \mid \mathbf{U}^\perp \right] \left[\mathbf{S} \oplus \mathbf{S}^\perp \right] \left[\mathbf{V} \mid \mathbf{V}^\perp \right]^\top$ where $\mathbf{S}$ is the $d \times d$ diagonal matrix of the d largest singular values and $\mathbf{U}$ and $\mathbf{V}$ are the matrix of corresponding left and right singular vectors, and we embed the graph as n points in $\mathbb{R}^{2d}$ via the concatenation

$$\hat{\mathbf{X}} = \left[\mathbf{U}\mathbf{S}^{1/2} \mid \mathbf{V}\mathbf{S}^{1/2} \right] \in \mathbb{R}^{n \times 2d}.$$

(The scaled left-singular vectors $\mathbf{U}\mathbf{S}^{1/2}$ are interpreted as the "out-vector" representation of the digraph, modeling vertices' propensity to originate directed edges; similarly, $\mathbf{V}\mathbf{S}^{1/2}$ are interpreted as the "in-vectors".) Gaussian mixture modeling (GMM) then fits a K-component $2d$-dimensional Gaussian mixture model to the points $\hat{\mathbf{X}}_1, \cdots, \hat{\mathbf{X}}_n$ given by the rows of $\hat{\mathbf{X}}$. If the graph is a stochastic block model, then as we have discussed previously in Section 4.2, clustering the rows of the adjacency spectral embedding via Gaussian mixture modeling gives us consistent estimates for the latent positions.

Applying this procedure to our MB connectome yields the clustered embedding depicted via the pairs plot presented in Figure 13, with the associated cluster confusion matrix with respect to true neuron types presented in Table 2. The clusters are clearly coincident with the four true neuron types. (For ease of illustration, Figure 13 presents just the Out1 vs. Out2 subspace.)

There are two model selection problems inherent in spectral clustering in general, and in obtaining our clustered embedding (Figure 13) in particular: choice of embedding dimension ($\hat{d}$), and choice of mixture complexity ($\hat{K}$). A ubiquitous and principled method for choosing the number of dimensions in eigendecompositions and singular value decompositions is to examine the so-called *scree plot* and look for "elbows" or "knees" defining the cut-off between the top signal dimensions and the noise dimensions. Identifying a "best" method is, in general, impossible, as the bias-variance tradeoff demonstrates that, for small n, subsequent inference may be optimized by choosing a dimension *smaller than* the true signal dimension; see Section 3 of Jain et al. (2000) for a clear and concise illustration of this phenomenon. There are a plethora of variations for automating this singular value thresholding (SVT); Section 2.8 of Jackson (2004) provides a comprehensive discussion in
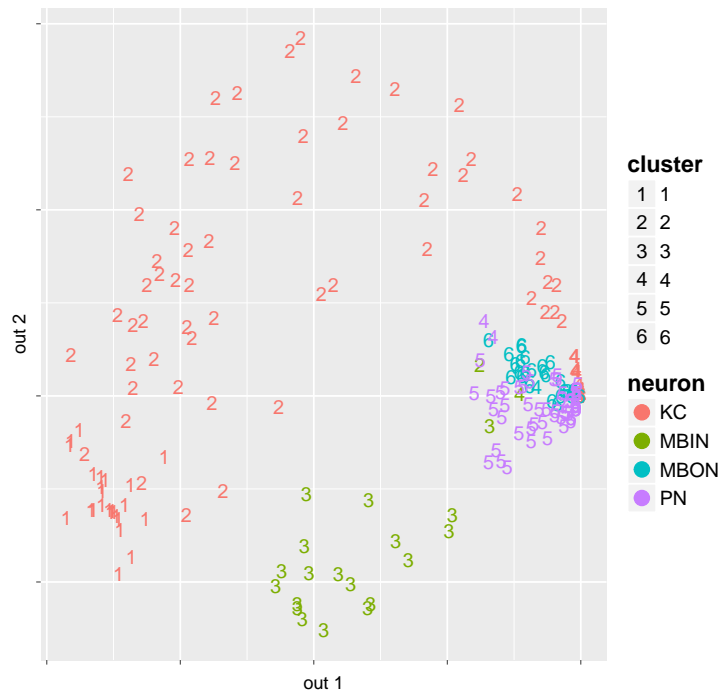
Figure 13: Plot for the clustered embedding of our MB connectome in the Out1 vs. Out2 dimensions. For ease of illustration, we present embedding results in this two-dimensional subspace. Recall that this is a two-dimensional visualization of six-dimensional structure. Figure duplicated from Priebe et al. (2017).

the context of principal components, and Chatterjee (2015) provides a theoretically-justified (but perhaps practically suspect, for small n) universal SVT. Using the profile likelihood SVT method of Zhu and Ghodsi (2006) yields a cut-off at three singular values, as depicted in Figure 14. Because this is a directed graph, we have both left- and right-singular vectors for each vertex; thus the SVT choice of three singular values results in $\hat{d} = 6$.

Similarly, a ubiquitous and principled method for choosing the number of clusters in, for example, Gaussian mixture models, is to maximize a fitness criterion penalized by model complexity. Common approaches include Akaike Information Criterion (AIC) (Akaike, 1974), Bayesian Information Criterion (BIC) (Schwarz, 1978), and Minimum Description Length (MDL) (Rissanen, 1978), to name a few. Again, identifying a "best" method is, in general, impossible, as the bias-variance tradeoff demonstrates that, for small n, inference performance may be optimized by choosing a number of clusters *smaller than* the true cluster complexity. The MCLUST algorithm of Fraley and Raftery (2002), as implemented in R, and its associated BIC applied to our MB connectome embedded via ASE into $\mathbb{R}^{\hat{d}=6}$, is maximized at six clusters, as depicted in Figure 15, and hence $\hat{K} = 6$.
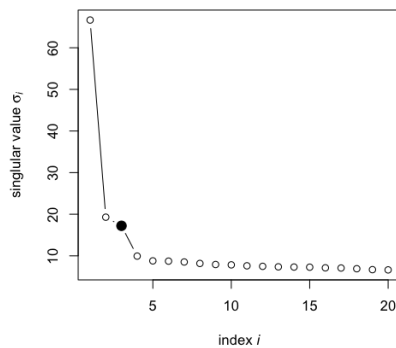

Figure 14: Model Selection: embedding dimension $\hat{d} = 6$ – the top 3 singular values and their associated left- and right-singular vectors – is chosen by SVT. Figure duplicated from Priebe et al. (2017).

	1	2	3	4	5	6
KC	**25**	**57**	0	**16**	2	0
MBIN	0	1	**19**	1	0	0
MBON	0	0	0	1	0	**28**
PN	0	0	0	2	**61**	0

Table 2: $GMM \circ ASE$ for our MB connectome yields $\hat{K} = 6$ clusters. The clusters are clearly coherent with but not perfectly aligned with the four true neuron types, as presented in this confusion matrix. Table duplicated from Priebe et al. (2017).
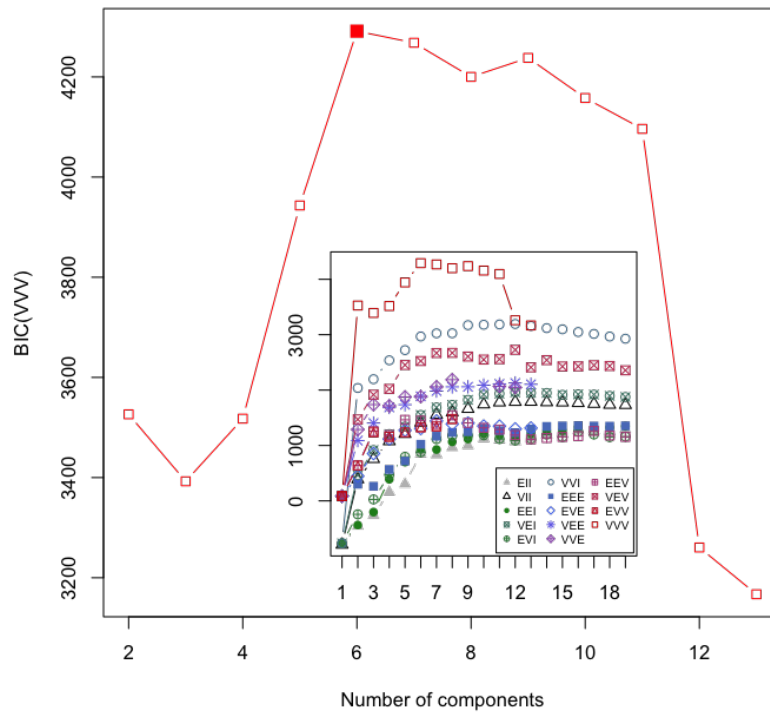
Figure 15: Model Selection: mixture complexity $\hat{K} = 6$ is chosen by BIC. (The inset shows that the main curve – BIC for dimensions 2 through 13 for MCLUST's most general covariance model, in red – dominates all other dimensions and all other models.) Figure duplicated from Priebe et al. (2017).

While BIC chooses $\hat{K} = 6$ clusters, it is natural to ask whether the distribution of KC across multiple clusters is an artifact of insufficiently parsimonious model selection. However, choosing four or five clusters not only (substantially) decreases BIC, but in fact leaves KC distributed across multiple clusters while splitting and/or merging other neuron types. In the direction of less parsimony, Figure 15 suggests that any choice from 7 to 11 clusters is competitive, in terms of BIC, with the maximizer $\hat{K} = 6$. Moreover, any of these choices only slightly decreases BIC, while leaving PN, MBIN, and MBON clustered (mostly) singularly and (mostly) purely and distributing KC across more clusters. Tables 3, 4, and 5 show cluster confusion matrices for other choices of K.

	1	2	3	4
KC	26	56	16	2
MBIN	0	20	1	0
MBON	0	28	1	0
PN	0	0	16	47

Table 3: Cluster confusion matrix for $GMM \circ ASE$ with 4 clusters. Choosing four or five clusters not only (substantially) decreases BIC (compared to $\hat{K} = 6$), but in fact leaves KC distributed across multiple clusters while splitting and/or merging other neuron types.

	1	2	3	4	5
KC	26	56	16	2	0
MBIN	0	20	1	0	0
MBON	0	0	1	0	28
PN	0	0	16	47	0

Table 4: Cluster confusion matrix for $GMM \circ ASE$ with 5 clusters. Choosing four or five clusters not only (substantially) decreases BIC (compared to $\hat{K} = 6$), but in fact leaves KC distributed across multiple clusters while splitting and/or merging other neuron types.

	1	2	3	4	5	6	7
KC	25	42	15	0	16	2	0
MBIN	0	0	1	19	1	0	0
MBON	0	0	0	0	1	0	28
PN	0	0	0	0	2	61	0

Table 5: Cluster confusion matrix for $GMM \circ ASE$ with 7 clusters. Any choice from 7 to 11 clusters only slightly decreases BIC (compared to $\hat{K} = 6$), while leaving PN, MBIN, and MBON clustered (mostly) singularly and (mostly) purely and distributing KC across more clusters. All tables reproduced from Priebe et al. (2017).

We see that our spectral clustering of the MB connectome via $GMM \circ ASE$, with principled model selection for choosing embedding dimension and mixture complexity, yields meaningful results: a single Gaussian cluster for each of MBIN, MBON, and PN, and multiple clusters for KC. That is, we have one substantial revision to Figure 11's illustration of the larval *Drosophila* mushroom body connectome as a directed graph on four neuron types:
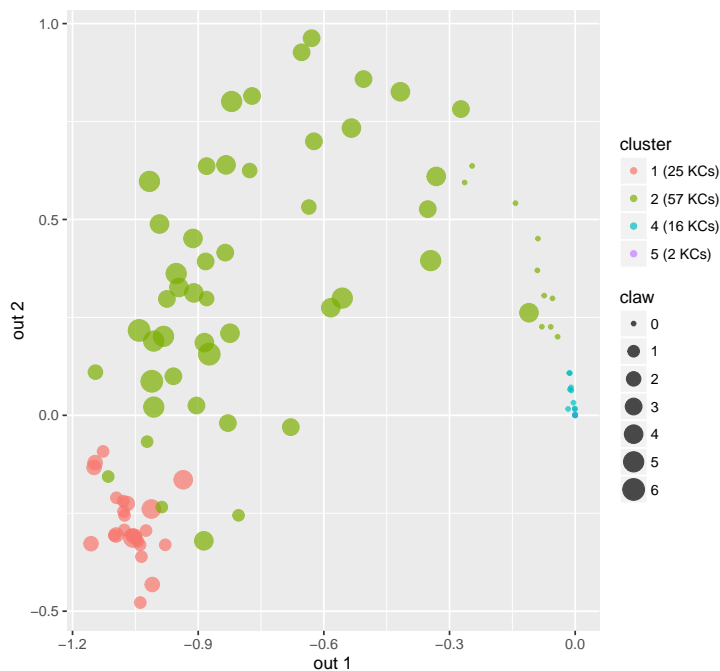
Figure 16: The multiple clusters for the KC neurons are capturing neuron age. Depicted are the first two dimensions for the KC neuron out-vectors, with color representing $\hat{K} = 6$ cluster membership – recall from Table 2 that the $n_{KC} = 100$ KCs are distributed across multiple clusters, with 25 neurons in cluster #1, 57 in #2, 0 in #3, 16 in #4, 2 in #5, and 0 in #6. The size of the dots represent the number of claws associated with the neurons. We see from the scatter plot that the embedded KC neurons arc from oldest (one-claw, lower left, cluster 1, in red), up and younger (more claws) through cluster 2 in green, and back down to youngest (zero-claw, clusters 4 and 5). See also Table 6. Recall that this is a two-dimensional visualization of six-dimensional structure. Figure duplicated from Priebe et al. (2017).

cluster	1	2	3	4	5	6
#KCs	25	57	0	16	2	0
claw: 1 (oldest)	15	4	—	0	0	—
claw: 2	7	4	—	0	0	—
claw: 3	0	15	—	0	0	—
claw: 4	3	13	—	0	0	—
claw: 5	0	8	—	0	0	—
claw: 6	0	3	—	0	0	—
claw: 0 (youngest)	0	10	—	16	2	—

Table 6: The multiple clusters for the KC neurons are capturing neuron age via the number of claws associated with the neuron. We see from the $\hat{K} = 6$ clustering table, for the $n_{KC} = 100$ KC neurons, that cluster 1 captures predominantly older neurons, cluster 2 captures both old and young neurons, and clusters 4 and 5 capture only the youngest neurons. See also Figure 16. Table reproduced from Priebe et al. (2017).

significant *substructure* associated with the KC neurons. Indeed, this hints at the possibility of a continuous, rather than discrete, structure for the KC. The paper Eichler et al. (2017) describes so-called "claws" associated with each KC neuron, and posits that KCs with only 1 claw are the oldest, followed in decreasing age by multi-claw KCs (from 2 to 6 claws), with finally the youngest KCs being those with 0 claws.

Figure 16 and Table 6 use this additional neuronal information to show that the multiple clusters for the KC neurons are capturing neuron age – and in a seemingly coherent geometry. Indeed, precisely because the clusters for the KC neurons are capturing neuron age – a continuous vertex attribute – again, in a seemingly coherent geometry, we define a "latent structure model" (LSM), a generalization of the SBM, together with a principled semiparametric spectral clustering methodology $SemiparGMM \circ ASE$ associated thereto. Specifically, we fit a continuous curve to (the KC subset of) the data in latent space and show that traversal of this curve corresponds monotonically to neuron age. To make this precise, we begin with a directed stochastic block model:

Definition 46 (Directed Stochastic Blockmodel (SBM)) *Let $d_{\text{out}} = d_{\text{in}}$, with $d = d_{\text{out}} + d_{\text{in}}$. We say that an n vertex graph $(\mathbf{A}, \mathbf{X}) \sim \text{RDPG}(F)$ is a directed stochastic blockmodel (SBM) with K blocks if the distribution F is a mixture of K point masses,*

$$dF = \sum_{k=1}^{K} \rho_k \delta_{x_k},$$

with block membership probability vector $\vec{\rho}$ in the unit $(K-1)$-simplex and distinct latent positions given by $\boldsymbol{\nu} = [\nu_1, \nu_2, \ldots, \nu_K]^\top \in \mathbb{R}^{K \times d}$. The first d_{out} entries of each latent position ν_k are the out-vectors, denoted $\xi_k \in \mathbb{R}^{d_{\text{out}}}$, and the remaining d_{in} elements are the in-vectors ζ_k. We write $G \sim SBM(n, \vec{\rho}, \boldsymbol{\xi}\boldsymbol{\zeta}^\top)$, and we refer to $\boldsymbol{\xi}\boldsymbol{\zeta}^\top \in \mathbb{R}^{K \times K}$ as the block connectivity probability matrix for the model.

