Statistical Connectomics

Jaewon Chung[1], Eric Bridgeford[1], Jesús Arroyo[1], Benjamin D. Pedigo[1], Ali Saad-Eldin[1], Vivek Gopalakrishnan[1], Liang Xiang[1], Carey E. Priebe[1], and Joshua Vogelstein[1*]

Abstract. The data science of networks is a rapidly developing field with myriad applications. In neuroscience, the brain is commonly modeled as a connectome, a network of nodes connected by edges. While there have been thousands of papers on connectomics, the statistics of networks remains limited and poorly understood. Here, we provide an overview from the perspective of statistical network science of the kinds of models, assumptions, problems, and applications that are theoretically and empirically justified for analysis of connectome data. We hope this review spurs further development and application of statistically grounded methods in connectomics.

Key words. Connectomics, networks, graphs, statistical models

1 Introduction The idea of the brain as a network of interconnected neuronal elements has existed since the late 19th century. These neuronal elements (e.g. long-range fibers, synapses, subcellular processes) are anatomically organized in multiple scales of space to allow communications over multiple scales of time enabling perception, cognition and action [79, 85, 90]. Recent advances in neuroimaging [14, 24, 47] along with large-scale projects opened new frameworks for studying the brain by modeling brain connectivity as networks, or connectomes [4, 103, 123]. One of the main challenges in connectomics is to understand the network structures that link individual histories, such as the genome, developmental stage, or experience, to cognitive phenotypes, such as personality traits, behaviors, or disorders, which has been dubbed "connectal coding" [107].

A connectome is defined as an abstract mathematical model of brain structure as a network, composed of two sets: vertices (or nodes) that represents a biophysical entity of the brain, and edges that represent connections, or communication, between pairs of vertices [47, 93, 107]. Connectomes can have additional structures. For example, edges can have weights that describe the strength of connection, and have other attributes, such as physical location of the edge. Similarly, nodes can also have attributes, such as anatomical labels, shape and size. This capacity of connectomes as a brain model comes with challenges in their analysis.

The first challenge is the choice of the representation of a connectome. Figure 1(A) and (B) shows two valid, but different representations of a human connectome. In Figure 1(A), the connectome is shown as a collection of vertices and edges in the classical graph theory perspective. The vertices are organized by their location in the human brain, but this is only one choice of layout. There are infinitely many layouts that are equally valid, and, potentially, useful. In Figure 1(B), the connectome is shown as a collection of numbers laid out in rows and columns as an "adjacency matrix" in the computer science perspective. In this view, a row/column pair is a vertex, and edges between vertices u and v are depicted by a non-zero entry in the corresponding element of the matrix. Consequently, the row identities are linked to column identities. Permuting both rows and columns together results in a "different" matrix, but they represent the same connectome. Nonetheless, the adjacency matrix is a useful representation of connectomes.

The second challenge is that connectomics data are different from typical Euclidean data in many ways. Some operations, such as addition and multiplication, are not well defined. What would it mean to add two connectomes together? Distance metrics are also not well defined, making comparisons between connectomes difficult. In the view of adjacency matrices, each entry is potentially related and dependent on other entries.

The third challenge is that connectomics data can be highly variable. For a graph with n vertices,

*Corresponding author; [1]Johns Hopkins University

there are $\binom{n}{2}$ possible edges so the number of unique graphs is $2^{\binom{n}{2}}$. Figure 1(C) shows the exponential growth in the number of unique graphs as the number of vertices increase. The large number of possible graphs makes characterizing and describing the graphs is difficult without statistical analysis of connectomics data.

Current connectomics analysis frameworks can be organized into four categories, each of which address the above challenges to various extents. The first approach, and by far the most popular, is dubbed the bag of features. In this approach, a set of graph-wise or vertex-wise statistics that capture the structural aspects of networks are computed and compared [17, 69]. One major drawback to this method is that features are not independent of one another, making results from subsequent inference using these features difficult to interpret. In the second approach, the bag of edges, each edge is studied individually. As a consequence, edges are treated independently, ignoring the other potential interactions [26, 106]. In the third approach, the bag of vertices, the vertices are studied while leveraging some structural information of the connectomes. In the fourth approach, the bag of communities, the vertices are first organized into (typically) disjoint groups to form communities, and then edges within and across communities are studied. The last approach, the bag of networks, studies the connectomes as a whole to test for differences across groups or to classify connectomes.

While each of the frameworks provide complementary and meaningful insights into the connectomes, the underlying methodologies, and, thus, the interpretation of results can vary significantly. Statistical modeling of connectomes bridges the gap by providing a unified framework for studying connectomes. Conceptually, statistical models capture important differences within or among networks while considering the built-in structures and heterogeneity in networks [7, 11, 119, 121]. These differences are summarized by model parameters that can be used in a variety of subsequent inference tasks.

This article is intended as a quantitative review of current connectomics analysis methods, and how statistical models can be incorporated to improve current analysis methods. We perform empirical investigations to demonstrate to what extent conclusions can be trusted as a function of the analysis method and the hypothesis in consideration. We vary parameters for the data, such as the generative model, sample size, and effect size, and hypothesis testing frameworks. Ultimately, the statistical modeling of networks uniquely provides a framework for meaningful and accurate testing and estimation for connectomics.

2 Representations

Due to the flexibility of networks, different representations of the connectomes can be studied, which we organize into four categories. In the following sections, we first formally define a network and then describe the four different frameworks of studying connectomics data. All frameworks provide complementary insights and understanding of the connectomes.

2.1 Graph/Network

A graph, or network, $\mathcal{G}$, is defined as an ordered set of vertices and edges (V, E) where V is the vertex set, and E, the set of edges, is a subset of the Cartesian product of $V \times V$. That is, a graph has at most a single edge for each pair of unique vertices. A vertex set is represented as $V = \{1, 2, \ldots, n\}$ where $|V| = n$, and an edge exists between vertices i and j if $(i, j) \in E$. An unweighted graph is a graph in which we are only concerned with the presence (or absence) of an edge. Each graph has an associated adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$ where $\mathbf{A}_{ij}$ represents the presence (or absence) of the edge between nodes i and j. Note that $\mathbf{A}$ provides a unique representation of $\mathcal{G}$; that is, there exists a 1-to-1 relationship between a graph and its adjacency matrix.

The above definition can be further extended in two ways:

1. Weighted graphs - the edges can take on arbitrary values, typically a real number. For example, the edge weight in human structural connectomes are non-negative integers that represent the number of estimated neuronal fibers that traverse from one region of the brain to another. Thus, each weighted graph has an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ where $\mathbf{A}_{ij}$ represents the edge weight.
2. Directed graphs - E is now an *ordered* set of edges. Each edge has an associated direction,

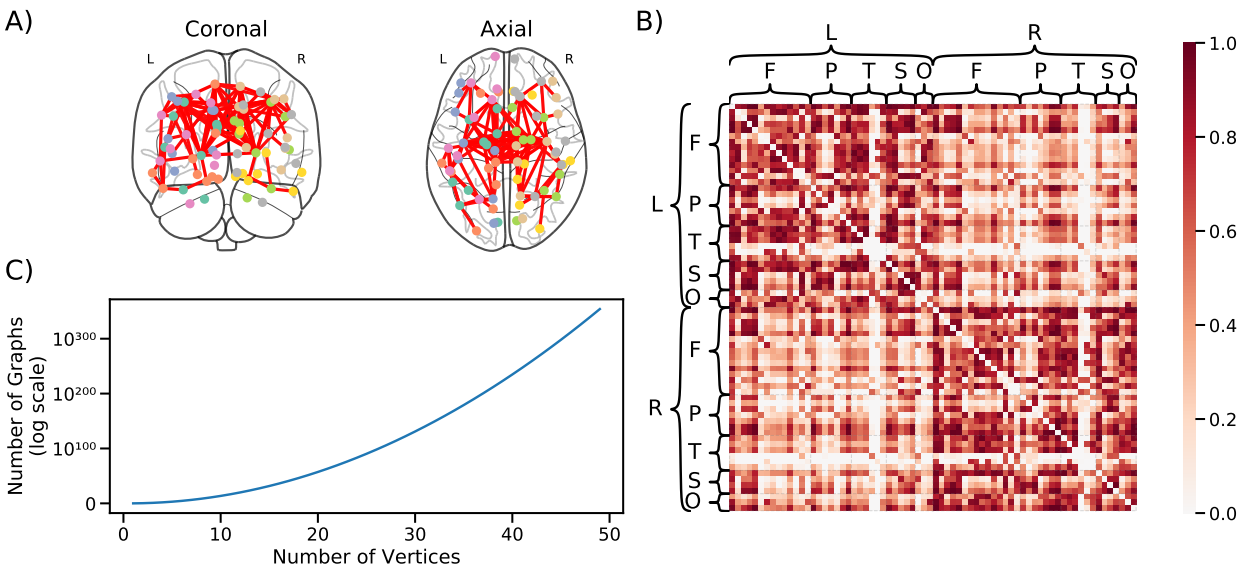


Figure 1: **Different Representations of a Connectome.** Human structural connectome estimated from averaging 1059 human connectomes from the Human Connectome Project [103]. Vertices represent regions of the brain, and are assigned into right (R) and left (L) hemispheres and then further assigned into frontal (F), occipital (O), parietal (P), and temporal (T), and subcortical structures (S). **(A)** Connectivity shown in the coronal and axial views. Dots corresponds to the center-of-mass of the a region, lines correspond to connections, and line thickness corresponds to magnitude of the connection. Only the largest 5% of edges are shown for visualization purposes. Note that infinitely many spatial arrangement of the vertices exist, and only one particular arrangement is being shown. **(B)** Connectivity of the average structural connectome shown as an adjacency matrix, $\mathbf{A}$. The rows and columns are organized by hemisphere then further organized by sub-structures. However, given any permutation matrix $\mathbf{P}$, the permuted adjacency matrix $\mathbf{P}\,\mathbf{A}\,\mathbf{P}^\top$ is still a valid matrix of original connectome. For a graph with n vertices, there are n^2 permutations. **(C)** The number of unique graphs grows exponentially as the number of vertices increases. The large number of graphs motivate statistical analysis to characterize and describe connectomes.

and a directed edge exists between vertices i and j if $(i, j) \in E$. In undirected graphs, the associated adjacency matrix $\mathbf{A}$ is symmetric, but in directed graphs, $\mathbf{A}$ is not necessarily symmetric, that is, it is possible that $\mathbf{A}_{ij} \neq \mathbf{A}_{ji}$, for any $i, j \in V$.

For the remainder of the paper, graphs are considered undirected and unweighted and with no self-loops, that is $\text{diag}(\mathbf{A}) = \vec{0}$, unless specified otherwise.

2.2 Bag of Features Network statistics, or features, are abstract representations that capture either global or local structures of a network [69, 75]. This method computes a set of network statistics for each network, and analyzes differences between, or among, populations. For example, when comparing populations of networks from healthy and individuals with depression, the difference in global clustering coefficient, which measures how likely vertices tend to cluster together, can be computed [19]. These network statistics have enjoyed applications in many connectomics studies that compare different populations of networks [18, 41]. However, there are infinitely number of such statistics, and we lack general guidance in which statistics to compute. Furthermore, no set of network statistics can adequately characterize a network [22, 67]. These considerable shortcomings further motivates the use of other representations of networks, and below examples demonstrate the shortcomings of studying bags of features.

2.2.1 Non-identifiability of graph features Summary statistics, such as the mean, variance, and correlation, are often used to describe real valued datasets, which can be insightful in understanding

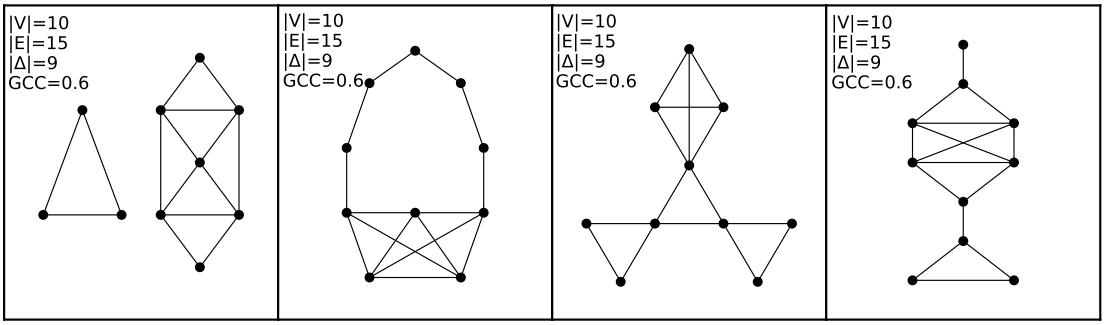


Figure 2: **Four Networks with Same Network Statistics.** Each network has $|V| = 10$, $|E| = 15$, number of triangles is 9, and the global clustering coefficient is 0.6. However these graphs have distinctive topologies. For example, the left-most network is disconnected, while others are connected. This suggest that given a small set of network statistics, one cannot identify from which network the features are computed.

the data. However, the Anscombe's quartet illustrates four drastically different distributions of eleven points that have the same summary statistics [6]. This suggest that any small number of summary statistics can fail to meaningfully characterize the data.