We model the MB connectome as a four-component latent structure model (LSM), where LSM denotes the "generalized SBM" where each "block" may be generalized from point mass latent position distribution to latent position distribution with support on some curve (with the "block" curves disconnected, as (of course) are SBM's point masses). So LSM does have block structure, albeit not as simple as an SBM; and LSM will exhibit clustering, albeit just as transparently as an SBMs. As such, it is similar to other generalizations of SBMs, including the degree-corrected and hierarchical variants.

Definition 47 (Directed Latent Structure Model (LSM)) *Let $d_{\text{out}} = d_{\text{in}}$, and let F be a distribution on a set $\mathcal{X} = \mathcal{Y} \times \mathcal{Z} \subset \mathbb{R}^{d_{\text{out}}} \times \mathbb{R}^{d_{\text{in}}}$ such that $\langle y, z \rangle \in [0, 1]$ for all $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$. We say that an n vertex graph $(\mathbf{A}, \mathbf{X}) \sim \text{RDPG}(F)$ is a directed latent structure model (LSM) with K "structure components" if the support of distribution F is a mixture of K (disjoint) curves,*

$$dF = \sum_{k=1}^{K} \rho_k dF_k(x),$$

with block membership probability vector $\vec{\rho}$ in the unit $(K-1)$-simplex and F_k supported on $\mathcal{C}_k$ and $\mathcal{C}_1, \cdots, \mathcal{C}_K$ disjoint. We write $G \sim LSM(n, \vec{\rho}, (F_1, \cdots, F_K))$.

We now investigate our MB connectome as an LSM with latent positions $X_i \overset{\text{i.i.d.}}{\sim} F$ where F is no longer a mixture of four point masses with one point mass per neuron type but instead $\text{supp}(F)$ is three points and a continuous curve $\mathcal{C}_{KC}$.

Motivated by approximate normality of the adjacency spectral embedding of an RDPG, we consider estimating F via a semiparametric Gaussian mixture model for the $\hat{\mathbf{X}}_i$'s. Let H be a probability measure on a parameter space $\Theta \subset \mathbb{R}^d \times S_{d \times d}$, where $S_{d \times d}$ is the space of d-dimensional covariance matrices, and let $\{\varphi(\cdot; \theta) : \theta \in \Theta\}$ be a family of normal densities. Then the function given by

$$\alpha(\cdot; H) = \int_\Theta \varphi(\cdot; \theta) dH(\theta)$$

is a semiparametric GMM. $H \in \mathcal{M}$ is referred to as the mixing distribution of the mixture, where $\mathcal{M}$ is the class of all probability measures on Θ. If H consists of a finite number of atoms, then $\alpha(\cdot; H)$ is a finite normal mixture model with means, variances and proportions determined by the locations and weights of the point masses, and in Lindsay (1983), the author provides theory for maximum likelihood estimation (MLE) in this context.

Thus (ignoring covariances for presentation simplicity, so that $\theta \in \mathbb{R}^d$ is the component mean vector) we see that the central limit theorem suggests that we estimate the probability density function of the embedded MB connectome $\hat{\mathbf{X}}_1, \cdots, \hat{\mathbf{X}}_{n=213}$, under the LSM assumption, as the semiparametric GMM $\alpha(\cdot; H)$ with $\Theta = \mathbb{R}^6$ and where $H = F$ is supported by three points and a continuous curve $\mathcal{C}_{KC}$. Note that in the general case, where Θ includes both means and covariance matrices, we have $H = H_{F,n}$. The central limit theorem for the adjacency spectral embedding provides a large-sample approximation for $H_{F,n}$, and provides a mean-covariance constraint so that if we knew the latent position distribution F we would have no extra degrees of freedom (though perhaps a more challenging MLE optimization problem). As it is, we do our fitting in the general case, with simplifying constraints on the covariance structure associated with $\mathcal{C}_{KC}$.

Our MLE (continuing to ignore covariances for presentation simplicity) is given by

$$d\hat{H}(\theta) = \sum_{k=1}^{3} \widehat{\rho}_k I\{\theta = \widehat{\theta}_k\} + \left(1 - \sum_{k=1}^{3} \widehat{\rho}_k\right) \widehat{\rho}_{KC}(\theta) I\{\theta \in \widehat{\mathcal{C}}_{KC}\}$$

where $\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3$ are given by the means of the $GMM \circ ASE$ Gaussian mixture components for MBIN, MBON, and PN, and $\widehat{\mathcal{C}}_{KC} \subset \mathbb{R}^d$ is a one-dimensional curve. Figure 17 displays the MLE results from an EM optimization for the curve $\widehat{\mathcal{C}}_{KC}$ constrained to be quadratic, as detailed in the Appendix. (Model testing for $\mathcal{C}_{KC}$ in $\mathbb{R}^6$ does yield quadratic: testing the null hypothesis of linear against the alternative of quadratic yields clear rejection ($p < 0.001$), while there is insufficient evidence to favor H_A cubic over H_0 quadratic ($p \approx 0.1$).) That is, (continuing to ignore covariances for presentation simplicity) our structure discovery via $SemiparGMM \circ ASE$ yields an $\mathbb{R}^6$ latent position estimate for the MB connectome – a *connectome code* for the larval *Drosophila* mushroom body – as a semiparametric Gaussian mixture of three point masses and a continuous parameterized curve $\widehat{\mathcal{C}}_{KC}$; the three Gaussians correspond to three of the four neuron types, and the curve corresponds to the fourth neuron type (KC) with the parameterization capturing neuron age (see Figure 18). We note that in Eichler et al. (2017) , the authors suggest distance-to-neuropile δ_i – the
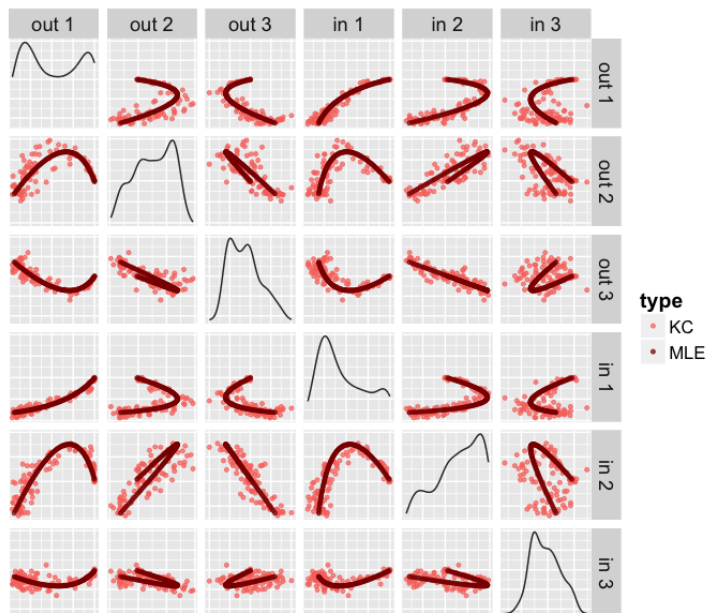
Figure 17: Semiparametric MLE $\widehat{\mathcal{C}}_{KC}$ for the KC latent-space curve in $\mathbb{R}^6$. Figure duplicated from Priebe et al. (2017).

distance to the MB neuropile from the bundle entry point of each KC neuron i – as a proxy for neuron age, and analyzes this distance in terms of number of claws for neuron i (see Figure 19). We now demonstrate that the correlation of this distance with the KC neurons' projection onto the parameterized curve $\widehat{\mathcal{C}}_{KC}$ is highly significant – this semiparametric spectral model captures neuroscientifically important structure in the connectome. To wit, we project each KC neuron's embedding onto our parameterized $\widehat{\mathcal{C}}_{KC}$ and study the relationship between the projection's position on the curve, t_i, and the neuron's age through the distance proxy δ_i (see Figures 20 and 21). We find significant correlation of δ_i with t_i – Spearman's $s = -0.271$, Kendall's $\tau = -0.205$, Pearson's $\rho = -0.304$, with $p < 0.01$ in each case – demonstrating that our semiparametric spectral modeling captures biologically relevant neuronal properties.

In summary, motivated by the results of a spectral clustering investigation of the recently-reconstructed synapse-level larval *Drosophila* mushroom body structural connectome, which demonstrate conclusively that modeling the Kenyon Cells (KC) demands additional latent space structure, we have developed semiparametric spectral modeling. Exploratory data analysis suggests that the MB connectome can be productively approximated by a four-component latent structure model (LSM), and the resulting MB connectome code derived via $SemiparGMM \circ ASE$ captures biologically relevant neuronal properties. Data and code for all our analyses are available at `http://www.cis.jhu.edu/~parky/MBstructure.html`.

Of course, the true connectome code is more elaborate, and cannot be completely encompassed by any simple latent position model – such a model precludes the propensity for transitivity, for example – but our semiparametric spectral modeling provides another step
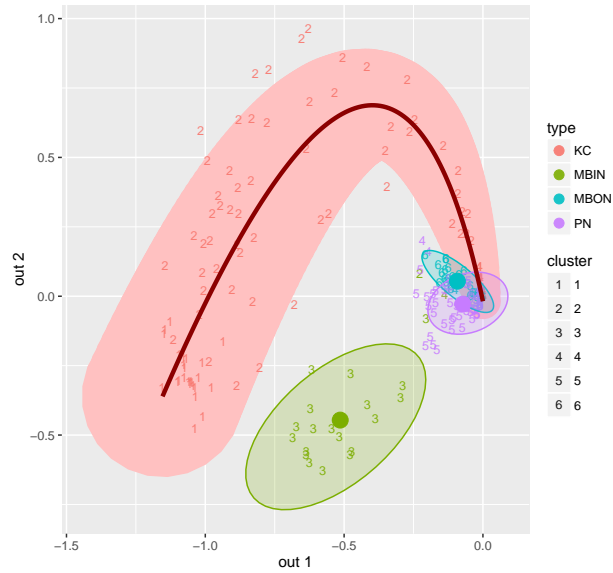
Figure 18: Semiparametric spectral latent space estimate of our MB connectome as three Gaussians and a KC curve: colors distinguish the four neuron types and numbers distinguish the original $\hat{K} = 6$ clusters. Recall that this is a two-dimensional visualization of six-dimensional structure. Figure duplicated from Priebe et al. (2017).
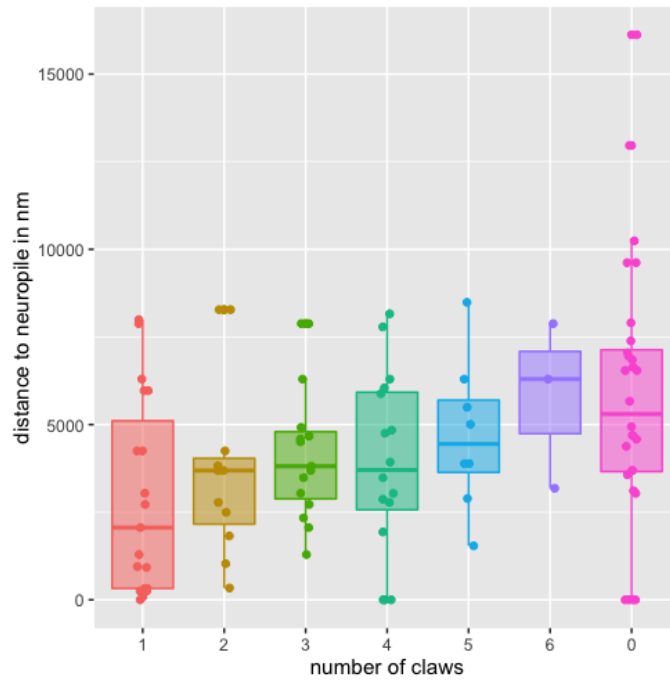

Figure 19: Relationship between number of claws and distance δ_i (a proxy for age) for the KC neurons, from Eichler et al. (2017). Figure duplicated from Priebe et al. (2017).

67
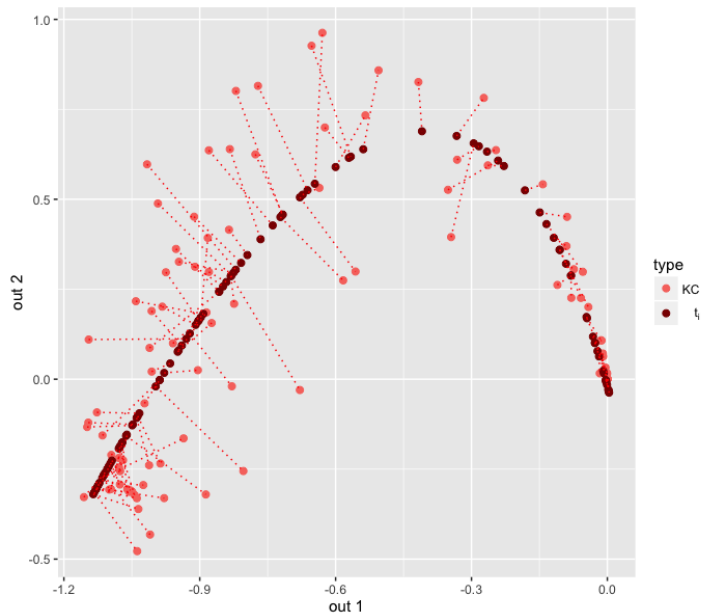
Figure 20: Projection of KC neurons onto the quadratic curve $\widehat{\mathcal{C}}_{KC}$, yielding projection point t_i for each neuron. Recall that this is a two-dimensional visualization of six-dimensional structure. Figure duplicated from Priebe et al. (2017).
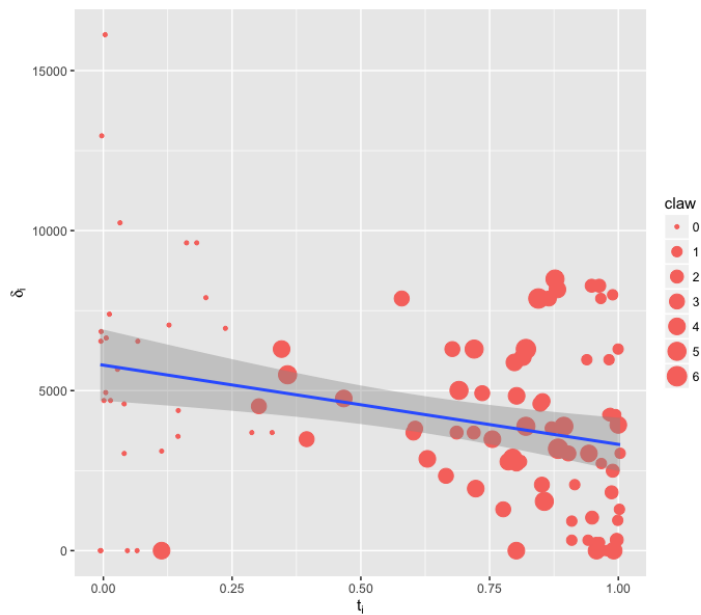
Figure 21: The correlation between the projection points t_i on the quadratic curve $\widehat{\mathcal{C}}_{KC}$ and distance δ_i (a proxy for age) for the KC neurons is highly significant, demonstrating that our semiparametric spectral modeling captures biologically relevant neuronal properties. Figure duplicated from Priebe et al. (2017).

along the path. In terms of a (partial) ladder of biological scales – for example, *C. elegans*, *Drosophila*, zebrafish, mouse, primate, and humans – this works moves us off the first rung for analysis of a complete neurons-as-vertices and synapses-as-edges connectome.