In network analysis, variety of network level statistics can be computed to summarize networks. Similar to the Anscombe's quartet, networks with different topologies can have the same network features as shown in Figure 2. These four networks have the same number of vertices, edges, triangles, and global clustering coefficient, but have different properties such as connectedness and symmetry. Other works have also explored the distributions of network statistics [22, 67].

2.2.2 Network features are correlated and relatively uninformative We consider all non-isomorphic, undirected, binary networks with 10 vertices, which results in ≈ 12 million networks. Formally, $\mathcal{G}$ and $\mathcal{H}$ are isomorphic networks when there exists a vertex permutation function $f : V(\mathcal{G}) \to V(\mathcal{H})$ such that if edge $(u, v) \in E(\mathcal{G})$, then $(f(u), f(v)) \in E(\mathcal{H})$. Only non-isomorphic networks are considered since isomorphic networks have identical network features.

For each network, the following six graph network statistics are computed: 1) average path length (APL), 2) global clustering coefficient (GCC), 3) average clustering coefficient (ACC), 4) global efficiency (GE), 5) local efficiency (LE), and 6) modularity. These statistics are some of the most commonly computed statistics [19, 93]. The distribution of network statistics are plotted against modularity. The top row of Figure 3 shows that all of the network features are highly correlated with modularity. We then constrain the networks in two different ways. First, we consider all networks with 20 ± 2 edges. Second, we choose a "base" network at random with 20 edges, and then identify all networks with no more than 3 edges different from the base network. The distribution of each of the above network statistics on this subset of networks are computed for both constraints. The middle and bottom rows of Figure 3 show that constraining the networks in these ways hardly constrains the network features at all. Changing only a few edges on a network can yield a network with almost any possible configuration according to these statistics, and therefore are inadequate to characterize these populations. Thus, when any given metric is correlated with a covariate of interest, so are many other metrics. Thus, claiming that a particular property of the brain "explains" a given phenotypic property of a person is spurious reasoning.

The experiment is repeated using the binarized structural connectomes from HCP dataset. For all 1059 connectomes, which have 70 vertices, the network features are computed. Figure 4 *top row* shows the distributions for all connectomes, and *middle row* and *bottom row* show the distributions after constraining by considering all connectomes with number of edges between 1010 and 1210, and then by choosing a network with 1100edges at random and choosing all networks with at most 300

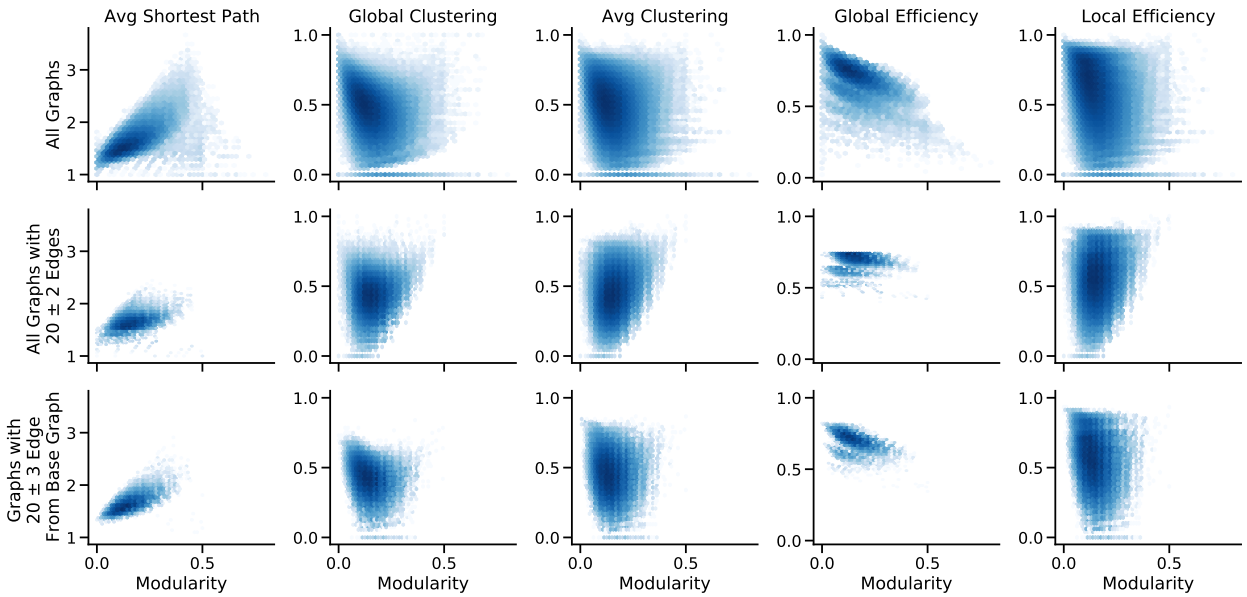


Figure 3: **Density Plots of Network Statistics.** *(Top Row)* The distributions of networks statistics for all possible 10-node networks are shown. *(Middle Row)* Networks are constrained by only considering all networks with 20 ± 2 edges. *(Bottom Row)* A base graph with 20 edges is chosen at random, and only networks that have differences up to 3 edges are considered. In both constrained set of networks, the distribution of these network statistics remains essentially unchanged. In other words, changing only a few edges on a network can yield a network with almost any possible configuration according to these statistics.

edge differences. Even in real data, constraining the networks produce similar distributions of network statistics.

2.3 Bag of Edges In this approach, the edges of connectomes are studied. Most commonly, each edge is studied independently, while ignoring any interactions between edges [26, 105, 120]. Univariate edge-wise testing can reveal easily interpretable relationships between specific edges and covariates through hypothesis testing. However, edge-wise testing requires performing multiple hypothesis tests, and multiple comparisons must be corrected to control the false positive rate [31, 40]. While certain methods, such as Benjamini–Hochberg corrections, have strong theoretical guarantees, they require assumptions about the data, such as independence, that connectomics data do not satisfy [13, 91, 118]. On the other hand, Bonferroni corrections are considered too conservative, and, therefore, lack the sensitivity for connectomics [91].

More intricate methods represent each connectome as a long vector containing all of its edges [5, 78]. Vector representations can allow for correlation of edges and direct application of common machine learning algorithms, but still discards the structural information in networks.

2.4 Bag of Vertices In this approach, the vertices of connectomes are analyzed while leveraging structural information, typically global structures, of the graphs. A common approach embeds the connectomes to learn a low-dimensional and Euclidean representation of the vertices [7, 11, 45]. Algorithms that operate on Euclidean data (e.g. Gaussian Mixture Model (`GMM`) for clustering vertices, random forests for classifying vertices, multivariate hypothesis tests for testing for differences between vertices) can be employed for subsequent analysis [76, 101].

2.5 Bag of Communities Networks often contain structural information such as communities, which are subsets of vertices that behave similarly. For example, similar vertices can be defined by those that are more likely to be connected with each other than to other vertices. The set of communities

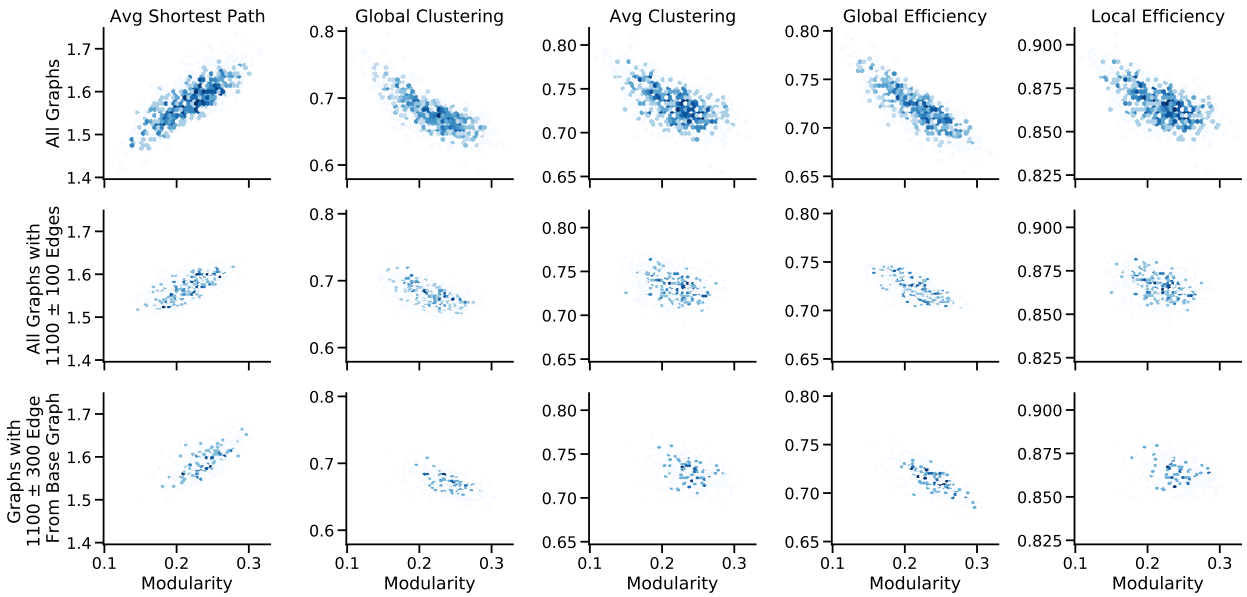


Figure 4: **Density plots of network statistics on HCP connectomes.** All connectomes ($N = 1059$) have 70 vertices defined by the Deskian parcellation. *(Top row)* The distributions of networks statistics for all HCP connectomes are shown. *(Middle Row)* connectomes are constrained by only considering all networks with 1100 ± 100 edges. and *(Bottom Row)* A base graph with 1100 edges is chosen at random, and only networks that have differences up to 300 edges are considered. Similar to the simulated examples, the distributions are qualitatively similar.

that comprise a network, called community structure, can describe both the local and global patterns of the network. At local-scale, we can examine the properties of vertices that are within the same community. At global-scale, we can measure associations between connectivity patterns of communities across groups or other covariates [8, 34, 54]. Furthermore, the community structure in spatial resolution connectomes from human MRI can be used to delineate regions of the brain, called parcellations [102].

Community detection in networks have been studied extensively [37, 70]. Typically, the community structure is identified by modularity optimization methods [15, 25]. In this paper, we present spectral methods that rely on statistical models for community detection, which have strong statistical guarantees for recovering true communities [7, 11, 64, 94]. It is important to note that analysis of communities depends on the performance of the community detection algorithms.

2.6 Bag of Networks In this approach, one or more groups of networks are studied in various settings, such as one- and two-sample hypothesis testing, and classification, using some representation of networks. For example, bag of vertices representation can be used to test whether two networks are different [97, 98]. For studying more than two networks, geometry in the space of the networks is defined and are represented in that geometry, which are then used for finding differences across groups [7, 42, 116].

Another group of methods finds subsets of vertices, or a subgraph, that contain the most information about certain covariates [9, 46, 109, 110, 113]. Estimating signal subgraphs is useful since networks can be extremely large (i.e. millions of vertices), which present computational challenges, and can potentially improve the performance of subsequent inference tasks, such as classification. Different approaches for finding the subgraph have been proposed, but all approaches leverage the network topologies inherent in connectomics data.