7. Conclusion: complexities, open questions, and future work

Our paradigm for statistical inference on random graphs is anchored by the familiar pillars of classical Euclidean inference. We exhibit estimates for graph parameters that satisfy (uniform) consistency and asymptotic normality, and we demonstrate how these estimates can be exploited in a bevy of subsequent inference tasks: community detection in heterogeneous networks, multi-sample graph hypothesis testing, and exploratory data analysis in connectomics. The random dot product graph model in which we ground our techniques has both linear algebraic transparency and wide applicability, since a hefty class of independent-edge graphs is well-approximated by RDPGs. The lynchpins of our approach are spectral decompositions of adjacency and Laplacian matrices, but many of our results and proof techniques can be applied to more general random matrices. In recent work, for example, we examine eigenvalue concentration for certain classes of random matrices (Cape et al., 2017a), and accurate estimation of covariance matrices (Cape et al., 2017b). As such, our methodology is a robust edifice from which to explore questions in graph inference, data analysis, and random matrix theory.

The results we summarize here are but the tip of the iceberg, though, and there persist myriad open problems in spectral graph inference—for random dot product graphs in particular, of course, but also for random graphs writ large. In this section, we outline some current questions of interest and describe future directions for research.

Our consistency results for the adjacency spectral embedding depend on knowing the correct embedding dimension. In real data, this optimal embedding dimension is typically not only not known, but, since the RDPG model is only an approximation to any true model, may well depend on the desired subsequent inference task. As we have pointed out in Section 6.3, multiple methods exist for estimating embedding dimension, and the universal singular value thresholding of Chatterjee (2015) and other thresholding methods (such as Fishkind et al., 2013) are theoretically justified in the large-n limit. For finite n, however, model selection is considerably trickier. Underestimation of the embedding dimension can markedly—and provably—bias subsequent inference. While we do not address it here, we remark that asymptotic results can be shown for the adjacency spectral embedding of a d-dimensional RDPG when the chosen embedding dimension is $d' < d$. On the other hand, if the embedding dimension is overestimated, no real signal is lost; therefore, most embedding methods continue to perform well, albeit with some loss of efficiency due to increased variance. Precisely quantifying this loss of efficiency is very much an open question and is related to analogous classical questions in random matrix theory (see Tao and Vu, 2012).

In our RDPG model, an important assumption is that the $\mathbf{P}$ matrix be positive semidefinite. While this limits the model, as we mentioned in the introduction, the *generalized random dot product graph model* (gRDPG) of Rubin-Delanchy et al. (2017) is not similarly restricted, but nevertheless shares many important useful properties of the RDPG. Considering a matrix

of latent positions $\mathbf{X} \in \mathbb{R}^{p+q}$, one can set $\mathbf{P} = \mathbf{X}\mathbf{I}_{p,q}\mathbf{X}^\top$ where $\mathbf{I}_{p,q}$ is the diagonal matrix of size $(p+q) \times (p+q)$ with p diagonal entries being 1 and q diagonal entries being -1. Under this generalization, any $\mathbf{P}$ can be obtained, provided that $p+q$ is appropriately chosen. This then implies that any latent position graph model, even a non-positive-semidefinite one, can be approximated arbitrarily closely by this generalized RDPG. We are currently investigation extensions of our results on consistency, distributional limits, and testing for estimates of gRDPGs parameters, including, for instance, suitable reformulations of the two-sample hypothesis tests for, say, mixed membership stochastic blockmodels.

We also wish to adapt our procedures to weighted graphs. For simple weighting, such as Poisson-distributed weights, our existing methodology applies quite naturally to the weighted adjacency matrix. More intricate questions arise when the weights are contaminated or their distribution is heavily skewed. In such cases, one can "pass to ranks"; that, is replace the nonzero weight by its normalized rank among all the edge weights. This mitigates skew and works well in many examples, but a deeper theoretical analysis of this procedure, as well as other approaches to weighted graph inference, remain open.

To address graphs with self-edges, note that the random dot product graph model does not preclude such edges. Indeed, since $\mathbf{P}$ is not constrained to be hollow, the adjacency matrix $\mathbf{A}$ thus generated need not be hollow, either. However, the generative mechanism for self-edges in a graph may be different from the mechanism for edges between two different vertices. One approach to addressing this is to set the diagonal entries of $\mathbf{A}$ to zero and then *augment* the diagonal artificially with imputed values. In fact, even when there are no self loops, such procedures can improve finite-sample inferential performance. In Marchette et al. (2011), it is demonstrated that augmenting the diagonal via $\mathbf{A}_{ii} = d_i/(n+1)$, where d_i is the degree of vertex i, can improve inference in certain cases. Similarly, in Scheinerman and Tucker (2010), we find an iterative procedure to find a diagonal augmentation consistent with the low-rank structure of $\mathbf{P}$. It is still unclear exactly what augmentation of the diagonal, if any, might be optimal for each particular inference task.

In the case when the vertices or edges are corrupted by occlusion or noise, the work of Priebe et al. (2015) and Levin and Lyzinski (2017) consider approaches to vertex classification and graph recovery, demonstrating that spectral embeddings are robust to certain sources of graph error. Violations of the independent edge assumption, though, can lead to more significant hurdles, both theoretical practical, since it is a requirement for a number of the concentration inequalities on which we depend.

For joint graph inference and testing, open problems abound. We mention, for instance, the analysis of the omnibus embedding when the m graphs are correlated, or when some are corrupted; a closer examination of the impact of the Procrustes alignment on power; the development of an analogue to a Tukey test for determining which graphs differ when we test for the equality of distribution for more than two graphs; the comparative efficiency of the omnibus embedding relative to other spectral decompositions; and a quantification of the trade-off for subsequent inference between a large number of independent graphs and large graph size (Tang et al., 2016).

In sum, the random dot product graph is a compact, manageable, and applicable model. The Euclidean nature of the adjacency and Laplacian spectral embedding for a random dot product graph allows us to approach statistical inference in this setting from a familiar Euclidean perspective. Both the adjacency spectral embedding and the Laplacian spectral embedding can be profitably leveraged for latent position estimation and single- and multi-sample graph hypothesis testing. Moreover, our distributional results for these spectral embeddings provide reassuring classical analogues of asymptotic normality for estimators, and in current ongoing work, we consider how to compare asymptotic relative efficiency of different estimators for graph parameters. While spectral methods may not always be optimal for a given task, they are often feasible and can provide a way to accurately initialize more complex procedures. Moreover, these Euclidean representations of graph data render possible the application of techniques for analysis of Euclidean data—clustering, classification, and density estimation, for instance—to graphs. As we have outlined above, while many important theoretical and practical challenges remain, spectral embeddings for random dot product graphs constitute an important piece of the greater puzzle of random graph inference.

Acknowledgments

Appendix A. Appendix

In the appendix, we provide details for the proofs of our results on consistency and asymptotic normality of the adjacency spectral embedding, as well as an outline of the proof of the central limit theorem for the Laplacian spectral embedding. Our notation remains as stated in Section 4. We begin with a detailed proof of our result on the consistency, in the $2 \to \infty$ norm, of the ASE for latent position recovery in RDPGs.

A.1 Proof of Theorem 26

Let us recall **Theorem 26**: Let $\mathbf{A}_n \sim \mathrm{RDPG}(\mathbf{X}_n)$ for $n \geq 1$ be a sequence of random dot product graphs where the $\mathbf{X}_n$ is assumed to be of rank d for all n sufficiently large. Denote by $\hat{\mathbf{X}}_n$ the adjacency spectral embedding of $\mathbf{A}_n$ and let $(\hat{\mathbf{X}}_n)_i$ and $(\mathbf{X}_n)_i$ be the i-th row of $\hat{\mathbf{X}}_n$ and $\mathbf{X}_n$, respectively. Let E_n be the event that there exists an orthogonal transformation $\mathbf{W}_n \in \mathbb{R}^{d \times d}$ such that

$$\max_i \|(\hat{\mathbf{X}})_i - \mathbf{W}_n(\mathbf{X}_n)_i\| \leq \frac{Cd^{1/2}\log^2 n}{\delta^{1/2}(\mathbf{P}_n)}$$

where $C > 0$ is some fixed constant and $\mathbf{P}_n = \mathbf{X}_n \mathbf{X}_n^\top$. Then E_n occurs asymptotically almost surely; that is, $\Pr(E_n) \to 1$ as $n \to \infty$.

The proof of Theorem 26 will follow from a succession of supporting results. We note that Theorem 50, which deals with the accuracy of spectral embedding estimates in Frobenius norm, may be of independent interest. We begin with the following simple but essential proposition, in which we show that $\mathbf{U_P^\top U_A}$ is close to an orthogonal transformation. For ease of exposition, in the remainder of this subsection we shall suppress the subscript index n from our matrices $\mathbf{X}_n$, $\mathbf{A}_n$ and $\hat{\mathbf{X}}_n$.

Proposition 48 *Let* $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X})$ *and let* $\mathbf{W}_1 \boldsymbol{\Sigma} \mathbf{W}_2^\top$ *be the singular value decomposition of* $\mathbf{U_P^\top U_A}$. *Then with high probability,*

$$\|\mathbf{U_P^\top U_A} - \mathbf{W}_1 \mathbf{W}_2^\top\| = O(\delta^{-1}(\mathbf{P})).$$

Proof Let $\sigma_1, \sigma_2, \ldots, \sigma_d$ denote the singular values of $\mathbf{U_P^\top U_A}$ (the diagonal entries of $\boldsymbol{\Sigma}$). Then $\sigma_i = \cos(\theta_i)$ where the θ_i are the principal angles between the subspaces spanned by $\mathbf{U_A}$ and $\mathbf{U_P}$. Furthermore, by the Davis-Kahan Theorem,

$$\|\mathbf{U_A U_A^\top} - \mathbf{U_P U_P^\top}\| = \max_i |\sin(\theta_i)| \leq \frac{C\sqrt{d}\|\mathbf{A} - \mathbf{P}\|}{\lambda_d(\mathbf{P})}$$

for sufficiently large n. Recall here $\lambda_d(\mathbf{P})$ denotes the d-th largest eigenvalue of $\mathbf{P}$. The spectral norm bound for $\mathbf{A} - \mathbf{P}$ from Theorem 21 along with the assumption that $\lambda_d(\mathbf{P})/\delta(\mathbf{P}) \geq c_0$ for some constant c_0 in Assumption 1 yield

$$\|\mathbf{U_A U_A^\top} - \mathbf{U_P U_P^\top}\| \leq \frac{C\sqrt{d}}{\delta^{1/2}(\mathbf{P})}.$$

We thus have

$$\|\mathbf{U_P^\top U_A} - \mathbf{W}_1 \mathbf{W}_2^\top\| = \|\boldsymbol{\Sigma} - I\| = \max_i |1 - \sigma_i| \leq \max_i (1 - \sigma_i^2)$$
$$= \max_i \sin^2(\theta_i) = \|\mathbf{U_A U_A^T} - \mathbf{U_P U_P^\top}\|^2 = O(\delta^{-1}(\mathbf{P}))$$

as desired. ∎

Denote by $\mathbf{W}^*$ the orthogonal matrix $\mathbf{W}_1 \mathbf{W}_2^\top$ as defined in the above proposition. We now establish the following key lemma. The lemma allows us to exchange the order of the orthogonal transformation $\mathbf{W}^*$ and the diagonal scaling transformation $\mathbf{S_A}$ or $\mathbf{S_P}$.

Lemma 49 *Let* $(\mathbf{A}, \mathbf{X}) \sim \mathrm{RDPG}(F)$ *with sparsity factor* ρ_n. *Then asymptotically almost surely,*

$$\|\mathbf{W}^* \mathbf{S_A} - \mathbf{S_P} \mathbf{W}^*\|_F = O(\log n) \tag{48}$$

and

$$\|\mathbf{W}^* \mathbf{S_A}^{1/2} - \mathbf{S_P}^{1/2} \mathbf{W}^*\|_F = O((\log n)\, \delta^{-1/2}(\mathbf{P})) \tag{49}$$

72

Proof Let $\mathbf{R} = \mathbf{U_A} - \mathbf{U_P}\mathbf{U_P^\top}\mathbf{U_A}$. We note that $\mathbf{R}$ is the residual after projecting $\mathbf{U_A}$ orthogonally onto the column space of $\mathbf{U_P}$, and thus

$$\|\mathbf{U_A} - \mathbf{U_P}\mathbf{U_P^\top}\mathbf{U_A}\|_F \le \min_{\mathbf{W}} \|\mathbf{U_A} - \mathbf{U_P}\mathbf{W}\|_F$$

where the minimization is over all orthogonal matrices $\mathbf{W}$. The variant of the Davis-Kahan Theorem given in Eq. (9) then implies

$$\|\mathbf{U_A} - \mathbf{U_P}\mathbf{U_P^\top}\mathbf{U_A}\|_F \le O(\delta^{-1/2}(\mathbf{P})).$$

We next derive that

$$
\begin{aligned}
\mathbf{W}^*\mathbf{S_A} &= (\mathbf{W}^* - \mathbf{U_P^\top}\mathbf{U_A})\mathbf{S_A} + \mathbf{U_P^\top}\mathbf{U_A}\mathbf{S_A} = (\mathbf{W}^* - \mathbf{U_P^\top}\mathbf{U_A})\mathbf{S_A} + \mathbf{U_P^\top}\mathbf{A}\mathbf{U_A} \\
&= (\mathbf{W}^* - \mathbf{U_P^\top}\mathbf{U_A})\mathbf{S_A} + \mathbf{U_P^\top}(\mathbf{A} - \mathbf{P})\mathbf{U_A} + \mathbf{U_P^\top}\mathbf{P}\mathbf{U_A} \\
&= (\mathbf{W}^* - \mathbf{U_P^\top}\mathbf{U_A})\mathbf{S_A} + \mathbf{U_P^\top}(\mathbf{A} - \mathbf{P})\mathbf{R} + \mathbf{U_P^\top}(\mathbf{A} - \mathbf{P})\mathbf{U_P}\mathbf{U_P^\top}\mathbf{U_A} + \mathbf{U_P^\top}\mathbf{P}\mathbf{U_A} \\
&= (\mathbf{W}^* - \mathbf{U_P^\top}\mathbf{U_A})\mathbf{S_A} + \mathbf{U_P^\top}(\mathbf{A} - \mathbf{P})\mathbf{R} + \mathbf{U_P^\top}(\mathbf{A} - \mathbf{P})\mathbf{U_P}\mathbf{U_P^\top}\mathbf{U_A} + \mathbf{S_P}\mathbf{U_P^\top}\mathbf{U_A}
\end{aligned}
$$

Writing $\mathbf{S_P}\mathbf{U_P^\top}\mathbf{U_A} = \mathbf{S_P}(\mathbf{U_P^\top}\mathbf{U_A} - \mathbf{W}^*) + \mathbf{S_P}\mathbf{W}^*$ and rearranging terms, we obtain

$$
\begin{aligned}
\|\mathbf{W}^*\mathbf{S_A} - \mathbf{S_P}\mathbf{W}^*\|_F &\le \|\mathbf{W}^* - \mathbf{U_P^\top}\mathbf{U_A}\|_F(\|\mathbf{S_A}\| + \|\mathbf{S_P}\|) + \|\mathbf{U_P^\top}(\mathbf{A} - \mathbf{P})\mathbf{R}\|_F \\
&\quad + \|\mathbf{U_P^\top}(\mathbf{A} - \mathbf{P})\mathbf{U_P}\mathbf{U_P^\top}\mathbf{U_A}\|_F \\
&\le O(1) + O(1) + \|\mathbf{U_P^\top}(\mathbf{A} - \mathbf{P})\mathbf{U_P}\|_F\|\mathbf{U_P^\top}\mathbf{U_A}\|
\end{aligned}
$$

asymptotically almost surely. Now, $\|\mathbf{U_P^\top}\mathbf{U_A}\| \le 1$. Hence we can focus on the term $\mathbf{U_P^\top}(\mathbf{A} - \mathbf{P})\mathbf{U_P}$, which is a $d \times d$ matrix whose ij-th entry is of the form

$$
\begin{aligned}
(\mathbf{U_P^\top}(\mathbf{A} - \mathbf{P})\mathbf{U_P})_{ij} &= \sum_{k=1}^{n}\sum_{l=1}^{n}(\mathbf{A}_{kl} - \mathbf{P}_{kl})\mathbf{U}_{ki}\mathbf{U}_{lj} \\
&= 2\sum_{k,l:k<l}(\mathbf{A}_{kl} - \mathbf{P}_{kl})\mathbf{U}_{ki}\mathbf{U}_{lj} - \sum_{k}\mathbf{P}_{kk}\mathbf{U}_{ki}\mathbf{U}_{kj}
\end{aligned}
$$

where $\mathbf{U}_{\cdot i}$ and $\mathbf{U}_{\cdot j}$ are the i-th and j-th columns of $\mathbf{U_P}$. Thus, conditioned on $\mathbf{P}$, $(\mathbf{U_P^\top}(\mathbf{A} - \mathbf{P})\mathbf{U_P})_{ij}$ is a sum of independent mean 0 random variables and a term of order $O(1)$. Now, by Hoeffding's inequality,

$$\Pr\left[\left|\sum_{k,l:k<l}2(\mathbf{A}_{kl} - \mathbf{P}_{kl})\mathbf{U}_{ki}u_{lj}\right| \ge t\right] \le 2\exp\left(\frac{-2t^2}{\sum_{k,l:k<l}(2\mathbf{U}_{ki}\mathbf{U}_{lj})^2}\right) \le 2\exp(-t^2).$$

Therefore, each entry of $\mathbf{U_P^\top}(\mathbf{A} - \mathbf{P})\mathbf{U_P}$ is of order $O(\log n)$ with high probability, and as a consequence, since $\mathbf{U_P^\top}(\mathbf{A} - \mathbf{P})\mathbf{U_P}$ is a $d \times d$ matrix,

$$\|\mathbf{U_P^\top}(\mathbf{A} - \mathbf{P})\mathbf{U_P}\|_F = O(\log n) \tag{50}$$

with high probability. We establish that

$$\|\mathbf{W}^*\mathbf{S_A}^{1/2} - \mathbf{S_P}^{1/2}\mathbf{W}^*\|_F = O((\log n)\lambda_d^{-1/2}(\mathbf{P}))$$

73

by noting that the ij-th entry of $\mathbf{W}^*\mathbf{S}_{\mathbf{A}}^{1/2} - \mathbf{S}_{\mathbf{P}}^{1/2}\mathbf{W}^*$ can be written as

$$\mathbf{W}_{ij}^*(\lambda_i^{1/2}(\mathbf{A}) - \lambda_j^{1/2}(\mathbf{P})) = \mathbf{W}_{ij}^*\frac{\lambda_i(\mathbf{A}) - \lambda_j(\mathbf{P})}{\lambda_i^{1/2}(\mathbf{A}) + \lambda_j^{1/2}(\mathbf{P})},$$

and the desired bound follows from the above after bounding $\lambda_i(\mathbf{A})$, either by Weyl's inequality and Theorem 21, or, alternatively, by a Kato-Temple inequality from Cape et al. (2017a). ∎

We next present Theorem 50, which allows us to write the Frobenius norm difference of the adjacency spectral embedding $\hat{\mathbf{X}}$ and the true latent position $\mathbf{X}$ in terms of the Frobenius norm difference of $\mathbf{A} - \mathbf{P}$ and smaller order terms.

Theorem 50 *Let $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X})$. Then there exists an orthogonal matrix $\mathbf{W}$ such that, with high probability,*

$$\|\hat{\mathbf{X}} - \mathbf{X}\mathbf{W}\|_F = \|(\mathbf{A} - \mathbf{P})\mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{-1/2}\|_F + O((\log n)\delta^{-1/2}(\mathbf{P}))$$

Proof Let

$$\mathbf{R}_1 = \mathbf{U}_{\mathbf{P}}\mathbf{U}_{\mathbf{P}}^{\top}\mathbf{U}_{\mathbf{A}} - \mathbf{U}_{\mathbf{P}}\mathbf{W}^*$$
$$\mathbf{R}_2 = (\mathbf{W}^*\mathbf{S}_{\mathbf{A}}^{1/2} - \mathbf{S}_{\mathbf{P}}^{1/2}\mathbf{W}^*).$$

We deduce that

$$\begin{aligned}
\hat{\mathbf{X}} - \mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{1/2}\mathbf{W}^* &= \mathbf{U}_{\mathbf{A}}\mathbf{S}_{\mathbf{A}}^{1/2} - \mathbf{U}_{\mathbf{P}}W^*\mathbf{S}_{\mathbf{A}}^{1/2} + \mathbf{U}_{\mathbf{P}}(\mathbf{W}^*\mathbf{S}_{\mathbf{A}}^{1/2} - \mathbf{S}_{\mathbf{P}}^{1/2}W^*) \\
&= (\mathbf{U}_{\mathbf{A}} - \mathbf{U}_{\mathbf{P}}\mathbf{U}_{\mathbf{P}}^{\top}\mathbf{U}_{\mathbf{A}})\mathbf{S}_{\mathbf{A}}^{1/2} + \mathbf{R}_1\mathbf{S}_{\mathbf{A}}^{1/2} + \mathbf{U}_{\mathbf{P}}\mathbf{R}_2 \\
&= \mathbf{U}_{\mathbf{A}}\mathbf{S}_{\mathbf{A}}^{1/2} - \mathbf{U}_{\mathbf{P}}\mathbf{U}_{\mathbf{P}}^{\top}\mathbf{U}_{\mathbf{A}}\mathbf{S}_{\mathbf{A}}^{1/2} + \mathbf{R}_1\mathbf{S}_{\mathbf{A}}^{1/2} + \mathbf{U}_{\mathbf{P}}\mathbf{R}_2
\end{aligned}$$

Observe that $\mathbf{U}_{\mathbf{P}}\mathbf{U}_{\mathbf{P}}^{\top}\mathbf{P} = \mathbf{P}$ and $\mathbf{U}_{\mathbf{A}}\mathbf{S}_{\mathbf{A}}^{1/2} = \mathbf{A}\mathbf{U}_{\mathbf{A}}\mathbf{S}_{\mathbf{A}}^{-1/2}$. Hence

$$\hat{\mathbf{X}} - \mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{1/2}\mathbf{W}^* = (\mathbf{A} - \mathbf{P})\mathbf{U}_{\mathbf{A}}\mathbf{S}_{\mathbf{A}}^{-1/2} - \mathbf{U}_{\mathbf{P}}\mathbf{U}_{\mathbf{P}}^{\top}(\mathbf{A} - \mathbf{P})\mathbf{U}_{\mathbf{A}}\mathbf{S}_{\mathbf{A}}^{-1/2} + \mathbf{R}_1\mathbf{S}_{\mathbf{A}}^{1/2} + \mathbf{U}_{\mathbf{P}}\mathbf{R}_2$$

Writing

$$\mathbf{R}_3 = \mathbf{U}_{\mathbf{A}} - \mathbf{U}_{\mathbf{P}}\mathbf{W}^* = \mathbf{U}_{\mathbf{A}} - \mathbf{U}_{\mathbf{P}}\mathbf{U}_{\mathbf{P}}^{\top}\mathbf{U}_{\mathbf{A}} + \mathbf{R}_1,$$

we derive that

$$\begin{aligned}
\hat{\mathbf{X}} - \mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{1/2}\mathbf{W}^* =& (\mathbf{A} - \mathbf{P})\mathbf{U}_{\mathbf{P}}\mathbf{W}^*\mathbf{S}_{\mathbf{A}}^{-1/2} - \mathbf{U}_{\mathbf{P}}\mathbf{U}_{\mathbf{P}}^{\top}(\mathbf{A} - \mathbf{P})\mathbf{U}_{\mathbf{P}}\mathbf{W}^*\mathbf{S}_{\mathbf{A}}^{-1/2} \\
&+ (\mathbf{I} - \mathbf{U}_{\mathbf{P}}\mathbf{U}_{\mathbf{P}}^{\top})(\mathbf{A} - \mathbf{P})\mathbf{R}_3\mathbf{S}_{\mathbf{A}}^{-1/2} + \mathbf{R}_1\mathbf{S}_{\mathbf{A}}^{1/2} + \mathbf{U}_{\mathbf{P}}\mathbf{R}_2
\end{aligned}$$

Recalling that $\|\mathbf{U}_{\mathbf{A}} - \mathbf{U}_{\mathbf{P}}\mathbf{U}_{\mathbf{P}}^{\top}\mathbf{U}_{\mathbf{A}}\|_F = O(\delta^{-1/2}(\mathbf{P}))$ with high probability, we have

$$\|\mathbf{R}_1\|_F = O(\delta^{-1}(\mathbf{P})), \quad \|\mathbf{R}_2\|_F = O((\log n)\,\delta^{-1/2}(\mathbf{P})), \quad \text{and } \|\mathbf{R}_3\|_F = O(\delta^{-1/2}(\mathbf{P}))$$

74

with high probability. Furthermore, a similar application of Hoeffding's inequality to that in the proof of Lemma 49, along with an application of Weyl's inequality and Theorem 21 to bound $\lambda_i(\mathbf{A})$, ensures that

$$\|\mathbf{U_P}\mathbf{U_P^\top}(\mathbf{A}-\mathbf{P})\mathbf{U_P}\mathbf{W}^*\mathbf{S_A}^{-1/2}\|_F \le \|\mathbf{U_P^\top}(\mathbf{A}-\mathbf{P})\mathbf{U_P}\|_F\|\mathbf{S_A}^{-1/2}\|_F = O(\log n \delta^{-1/2}(\mathbf{P}))/$$

As a consequence, with high probability

$$\begin{aligned}\|\hat{\mathbf{X}}-\mathbf{U_P}\mathbf{S_P}^{1/2}\mathbf{W}^*\|_F &= \|(\mathbf{A}-\mathbf{P})\mathbf{U_P}\mathbf{W}^*\mathbf{S_A}^{-1/2}\|_F + O((\log n)\,\delta^{-1/2}(\mathbf{P}))\\ &= \|(\mathbf{A}-\mathbf{P})\mathbf{U_P}\mathbf{S_P}^{-1/2}\mathbf{W}^* - (\mathbf{A}-\mathbf{P})\mathbf{U_P}(\mathbf{S_P}^{-1/2}\mathbf{W}^*-\mathbf{W}^*\mathbf{S_A}^{-1/2})\|_F\\ &\quad + O((\log n)\,\delta^{-1/2}(\mathbf{P}))\end{aligned}$$

A very similar argument to that employed in the proof of Lemma 49 implies that

$$\|\mathbf{S_P}^{-1/2}\mathbf{W}^* - \mathbf{W}^*\mathbf{S_A}^{-1/2}\|_F = O((\log n)\,\delta^{-3/2}(\mathbf{P}))$$

with high probability. We thus obtain

$$\begin{aligned}\|\hat{\mathbf{X}}-\mathbf{U_P}\mathbf{S_P}^{1/2}\mathbf{W}^*\|_F &= \|(\mathbf{A}-\mathbf{P})\mathbf{U_P}\mathbf{S_P}^{-1/2}\mathbf{W}^*\|_F + O((\log n)\,\delta^{-1/2}(\mathbf{P}))\\ &= \|(\mathbf{A}-\mathbf{P})\mathbf{U_P}\mathbf{S_P}^{-1/2}\|_F + O((\log n)\,\delta^{-1/2}(\mathbf{P}))\end{aligned} \tag{51}$$

with high probability. Finally, to complete the proof, we note that $\mathbf{X}=\mathbf{U_P}\mathbf{S_P}^{1/2}\tilde{\mathbf{W}}$ for some orthogonal matrix $\mathbf{W}$. Since $\mathbf{W}^*$ is also orthogonal, we conclude that there exists some orthogonal $\mathbf{W}$ for which $\mathbf{XW}=\mathbf{U_P}\mathbf{S_P}^{1/2}\mathbf{W}^*$, as desired. ∎

We are now ready to prove the $2\to\infty$ consistency we assert in Theorem 26.

Proof

To establish Theorem 26, we note that from Theorem 50

$$\|\hat{\mathbf{X}}-\mathbf{XW}\|_F = \|(\mathbf{A}-\mathbf{P})\mathbf{U_P}\mathbf{S_P}^{-1/2}\|_F + O((\log n)\,\delta^{-1/2}(\mathbf{P}))$$

and hence

$$\begin{aligned}\max_i \|\hat{\mathbf{X}}_i - \rho_n^{1/2}\mathbf{WX}_i\| &\le \frac{1}{\lambda_d^{1/2}(\mathbf{P})}\max_i \|((\mathbf{A}-\mathbf{P})\mathbf{U_P})_i\| + O((\log n)\,\delta^{-1/2}(\mathbf{P}))\\ &\le \frac{d^{1/2}}{\lambda_d^{1/2}(\mathbf{P})}\max_j \|(\mathbf{A}-\mathbf{P})\mathbf{U}_{\cdot j}\|_\infty + O((\log n)\,\delta^{-1/2}(\mathbf{P}))\end{aligned}$$

where $\mathbf{U}_{\cdot j}$ denotes the j-th column of $\mathbf{U_P}$. Now, for a given j and a given index i, the i-th element of the vector $(\mathbf{A}-\mathbf{P})\mathbf{U}_{\cdot j}$ is of the form

$$\sum_k (\mathbf{A}_{ik}-\mathbf{P}_{ik})\mathbf{U}_{kj}$$

and once again, by Hoeffding's inequality, the above term is $O(\log n)$ asymptotically almost surely. Taking the union bound over all indices i and all columns j of $\mathbf{U_P}$, we conclude that with high probability,

$$\max_i \|\hat{\mathbf{X}}_i - \rho_n^{1/2}\mathbf{W}\mathbf{X}_i\| \le \frac{Cd^{1/2}}{\lambda_d^{1/2}(P)} \log^2 n + O((\log n)\, \delta^{-1/2}(\mathbf{P}))$$

$$\le \frac{Cd^{1/2}\log^2 n}{\delta^{1/2}(\mathbf{P})}$$

as desired. ∎

Now, we move to our distributional results.

A.2 Proof of the central limit theorem for the adjacency spectral embedding

.

Recall **Theorem 27**: Let $(\mathbf{A}_n, \mathbf{X}_n) \sim \text{RDPG}(F)$ be a sequence of adjacency matrices and associated latent positions of a d-dimensional random dot product graph according to an inner product distribution F. Let $\Phi(\mathbf{x}, \boldsymbol{\Sigma})$ denote the cdf of a (multivariate) Gaussian with mean zero and covariance matrix $\boldsymbol{\Sigma}$, evaluated at $\mathbf{x} \in \mathbb{R}^d$. Then there exists a sequence of orthogonal d-by-d matrices $(\mathbf{W}_n)_{n=1}^{\infty}$ such that for all $\boldsymbol{z} \in \mathbb{R}^d$ and for any fixed index i,

$$\lim_{n\to\infty} \Pr\left[n^{1/2}\left(\hat{\mathbf{X}}_n\mathbf{W}_n - \mathbf{X}_n\right)_i \le \boldsymbol{z}\right] = \int_{\text{supp}\,F} \Phi\left(\boldsymbol{z}, \boldsymbol{\Sigma}(\mathbf{x})\right) dF(\mathbf{x}),$$

where
$$\boldsymbol{\Sigma}(\mathbf{x}) = \Delta^{-1}\mathbb{E}\left[(\mathbf{x}^\top\mathbf{X}_1 - (\mathbf{x}^\top\mathbf{X}_1)^2)\mathbf{X}_1\mathbf{X}_1^\top\right]\Delta^{-1}; \quad \text{and} \quad \Delta = \mathbb{E}[\mathbf{X}_1\mathbf{X}_1^\top]. \tag{52}$$

To prove this, we begin with the following simple lemma which indicates that when the rows of $\mathbf{X}$ are sampled i.i.d. from some distribution F, that the eigenvalues of $\mathbf{X}$ grows proportionally with n.