Table 1: Notations and symbols used in this paper

Symbols	Description	Symbols	Description
$[n]$	$\{1, 2, \ldots, n\}$	$\mathbf{P}$	Edge connectivity probability matrix
$\mathcal{G}$	Graph	$\mathbf{B}$	Block connectivity probability matrix
n	Number of nodes	$\vec{\tau}$	Vertex community assignment vector
$\mathbf{A}$	Adjacency matrix	$\mathbf{M}$	Edge community assignment matrix
$\mathbf{A}_i$	i-th row of $\mathbf{A}$	$\mathbf{X}$	Latent position matrix
$\mathbf{A}_{ij}$	(i, j) entry of $\mathbf{A}$	$\hat{\mathbf{X}}$	Estimated latent position matrix
$\mathbf{A}^{(l)}$	l-th element in sequence of $\mathbf{A}$		

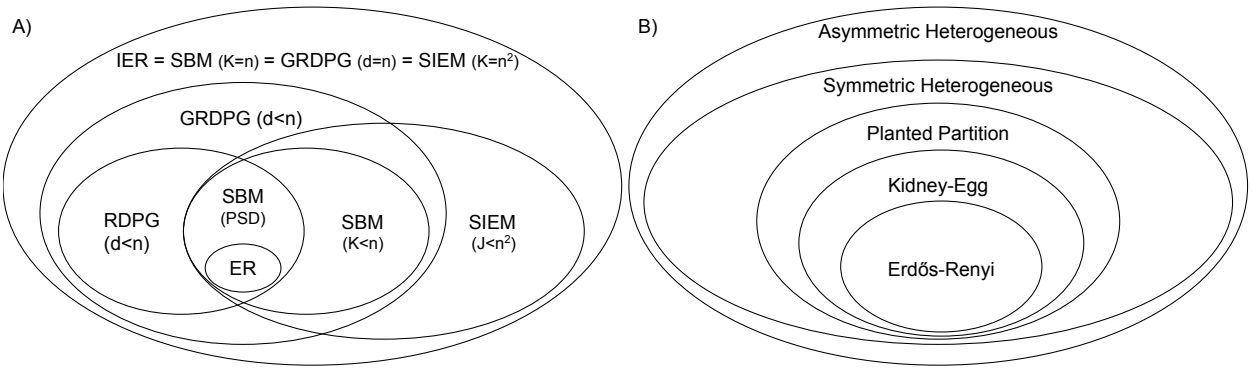


Figure 5: **Hierarchical Relationships of Statistical Models (A)** Relationships among all the single-graph statistical models. ER is a SBM with one community. SBM with a positive semidefinite block probability matrix $\mathbf{B}$ is also a RDPG. Any SBM, RDPG, and some SIEM can be represented as d-dimensional GRDPG with d less than number of vertices n. IER graphs are equivalent to a n-block SBM, n-dimensional GRDPG, and n^2-group SIEM. **(B)** Relationships among the two-block SBM models. The most complex model is the asymmetric heterogeneous SBM, and the simplest model is the Erdös – Rényi (ER), which is a degenerate case of 2-block SBM.

3 Statistical Models

Connectomes can be modelled using statistical models designed for network data [43, 56]. Statistical models consider the entire network as a random variable, including the inherent structure, dependencies within networks, and the noise in observed data. Thus, statistical models can formalize detecting similarities or differences for each of the representations in Section 2. This section provides an overview of many statistical models for network data, including those designed for representing single and multiple networks.

Section 3.1 provides an overview of single graph models that have been extensively studied as well as recently introduced models in the order of least to greatest complexity. Figure 5 shows the relationship between all the single graph models presented in this paper. Section 3.2 provides an overview of some models for multiple networks. While other statistical models for multiple network data exist [30, 72, 112, 119], we focus on some recent models that are used in spectral inference for connectomics data.

3.1 Single Graph Models

3.1.1 Erdős-Rényi Random Graphs (ER)

The simplest random graph model is the Erdös – Rényi (ER) model [33]. For a given set of n vertices, each distinct pair of vertices are connected independently with probability $p \in [0, 1]$. Specifically, $\mathbf{A} \sim \text{ER}_n(p)$ if $\mathbf{A}$ has entries $\mathbf{A}_{ij} \sim \text{Bernoulli}(p)$ for $i, j \in [n]$. While the ER model is not representative of real data, it has been studied extensively since many of its properties can be solved exactly [71, 84].

3.1.2 Stochastic Block Model (SBM) First introduced in [49], SBM is a model that can produce graphs with vertices grouped into K communities [81, 94, 114]. There are two simple variations of the SBM in which the vertex assignment vector $\vec{\tau} \in \{1, \cdots, K\}^n$ is known *a priori*, and where $\vec{\tau}$ is not known. In both cases, a symmetric $K \times K$ block connectivity probability matrix $\mathbf{B}$ with entries in $[0,1]^{K \times K}$ governs the probability of an edge between vertices given their block memberships.

If $\vec{\tau} \in \{1, \cdots, K\}^n$ is known *a priori*, the *a priori* SBM is parametrized only by the block connectivity matrix $\mathbf{B}$, and the model is $\mathbf{A} \sim \text{SBM}_n(\vec{\tau}, \mathbf{B})$ if $\mathbf{A}$ has entries $\mathbf{A}_{ij} \sim \text{Bernoulli}(\mathbf{B}_{kl})$ where $\tau_i = k, \tau_j = l$, for $i, j \in [n]$, and $k, l \in [K]$. In the case where $\vec{\tau}$ is not known, the *a posteriori* SBM is additionally parameterized by a block membership probability vector $\vec{\pi} = [\pi_1, \ldots, \pi_K]^\top$ on the probability simplex. The model is $\mathbf{A} \sim \text{SBM}_n(\vec{\pi}, \mathbf{B})$ if $\mathbf{A}$ has entries $\mathbf{A}_{ij} \big| k = \tau_i, l = \tau_j \sim \text{Bernoulli}(\mathbf{B}_{kl})$, where $\tau_i \sim \text{Multinomial}(\vec{\pi})$ for $i = 1, \cdots, n$.

Throughout the context of this paper, we will focus particularly on a few variations of the two-block SBM ($K = 2$) with block connectivity matrix $\mathbf{B} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, abbreviated as $\mathbf{B} = [a, b; c, d]$. The common variants include:

1. Kidney – Egg: $b = c = d$. In this model, one of the blocks has edges with a different probability than the others, but the remaining blocks are homogeneous, where $a \neq b$. Furthermore, when $b > a$, the model is referred to as core-periphery SBM.
2. Planted Partition: $a = d$ and $b = c$. In this model, the within-block edges share a common probability a, and the between-block edges share a common probability b, where $a \neq b$.
3. Symmetric Heterogeneous: $b = c$. In this model, the between-block edges share a common probability b, but the within-block edges have a disparate probabilities, where $a \neq b \neq d$.
4. Asymmetric Heterogeneous: $a \neq b \neq c \neq d$. In this directed model, every block has a unique probability.
5. Erdös – Rényi: $a = b = c = d$. In this degenerate model, all blocks have a common probability, and the partitioning is irrelevant.
6. Homophilic/Assortative/Affinity: $a, d > b, c$. In this model, the within-block probabilities are greater than cross-block probabilities.
7. Disassortative: $b, c > a, d$. In this model, the cross-block probabilities are greater than the within-block probabilities.

Figure 5**(B)** summarizes the relationships of SBM models.

3.1.3 Structured Independent Edge Model (SIEM) SIEM is a generalization of SBM that produces graphs in which edges are grouped into one of K clusters. Analogous to the vertex assignment vector of the *a priori* SBM, the SIEM features an edge community assignment matrix $\mathbf{M} \in \{1, \cdots, K\}^{n \times n}$ which is known *a priori*. Given the community assignment matrix $\mathbf{M}$, the SIEM is $\mathbf{A} \sim \text{SIEM}_n(\mathbf{M}, \vec{p})$ if $\mathbf{A}_{ij} \sim \text{Bernoulli}(p_k)$ where $\mathbf{M}_{ij} = k$, for $i, j \in [n]$ and $k \in [K]$. $\vec{p} = [p_1, \cdots, p_K]^\top \in [0,1]^K$ is the edge probability vector which governs the probability of an edge between vertices.

The *a priori* SBM is a special case of SIEM in which edges are assigned to blocks $\mathbf{M}$ which respect the vertex assignment vector $\vec{\tau}$. For the purposes of this paper, we will consider a case that frequently comes up in neuroimaging, the Homotopic SIEM, in which each vertex has a matched "pair" amongst other vertices. The edges corresponding to a pair $\mathbf{M}_{ij} = 2$ where (v_i, v_j) are a pair of vertices sharing a property, and the edges corresponding to a non-pair are $\mathbf{M}_{ij} = 1$. A matched pair of vertices, for instance, could be homotopic brain regions (two brain regions with similar function but in opposing hemispheres of the brain).

3.1.4 Random Dot Product Graphs (RDPG) RDPG belongs to the class of latent position random graphs [48]. In a latent position graph, every vertex has associated to it a (typically unobserved) *latent position* in some space $\mathcal{X}$, and the probability of connection between vertices i and j are given by a link function. In RDPG, the space $\mathcal{X}$ is a constrained subspace of Euclidean space $\mathbb{R}^d$ and the link

function is the dot product [87, 95, 117]. Thus, in a d–dimensional RDPG with n vertices, the matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ whose rows are the latent positions, and the matrix of connection probabilities is given by $\mathbf{P} = \mathbf{X}\mathbf{X}^\top$, which is positive semidefinite. The model is $\mathbf{A} \sim \text{RDPG}_n(\mathbf{X})$ if the adjacency matrix $\mathbf{A}$ has entries $\mathbf{A}_{ij} \sim \text{Bernoulli}(\mathbf{X}_i \mathbf{X}_j^\top)$. Subsequent inference tasks include community detection [94], vertex classification [100], or two-sample hypothesis testing for graphs with matched and non-matched vertices for a pair of graphs [77, 97, 98].

The RDPG is a flexible model, and other models of interest can be seen as special cases of the RDPG. A SBM whose block connectivity matrix $\mathbf{B}$ is positive semi-definite is a RDPG with K distinct latent positions. Thus, a SBM with K blocks can be represented with a latent position matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, with $d \leq K$, where there are only K different rows of $\mathbf{X}$, and letting $\mathbf{X}_{\mathcal{U}} \in \mathbb{R}^{K \times d}$ be the matrix with the subset of the rows $\mathcal{U}$ where each row is the latent position for a block, then the block connectivity matrix is $\mathbf{B} = \mathbf{X}_{\mathcal{U}} \mathbf{X}_{\mathcal{U}}^\top \in \mathbb{R}^{K \times K}$. More generally, the RDPG can represent other models with more complex structures, such as mixed memberships [2] or hierarchical communities [65].

3.1.5 Generalized Random Dot Product Graphs (GRDPG) Unlike RDPG model, GRDPG does not assume that $\mathbf{P}$ is a positive semidefinite probability matrix [83]. In this model, the edge probability matrix is given by $\mathbf{P} = \mathbf{X}\mathbf{I}_{pq}\mathbf{X}^\top$, and $\mathbf{A} \sim \text{GRDPG}_n(\mathbf{X}, p, q)$ if $\mathbf{A}_{ij} \sim \text{Bernoulli}(\mathbf{X}_i \mathbf{I}_{pq} \mathbf{X}_j^\top)$ where $\mathbf{I}_{pq} = \text{diag}(1, \ldots, 1, -1, \ldots, -1)$ with p ones followed by q minus ones on its diagonal, and where $p \geq 1$ and $q \geq 0$ are two integers satisfying $p + q = d$.

The GRDPG generalizes all of the previous models. When $q = 0$, GRDPG reduces to a RDPG model. To represent any SBM as GRDPG, let $p \geq 1, q \geq 0$ be the number of positive and negative eigenvalues of the block connectivity matrix $\mathbf{B} \in \mathbb{R}^{K \times K}$, respectively. The block matrix can be represented as $\mathbf{B} = \mathbf{X}_{\mathcal{U}} \mathbf{I}_{pq} \mathbf{X}_{\mathcal{U}}^\top$.