Lemma 51 *With notation as above, let F be an inner product distribution and suppose $\mathbf{X}_1, \cdots, \mathbf{X}_n, \mathbf{Y}$ be i.i.d F. Suppose also that $\Delta = \mathbb{E}[\mathbf{X}_1\mathbf{X}_1^\top]$ is of rank d. Then for $1 \le i \le d$, $\lambda_i(\mathbf{P}) = \Omega(n\lambda_i(\Delta))$ almost surely.*

Proof Recall that for any matrix $\mathbf{H}$, the nonzero eigenvalues of $\mathbf{H}^\top\mathbf{H}$ are the same as those of $\mathbf{H}\mathbf{H}^\top$, so $\lambda_i(\mathbf{X}\mathbf{X}^\top) = \lambda_i(\mathbf{X}^\top\mathbf{X})$. In what follows, we remind the reader that $\mathbf{X}$ is a matrix whose rows are the tranposes of the column vectors $\mathbf{X}_i$, and $\mathbf{Y}$ is a d-dimensional vector that is independent from and has the same distribution as that of the $\mathbf{X}_i$. We observe that $(\mathbf{X}^\top\mathbf{X} - n\mathbb{E}\mathbf{Y}\mathbf{Y}^\top)_{ij} = \sum_{k=1}^n (\mathbf{X}_{ki}\mathbf{X}_{kj} - \mathbb{E}\mathbf{Y}_i\mathbf{Y}_j)$ is a sum of n independent zero-mean random variables, each contained in $[-1, 1]$. Thus, Hoeffding's inequality yields, for all $i, j \in [d]$,

$$\Pr\left[|(\mathbf{X}^\top\mathbf{X}) - n\mathbb{E}\mathbf{Y}\mathbf{Y}^\top|_{ij} \ge 2\sqrt{n\log n}\right] \le \frac{2}{n^2}.$$

A union bound over all $i, j \in [d]$ implies that $\|\mathbf{X}^\top \mathbf{X} - n\mathbb{E}\mathbf{Y}\mathbf{Y}^\top\|_F^2 \leq 4d^2 n \log n$ with probability at least $1 - 2d^2/n^2$. Taking square roots and noting that the Frobenius norm is an upper bound on the spectral norm, we have that $\|\mathbf{X}^\top \mathbf{X} - n\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)\| \leq 2d\sqrt{n \log n}$ with probability at least $1 - 2d^2/n^2$, and Weyl's inequality yields that for all $1 \leq i \leq d$, $|\lambda_i(\mathbf{X}\mathbf{X}^\top) - n\lambda_i(\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)| \leq 2d\sqrt{n \log n}$ with probability at least $1 - 2d^2/n^2$. Of course, since the vector $\mathbf{Y}$ has the same distribution as any of the latent positions $\mathbf{X}_i$, we see that $\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top) = \Delta$. By the reverse triangle inequality, for any $1 \leq i \leq d$, we have

$$\lambda_i(\mathbf{X}\mathbf{X}^\top) \geq \lambda_d(\mathbf{X}\mathbf{X}^\top) \geq |n\lambda_d(\Delta) - 2d\sqrt{n \log n}| = \Omega(n).$$

Multiplying through by ρ_n, we find that there exists some constant C so that for all n sufficiently large, $\lambda_d(\rho_n \mathbf{X}\mathbf{X}^\top) \geq n\rho_n\lambda_d(\Delta)$ with probability at least $1 - 2d^2/n^2$. ∎

As described previously, to prove our central limit theorem, we require somewhat more precise control on certain residual terms, which we establish in the following key lemma. In the lemmas and proofs that follow, we frequently suppress the dependence of the sequence of graphs and embeddings on n.

Lemma 52 *Let* $\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3$ *be defined, as above, by*

$$\mathbf{R}_1 = \mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top\mathbf{U}_\mathbf{A} - \mathbf{U}_\mathbf{P}\mathbf{W}^*$$

$$\mathbf{R}_2 = \mathbf{W}^*\mathbf{S}_\mathbf{A}^{1/2} - \mathbf{S}_\mathbf{P}^{1/2}\mathbf{W}^*$$

$$\mathbf{R}_3 = \mathbf{U}_\mathbf{A} - \mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top\mathbf{U}_\mathbf{A} + \mathbf{R}_1 = \mathbf{U}_\mathbf{A} - \mathbf{U}_\mathbf{P}\mathbf{W}^*.$$

where, as before, $\mathbf{W}^*$ *is the orthogonal transformation* $\mathbf{W}^* = \mathbf{W}_1\mathbf{W}_2^\top$ *with* $\mathbf{W}_1\Sigma\mathbf{W}_2$ *being the singular value decomposition of* $\mathbf{U}_\mathbf{P}^\top\mathbf{U}_\mathbf{A}$. *Then the following convergences in probability hold:*

$$\sqrt{n}\left[(\mathbf{A} - \mathbf{P})\mathbf{U}_\mathbf{P}(\mathbf{W}^*\mathbf{S}_\mathbf{A}^{-1/2} - \mathbf{S}_\mathbf{P}^{-1/2}\mathbf{W}^*)\right]_h \xrightarrow{P} \mathbf{0}, \tag{53}$$

$$\sqrt{n}\left[\mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top(\mathbf{A} - \mathbf{P})\mathbf{U}_\mathbf{P}\mathbf{W}^*\mathbf{S}_\mathbf{A}^{-1/2}\right]_h \xrightarrow{P} \mathbf{0}, \tag{54}$$

$$\sqrt{n}\left[(\mathbf{I} - \mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top)(\mathbf{A} - \mathbf{P})\mathbf{R}_3\mathbf{S}_\mathbf{A}^{-1/2}\right]_h \xrightarrow{P} \mathbf{0}, \tag{55}$$

and with high probability,

$$\|\mathbf{R}_1\mathbf{S}_\mathbf{A}^{1/2} + \mathbf{U}_\mathbf{P}\mathbf{R}_2\|_F \leq \frac{C \log n}{n^{1/2}}.$$

Proof We begin by observing that

$$\|\mathbf{R}_1\mathbf{S}_\mathbf{A}^{1/2} + \mathbf{U}_\mathbf{P}\mathbf{R}_2\|_F \leq \|\mathbf{R}_1\|_F\|\mathbf{S}_\mathbf{A}^{1/2}\| + \|\mathbf{R}_2\|_F.$$

Proposition 48 and the trivial upper bound on the eigenvalues of $\mathbf{A}$ ensures that

$$\|\mathbf{R}_1\|_F\|\mathbf{S}_\mathbf{A}^{1/2}\| \leq \frac{C \log n}{n^{1/2}} \text{ w.h.p. },$$

77

Combining this with Eq. (49) in Lemma 49, we conclude that

$$\|\mathbf{R}_1\mathbf{S}_\mathbf{A}^{1/2} + \mathbf{U}_\mathbf{P}\mathbf{R}_2\|_F \le \frac{C\log n}{n^{1/2}} \text{ w.h.p.}$$

We will establish (53), (54) and (55) order. To see (53), observe that

$$\sqrt{n}\|(\mathbf{A} - \mathbf{P})\mathbf{U}_\mathbf{P}(\mathbf{W}^*\mathbf{S}_\mathbf{A}^{-1/2} - \mathbf{S}_\mathbf{P}^{-1/2}\mathbf{W}^*)\|_F \le \sqrt{n}\|(\mathbf{A} - \mathbf{P})\mathbf{U}_\mathbf{P}\|\|\mathbf{W}^*\mathbf{S}_\mathbf{A}^{-1/2} - \mathbf{S}_\mathbf{P}^{-1/2}\mathbf{W}^*\|_F,$$

and Lemma 49 imply that with high probability

$$\sqrt{n}\|(\mathbf{A} - \mathbf{P})\mathbf{U}_\mathbf{P}(\mathbf{W}^*\mathbf{S}_\mathbf{A}^{-1/2} - \mathbf{S}_\mathbf{P}^{-1/2}\mathbf{W}^*)\|_F \le \frac{C\log n}{\sqrt{n}},$$

which goes to 0 as $n \to \infty$.

To show the convergence in (54), we recall that $\mathbf{U}_\mathbf{P}\mathbf{S}_\mathbf{P}^{1/2}\mathbf{W} = \mathbf{X}$ for some orthogonal matrix $\mathbf{W}$ and observe that since the rows of the latent position matrix $\mathbf{X}$ are necessarily bounded in Euclidean norm by 1, and since the top d eigenvalues of $\mathbf{P}$ are of order n (recall Lemma 51), it follows that

$$\|\mathbf{U}_\mathbf{P}\|_{2\to\infty} \le Cn^{-1/2} \text{ w.h.p.} \tag{56}$$

Next, (50) and Lemma 51 imply that

$$\|(\mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top(\mathbf{A} - \mathbf{P})\mathbf{U}_\mathbf{P}\mathbf{W}^*\mathbf{S}_\mathbf{A}^{-1/2})_h\| \le \|\mathbf{U}_\mathbf{P}\|_{2\to\infty}\|\mathbf{U}_\mathbf{P}^\top(\mathbf{A} - \mathbf{P})\mathbf{U}_\mathbf{P}\|\|\mathbf{S}_\mathbf{A}^{-1/2}\|$$
$$\le \frac{C\log n}{n} \text{ w.h.p.,}$$

which implies (54). Finally, to establish (55), we must bound the Euclidean norm of the vector

$$\left[(\mathbf{I} - \mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top)(\mathbf{A} - \mathbf{P})\mathbf{R}_3\mathbf{S}_\mathbf{A}^{-1/2}\right]_h, \tag{57}$$

where, as defined above, $\mathbf{R}_3 = \mathbf{U}_\mathbf{A} - \mathbf{U}_\mathbf{P}\mathbf{W}^*$. Let $\mathbf{B}_1$ and $\mathbf{B}_2$ be defined as follows:

$$\begin{aligned}
\mathbf{B}_1 &= (\mathbf{I} - \mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top)(\mathbf{A} - \mathbf{P})(\mathbf{I} - \mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top)\mathbf{U}_\mathbf{A}\mathbf{S}_\mathbf{A}^{-1/2} \\
\mathbf{B}_2 &= (\mathbf{I} - \mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top)(\mathbf{A} - \mathbf{P})\mathbf{U}_\mathbf{P}(\mathbf{U}_\mathbf{P}^\top\mathbf{U}_\mathbf{A} - \mathbf{W}^*)\mathbf{S}_\mathbf{A}^{-1/2}
\end{aligned} \tag{58}$$

Recalling that $\mathbf{R}_3 = \mathbf{U}_\mathbf{A} - \mathbf{U}_\mathbf{P}\mathbf{W}^*$, we have

$$\begin{aligned}
(\mathbf{I} - \mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top)(\mathbf{A} - \mathbf{P})\mathbf{R}_3\mathbf{S}_\mathbf{A}^{-1/2} &= (\mathbf{I} - \mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top)(\mathbf{A} - \mathbf{P})(\mathbf{U}_\mathbf{A} - \mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top\mathbf{U}_\mathbf{A})\mathbf{S}_\mathbf{A}^{-1/2} \\
&\quad + (\mathbf{I} - \mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top)(\mathbf{A} - \mathbf{P})(\mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top\mathbf{U}_\mathbf{A} - \mathbf{U}_\mathbf{P}\mathbf{W}^*)\mathbf{S}_\mathbf{A}^{-1/2} \\
&= \mathbf{B}_1 + \mathbf{B}_2.
\end{aligned}$$

We will bound the Euclidean norm of the h-th row of each of these two matrices on the right-hand side, from which a triangle inequality will yield our desired bound on the quantity in Equation (57). Recall that we use C to denote a positive constant, independent of n and m, which may change from line to line.

Let us first consider $\mathbf{B}_2 = (\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top})(\mathbf{A} - \mathbf{P})\mathbf{U_P}(\mathbf{U_P^\top}\mathbf{U_A} - \mathbf{W}^*)\mathbf{S_A^{-1/2}}$. We have

$$\|\mathbf{B}_2\|_F \le \|(\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top})(\mathbf{A} - \mathbf{P})\mathbf{U_P}\|\|\mathbf{U_P^\top}\mathbf{U_A} - \mathbf{W}^*\|_F\|\mathbf{S_A^{-1/2}}\|.$$

By submultiplicativity of the spectral norm and Theorem 20, $\|(\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top})(\mathbf{A} - \mathbf{P})\mathbf{U_P}\| \le Cn^{1/2}\log^{1/2} n$ with high probability (and indeed, under our degree assumptions and Theorem 21, the $\log n$ factor can be dropped). From Prop. 48 and Lemma 51, respectively, we have with high probability

$$\|\mathbf{U_P^\top}\mathbf{U_A} - \mathbf{W}^*\|_F \le Cn^{-1}\log n \quad \text{and} \quad \|\mathbf{S_A^{-1/2}}\| \le Cn^{-1/2}$$

Thus, we deduce that with high probability,

$$\|\mathbf{B}_2\|_F \le \frac{C\log^{3/2} n}{n} \tag{59}$$

from which it follows that $\|\sqrt{n}\mathbf{B}_2\|_F \xrightarrow{P} 0$, and hence $\|\sqrt{n}(\mathbf{B}_2)_h\| \xrightarrow{P} 0$.