3.1.6 Inhomogenous Erdős-Rényi Random Graphs (IER) The Inhomogenous Erdös – Rényi (IER) is a model where each pair of nodes has a unique probability of an edge existing between the two, and is therefore the most general independent edge model. For a given set of n vertices, the IER is parametrized by a matrix $\mathbf{P} \in [0,1]^{n \times n}$, where $\mathbf{P}_{ij}$ is the probability of an edge connecting vertices v_i, v_j where $i, j \in [n]$. That is, $\mathbf{A} \sim \text{IER}_n(\mathbf{P})$ if $\mathbf{A}$ has entries $\mathbf{A}_{ij} \sim \text{Bernoulli}(\mathbf{P}_{ij})$ for $i, j \in [n]$. IER cannot be estimated from a single graph, as there are $\binom{n}{2}$ unknowns (the probabilities) with $\binom{n}{2}$ total observations (the edges).

Note that all single graph models are special cases of IER. Additionally, SBM with $K = n$, SIEM with $K = n^2$, and GRDPG with $d = n$ are equivalent to an IER model.

3.2 Multiple Graph Models A common idea in statistical models for multiple graphs is a shared latent space that contain structural information common to all graphs. The two models presented in this section constrain the shared latent space in different ways to describe the heterogeneity in graphs, which results in sensitivity to different kinds of heterogeneity. The advantages and disadvantages of each model are highlighted in Section 9.

In the following models, consider a sample of m observed graphs $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \ldots, \mathcal{G}^{(m)}$ and their associated adjacency matrices, $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(m)} \in \mathbb{R}^{n \times n}$ with n vertices that are identical and shared across all graphs.

3.2.1 Joint Random Dot Product Graphs (JRDPG) In this model, we consider a collection of m RDPGs all with the same generating latent positions. Similar to a RDPG, given an appropriately constrained Euclidean subspace $\mathbb{R}^d$, the model is parameterized by a latent positions matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ where $d \ll n$. The model is $\left(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(m)}\right) \sim \text{JRDPG}(\mathbf{X})$ where $\mathbf{A}_{ij}^{(l)} \sim \text{Bernoulli}(\mathbf{X}_i \mathbf{X}_j^\top)$ for all $i, j \in [n]$ and $l \in [m]$. Each graph has marginal distribution $\mathbf{A}^{(l)} \sim \text{RDPG}(\mathbf{X})$ for all $l \in [m]$, meaning that the matrices $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}$ are conditionally independent given $\mathbf{X}$ [11, 58]. While the model assumes that the latent positions for the graphs are the same, we note that this assumption is likely violated in heterogeneous networks, but still remains a very useful model as shown in Section 9.

3.2.2 Common Subspace Independent-Edge Model (COSIE) In this model, the heterogeneous networks are described via a shared latent structure on the vertices, but also permits sufficient heterogeneity via individual matrices for each graph [7]. The model is parameterized by a matrix $\mathbf{V} \in \mathbb{R}^{n \times d}$ with orthonormal columns, where n is the number of vertices and $d \ll n$, and symmetric individual score matrices $\mathbf{R}^{(i)} \in \mathbb{R}^{d \times d}$. The matrix $\mathbf{V}$ characterizes a low-rank common subspace, and is related to the latent positions for the vertices, and the score matrices incorporate individual differences to model the heterogeneity of the graphs. The model is denoted by $\left(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}\right) \sim \text{COSIE}(\mathbf{V}; \mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)})$ where $\mathbf{A}^{(l)}_{ij} \sim \text{Bernoulli}(\mathbf{P}^{(l)}_{ij})$ for all $i, j \in [n], i < j$, and $\mathbf{P}^{(l)} = \mathbf{V}\,\mathbf{R}^{(l)}\,\mathbf{V}^\top$. This factorization of the expected adjacency matrices is related to other decompositions for multiple matrices into population singular vectors or eigenvectors and individual parameters [1, 27, 59, 111].

3.2.3 Correlated Models Finally, we are interested in graph models for a pair of graphs, $\mathcal{G}_1$ and $\mathcal{G}_2$, where the two graphs are said to be correlated; that is, the edges adjoining incident vertices have a non-zero correlation. Correlated graph models have numerous applications, such as when a graph is estimated repeatedly for the same source at different points in time.

Correlated $(\mathbf{P}, \mathbf{Q})$ The $\mathbf{R}$-correlated $(\mathbf{P}, \mathbf{Q})$ model [61] with parameters $\mathbf{R}, \mathbf{P}, \mathbf{Q} \in [0, 1]^{n \times n}$, denoted as $\text{CorrER}(\mathbf{P}, \mathbf{Q}, \mathbf{R})$, produces two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ with adjacency matrices $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}$ such that each graph is marginally an inhomogeneous Erdős-Rényi with $\mathbf{A}^{(1)} \sim \text{IER}(\mathbf{P})$, $\mathbf{A}^{(2)} \sim \text{IER}(\mathbf{Q})$, but the pairs of corresponding edges have Pearson correlation encoded in the matrix $\mathbf{R}$ such that

$$\mathbf{R}_{ij} = \text{Corr}(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}) = \frac{\mathbb{P}(\mathbf{A}^{(1)}_{ij} = \mathbf{A}^{(2)}_{ij} = 1) - \mathbf{P}_{ij}\,\mathbf{Q}_{ij}}{\sqrt{\mathbf{P}_{ij}(1 - \mathbf{P}_{ij})\,\mathbf{Q}_{ij}(1 - \mathbf{Q}_{ij})}}.$$

When $\mathbf{P}$ and $\mathbf{Q}$ are different, there are restrictions in the values that the correlation matrix $\mathbf{R}$ can take. In particular, if $\mathbf{P}_{ij} \neq \mathbf{Q}_{ij}$ and $\mathbf{P} > \mathbf{Q}$, then $\mathbf{R}_{ij} \leq \sqrt{\frac{\mathbf{Q}_{ij}(1-\mathbf{Q}_{ij})}{\mathbf{P}_{ij}(1-\mathbf{P}_{ij})}}$ [61].

We are interested particularly in two special cases of the $\text{CorrER}(\mathbf{P}, \boldsymbol{Q}, \boldsymbol{R})$:

1. The ρ-correlated RDPG model arises when $\mathbf{P} = \mathbf{Q} = \mathbf{X}\,\mathbf{X}^\top$ for some latent position matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ as in Section 3.1.4, and $\boldsymbol{R} = \rho \mathbf{1}_{n \times n}$ (that is, the matrix of edge correlations $\boldsymbol{R}$ has only a single unique entry $\rho \geq 0$). We say that $\mathbf{A}_1, \mathbf{A}_2 \sim \rho\,\text{RDPG}(\mathbf{X})$.
2. The ρ-correlated ER model arises in the case where $\mathbf{P} = \mathbf{Q} = p\mathbf{1}_{n \times n}$ (i.e., the probability matrix has a single unique entry $p > 0$), and $\boldsymbol{R} = \rho \mathbf{1}_{n \times n}$ (as above, the matrix of correlations has a single unique entry). We say that $\mathbf{A}_1, \mathbf{A}_2 \sim \rho\,\text{ER}(p)$.

4 Model Extensions

4.1 Weighted Models The single graph models in Section 3.1 can be extended to weighted graphs trivially. For example, in the *a priori* SBM, each distinct community of edges within the graph, simply take the corresponding entries of the adjacency matrix $\mathbf{A}_{ij}$ to take distribution F_{ij} with parameters $\boldsymbol{\theta}_{ij}$. Adding additional structure to F_{ij} allows the parameters $\theta_{i,j}$ to be estimable for a single graph. In particular, we will be concerned with the Truncated-Normal SBM, where:

$$\mathbf{A}_{ij}; \vec{\tau}_i = k, \vec{\tau}_j = l, \boldsymbol{\theta}_{kl} \overset{ind}{\sim} \text{TN}(\boldsymbol{\theta}_{kl})$$

Where $\boldsymbol{\theta}_{kl} = (\mu_{k,l}, \sigma^2_{k,l}, \min_{k,l}, \max_{k,l})$ are the parameters associated with the k, l block of edge weights.

4.2 Degree-Corrected Models In the standard SBM defined in Section 3.1.2, the degree of a vertex, or the expected number of edges incident to a vertex, is constant within each block. Thus, vertices with same block assignment are stochastically equivalent to each other, which can limit practical applications [52]. In degree – corrected SBM (DCSBM), there is an additional "promiscuity" parameter that allows vertices within blocks to have heterogeneous expected degree distributions.

Similar to the standard SBM, the *a priori* DCSBM is parameterized by a vertex assignment vector $\vec{\tau} \in \{1, \cdots, K\}^n$, a symmetric $K \times K$ block connectivity probability matrix $\mathbf{B}$ with entries in $[0,1]^{K \times K}$, and the degree correction ("promiscuity") vector $\vec{\theta} \in \mathbb{R}^n$. The degree correction vector is constrained such that $\sum_i^n \vec{\theta}_i \mathbb{I}(\tau_i = k) = 1$ for $k \in [K]$ where $\mathbb{I}$ is an indicator function. The model is $\mathbf{A} \sim \text{DCSBM}_n(\vec{\tau}, \mathbf{B}, \vec{\theta})$ if $\mathbf{A}$ has entries $\mathbf{A}_{ij} \sim \text{Bernoulli}(\vec{\theta}_i \vec{\theta}_j \mathbf{B}_{kl})$ where $k = \tau_i, l = \tau_j$, for $i, j \in [n]$, and $k, l \in [K]$. The *a posteriori* DCSBM model is additionally parameterized by a block membership probability vector $\vec{\pi} = [\pi_1, \ldots, \pi_K]^\top$. The model is $\mathbf{A} \sim \text{SBM}_n(\vec{\pi}, \mathbf{B}, \vec{\theta})$ if $\mathbf{A}$ has entries $\mathbf{A}_{ij} \left| k = \tau_i, l = \tau_j \sim \text{Bernoulli}(\vec{\theta}_i \vec{\theta}_j \mathbf{B}_{kl}) \right.$, where $\tau_i \sim \text{Multinomial}(\vec{\pi})$ for $i \in [n]$.

5 Algorithms In this section, we introduce algorithms for statistical analysis of networks. Section 5.1 provides an overview of algorithms for a single graph and Section 5.2 provides an overview of algorithms designed for multiple graphs.

5.1 Single Graph Algorithms

5.1.1 Adjacency and Laplacian Spectral Embedding (ASE, LSE) Given an undirected graph with adjacency matrix $\mathbf{A}$, the adjacency spectral embedding (ASE) and Laplacian spectral embedding (LSE) construct a representation of the vertices of the graphs into d dimensions via its eigendecomposition, given by $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{U}^\top$ where $\mathbf{U} \in \mathbb{R}^{n \times n}$ is the orthogonal matrix of eigenvectors and $\mathbf{S} \in \mathbb{R}^{n \times n}$ is a diagonal matrix containing the eigenvalues of $\mathbf{A}$ ordered by magnitude, such that $|\mathbf{S}_{11}| \geq |\mathbf{S}_{22}| \geq \ldots \geq |\mathbf{S}_{nn}|$. The ASE of the graph into $\mathbb{R}^d$ is defined as $\text{ASE}(\mathbf{A}) = \hat{\mathbf{X}} = \hat{\mathbf{U}}|\hat{\mathbf{S}}|^{1/2}$, where $\hat{\mathbf{U}} \in \mathbb{R}^{n \times d}$ contains the first d columns of $\mathbf{U}$, which correspond to the largest eigenvectors, and $\hat{\mathbf{S}} \in \mathbb{R}^{d \times d}$ is the submatrix of $\mathbf{S}$ corresponding to the d largest eigenvalues in magnitude. The LSE, of $\mathbf{A}$ is defined in a similar manner using the normalized Laplacian of the graph defined as $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ where $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with entries $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$. Then, the LSE of the graph is given by $\text{LSE}(\mathbf{A}) = \text{ASE}(\mathbf{L}) = \tilde{\mathbf{X}} \in \mathbb{R}^{n \times d}$.

In the case of directed graphs, the eigendecomposition is not available since adjacency matrix is not symmetric, so instead we use the singular value decomposition of the adjacency matrix as $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n \times n}$ are orthogonal matrices containing the left and right singular vectors, and $\mathbf{S} \in \mathbb{R}^{n \times n}$ is a non-negative diagonal matrix with the singular values. The ASE of a directed graph results in two different latent position matrices $\hat{\mathbf{X}} = \hat{\mathbf{U}}\hat{\mathbf{S}}^{1/2}$ and $\hat{\mathbf{Y}} = \hat{\mathbf{V}}\hat{\mathbf{S}}^{1/2}$, denoted *in* and *out* latent positions, respectively, where $\hat{\mathbf{U}}, \hat{\mathbf{V}} \in \mathbb{R}^{n \times d}$ contain the d columns of $\mathbf{U}$ and $\mathbf{V}$ corresponding to the d leading singular vectors, and $\hat{\mathbf{S}}$ is the submatrix of $\mathbf{S}$ containing the d leading singular values. While there exists many definitions for directed normalized Laplacian, we define it as $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{O}^{-1/2}$ where $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$ and $\mathbf{O}_{ii} = \sum_j \mathbf{A}_{ji}$ are the in and out degree diagonal matrices [82]. The LSE of directed graph processed similarly to that of directed ASE.