Turning our attention to $\mathbf{B}_1$, and recalling that $\mathbf{U_A^\top}\mathbf{U_A} = \mathbf{I}$, we note that

$$\|(\mathbf{B}_1)_h\| = \left\|\left[(\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top})(\mathbf{A} - \mathbf{P})(\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top})\mathbf{U_A}\mathbf{S_A^{-1/2}}\right]_h\right\|$$

$$= \left\|\left[(\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top})(\mathbf{A} - \mathbf{P})(\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top})\mathbf{U_A}\mathbf{U_A^\top}\mathbf{U_A}\mathbf{S_A^{-1/2}}\right]_h\right\|$$

$$\le \left\|\mathbf{U_A}\mathbf{S_A^{-1/2}}\right\|\left\|\left[(\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top})(\mathbf{A} - \mathbf{P})(\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top})\mathbf{U_A}\mathbf{U_A^\top}\right]_h\right\|.$$

Let $\epsilon > 0$ be a constant. We will show that

$$\lim_{n\to\infty} \Pr\left[\|\sqrt{n}(\mathbf{B}_1)_h\| > \epsilon\right] = 0. \tag{60}$$

For ease of notation, define

$$\mathbf{E}_1 = (\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top})(\mathbf{A} - \mathbf{P})(\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top})\mathbf{U_A}\mathbf{U_A^\top}.$$

We will show that

$$\lim_{n\to\infty} \Pr\left[\sqrt{n}\,\|[\mathbf{E}_1]_h\| > n^{1/4}\right] = 0, \tag{61}$$

which will imply (60) since, by Lemma 51, $\|\mathbf{U_A}\mathbf{S_A^{-1/2}}\| \le Cn^{-1/2}$ with high probability. Let $\mathbf{Q} \in \mathbb{R}^{n\times n}$ be any permutation matrix. We observe that

$$\mathbf{Q}\mathbf{U_P}\mathbf{U_P^\top}\mathbf{Q}^\top\mathbf{Q}\mathbf{P}\mathbf{Q}^\top = \mathbf{Q}\mathbf{P}\mathbf{Q}^\top,$$

and thus $\mathbf{Q}\mathbf{U_P}\mathbf{U_P^\top}\mathbf{Q}^\top$ is a projection matrix for $\mathbf{Q}\mathbf{P}\mathbf{Q}^\top$ if and only if $\mathbf{U_P}\mathbf{U_P^\top}$ is a projection matrix for $\mathbf{P}$. A similar argument applies to the matrix $\mathbf{U_A}\mathbf{U_A^\top}$. Combining this with the exchangeability structure of the matrix $\mathbf{A} - \mathbf{P}$, it follows that the Frobenius norms of the rows of $\mathbf{E}_1$ are equidistributed. This row-exchangeability for $\mathbf{E}_1$ implies that $n\mathbb{E}\|(\mathbf{E}_1)_h\|^2 = \|\mathbf{E}_1\|_F^2$. Applying Markov's inequality,

$$\Pr\left[\|\sqrt{n}\,[\mathbf{E}_1]_h\| > t\right] \le \frac{n\mathbb{E}\left\|\left[(\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top})(\mathbf{A} - \mathbf{P})(\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top})\mathbf{U_A}\mathbf{U_A^\top}\right]_h\right\|^2}{t^2}$$

$$= \frac{\mathbb{E}\left\|(\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top})(\mathbf{A} - \mathbf{P})(\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top})\mathbf{U_A}\mathbf{U_A^\top}\right\|_F^2}{t^2}. \tag{62}$$

We will proceed by showing that with high probability,

$$\left\| (\mathbf{I} - \mathbf{U_P U_P^\top})(\mathbf{A} - \mathbf{P})(\mathbf{I} - \mathbf{U_P U_P^\top})\mathbf{U_A U_A^\top} \right\|_F \le C \log n, \tag{63}$$

whence choosing $t = n^{1/4}$ in (62) yields that

$$\lim_{n \to \infty} \Pr\left[\left\| \sqrt{n} \left[(\mathbf{I} - \mathbf{U_P U_P^\top})(\mathbf{A} - \mathbf{P})(\mathbf{I} - \mathbf{U_P U_P^\top})\mathbf{U_A U_A^\top} \right]_h \right\| > n^{1/4} \right] = 0,$$

and (60) will follow. We have

$$\left\| (\mathbf{I} - \mathbf{U_P U_P^\top})(\mathbf{A} - \mathbf{P})(\mathbf{I} - \mathbf{U_P U_P^\top})\mathbf{U_A U_A^\top} \right\|_F \le \|\mathbf{A} - \mathbf{P}\| \|\mathbf{U_A} - \mathbf{U_P U_P^\top U_A}\|_F \|\mathbf{U_A}\|$$

Theorem 21 and Lemma 51 implies that the first term in this product is at most $Cn^{1/2}$ high probability, and the final term in this product is, trivially, at most 1. To bound the second term, we will follow reasoning similar to that in Lemma 49, combined with the Davis-Kahan theorem. The Davis-Kahan Theorem implies that for a suitable constant $C > 0$,

$$\|\mathbf{U_A U_A^\top} - \mathbf{U_P U_P^\top}\| \le \frac{C\|\mathbf{A} - \mathbf{P}\|}{\lambda_d(\mathbf{P})}.$$

By Theorem 2 in Yu et al. (2015), there exists orthonormal $\mathbf{W} \in \mathbb{R}^{d \times d}$ such that

$$\|\mathbf{U_A} - \mathbf{U_P W}\|_F \le C\|\mathbf{U_A U_A^\top} - \mathbf{U_P U_P^\top}\|_F.$$

We observe further that the multivariate linear least squares problem

$$\min_{\mathbf{T} \in \mathbb{R}^{d \times d}} \|\mathbf{U_A} - \mathbf{U_P T}\|_F^2$$

is solved by $\mathbf{T} = \mathbf{U_P^\top U_A}$. Combining all of the above, we find that

$$\|\mathbf{U_A} - \mathbf{U_P U_P^\top U_A}\|_F^2 \le \|\mathbf{U_A} - \mathbf{U_P W}\|_F^2 \le C\|\mathbf{U_A U_A^\top} - \mathbf{U_P U_P^\top}\|_F^2$$

$$\le C\|\mathbf{U_A U_A^\top} - \mathbf{U_P U_P^\top}\|^2 \le \left(\frac{C\|\mathbf{A} - \mathbf{P}\|}{\lambda_d(\mathbf{P})} \right)^2 \le \frac{C}{n} \quad \text{w.h.p.}$$

We conclude that

$$\left\| (\mathbf{I} - \mathbf{U_P U_P^\top})(\mathbf{A} - \mathbf{P})(\mathbf{I} - \mathbf{U_P U_P^\top})\mathbf{U_A U_A^\top} \right\|_F \le C\|\mathbf{A} - \mathbf{P}\| \|\mathbf{U_A} - \mathbf{U_P U_P^\top U_A}\|_F \|\mathbf{U_A}\|_S$$

$$\le C \log n \text{ w.h.p. },$$

which implies (63), as required, and thus the convergence in (55) is established, completing the proof. ∎

Next, recall the inherent nonidentifiability in the RDPG model: suppose the "true" latent positions are some matrix $\mathbf{X}$. Then $\mathbf{X} = \mathbf{U_P S_P^{1/2} W}$ for some suitably-chosen orthogonal matrix $\mathbf{W}$. We now consider the asymptotic distribution of

$$n^{1/2} \mathbf{W}_n^\top \left[(\mathbf{A} - \mathbf{P})\mathbf{U_P S_P^{-1/2}} \right]_h$$

conditional on $\mathbf{X}_i = \mathbf{x}_i$. Because we can write suitable terms of $\mathbf{A} - \mathbf{P}$ as the sum of independent random variables, we can invoke the Lindeberg-Feller Central Limit Theorem to establish the asymptotic normality of

$$n^{1/2}\mathbf{W}_n^\top\left[(\mathbf{A} - \mathbf{P})\mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{-1/2}\right]_h$$

as follows.

Lemma 53 *Fix some $i \in [n]$. Conditional on $\mathbf{X}_i = \mathbf{x}_i \in \mathbb{R}^d$, there exists a sequence of d-by-d orthogonal matrices $\{\mathbf{W}_n\}$ such that*

$$n^{1/2}\mathbf{W}_n^\top\left[(\mathbf{A} - \mathbf{P})\mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{-1/2}\right]_h \xrightarrow{\mathcal{L}} \mathcal{N}(0, \boldsymbol{\Sigma}(\mathbf{x}_i)),$$

where $\boldsymbol{\Sigma}(\mathbf{x}_i) \in \mathbb{R}^{d \times d}$ is a covariance matrix that depends on $\mathbf{x}_i$.

Proof For each n, choose orthogonal $\mathbf{W}_n \in \mathbb{R}^{d \times d}$ so that $\mathbf{X} = \mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{1/2}\mathbf{W}_n$. At least one such $\mathbf{W}_n$ exists for each value of n, since, as discussed previously, the true latent positions $\mathbf{X}$ are specified only up to some orthogonal transformation. We then have

$$n^{1/2}\mathbf{W}_n^\top\left[(\mathbf{A} - \mathbf{P})\mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{-1/2}\right]_i = n^{1/2}\mathbf{W}_n^\top\mathbf{S}_{\mathbf{P}}^{-1}\mathbf{W}_n\left[\mathbf{A}\mathbf{X} - \mathbf{P}\mathbf{X}\right]_i$$

$$= n^{1/2}\mathbf{W}_n^\top\mathbf{S}_{\mathbf{P}}^{-1}\mathbf{W}_n\left(\sum_{j=1}^n (\mathbf{A}_{ij} - \mathbf{P}_{ij})\mathbf{X}_j\right)$$

$$= n^{1/2}\mathbf{W}_n^\top\mathbf{S}_{\mathbf{P}}^{-1}\mathbf{W}_n\left(\sum_{j \neq i} (\mathbf{A}_{ij} - \mathbf{P}_{ij})\mathbf{X}_j\right) - n^{1/2}\mathbf{W}_n^\top\mathbf{S}_{\mathbf{P}}^{-1}\mathbf{W}_n\mathbf{P}_{ii}\mathbf{X}_i$$

$$= \left(n\mathbf{W}_n^\top\mathbf{S}_{\mathbf{P}}^{-1}\mathbf{W}_n\right)\left[n^{-1/2}\sum_{j \neq i} (\mathbf{A}_{ij} - \mathbf{P}_{ij})\mathbf{X}_j\right] - n\mathbf{W}_n^\top\mathbf{S}_{\mathbf{P}}^{-1}\mathbf{W}_n\frac{\mathbf{P}_{ii}\mathbf{X}_i}{n^{1/2}}.$$

Conditioning on $\mathbf{X}_i = \mathbf{x}_i \in \mathbb{R}^d$, we first observe that $\frac{\mathbf{P}_{ii}}{n^{1/2}}\mathbf{X}_i = \frac{\mathbf{x}_i^\top\mathbf{x}_i}{n^{1/2}}x_i \to 0$ almost surely. Furthermore,

$$n^{-1/2}\sum_{j \neq i} (\mathbf{A}_{ij} - \mathbf{P}_{ij})\mathbf{X}_j = n^{-1/2}\sum_{j \neq i} (\mathbf{A}_{ij} - \mathbf{X}_j^\top\mathbf{x}_i)\mathbf{X}_j$$

is a scaled sum of $n - 1$ independent 0-mean random variables, each with covariance matrix given by

$$\tilde{\boldsymbol{\Sigma}}(\mathbf{x}_i) = \mathbb{E}\left[\left(\mathbf{x}_i^\top\mathbf{X}_j - (\mathbf{x}_i^\top\mathbf{X}_j)^2\right)\mathbf{X}_j\mathbf{X}_j^\top\right].$$

The multivariate central limit theorem thus implies that

$$n^{-1/2}\sum_{j \neq i} \left(\mathbf{A}_{ij} - \mathbf{X}_j\mathbf{x}_i^\top\right)\mathbf{X}_j \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}(\mathbf{x}_i)). \tag{64}$$

Finally, by the strong law of large numbers,

$$n^{-1}\mathbf{W}_n^\top\mathbf{S}_{\mathbf{P}_n}\mathbf{W}_n = \frac{1}{n}\mathbf{X}^\top\mathbf{X} \to \boldsymbol{\Delta} \text{ a.s.}$$

and thus $(n\mathbf{W}_n^\top \mathbf{S}_{\mathbf{P}_n}^{-1} \mathbf{W}_n) \to \mathbf{\Delta}^{-1}$ almost surely. Combining this fact with (64), the multivariate version of Slutsky's theorem yields

$$n^{1/2}\mathbf{W}_n^\top \left[(\mathbf{A}-\mathbf{P})\mathbf{U_P}\mathbf{S_P}^{-1/2} \right]_h \xrightarrow{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}(\mathbf{x}_i)\right)$$

where $\mathbf{\Sigma}(\mathbf{x}_i) = \mathbf{\Delta}^{-1}\tilde{\mathbf{\Sigma}}(\mathbf{x}_i)\mathbf{\Delta}^{-1}$. Integrating over the possible values of $\mathbf{x}_i$ with respect to distribution F completes the proof. ∎

Lemmas 52 and 53 are the main ingredients in the proof of Theorem 27, whose proof now follows easily:

Proof [Proof of Theorem 27]. We start with the following decomposition that was originally used in the proof of Theorem 50.

$$\sqrt{n}\left(\mathbf{U_A}\mathbf{S_A}^{1/2} - \mathbf{U_P}\mathbf{S_P}^{1/2}\mathbf{W}^*\right) = \sqrt{n}(\mathbf{A}-\mathbf{P})\mathbf{U_P}\mathbf{S_P}^{-1/2}\mathbf{W}^* + \sqrt{n}(\mathbf{A}-\mathbf{P})\mathbf{U_P}(\mathbf{W}^*\mathbf{S_A}^{-1/2} - \mathbf{S_P}^{-1/2}\mathbf{W}^*)$$
$$- \sqrt{n}\mathbf{U_P}\mathbf{U_P}^\top(\mathbf{A}-\mathbf{P})\mathbf{U_P}\mathbf{W}^*\mathbf{S_A}^{-1/2}$$
$$+ \sqrt{n}(\mathbf{I}-\mathbf{U_P}\mathbf{U_P}^\top)(\mathbf{A}-\mathbf{P})\mathbf{R}_3\mathbf{S_A}^{-1/2} + \sqrt{n}\mathbf{R}_1\mathbf{S_A}^{1/2} + \sqrt{n}\mathbf{U_P}\mathbf{R}_2.$$
$$(65)$$

Now given any index i, Lemma 52 can be used to bound the Euclidean norms of the i-th rows of

$$(\mathbf{A}-\mathbf{P})\mathbf{U_P}(\mathbf{W}^*\mathbf{S_A}^{-1/2} - \mathbf{S_P}^{-1/2}\mathbf{W}^*),$$
$$\mathbf{U_P}\mathbf{U_P}^\top(\mathbf{A}-\mathbf{P})\mathbf{U_P}\mathbf{W}^*\mathbf{S_A}^{-1/2}$$

and

$$(\mathbf{I}-\mathbf{U_P}\mathbf{U_P}^\top)(\mathbf{A}-\mathbf{P})\mathbf{R}_3\mathbf{S_A}^{-1/2}.$$

The Euclidean norms of the i-th row of $\mathbf{R}_1\mathbf{S_A}^{1/2}$ and $\mathbf{U_P}\mathbf{R}_2$ can be bounded from above by the bounds on $\|\mathbf{R}_1\mathbf{S_A}^{1/2}\|_{2\to\infty}$ and $\|\mathbf{U_P}\mathbf{R}_2\|_{2\to\infty}$. Since

$$\|\mathbf{U_P}\|_{2\to\infty} = O(n^{-1/2})$$

by Eq. (56), we conclude that

$$\sqrt{n}\|\mathbf{R}_1\mathbf{S_A}^{1/2}\|_{2\to\infty} = O(n^{-1/2}\log^{1/2} n)$$

and

$$\sqrt{n}\|\mathbf{U_P}\mathbf{R}_2\|_{2\to\infty} = O(n^{-1/2}\log^{1/2} n)$$

Therefore, for any fixed index i, we have

$$\sqrt{n}\left(\mathbf{U_A}\mathbf{S_A}^{1/2} - \mathbf{U_P}\mathbf{S_P}^{1/2}\mathbf{W}^*\right)_i = \sqrt{n}((\mathbf{A}-\mathbf{P})\mathbf{U_P}\mathbf{S_P}^{-1/2}\mathbf{W}^*)_i + O(n^{-1/2}\log^{1/2} n)$$

with high probability. Since $\mathbf{X} = \mathbf{U_P}\mathbf{S_P}\mathbf{W}$, we can rewrite the above expression as

$$\sqrt{n}\left(\mathbf{U_A}\mathbf{S_A}^{1/2}(\mathbf{W}^*)^\top\mathbf{W} - \mathbf{U_P}\mathbf{S_P}^{1/2}\mathbf{W}\right)_i = \sqrt{n}((\mathbf{A}-\mathbf{P})\mathbf{U_P}\mathbf{S_P}^{-1/2}\mathbf{W})_i + O(n^{-1/2}\log^{1/2} n).$$

Lemma 53 then establishes the asymptotic normality of $\sqrt{n}((\mathbf{A}-\mathbf{P})\mathbf{U_P}\mathbf{S_P}^{-1/2}\mathbf{W})_i$ as desired. ∎

We now turn out attention a brief sketch of the proof of the central limit theorem for the Laplacian spectral embedding.

A.3 Sketch of proof of Theorem 29

We present in this subsection a sketch of the main ideas in the proof of Theorem 29; detailed proofs are given in Tang and Priebe (2018). We first introduce some additional notation. For $(\mathbf{X}_n, \mathbf{A}_n) \sim \mathrm{RDPG}(F)$, let $\mathbf{T}_n = \mathrm{diag}(\mathbf{P}_n \mathbf{1})$ be the $n \times n$ diagonal matrices whose diagonal entries are the *expected* vertex degrees. Then defining $\tilde{\mathbf{X}}_n = \mathbf{T}_n^{-1/2} \mathbf{X}_n$, and noting that $\tilde{\mathbf{X}}_n \tilde{\mathbf{X}}_n^\top = \mathcal{L}(\mathbf{P}_n) = \mathbf{T}_n^{-1/2} \mathbf{P}_n \mathbf{T}_n^{-1/2}$, Theorem 29 depends on showing that there exists an orthogonal matrix $\mathbf{W}_n$ such that

$$\breve{\mathbf{X}}_n \mathbf{W}_n - \tilde{\mathbf{X}}_n = \mathbf{T}_n^{-1/2} (\mathbf{A}_n - \mathbf{P}_n) \mathbf{T}_n^{-1/2} \tilde{\mathbf{X}}_n (\tilde{\mathbf{X}}_n^\top \tilde{\mathbf{X}}_n)^{-1} + \tfrac{1}{2} (\mathbf{I} - \mathbf{D}_n \mathbf{T}_n^{-1}) \tilde{\mathbf{X}}_n + \mathbf{R}_n \qquad (66)$$

where $\|\mathbf{R}_n\|_F = O(n^{-1})$ with high probability. The motivation behind Eq. (66) is as follows. Given $\tilde{\mathbf{X}}_n$, the entries of the right hand side of Eq. (66), except for the term $\mathbf{R}_n$, can be expressed explicitly in terms of linear combinations of the entries $a_{ij} - p_{ij}$ of $\mathbf{A}_n - \mathbf{P}_n$. This is in contrast with the left hand side of Eq. (66), which depends on the quantities $\mathbf{U}_{\mathcal{L}(\mathbf{A}_n)}$ and $\mathbf{S}_{\mathcal{L}(\mathbf{A}_n)}$ (recall Definition 18). Since the quantities $\mathbf{U}_{\mathcal{L}(\mathbf{A}_n)}$ and $\mathbf{S}_{\mathcal{L}(\mathbf{A}_n)}$ can not be expressed explicitly in terms of the entries of $\mathbf{A}_n$ and $\mathbf{P}_n$, we conclude that the right hand side of Eq. (66) is simpler to analyze.

Once Eq. (66) is established, we can derive Theorem 29 as follows. Let ξ_i denote the i-th row of $n(\breve{\mathbf{X}}_n \mathbf{W}_n - \tilde{\mathbf{X}}_n)$ and let r_i denote the i-th row of $\mathbf{R}_n$. Eq. (66) then implies

$$\begin{aligned}
\xi_i &= (\tilde{\mathbf{X}}_n^\top \tilde{\mathbf{X}}_n)^{-1} \frac{n}{\sqrt{t_i}} \Big(\sum_j \frac{a_{ij} - p_{ij}}{\sqrt{t_j}} (\tilde{\mathbf{X}}_n)_j \Big) + \frac{n(t_i - d_i)}{2 t_i} (\tilde{\mathbf{X}}_n)_i + n r_i \\
&= (\tilde{\mathbf{X}}_n^\top \tilde{\mathbf{X}}_n)^{-1} \frac{\sqrt{n}}{\sqrt{t_i}} \Big(\sum_j \frac{\sqrt{n \rho_n}(a_{ij} - p_{ij})(\mathbf{X}_n)_j}{t_j} \Big) - \frac{n(\mathbf{X}_n)_i}{2 t_i^{3/2}} \sum_j (a_{ij} - p_{ij}) + n r_i \\
&= \frac{\sqrt{n}}{\sqrt{t_i}} \sum_j \frac{(a_{ij} - p_{ij})}{\sqrt{n}} \Big(\frac{(\tilde{\mathbf{X}}_n^\top \tilde{\mathbf{X}}_n)^{-1}(\mathbf{X}_n)_j}{t_j/n} - \frac{(\mathbf{X}_n)_i}{2 t_i/n} \Big) + n r_i
\end{aligned}$$

where a_{ij} and p_{ij} are the ij-th entries of $\mathbf{A}$ and $\mathbf{P}$, respectively, and t_i is the i-th diagonal entry of $\mathbf{T}_n$. We can then show that $n r_i \xrightarrow{\mathrm{d}} 0$. Indeed, there are n rows in $\mathbf{R}_n$ and $\|\mathbf{R}_n\|_F = O(n^{-1})$; hence, on average, for each index i, $\|r_i\|^2 = O(n^{-3})$ with high probability (a more precise argument similar to that used in proving Lemma 52 is needed to establish this rigorously). Furthermore,

$$t_i/n = \sum_j (\mathbf{X}_n)_i^\top (\mathbf{X}_n)_j / n \xrightarrow{\mathrm{a.s.}} (\mathbf{X}_n)_i^\top \boldsymbol{\mu}$$

as $n \to \infty$. Finally,

$$\tilde{\mathbf{X}}_n^\top \tilde{\mathbf{X}}_n = \sum_i \big((\mathbf{X}_n)_i (\mathbf{X}_n)_i^\top / (\sum_j (\mathbf{X}_n)_i^\top (\mathbf{X}_n)_j) \big),$$

and this can be shown to converge to $\tilde{\Delta} = \mathbb{E}\left[\frac{\mathbf{X}_1 \mathbf{X}_1^\top}{\mathbf{X}_1^\top \boldsymbol{\mu}}\right]$ as $n \to \infty$. We therefore have, after additional algebraic manipulations, that

$$\xi_i = \frac{\sqrt{n}}{\sqrt{t_i}} \sum_j \frac{(a_{ij} - p_{ij})}{\sqrt{n}} \left(\frac{\tilde{\Delta}^{-1}(\mathbf{X}_n)_j}{(\mathbf{X}_n)_j^\top \boldsymbol{\mu}} - \frac{(\mathbf{X}_n)_i}{2(\mathbf{X}_n)_i^\top \boldsymbol{\mu}}\right) + o(1)$$

$$= \frac{\sqrt{n}}{\sqrt{t_i}} \sum_j \frac{(a_{ij} - (\mathbf{X}_n)_i^\top (\mathbf{X}_n)_j)}{\sqrt{n}} \left(\frac{\tilde{\Delta}^{-1}(\mathbf{X}_n)_j}{(\mathbf{X}_n)_j^\top \boldsymbol{\mu}} - \frac{(\mathbf{X}_n)_i}{2(\mathbf{X}_n)_i^\top \boldsymbol{\mu}}\right) + o(1)$$

with high probability. Conditioning on $(\mathbf{X}_n)_i = \boldsymbol{x}$, the above expression for ξ_i is roughly a sum of independent and identically distributed mean 0 random variables. The multivariate central limit theorem can then be applied to the above expression for ξ_i, thereby yielding Theorem 29.

We now sketch the derivation of Eq. (66). For simplicity of notation, we shall ignore the subscript n in the matrices $\mathbf{A}_n$, $\mathbf{X}_n$, $\mathbf{P}_n$ and related matrices. First, consider the following expression.

$$\mathbf{U}_{\mathcal{L}(\mathbf{A})}\mathbf{S}_{\mathcal{L}(\mathbf{A})}^{1/2} - \mathbf{U}_{\mathcal{L}(\mathbf{P})}\mathbf{S}_{\mathcal{L}(\mathbf{P})}^{1/2}\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top \mathbf{U}_{\mathcal{L}(\mathbf{A})} = \mathcal{L}(\mathbf{A})\tilde{\mathbf{U}}_{\mathcal{L}(\mathbf{A})}\tilde{\mathbf{S}}_{\mathcal{L}(\mathbf{A})}^{-1/2} - \mathcal{L}(\mathbf{P})\tilde{\mathbf{U}}_{\mathcal{L}(\mathbf{P})}\tilde{\mathbf{S}}_{\mathcal{L}(\mathbf{P})}^{-1/2}\tilde{\mathbf{U}}_{\mathcal{L}(\mathbf{P})}^\top \tilde{\mathbf{U}}_{\mathcal{L}(\mathbf{A})}$$

$$= \mathcal{L}(\mathbf{A})\mathbf{U}_{\mathcal{L}(\mathbf{A})}\mathbf{U}_{\mathcal{L}(\mathbf{A})}^\top \mathbf{U}_{\mathcal{L}(\mathbf{A})}\mathbf{S}_{\mathcal{L}(\mathbf{A})}^{-1/2} - \mathcal{L}(\mathbf{P})\mathbf{U}_{\mathcal{L}(\mathbf{P})}\mathbf{S}_{\mathcal{L}(\mathbf{P})}^{-1/2}\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top \mathbf{U}_{\mathcal{L}(\mathbf{A})}$$
$$\tag{67}$$

Now $\mathcal{L}(\mathbf{A})$ is concentrated around $\mathcal{L}(\mathbf{P})$: namely, in the current setting,

$$\|\mathcal{L}(\mathbf{A}) - \mathcal{L}(\mathbf{P})\| = O(n^{-1/2})$$

with high probability (see Theorem 2 in Lu and Peng, 2013). Since $\|\mathcal{L}(\mathbf{P})\| = \Theta(1)$ and the non-zero eigenvalues of $\mathcal{L}(\mathbf{P})$ are all of order $\Theta(1)$, this again implies, by the Davis-Kahan theorem, that the eigenspace spanned by the d largest eigenvalues of $\mathcal{L}(\mathbf{A})$ is "close" to that spanned by the d largest eigenvalues of $\mathcal{L}(\mathbf{P})$. More precisely, $\|\mathbf{U}_{\mathcal{L}(\mathbf{A})}\mathbf{U}_{\mathcal{L}(\mathbf{A})} - \mathbf{U}_{\mathcal{L}(\mathbf{P})}\mathbf{U}_{\mathcal{L}(\mathbf{P})}\| = O(n^{-1/2})$ with high probability, and

$$\mathbf{U}_{\mathcal{L}(\mathbf{A})}\mathbf{S}_{\mathcal{L}(\mathbf{A})}^{1/2} - \mathbf{U}_{\mathcal{L}(\mathbf{P})}\mathbf{S}_{\mathcal{L}(\mathbf{P})}^{1/2}\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top \mathbf{U}_{\mathcal{L}(\mathbf{A})} = \mathcal{L}(\mathbf{A})\mathbf{U}_{\mathcal{L}(\mathbf{P})}\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top \mathbf{U}_{\mathcal{L}(\mathbf{A})}\mathbf{S}_{\mathcal{L}(\mathbf{A})}^{-1/2}$$
$$- \mathcal{L}(\mathbf{P})\mathbf{U}_{\mathcal{L}(\mathbf{P})}\mathbf{S}_{\mathcal{L}(\mathbf{P})}^{-1/2}\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top \mathbf{U}_{\mathcal{L}(\mathbf{A})} + \mathbf{R}_n$$

where $\|\mathbf{R}_n\| = O(n^{-1})$ with high probability. In addition, $\|\mathbf{U}_{\mathcal{L}(\mathbf{A})}\mathbf{U}_{\mathcal{L}(\mathbf{A})} - \mathbf{U}_{\mathcal{L}(\mathbf{P})}\mathbf{U}_{\mathcal{L}(\mathbf{P})}\| = O(n^{-1/2})$ also implies that there exists an orthogonal matrix $\mathbf{W}^*$ such that $\|\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top \mathbf{U}_{\mathcal{L}(\mathbf{A})} - \mathbf{W}^*\| = O(n^{-1})$ with high probability.

We next consider the terms $\mathbf{S}_{\mathcal{L}(\mathbf{P})}^{-1/2}\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top \mathbf{U}_{\mathcal{L}(\mathbf{A})}$ and $\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top \mathbf{U}_{\mathcal{L}(\mathbf{A})}\mathbf{S}_{\mathcal{L}(\mathbf{A})}^{-1/2}$. Note that the both are $d \times d$ matrices; furthermore, since $\mathbf{S}_{\mathcal{L}(\mathbf{A})}$ and $\mathbf{S}_{\mathcal{L}(\mathbf{P})}$ are diagonal matrices, the ij-th entry of $\mathbf{S}_{\mathcal{L}(\mathbf{P})}^{-1/2}\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top \mathbf{U}_{\mathcal{L}(\mathbf{A})} - \mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top \mathbf{U}_{\mathcal{L}(\mathbf{A})}\mathbf{S}_{\mathcal{L}(\mathbf{A})}^{-1/2}$ can be written as the $\zeta_{ij} \times h_{ij}$ where ζ_{ij} is the ij-th entry of $\mathbf{S}_{\mathcal{L}(\mathbf{P})}\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top \mathbf{U}_{\mathcal{L}(\mathbf{A})} - \mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top \mathbf{U}_{\mathcal{L}(\mathbf{A})}\mathbf{S}_{\mathcal{L}(\mathbf{A})}$ and the h_{ij} are functions of $\lambda_i(\mathcal{L}(\mathbf{A}))$

and $\lambda_j(\mathcal{L}(\mathbf{P}))$. In particular, $|h_{ij}| < C$ for some positive constant C for all n. We then have that

$$\mathbf{S}_{\mathcal{L}(\mathbf{P})}\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top\mathbf{U}_{\mathcal{L}(\mathbf{A})} - \mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top\mathbf{U}_{\mathcal{L}(\mathbf{A})}\mathbf{S}_{\mathcal{L}(\mathbf{A})} = \mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top(\mathcal{L}(\mathbf{P}) - \mathcal{L}(\mathbf{A}))\mathbf{U}_{\mathcal{L}(\mathbf{A})}$$
$$= \mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top(\mathcal{L}(\mathbf{P}) - \mathcal{L}(\mathbf{A}))\mathbf{U}_{\mathcal{L}(\mathbf{P})}\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top\mathbf{U}_{\mathcal{L}(\mathbf{A})}$$
$$+ \mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top((\mathcal{L}(\mathbf{P}) - \mathcal{L}(\mathbf{A}))(\mathbf{I} - \mathbf{U}_{\mathcal{L}(\mathbf{P})}\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top)\mathbf{U}_{\mathcal{L}(\mathbf{A})}$$

Now, conditioning on $\mathbf{P}$, the ij-th entry of $\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top(\mathcal{L}(\mathbf{P}) - \mathcal{L}(\mathbf{A}))\mathbf{U}_{\mathcal{L}(\mathbf{P})}$ can be written as a linear combination of the entries of $\mathbf{A} - \mathbf{P}$ (which are independent) and the rows of $\mathbf{X}$; hence, it can be bounded using Hoeffding's inequality. Meanwhile, the term

$$\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top((\mathcal{L}(\mathbf{P}) - \mathcal{L}(\mathbf{A}))(\mathbf{I} - \mathbf{U}_{\mathcal{L}(\mathbf{P})}\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top)\mathbf{U}_{\mathcal{L}(\mathbf{A})}$$

can be bounded by the Davis-Kahan Theorem and the spectral norm difference of $\|\mathcal{L}(\mathbf{A}) - \mathcal{L}(\mathbf{P})\|$. We therefore arrive at the important fact that

$$\|\mathbf{S}_{\mathcal{L}(\mathbf{P})}\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top\mathbf{U}_{\mathcal{L}(\mathbf{A})} - \mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top\mathbf{U}_{\mathcal{L}(\mathbf{A})}\mathbf{S}_{\mathcal{L}(\mathbf{A})}\|_F = O(n^{-1})$$

with high probability, and hence

$$\|\mathbf{S}_{\mathcal{L}(\mathbf{P})}^{-1/2}\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top\mathbf{U}_{\mathcal{L}(\mathbf{A})} - \mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top\mathbf{U}_{\mathcal{L}(\mathbf{A})}\mathbf{S}_{\mathcal{L}(\mathbf{A})}^{-1/2}\| = O(n^{-1})$$

with high probability.