Spectral embedding is the first step in many subsequent inference tasks. For example, spectral clustering for community detection (Section 5.1.4) can be achieved via Gaussian Mixture modeling on $\hat{\mathbf{X}}$ from either ASE or LSE. The resulting cluster assignments can further be used to estimate the parameters for *a posteriori* SBM.

For real data, the true embedding dimension d is often not known and must be estimated. A general methodology for choosing the embedding dimension d is to examine the scree plot of the singular values of $\mathbf{A}$ and look for an "elbow" or a "big gap". While many methods for choosing the threshold exist [21, 50], we consider the method in [122] when applying any spectral embeddings in real data. Given $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{U}^\top$ for either ASE or LSE, the eigenvalues in $|\mathbf{S}|$ are used to estimate the embedding dimension $\hat{d}$ by maximizing the profile likelihood function, which determines the magnitude of the "gap" after first d largest eigenvalues. Multiple elbows can be found by discarding the $\hat{d}$ number of largest eigenvalues and repeating the process with the remaining eigenvalues. For applications in connectomics, we only consider $\lceil \log n \rceil$ largest eigenvalues as input to the profile likelihood function and take the second elbow as the estimate of $\hat{d}$.

5.1.2 Diagonal Augmentation

Many connectomes have no self-loops, resulting in all zero in the diagonal entries of the adjacency matrices. When computing spectral embeddings of graphs, the zero diagonal results in increased errors in estimation [101]. Furthermore, the sum of eigenvalues of the adjacency matrices is zero, leading to an indefinite matrix, which violate assumptions of the statistical models such as RDPG.

Diagonal augmentation (`diag-aug`) is a method for imputing the diagonals of adjacency matrices from graphs with no self-loops [66, 87, 101]. The diagonals are imputed with with the average of the non-diagonal entries of each row, which corresponds to the degree of each vertex divided by $n-1$. In the case of directed graphs, the average of in and out degree is used. Specifically, the diagonal augmented adjacency matrix is defined as $\tilde{\mathbf{A}} = \mathbf{A} + \tilde{\mathbf{D}}$ where $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\tilde{\mathbf{D}} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with entries $(\mathbf{A}\vec{1}^\top + \vec{1}\mathbf{A})/2(n-1)$ where $\vec{1} \in \mathbb{R}^n$ is a row vector of ones. To achieve best embedding estimation, the diagonal entries of adjacency matrices should be imputed prior to `ASE` (in `LSE`, the diagonals are imputed via the normalized Laplacian).

5.1.3 Pass-To-Ranks (PTR)

Connectomes have often weighted edges, which can take on arbitrary values. Rescaling and normalizing the edge weights has been shown to increase reliability and can improve estimation of spectral embeddings [53]. Pass-to-ranks (`PTR`) is a method for rescaling the *positive* edge weights such that all edge weights are between 0 and 1, inclusive.

Given an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, let $R(\mathbf{A}_{ij})$ be the "rank" of $\mathbf{A}_{ij}$, that is, $R(\mathbf{A}_{ij}) = k$ if $\mathbf{A}_{ij}$ is the k^{th} smallest number in $\mathbf{A}$. The rescaled adjacency matrix, $\tilde{\mathbf{A}}$, is defined as follows:

$$\tilde{\mathbf{A}}_{ij} = \begin{cases} \frac{R(\mathbf{A}_{ij})}{|E|} & \text{if } \mathbf{A}_{ij} > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $|E|$ is the number of non-zero edges. Ties in rank are broken by averaging the ranks. For spectral embedding of weighted connectomes, they are first normalized via `PTR`, then the diagonals are imputed via `diag-aug` prior to `ASE` (`diag-aug` is skipped for `LSE`).

5.1.4 Spectral Clustering for Community Detection

One of the most common uses of spectral clustering is for community detection, in which the vertices with similar connectivity patterns are grouped together. Given the embeddings of a graph from either `ASE` or `LSE`, classical Euclidean clustering of $\hat{\mathbf{X}}$ results in community structure. Central limit theorems for spectral embeddings of many statistical models (e.g. SBM, RDPG) suggest Gaussian Mixture modeling (`GMM`) for clustering (see Section 6.1).

The true number of clusters, K, is often not known in real data, but can be estimated by maximizing likelihood functions penalized by model complexity. Commonly used functions include Bayesian Information Criterion (`BIC`), Akaike Information Criterion (`AIC`), and Minimum Description Length (`MDL`) [3, 80, 88]. By default, we use penalized likelihood via `BIC` to estimate K [77]. In practice, various covariance types and initialization methods for `GMM`, and number of clusters are swept over to compute best estimated number of cluster, $\hat{K}$ [10, 89].

5.2 Multiple Graph Algorithms

5.2.1 Omnibus Embedding

Consider a sample of m observed graphs $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \ldots, \mathcal{G}^{(m)}$ and their associated adjacency matrices, $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(m)} \in \mathbb{R}^{n \times n}$ with n vertices that are identical and shared across all graphs. Under the JRDPG model, `OMNI` is a consistent method (see Section 6.2.1) for simultaneously estimating the latent position matrices for each graph by computing the spectral embedding into d-dimensions on the omnibus matrix, $\mathbf{O} \in \mathbb{R}^{nm \times nm}$, as defined below

$$\mathbf{O} = \begin{bmatrix} \mathbf{A}^{(1)} & \frac{1}{2}\left(\mathbf{A}^{(1)} + \mathbf{A}^{(2)}\right) & \cdots & \frac{1}{2}\left(\mathbf{A}^{(1)} + \mathbf{A}^{(m)}\right) \\ \frac{1}{2}\left(\mathbf{A}^{(2)} + \mathbf{A}^{(2)}\right) & \mathbf{A}^{(2)} & \cdots & \left(\mathbf{A}^{(2)} + \mathbf{A}^{(m)}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}\left(\mathbf{A}^{(m)} + \mathbf{A}^{(1)}\right) & \frac{1}{2}\left(\mathbf{A}^{(m)} + \mathbf{A}^{(2)}\right) & \cdots & \mathbf{A}^{(m)} \end{bmatrix}$$

The embeddings gives the matrix

$$\hat{\mathbf{Z}} = \texttt{ASE}(\mathbf{O}) = \begin{bmatrix} \hat{\mathbf{X}}^{(1)} \\ \hat{\mathbf{X}}^{(2)} \\ \vdots \\ \hat{\mathbf{X}}^{(m)} \end{bmatrix} \in \mathbb{R}^{mn \times d}$$

where the first n rows are the latent positions corresponding to $\mathbf{A}^{(1)}$, so on and so forth.

5.2.2 Multiple Adjacency Spectral Embedding (MASE) `MASE` is a consistent method for estimation (see Section 6.2.1) of underlying parameters for each graph under the COSIE model [7]. `MASE` is a is a three step process:

1. Each adjacency matrix, $\mathbf{A}^{(i)}$, is embedded into d dimensions via `ASE`, and the matrix $\hat{\mathbf{U}} = \left[\texttt{ASE}(\mathbf{A}^{(1)}), \texttt{ASE}(\mathbf{A}^{(2)}), \ldots, \texttt{ASE}(\mathbf{A}^{(m)})\right] \in \mathbb{R}^{n \times dm}$ is the concatenated matrix of spectral embeddings.
2. Calculate the singular value decomposition of $\hat{\mathbf{U}} = \mathbf{V}\mathbf{S}\mathbf{W}^\top$, and let $\hat{\mathbf{V}} \in \mathbb{R}^{n \times d}$ be the matrix containing the d singular vectors corresponding to d largest singular values. $\hat{\mathbf{V}}$ is the estimated shared common subspace matrix.
3. Individual matrices are estimated via $\hat{\mathbf{R}}^{(i)} = \hat{\mathbf{V}}^\top \mathbf{A}^{(i)} \hat{\mathbf{V}}$ where $\hat{\mathbf{R}}^{(i)} \in \mathbb{R}^{d \times d}$.

5.2.3 Spectral Clustering for Community Detection Similar to the procedure described in Section 5.1.4, one can also perform spectral clustering in the multi-graph setting. Clustering is performed on the the average latent position matrix, $\bar{\mathbf{X}} \coloneqq \frac{1}{m}\sum_{i=1}^{m} \hat{\mathbf{X}}^{(i)}$, in JRDPG model and the vertex subspace matrix, $\hat{\mathbf{V}}$ in COSIE model. The clustering procedure proceeds identically as described in Section 5.1.4.

5.2.4 Seeded Graph Matching (SGM) Consider two graphs $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ with n vertices and their associated adjacency matrices $\mathbf{A}$ and $\mathbf{B}$, respectively. The graph matching problem seeks to find an alignment of nodes between these two graphs that minimizes the number of edge disagreements. Formally, it is defined as the following optimization problem:

$$\begin{aligned} &\min \|\mathbf{A}\,\mathbf{P} - \mathbf{P}\,\mathbf{B}\|_F^2 \\ &\text{s.t. } \mathbf{P} \in \mathcal{P} \end{aligned} \tag{5.1}$$

where $\mathcal{P}$ is the set of permutation matrices in $\mathbb{R}^{n \times n}$. Seeded graph matching (`SGM`) is a modification of the graph matching algorithm, allowing for the specification of seed sets W_1, W_2 with seeding $\psi : W_1 \to W_2$, and solved via fast approximate quadratic assignment (`FAQ`) [108]. As the seeded graph matching problem is computationally intractable, `SGM` provides an approximate solution by relaxing the feasible region from $\mathcal{P}$ to $\mathcal{D}$, the set of doubly stochastic matrices. The algorithm is provided below:

1. Initialize at some $\mathbf{P}^{(0)} \in \mathcal{D}$, where $\mathcal{D}$ is the set of doubly stochastic matrices. Typically, initialization is chosen as $\mathbf{P}^{(0)} = \vec{1}\vec{1}^\top/n$, where $\vec{1}$ denotes the n-vector of all ones.
2. **while** stopping criteria not met **do**
   (a) Compute the gradient $\Delta f\left(\mathbf{P}^{(i)}\right)$
   (b) Compute the search direction $\mathbf{Q}^{(i)} \in \operatorname{argmax}\left(\operatorname{tr}\left(\mathbf{Q}^T \Delta f(\mathbf{P}^{(i)})\right)\right)$ via Hungarian Algorithm
   (c) Compute step size $\alpha^{(i)} \in \operatorname{argmax}(f(\alpha^{(i)}\,\mathbf{P}^{(i)} + (1-\alpha^{(i)})\,\mathbf{Q}^{(i)}))$
   (d) Update $\mathbf{P}^{(i+1)} \coloneqq \alpha^{(i)}\,\mathbf{P}^{(i)} + (1-\alpha^{(i)})\,\mathbf{Q}^{(i)}$
3. Compute $\hat{\mathbf{P}} \in \operatorname{argmax}\left(\operatorname{tr}\left(\mathbf{P}^\top \mathbf{P}^{(final)}\right)\right)$ via Hungarian Algorithm

6 Theory for Statistical Models In this section, we provide general outlines of the theorems and proofs for statistical models in Section 3 and algorithms in Section 5.

6.1 Theory for Single Graph Models Graph features, such as the ones described in Section 2.2, are popularly used to test hypothesis about a graph. However, the distribution of such features is usually unknown, and even in cases where the asymptotic distribution is available, one needs to proceed with caution as some of the asymptotic results might be misleading [74]. [84] studies the behavior of two simple graph features, namely, the number of edges and the maximum degree, for testing a simple hypothesis question about the distribution of a graph. While the statistic based on the number of edges achieves a higher power in the limit as the number of vertices grows, a comparative power analysis shows that even for large graphs with $n \leq 10^{24}$, the statistic based on the maximum degree dominates under certain cases.

A body of existing results in statistical inference for spectral embeddings is reviewed more deeply in [11]. We summarize next some of the main results related to the exposition in this paper. In this section, we assume that a sequence of random adjacency matrices $\{\mathbf{A}_n, n \geq 1\}$ generated from a sequence of latent positions $\{\mathbf{X}_n, n \geq 1\}$, where $\mathbf{A}_n \sim \text{RDPG}(\mathbf{X}_n)$, $n \geq 1$ is the adjacency matrix of a graph with n vertices, and $\mathbf{X}_n \in \mathbb{R}^{n \times d}$ are d-dimensional latent positions. We write $(\mathbf{X}_n)_i$ to represent the i-th row of $\mathbf{X}_n$, and we assume that the rows of $\mathbf{X}_n$, which correspond to the latent positions, are an i.i.d. sample $(\mathbf{X}_1), \ldots, (\mathbf{X}_n)_n \overset{\text{i.i.d.}}{\sim} F$, where F is a distribution with support $\mathcal{X} \subset \mathbb{R}^d$. We also assume that the second moment matrix $\mathbf{\Delta} = \mathbb{E}[(\mathbf{X}_n)_1 (\mathbf{X}_n)_1^\intercal] \in \mathbb{R}^{d \times d}$, has non-zero eigenvalues. We use $\widehat{\mathbf{X}}_n = \texttt{ASE}(\mathbf{A}_n) \in \mathbb{R}^{n \times d}$ to denote the d-dimensional adjacency spectral embedding of $\mathbf{A}_n$, and $\widetilde{\mathbf{X}}_n = \texttt{LSE}(\mathbf{A}_n)$ to denote its d-dimensional Laplacian spectral embedding.

The adjacency spectral embedding (ASE) method described in Section 5.1.1 is a consistent and asymptotically normal estimator for the latent positions of a random dot product graph. In [94], it is shown that clustering rows of the ASE of $\mathbf{A}_n$ can consistently recover the communities of an SBM. Consistency of the latent positions for an RDPG is studied in [63, 65, 95]. In particular, Theorem 5 of [65] shows that with probability tending to one, there exists some orthogonal rotation $\mathbf{W}_n \in \mathbb{R}^{d \times d}$ such that

$$\max_{i \in [n]} \| (\widehat{\mathbf{X}}_n)_i - \mathbf{W}_n (\mathbf{X}_n)_i \| \leq \frac{C d^{1/2} \log^2 n}{\sqrt{n}},$$

where $C > 0$ is a constant, and hence, the rows of of $\hat{\mathbf{X}}_n$ converge to the rows of $\mathbf{X}_n$, up to some orthogonal rotation, as the number of vertices n grows.

Distributional results on the rows of the adjacency spectral embedding show that the error in estimating the true latent positions is asymptotically normally distributed. In particular [12] showed a central limit theorem for the rows of the ASE of $\mathbf{A}_n$, in which the latent positions are shown to converge to a mixture of standard multivariate normal distributions, that is, for any $\mathbf{z} \in \mathbb{R}^d$,

$$\lim_{n \to \infty} \mathbb{P}\left(\sqrt{n} \left(\widehat{\mathbf{X}}_n \mathbf{W}_n - \mathbf{X} \right)_i \leq \mathbf{z} \right) = \int_{\mathcal{X}} \Phi(\mathbf{z}, \mathbf{\Sigma}(\mathbf{x}))\, dF(\mathbf{x}), \tag{6.1}$$

where $\Phi(\mathbf{z}, \mathbf{\Sigma}(\mathbf{x}))$ is the cumulative distribution function of a multivariate normal distribution with mean zero and a covariance matrix $\mathbf{\Sigma}(\mathbf{x}) \in \mathbb{R}^{d \times d}$ that is a function of $\mathbf{x} \in \mathcal{X}$ (see [12], Theorem 1, for an expression of this covariance matrix).

Similar results to the ones presented above are also available for the Laplacian spectral embedding (LSE). In particular, Theorem 3.1 of [99] provides an an asymptotic result on the estimation error of the rows of $\widetilde{\mathbf{X}}_n$ with respect to its population version, and Theorem 3.2 shows an analogous result to the one presented in Equation (6.1) to establish the asymptotic normality of the rows of this estimator, that is,

$$\lim_{n \to \infty} \mathbb{P}\left\{ \sqrt{n} \left(\mathbf{W}_n (\widetilde{\mathbf{X}}_n)_i - \frac{(\mathbf{X}_n)_i}{\sqrt{\sum_j (\mathbf{X}_n)_i^\intercal (\mathbf{X}_n)_j}} \right) \leq \mathbf{z} \right\} = \int_{\mathcal{X}} \Phi(\mathbf{z}, \widetilde{\mathbf{\Sigma}}(\mathbf{x}))\, dF(\mathbf{x}),$$

for some covariance matrix $\widetilde{\mathbf{\Sigma}}(\mathbf{x})$ which its exact form is presented in [99].

The consistency and asymptotic normality of ASE and LSE considered in this section have been recently extended to the GRDPG model (see Theorems 5-8 in [83]).

6.2 Theory for Multiple Graph Models

6.2.1 Spectral Embeddings The results discussed before have been used to develop valid statistical tests for two-graph hypothesis testing questions. The work of [97] studies a semiparametric graph hypothesis testing for the equivalence between the latent positions of the vertices of a pair of graphs. Formally, for each fixed n let $\mathbf{X}_n, \mathbf{Y}_n \in \mathbb{R}^{n\times d}$ be a sequence of latent positions matrices, and define $\mathbf{A}_n \sim \text{RDPG}(\mathbf{X}_n)$, $\mathbf{B}_n \sim \text{RDPG}(Y_n)$ as independent random adjacency matrices. The problem of testing the equality of the distributions of $\mathbf{A}_n$ and $\mathbf{B}_n$ is defined as

$$\mathcal{H}_0^n : \mathbf{X}_n =_{\mathbf{W}} \mathbf{Y}_n \qquad \text{vs.} \qquad \mathcal{H}_a^n : \mathbf{X}_n \neq_{\mathbf{W}} \mathbf{Y}_n,$$

where $\mathbf{X}_n =_{\mathbf{W}} \mathbf{Y}_n$ denotes that $\mathbf{X}_n$ and $\mathbf{Y}_n$ are equivalent up to an orthogonal transformation $\mathbf{W} \in \mathcal{O}_d$, and $\mathcal{O}_d$ is the set of $d \times d$ orthogonal matrices. To define the test statistic, denote $\widehat{\mathbf{X}}_n = \texttt{ASE}(\mathbf{A}_n)$, $\widehat{\mathbf{Y}}_n = \texttt{ASE}(\mathbf{B}_n)$, and for a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ with singular values $\sigma_1(\mathbf{A}) \geq \ldots \geq \sigma_n(\mathbf{A}) \geq 0$ and largest observed degree $\delta(\mathbf{A}) = \max_{i\in[n]} \sum_{j=1}^n \mathbf{A}_{ij}$, define

$$\gamma(\mathbf{A}) := \frac{\sigma_d(\mathbf{A}) - \sigma_{d+1}(\mathbf{A})}{\delta(\mathbf{A})}.$$

Define T_n as the test statistic

$$T_n := \frac{\min_{\mathbf{W}\in\mathcal{O}_d} \|\widehat{\mathbf{X}}_n \mathbf{W} - \widehat{\mathbf{Y}}_n\|_F}{\sqrt{d\gamma^{-1}(\mathbf{A}_n)} + \sqrt{d\gamma^{-1}(\mathbf{B}_n)}}.$$

It is shown in Theorem 3.1 of [97] that T_n is a consistent test for the hypothesis testing problem described above, in the sense that for any significance level α and $C > 1$, then $\mathbb{P}(T_n > C) \leq \alpha$ for n sufficiently large under $\mathcal{H}_0^n$ (type I error control), and if $\lim_{n\to\infty} \min_{\mathbf{W}\in\mathcal{O}_d} \|\widehat{\mathbf{X}}_n \mathbf{W} - \widehat{\mathbf{Y}}_n\|_F = \infty$, then $\mathbb{P}(T_n > C) \to 1$ under $\mathcal{H}_a^n$ (i.e., the type II error vanishes). For specific assumptions and some extensions to other hypothesis testing problems, the reader is referred to [97] and [11].

When the vertices of the graphs are not necessarily aligned (including cases in which the graphs do not have the same number of vertices), testing equality of latent positions is inappropriate. The work of [98] proposes a nonparametric test to determine whether the distribution of the latent positions of the graphs is the same. For a pair of matrices $\mathbf{X}_n \in \mathbb{R}^{n\times d}$ and $\mathbf{Y}_m \in^{m\times d}$ with their rows distributed as $(\mathbf{X}_n)_i \overset{\text{i.i.d.}}{\sim} F$ and $(\mathbf{Y}_m)_i \overset{\text{i.i.d.}}{\sim} G$ and a pair of independent adjacency matrices $\mathbf{A}_n \sim \text{RDPG}(\mathbf{X}_n)$, $\mathbf{B}_n \sim \text{RDPG}(\mathbf{Y}_n)$, the nonparametric graph hypothesis testing problem is given by

$$\mathcal{H}_0^n : F \perp\!\!\!\perp G \qquad \text{vs.} \qquad \mathcal{H}_a^n : F \not\perp\!\!\!\perp G,$$

where $F \perp\!\!\!\perp G$ indicates equality of the distributions up to an orthogonal transformation. To test such hypothesis, [98] proposes to use the following test statistic

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

where $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel. In [98], Theorem 1, it is shown that $U_{n,m}(\mathbf{X}, \mathbf{Y})$ is a consistent and unbiased estimate of the maximum mean discrepancy [44] between the distributions

F and G. Furthermore, under the null hypothesis, the quantity $(m+n)U_{n,m}(\mathbf{X},\mathbf{Y})$ converges in distribution to an infinite weighted sum of independent chi-squared random variables as $n,m\to\infty$, provided that $\frac{n}{n+m}\to\rho\in(0,1)$. Moreover, when the latent positions are used in place of the true latent positions, then Theorem 4 of [98] shows that the difference between $U_{n,m}(\widehat{\mathbf{X}},\widehat{\mathbf{Y}})$ and $U_{n,m}(\mathbf{X},Y)$ converges to zero sufficiently fast to yield a consistent test procedure.

The work of [58] studies the omnibus embedding described in Section 5.2.1 under the joint random dot product graph (JRDPG) model, where $(\mathbf{A}^{(1)},\dots,\mathbf{A}^{(m)})\sim\text{JRDPG}(\mathbf{X}_n)$, and the rows of $\mathbf{X}_n\in\mathbb{R}^{n\times d}$ are an i.i.d. sample from some distribution F. Let $\widehat{\mathbf{O}}\in\mathbb{R}^{mn\times mn}$ be the omnibus embedding of $\mathbf{A}^{(1)},\dots,\mathbf{A}^{(m)}$ and $\widehat{\mathbf{Z}}=\texttt{ASE}(\mathbf{O})\in\mathbb{R}^{mn\times d}$. Under this setting, it is shown in Lemma 1 of [58] that the rows of $\widehat{\mathbf{Z}}_n$ are a consistent estimator of the latent positions of each individual graph as $n\to\infty$, and that

$$\max_{i\in[n],j\in[m]}\|(\widehat{\mathbf{Z}}_n)_{(j-1)n+i}-\mathbf{W}_n(\mathbf{X}_n)_i\|\le\frac{C\sqrt{m}\log(mn)}{\sqrt{n}}. \tag{6.2}$$

Furthermore, a central limit theorem for the rows of the omnibus embedding asserts that

$$\lim_{n\to\infty}\mathbb{P}\left\{\sqrt{n}\left(\mathbf{W}_n(\widehat{\mathbf{Z}}_n)_{(j-1)n+i}-(\mathbf{X}_n)_i\right)\le\mathbf{z}\right\}=\int_{\mathcal{X}}\Phi(\mathbf{z},\widehat{\mathbf{\Sigma}}(\mathbf{x}))\,dF(\mathbf{x}), \tag{6.3}$$

for some covariance matrix $\widehat{\Sigma}(\mathbf{x})$ (see Theorem 1 of [58] for an exact expression). In recent work, [29] extended the study of the omnibus embedding and provided results analogous to the ones in Equations (6.2) and (6.3) under a more general model that allows for differences in the latent positions of each graph.