We can juxtapose $\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top\mathbf{U}_{\mathcal{L}(\mathbf{A})}$ and $\mathbf{S}_{\mathcal{L}(\mathbf{A})}^{-1/2}$ in the expression for Eq. (67) and then replace $\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top\mathbf{U}_{\mathcal{L}(\mathbf{A})}$ by the orthogonal matrix $\mathbf{W}^*$, thereby obtaining

$$\mathbf{U}_{\mathcal{L}(\mathbf{A})}\mathbf{S}_{\mathcal{L}(\mathbf{A})}^{1/2} - \mathbf{U}_{\mathcal{L}(\mathbf{P})}\mathbf{S}_{\mathcal{L}(\mathbf{P})}^{1/2}\mathbf{W}^* = (\mathcal{L}(\mathbf{A}) - \mathcal{L}(\mathbf{P}))\mathbf{U}_{\mathcal{L}(\mathbf{P})}\mathbf{S}_{\mathcal{L}(\mathbf{P})}^{-1/2}\mathbf{W}^* + \tilde{\mathbf{R}}_n$$

where $\|\tilde{\mathbf{R}}_n\| = O(n^{-1})$ with high probability. Since

$$\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top = \mathcal{L}(\mathbf{P}) = \mathbf{U}_{\mathcal{L}(\mathbf{P})}\mathbf{S}_{\mathcal{L}(\mathbf{P})}\mathbf{U}_{\mathcal{L}(\mathbf{P})}^\top,$$

we have $\tilde{\mathbf{X}} = \mathbf{U}_{\mathcal{L}(\mathbf{P})}\mathbf{S}_{\mathcal{L}(\mathbf{P})}^{1/2}\tilde{\mathbf{W}}$ for some orthogonal matrix $\tilde{\mathbf{W}}$. Therefore,

$$\mathbf{U}_{\mathcal{L}(\mathbf{A})}\mathbf{S}_{\mathcal{L}(\mathbf{A})}^{1/2} - \tilde{\mathbf{X}}\tilde{\mathbf{W}}^\top\mathbf{W}^* = (\mathcal{L}(\mathbf{A}) - \mathcal{L}(\mathbf{P}))\mathbf{U}_{\mathcal{L}(\mathbf{P})}\mathbf{S}_{\mathcal{L}(\mathbf{P})}^{-1/2}\mathbf{W}^* + \tilde{\mathbf{R}}_n$$
$$= (\mathcal{L}(\mathbf{A}) - \mathcal{L}(\mathbf{P}))\mathbf{U}_{\mathcal{L}(\mathbf{P})}\mathbf{S}_{\mathcal{L}(\mathbf{P})}^{1/2}\tilde{\mathbf{W}}\tilde{\mathbf{W}}^\top\mathbf{S}_{\mathcal{L}(\mathbf{P})}^{-1}\tilde{\mathbf{W}}\tilde{\mathbf{W}}^\top\mathbf{W}^* + \tilde{\mathbf{R}}_n$$
$$= (\mathcal{L}(\mathbf{A}) - \mathcal{L}(\mathbf{P}))\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{W}}^\top\mathbf{W}^* + \tilde{\mathbf{R}}_n.$$

Equivalently,

$$\mathbf{U}_{\mathcal{L}(\mathbf{A})}\mathbf{S}_{\mathcal{L}(\mathbf{A})}^{1/2}(\mathbf{W}^*)^\top\tilde{\mathbf{W}} - \tilde{\mathbf{X}} = (\mathcal{L}(\mathbf{A}) - \mathcal{L}(\mathbf{P}))\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1} + \tilde{\mathbf{R}}_n(\mathbf{W}^*)^\top\tilde{\mathbf{W}}. \qquad (68)$$

The right hand side of Eq. (68) — except for the residual term $\tilde{\mathbf{R}}_n(\mathbf{W}^*)^\top\tilde{\mathbf{W}}$ which has norm of order $O(n^{-1})$ with high probability — can now be written explicitly in terms of the entries of $\mathbf{A}$ and $\mathbf{P}$. However, since

$$\mathcal{L}(\mathbf{A}) = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2},$$

the entries of the right hand side of Eq. (68) are neither linear nor affine combinations of the entries of $\mathbf{A} - \mathbf{P}$. Nevertheless, a Taylor-series expansion of the entries of $\mathbf{D}^{-1/2}$ allows us to conclude that

$$\|\mathbf{D}^{-1/2} - \mathbf{T}^{-1/2} - \tfrac{1}{2}\mathbf{T}^{-3/2}(\mathbf{T} - \mathbf{D})\| = O(n^{-3/2})$$

with high probability. Substituting this into Eq. (68) yields, after a few further algebraic simplifications, Eq. (66).

References

E. Abbe and C. Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Proceedings of the 56th Annual Symposium on Foundations of Computer Science*, pages 670–688, 2015.

E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62:471–487, 2016.

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research*, 9:1981–2014, 2008.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

S. M. Ali and S. D. Shelvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B.*, 28:121–132, 1966.

N. Alon and J. Spencer. *The Probabilistic Method*. John Wiley, 3 edition, 2008.

E. Arias-Castro and N. Verzelen. Community detection in dense random networks. *Annals of Statistics*, 42:940–969, 2014.

A. Athreya, V. Lyzinski, D. J. Marchette, C. E. Priebe, D. L. Sussman, and M. Tang. A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A*, 78:1–18, 2016.

R. Bhatia. *Matrix Analysis*. Springer, 1997.

P. Bickel and P. Sarkar. Role of normalization for spectral clustering in stochastic blockmodels. *Annals of Statistics*, 43:962–990, 2015.

P. Bickel, D. Choi, X. Chang, and H. Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Annals of Statistics*, 41:1922–1943, 2013.

P. J. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Sciences of the United States of America*, 106:21068–21073, 2009.

P. J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *Annals of Statistics*, 39:38–59, 2011.

B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures and Algorithms*, 31:3–122, 2007.

J. Cape, M. Tang, and C. E. Priebe. The Kato-Temple inequality and eigenvalue concentration. *Electronic Journal of Statistics*, 11:3954–3978, 2017a.

J. Cape, M. Tang, and C. E. Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. Arxiv preprint at `https://arxiv.org/abs/1705.10735`, 2017b.

S. Chatterjee. Matrix estimation by universal singular value thresholding. *Annals of Statistics*, 43:177–214, 2015.

L. Chen, J.T. Vogelstein, V. L. Lyzinski, and C. E. Priebe. A joint graph inference case study: the C. Elegans chemical and electrical connectomes. *Worm*, 5:e1142041, 2016.

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

H. Chernoff. Large sample theory: Parametric case. *Annals of Mathematical Statistics*, 27:1–22, 1956.

F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

K. L. Chung. *A course in probability theory*. Academic Press, 3 edition, 2001.

G. A. Coppersmith. Vertex nomination. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6:144–153, 2014.

I. Csizár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

C. Davis and W. Kahan. The rotation of eigenvectors by a pertubation. III. *SIAM Journal on Numerical Analysis*, 7:1–46, 1970.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.

K. Eichler, F. Li, A. L. Kumar, Y. Park, I. Andrade, C. Schneider-Mizell, T. Saumweber, A. Huser, D. Bonnery, B. Gerber, R. D. Fetter, J. W. Truman, C. E. Priebe, L. F. Abbott, A. Thum, M. Zlatic, and A. Cardona. The complete wiring diagram of a high-order learning and memory center, the insect mushroom body. *Nature*, 548:175–182, 2017.

P. Erdős and A. Rényi. On the evolution of random graphs. *Proceedings of the Hungarian Academy of Sciences*, pages 17–61, 1960.

M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23:298–305, 1973.

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

D. E. Fishkind, V. Lyzinski, H. Pao, L. Chen, and C. E. Priebe. Vertex nomination schemes for membership prediction. *Annals of Applied Statistics*, 9:1510–1532, 2015a.

D. E. Fishkind, C. Shen, and C. E. Priebe. On the incommensurability phenomenon. *Journal of Classification*, 33:185–209, 2015b.

S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

C. Fraley and A. E. Raftery. MCLUST: Software for model-based cluster analysis. *Journal of Classification*, 16:297–306, 1999.

C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

Z. Füredi and J. Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1:233–241, 1981.

E. N. Gilbert. Random graphs. *Annals of Statistics*, 30:1141–1144, 1959.

A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2:129–233, 2010.

A. Gretton, K. M. Borgwadt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

P. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Proceedings of the 20th Conference on Neural Information Processing Systems*, pages 657–664, 2007.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97:1090–1098, 2002.

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

R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

J. E. Jackson. *A User's Guide to Principal Components*. John Wiley & Sons, Inc., 2004.

A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:4–37, 2000.

B. Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83, 2011.

T. Kato. On the upper and lower bounds of eigenvalues. *Physical Review Letters*, 77: 334–339, 1950.

T. Kawamoto and Y. Kabashima. Limitations in the spectral method for graph partitioning: detectability threshold and localization of eigenvectors. *Physical Review E*, 91, 2015.

E. D. Kolaczyk. *Statistical Analysis of Network Data.* Springer-Verlag, 2009.

F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences of the United States of America*, 110:20935–20940, 2013.

B. A. Landman, A. J. Huang, A. Gifford, D. S. Vikram, I. A. Lim, J. A. Farrell, et al. Multi-parametric neuroimaging reproducibility: a 3-t resource study. *Neuroimage*, 54: 2854–2866, 2011.

C. Le, E. Levina, and R. Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 51:538–561, 2017.

C. C. Leang and D. H. Johnson. On the asymptotics of M-hypothesis bayesian detection. *IEEE Transactions on Information Theory*, 43:280–282, 1997.

J. Lei. A goodness-of-fit test for stochastic block models. *Annals of Statistics*, 44:401–424, 2016.

J. Lei and A. Rinaldo. Consistency of spectral clustering in stochastic blockmodels. *Annals of Statistics*, 43:215–237, 2015.

K. Levin and V. Lyzinski. Laplacian eigenmaps from sparse, noisy similarity measurements. *IEEE Transactions on Signal Processing*, 65:1988–2003, 2017.

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

F. Liese and I. Vadja. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52:4394–4412, 2006.

B. G. Lindsay. The geometry of mixture likelihoods: A general theory. *Annals of Statistics*, 11:86–94, 1983.

L. Lu and X. Peng. Spectra of edge-independent random graphs. *Electronic Journal of Combinatorics*, 20, 2013.

R. Lyons. Distance covariance in metric spaces. *Annals of Probability*, 41:3284–3305, 2013.

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

V. Lyzinski, D. E. Fishkind, M. Fiori, J. T. Vogelstein, C. E. Priebe, and G. Sapiro. Graph matching: Relax at your own risk. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:60–73, 2016a.

V. Lyzinski, K. Levin, D. Fishkind, and C. E. Priebe. On the consistency of the likelihood maximization vertex nomination scheme: Bridging the gap between maximum likelihood estimation and graph matching. *Journal of Machine Learning Research*, 17, 2016b.

V. Lyzinski, M. Tang, A. Athreya, Y. Park, and C. E. Priebe. Community detection and classification in hierarchical stochastic blockmodels. *IEEE Transactions in Network Science and Engineering*, 4:13–26, 2017.

J.-F. Maa, D. K. Pearl, and R. Bartoszyński. Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *Annals of Statistics*, 24:1067–1074, 1996.

D. Marchette, C. E. Priebe, and G. Coppersmith. Vertex nomination via attributed random dot product graphs. In *Proceedings of the 57th ISI World Statistics Congress*, 2011.

E. Mossel, J. Neeman, and A. Sly. Belief propagation, robust reconstruction and optimal recovery of block models. In *Proceedings of the 27th Annual Conference on Learning Theory*, pages 356–370, 2014.

E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *Combinatorica*, 2018. In press.

F. Mosteller and R. A. Fisher. Questions and answer. *The American Statistician*, 2:30–31, 1948.

M. E. J. Newman. Modularity and community structure in networks. *Proccedings of the National Academy of Sciences of the United States of America*, 103:8577–8582, 2006.

T. Ohyama, C. M. Schneider-Mizell, R. D. Fetter J. V. Aleman, R. Franconville, M. Rivera-Alba, B. D. Mensh, K. M. Branson, J. H. Simpson, J. W. Truman, A. Cardona, and M. Zlatic. A multilevel multimodal circuit enhances action selection in Drosophila. *Nature*, 520:633–639, 2015.

R. I. Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. `http://arxiv.org/abs/0911.0600`, 2009.

D. Pollard. Strong consistency of k-means clustering. *Annals of Statistics*, 9:135–140, 1981.

C. E. Priebe, D. L. Sussman, M. Tang, and J. T. Vogelstein. Statistical inference on errorfully observed graphs. *Journal of Computational and Graphical Statistics*, 24:930–953, 2015.

C. E. Priebe, Y. Park, M. Tang, A. Athreya, V. Lyzinski, J. T. Vogelstein, Y. Qin, B. Cocanougher, K. Eichler, M. Zlatic, and A. Cardona. Semiparametric spectral modeling of the drosophila connectome. arXiv preprint at `https://arxiv.org/abs/1705.03297`, 2017.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

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

W. G. Roncal, Z. H. Koterba, D. Mhembere, D. M. Kleissas, J. T. Vogelstein, R. Burns, et al. Migraine: MRI graph reliability analysis and inference for connectomics. Arxiv preprint at `http://arxiv.org/abs/1312.4875`, 2013.

P. Rubin-Delanchy, C. E. Priebe, and M. Tang. The generalised random dot product graph. Arxiv preprint available at `http://arxiv.org/abs/1709.05506`, 2017.

E. R. Scheinerman and K. Tucker. Modeling graphs using dot product representations. *Computational statistics*, 25:1–16, 2010.

C. M. Schneider-Mizell, S. Gerhard, M. Longair, T. Kazimiers, F. Li, M. F. Zwart, A. Champion, F. M. Midgley, R. D. Fetter, S. Saalfeld, and A. Cardona. Quantitative neuroanatomy for connectomics in Drosophila. *Elife*, 5:e12059, 2016.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41:2263–2291, 2013.

T. A. B. Snijders and K.Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14:75–100, 1997.

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and RKHS embeddings of measures. *Journal of Machine Learning Research*, 12: 2389–2410, 2011.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

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

D. L. Sussman, M. Tang, and C. E. Priebe. Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:48–57, 2014.

S. Suwan, D. S Lee, R. Tang, D.L. Sussman, M. Tang, and C.E. Priebe. Empirical Bayes estimation for the stochastic blockmodel. *Electronic Journal of Statistics*, 10:761–782, 2016.

G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143:1249–1272, 2013.

M. Tang and C. E. Priebe. Limit theorems for eigenvectors of the normalized laplacian for random graphs. *Annals of Statistics*, 2018. In press.

M. Tang, D. L. Sussman, and C. E. Priebe. Universally consistent vertex classification for latent position graphs. *Annals of Statistics*, 41:1406 – 1430, 2013.

M. Tang, A. Athreya, D. L. Sussman, V. Lyzinski, Y. Park, and C. E. Priebe. A semiparametric two-sample hypothesis testing problem for random dot product graphs. *Journal of Computational and Graphical Statistics*, 26:344–354, 2017a.

M. Tang, A. Athreya, D. L. Sussman, V. Lyzinski, and C. E. Priebe. A nonparametric two-sample hypothesis testing problem for random dot product graphs. *Bernoulli*, 23: 1599–1630, 2017b.

R. Tang, M. Ketcha, J. T. Vogelstein, C. E. Priebe, and D. L. Sussman. Laws of large graphs. arXiv preprint at `http://arxiv.org/abs/1609.01672`, 2016.

T. Tao and V. Vu. Random matrices: Universal properties of eigenvectors. *Random Matrices: Theory and Applications*, 1, 2012.

J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8:1–230, 2015.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

Y. J. Wang and G. Y. Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82:8–19, 1987.

P. J. Wolfe and S. C. Olhede. Nonparametric graphon estimation. arXiv preprint at `http://arxiv.org/abs/1309/5936`, 2013.

S. Young and E. Scheinerman. Random dot product graph models for social networks. In *Proceedings of the 5th international conference on algorithms and models for the web-graph*, pages 138–149, 2007.

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

D. Zheng, D. Mhembere, R. Burns, J. T. Vogelstein, C. E. Priebe, and A. S. Szalay. Flashgraph: Processing billion-node graphs on an array of commodity SSDs. In *Proceedings of the 13th USENIX Conference on File and Storage Technologies*, 2015.

M. Zhu and A. Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics and Data Analysis*, 51:918–930, 2006.