The COSIE model described in Section 3.2.2 describes multiple networks with expected probability matrices that share the same common invariant subspace. It is shown in [7] that the MASE algorithm (see Section 5.2.2) is a consistent estimator for this common invariant subspace, and produces asymptotically normally distributed estimates for the individual symmetric matrices. Specifically, let $\mathbf{V}_n\in\mathbb{R}^{n\times d}$ be a sequence of orthonormal matrices and $\mathbf{R}_n^{(1)},\dots,\mathbf{R}_n^{(m)}\in\mathbb{R}^{d\times d}$ a sequence of score matrices such that $\mathbf{P}_n^{(l)}=\mathbf{V}_n\mathbf{R}_n^{(l)}\mathbf{V}_n^\top\in[0,1]^{n\times n}$, $(\mathbf{A}_n^{(1)},\dots,\mathbf{A}_n^{(m)})\sim\text{COSIE}(\mathbf{V}_n;,\mathbf{R}_n^{(1)},\dots,\mathbf{R}_n^{(m)})$, and $\widehat{\mathbf{V}},\widehat{\mathbf{R}}_n^{(1)},\dots,\widehat{\mathbf{R}}_n^{(1)}$ be the estimators obtained by MASE. Under appropriate regularity conditions (see Theorem 3 of [7]), the estimate for $\mathbf{V}$ is consistent as $n,m\to\infty$, and there exists some constant $C>0$ such that

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

In addition, the entries of $\widehat{\mathbf{R}}_n^{(l)}$, $l\in[m]$ are asymptotically normally distributed. Namely, there exists a sequence of orthogonal matrices $\mathbf{W}$ such that

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

as $n\to\infty$, where $\mathbb{E}[\|\mathbf{H}_m^{(l)}\|]=O\left(\frac{d}{\sqrt{m}}\right)$ and $\sigma_{l,j,k}^2=O(1)$. For a precise statement about the joint distribution of the entries of $\widehat{\mathbf{R}}_n^{(i)}$, see Theorem 7 in [7].

6.2.2 Graph Matching for Correlated Networks Given a pair of graphs $\mathbf{A}_n$ and $\mathbf{B}_n$ with n vertices each, the graph matching problem tries to find a correspondence between their vertices. A body of literature has studied the feasibility of finding the correct matching under different random graph models, including correlated Erdős-Rényi [28, 62] and Bernoulli graphs [60]. In this section we review some of the results for the correlated Erdős-Rényi model described in Section 3.2.3.

Formally, given parameters $\rho_n \in [0,1]$ and $q_n \in (0, 1-\xi_1)$ for some small $\xi_1 > 0$, the $n \times n$ adjacency matrices $\mathbf{A}_n$ and $\mathbf{B}_n$ are distributed as correlated Erdős-Rényi if their marginal distributions are $\mathbf{A}_n \sim \text{ER}_n(q_n)$, $\mathbf{B}_n \sim \text{ER}_n(q_n)$, but the edge pairs satisfy $\text{Corr}((\mathbf{A}_n)_{ij}, (\mathbf{Q}_n^\top \mathbf{B}_n \mathbf{Q}_n)_{ij}) = \rho_n$, where $\mathbf{Q}_n \in \mathcal{P}_n$ is a permutation matrix that gives the correct alignment between the vertices (here $\mathcal{P}_n$ denotes the set of $n \times n$ permutation matrices). The work of [62] studies the feasibility of finding $\mathbf{Q}_n$ by solving the optimization problem defined in Equation (5.1). In particular, it is shown that there exists positive constants c_1, c_2 such that if $\rho_n \geq c_1\sqrt{\frac{\log n}{n}}$ and $q_n \geq c_2\frac{\log n}{n}$, then $\mathbf{Q}_n$ can be correctly recovered with probability 1 for n sufficiently large (Theorem 1 of [62]).

While the solution of the quadratic assignment problem (5.1) can correctly recover the vertex alignment in theory, it is computationally challenging to solve the optimization problem. In the presence of s_n seed vertices with known correspondence between the graphs, [62] introduced an efficient polynomial algorithm to recover the alignment of the remaining $n - s_n$ vertices. Theorem 2 of [62] shows that this method can correctly recover $\mathbf{Q}_n$ in the setting where $\xi_2 < p_n < 1 - \xi_2 < 1$ and $\xi_2 < \rho_n < \xi_2$ for some $\xi_2 > 0$ in the presence of a logarithmic number of seeds (i.e. $s_n \geq c_3 \log n$ for some $c_3 > 0$).

7 Data Descriptions

The following two datasets are analyzed using the algorithms and models described Sections 3 and 5. Section 8 primarily focuses on the *Drosphila* connectome, while Section 9 primarily focuses on HCP connectomes.

7.1 *Drosphila* Larval Mushroom Body Data Description

The connectome was estimated from serial-section electron microscopy (EM) of an L1 *Drosophila* larva [32]. For the mushroom body (MB) subcircuit, the graph was defined by manually identifying synapses in the EM volume, and tracing the pre- and post-synaptic partners through the EM volume back to their cell bodies. Each node in this graph represents an individual neuron, and each edge consists of one or more synapses between those neurons. Thus, edge weights are the number of synapses between neurons.

Each node in the graph also has an associated cell type: Kenyon cell (KC), projection neuron (PN), MB input neuron (MBIN), and MB output neuron (MBON). Additionally, we can categorize neurons based on hemisphere (which side of the brain each neuron was on), and neuron pair (for most neurons, a homologous pair neuron in the other hemisphere was identified by morphological comparison).

7.2 HCP Data Description

We used publicly available diffusion MRI (dMRI) and structural MRI (sMRI) data from the S1200 (2017) release of the Human Connectome Project (HCP) Young Adult study, acquired by the Washington University in St. Louis (WUSTL) and the University of Minnesota (Minn) [103, 104]. Out of the 1206 participants released, 1059 had viable dMRI for processing.

Connectomes were estimated using the ndmg pipeline [53]. Briefly, the dMRI scans were preprocessed for eddy currents using FSL's `eddy-correct` [92]. FSL's "standard" linear registration pipeline was used to register the sMRI and dMRI images to the MNI152 atlas [51, 68, 92, 115].A tensor model is fit using DiPy [38] to obtain an estimated tensor at each voxel. A deterministic tractography algorithm is applied using DiPy's EuDX [38, 39] to obtain streamlines, which indicate the voxels connected by an axonal fiber tract. We used a modified version of Desikan–Killiany–Tourville (DKT) parcellation [55] to define the ROIs. Graphs are formed by counting the number of fibers between a pair of ROIs.

8 Applications for Single Graph Data

In this section, we explore the applications of the single graph models in Section 3.1 and the algorithms in Section 5.1. The *Drosophila* mushroom body connectome and HCP data are analyzed (see Sections 7.1 and 7.2 for description) along with simulated examples.

8.1 Testing for Differences between Communities of Edges

In Figure 6, we compare a number of different strategies using Fisher's exact test [35] for testing whether there exists a difference between $K = 2$ communities, or groups, of edges in a graph. Formally, let $e_{ij}^{(k)} \sim F_k$ be a single edge in the

graph, where $k \in \{1, 2\}$ is a community of edges, for $i, j \in [n]$. Our hypothesis test of interest is:

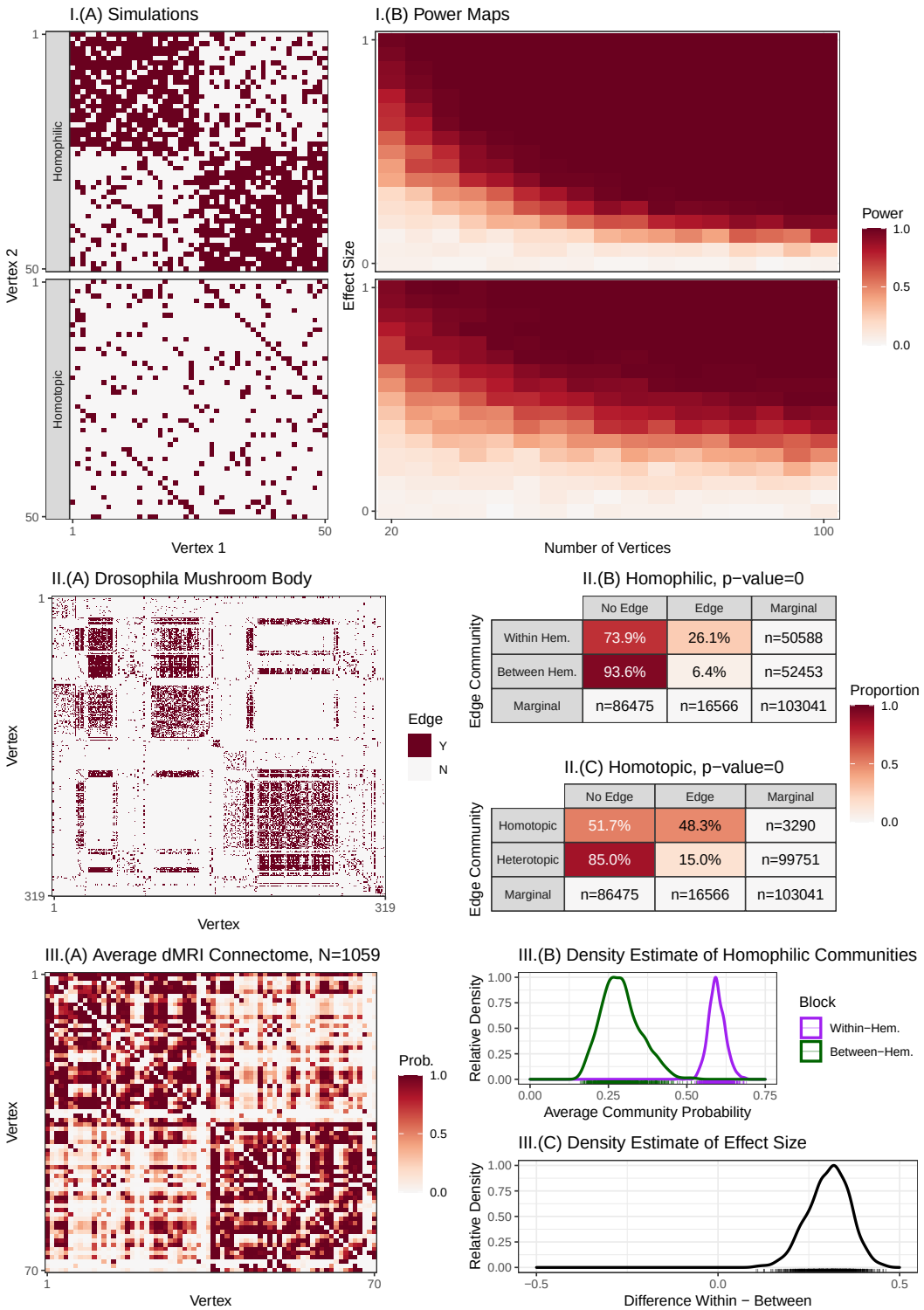
$$H_0 : F_1 = F_2, \quad H_a : F_1 \neq F_2$$

We simulate graphs from the homophilic planted partition SBM from Section 3.1.2 and symmetric homotopic SIEM from Section 3.1.3 in Figure 6I.(A). Under the given models, our hypotheses simplify to testing whether $p_1 = p_2$ against $p_1 \neq p_2$; that is, whether or not there exists a different probability for each edge community. Effect size, or the difference in probability between the two communities, for the SBM and the SIEM are varied linearly from 0 to 0.1, and from 0 to 0.5, respectively. A relative effect size of 0 corresponds to an ER graph, in which $F_1 = F_2$; at all other relative effect sizes, the alternative is true. We measure performance using the statistical power at $\alpha = 0.05$ in Figure 6I.(B). Across the simulation settings, we see that Fisher's exact test provides an appropriate statistical test and provides sufficiently high power with large enough effect size and graph. Importantly, Fisher's test displays both empirical validity (at an effect size of zero, the power is at most α) and empirical consistency (the test power converges to 1 as the effect size increases) in both simulations.

We demonstrate our techniques developed above on the *Drosophila* mushroom body, with $n = 319$ vertices in the left or right hemisphere (2 vertices located along the center of the brain are excluded). In Figure 6II we investigate the appropriateness of different unweighted independent edge models for the *Drosophila* mushroom body. Our goal is to identify whether the unweighted *Drosophila* mushroom body display homophilia (that is, the within hemisphere blocks have greater connectivity than between hemisphere blocks) or homotopia (that is, edges incident bilateral vertices have a different distribution from edges incident non-bilateral vertices). Figure 6II.(A) shows the unweighted *Drosophila* mushroom body. The within-hemisphere blocks appear to have a higher proportion of edges than the between-hemisphere edge blocks, shown in Figure 6II.(B). There is strong evidence that the within-hemisphere connectivity exceeds the between-hemisphere connectivity (Fisher's exact test, p-value=0.0). Next, we investigate whether the graph is homotopic; that is, whether bilateral (homotopic) connectivity exceeds non-bilateral (heterotopic) connectivity, in Figure 6II.(C). Strong evidence is present that homotopic connectivity exceeds heterotopic connectivity (Fisher's exact test, p-value=0.0).

Finally, we explore the appropriateness of various independent edge models for diffusion connectomes from the HCP Dataset. The diffusion connectomes are binarized according to whether an edge is present (the edge weight is greater than zero) or absent (the edge weight is zero). Figure 6III.(A) shows the average unweighted diffusion connectome over all participants in the study. Figure 6III.(B) shows the distribution of edge-weights within-hemisphere versus between-hemisphere. The diffusion connectomes appear to possess homophily; i.e., high within-hemisphere connectivity, with lower between-hemisphere connectivity. This is demonstrated by the fact that in all N=1059, the within-hemisphere connectivity exceeds the between-hemisphere connectivity. This effect can be observed by looking at the difference between within-hemisphere connectivity and between-hemisphere connectivity for each of the $N = 1059$ connectomes, shown in 6III.(C). All 1059 diffusion connectomes have significantly higher within-hemisphere connectivity than between-hemisphere connectivity at $\alpha = .05$ after Bonferroni correction (Fisher's exact test, $N = 1059$, maximum p-value$< 10^{-20}$).

8.2 Testing for Differences Between Communities of Weighted Edges In Figure 7, we investigate the appropriateness of different SIEM for the weighted *Drosophila* connectome, similar to Figure 6. Our goal is the same as previously; ie, to identify whether within-hemisphere connectivity exceeds between-hemisphere connectivity. Figure 7(A) shows a comparison of the within and the between-hemisphere edge blocks. The within-hemisphere edge blocks appear to have a higher proportion of non-zero edges than the between-hemisphere edge blocks. This effect is significant, with the interpretation that within-hemisphere connectivity exceeds between-hemisphere connectivity at $\alpha = .05$ (Mann-Whitney Wilcoxon Test, $n = 103041$, p-value$= 0.0$). Figure 7(B) shows a comparison of the homotopic and heterotopic edge blocks. The homotopic edges appear to have a higher proportion of non-zero edges with smaller edge



Edge Community	No Edge	Edge	Marginal
Within Hem.	73.9%	26.1%	n=50588
Between Hem.	93.6%	6.4%	n=52453
Marginal	n=86475	n=16566	n=103041

Edge Community	No Edge	Edge	Marginal
Homotopic	51.7%	48.3%	n=3290
Heterotopic	85.0%	15.0%	n=99751
Marginal	n=86475	n=16566	n=103041

Figure 6: **Comparing communities of edges in graphs**. **I.(B)** Fisher's exact test shows reasonable statistical power across both homophilic and homotopic block structures in **I.(A)**, with power converging to 1 as effect size and number of vertices grow. **II.(B)** and **II.(C)** the *Drosophila* mushroom body in **II.(A)** shows both homophilic planted partition and homotopic structure (Fisher's exact test, p-values$= 0$). **III.(B)** and **III.(C)** all $N = 1059$ HCP diffusion connectomes show homophilic planted partition structure, with within-hemisphere connectivity exceeding between-hemisphere connectivity (Fisher's exact test, $N = 1059$, Bonferroni corrected p-values$< 10^{-21}$).

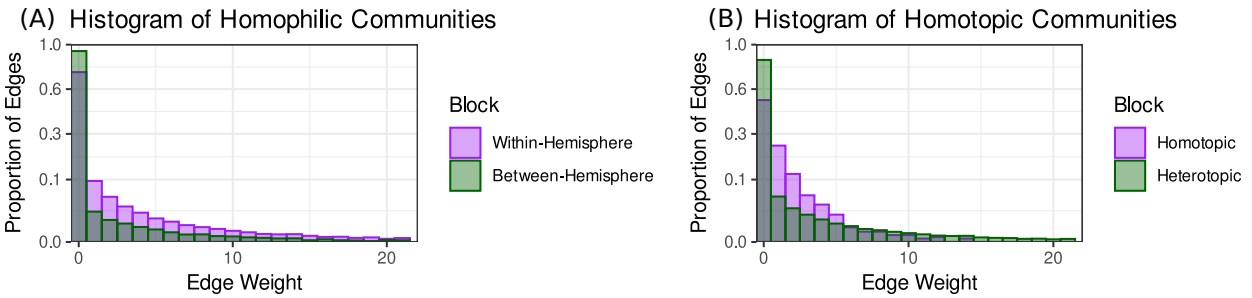


Figure 7: **Goodness of fit of homophilic and homotopic** SIEM **for weighted *Drosophila* mushroom body**. **(A)** A comparison of the homophilic communities as determined by the hemispheres of incident vertices. The relative heights of the bars are normalized by the square root of the proportion due to the fact that the substantial majority of edges in both communities have a weight of 0. Within-hemisphere edges appear to have greater connectivity than between-hemisphere edges. Within-hemisphere edges show significantly higher connectivity than between-hemisphere edges (Mann-Whitney Wilcoxon Test, $n = 103041$, p-value$= 0.0$). **(B)** A comparison of the homotopic edge communities, where homotopic edges are those that are incident a bilateral pair of vertices. Homotopic edges appear to have greater connectivity than heterotopic edges. Homotopic connectivity significantly exceeds heterotopic connectivity (Mann-Whitney Wilcoxon Test, $n = 103041$, p-value$= 0.0$).

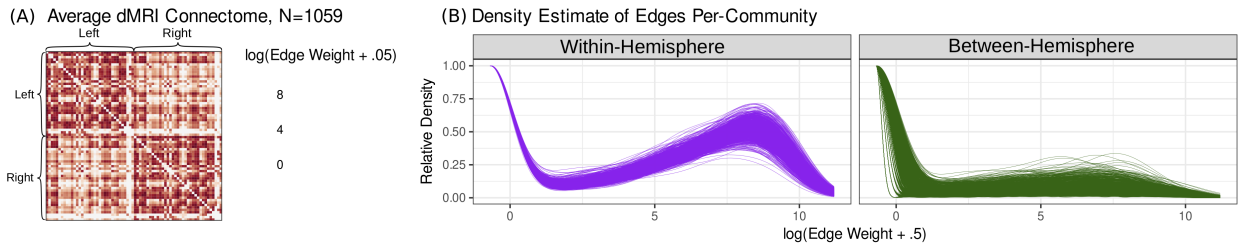


Figure 8: **Goodness of fit of homophilic** SIEM **for diffusion connectomes**. **(A)** The average diffusion connectome over $N = 1059$ connectomes with $n = 70$ vertices from the HCP dataset shows that diffusion connectomes appear to be homophilic, with higher within-hemisphere connectivity than between-hemisphere connectivity. Hemisphere annotations are provided for regions in the left and right hemispheres. Within-hemisphere edges are edges whose vertices are both located in the same hemisphere of the brain (the on-diagonal blocks). **(B)** Density estimates for the within-hemisphere and between-hemisphere edges for each of the $N = 1059$ connectomes. Homophily is tested per-graph using the Mann-Whitney Wilcoxon Test to detect whether the on-diagonal blocks have higher connectivity than the off-diagonal blocks. All $N = 1059$ diffusion connectomes have significantly higher within-hemisphere connectivity than between-hemisphere connectivity after Bonferroni correction at $\alpha = .05$, and the maximum corrected p-value is on the order of 10^{-21}.

weights, and a similar proportion of non-zero edges with larger edge weights. Homotopic connectivity significantly exceeds heterotopic connectivity at $\alpha = .05$ (Mann-Whitney Wilcoxon Test, $n = 103041$, p-value$= 0.0$).

In Figure 8 we explore the appropriateness of the SIEM for diffusion connectomes from the HCP Dataset. Figure 8(A) shows the average diffusion connectome over all participants in the study. Figure 8(B) shows the distribution of edge-weights within-hemisphere versus between-hemisphere. The diffusion connectomes appear to possess homophily; ie, high within-hemisphere connectivity, with lower between-hemisphere connectivity. To test this observation, we employ the `MWW` test. All 1059 diffusion connectomes have significantly higher within-hemisphere connectivity than between-hemisphere connectivity at $\alpha = .05$ after Bonferroni correction [16].

8.3 Model Selection for Appropriate Block Structure Recall that in Section 3.1.2, that for the case of a $K = 2$ SBM, the matrix $\mathbf{B}$ with entries $\mathbf{B}_{kl}$ defines the probability of an edge connecting a vertex in community k with a vertex in community l. By the bias-variance trade-off, simply supposing a unique entry for each block of $\mathbf{B}$ adds an additional level of complexity to the model, and may reduce the quality of inference, so the ability to make a principled decision when faced with numerous potential block structures is of importance. Formally, we are concerned with choosing one of the appropriate block structures from a subset of candidate block structures given in Section3.1.2, presenting a problem in model selection. Our hypotheses are the alternate candidate models, and our goal is to select the hypothesis corresponding to the candidate model that is most supported by the data by using the model with the lowest p-value.

In Figure 9I, we perform simulations where the true graph is either ER, Planted Partition, and Symmetric Heterogeneous, as shown in Figure 9I.(A). Effect size corresponds to the magnitude of the difference between disparate blocks in the model. We find that the χ^2 test is an appropriate test for identification of block structure in unweighted graphs, and successfully recovers the correct block structure as the effect size and the number of vertices increases. Figure 9I.(B) shows the test features both empirical validity and empirical consistency, as in Figure 6.

In Figure 9II, we investigate the appropriate block structure for the unweighted *Drosophila* mushroom body, which is shown in shows the probability of an edge existing within each block of $\mathbf{B}$, where the $n = 319$ vertices in either the left or right hemisphere are partitioned according to hemisphere. The on-diagonal (Left,Left) and (Right,Right) blocks share a similar distribution that is unique from the (Left, Right) and (Right, Left) blocks. Because the *Drosophila* mushroom body is inherently a directed graph, we investigate whether it is ER, Planted Partition, Asymmetric Homogeneous, Symmetric Heterogeneous, or Asymmetric Heterogeneous, using the χ^2 test. Testing indicates that the *Drosophila* mushroom body possesses a planted partition structure (χ^2 test, p-value=0.0). This has the interpretation that the optimal SBM includes a shared probability for the on-diagonal (Left,Left) and (Right,Right) blocks, and a different shared probability for the off-diagonal (Left,Right) and (Right,Left) blocks. An important considerations is that while the optimal SBM is symmetric, the graph itself is directed. This has the implication that while the SBM would posit that edges in the (Left,Right) and (Right,Left) blocks have the same probability, realizations of the (Left,Right) and (Right,Left) block will not necessarily be identical.

Figure 9III investigates the optimal block structure for the N=1059 diffusion connectomes from the HCP dataset. The figure shows the average connectivity for the 3 possible unique entries of the block probability matrix $\mathbf{B}$ for an SBM where vertices are segmented into communities according to hemisphere: Left-Hemisphere Connectivity, Right-Hemisphere Connectivity, and Contralateral (between-hemisphere) connectivity. Because the diffusion connectomes are inherently symmetric, the graph is directionless, and hence it is not possible for the *Left, Right* and *Right, Left* blocks to have different values. We consider 3 possible block structures for the diffusion connectome: ER, Planted Partition, and Symmetric Heterogeneous. On all N=1059 connectomes, the optimal block structures is Planted Partition, using the χ^2 test.

8.4 Model Selection for Appropriate Block Structure in Weighted Connectomes In Figure 10, we investigate the appropriate block structure for the weighted *Drosophila* mushroom body. 10(I) shows the distribution of edges associated with each block of $\mathbf{B}$, where the $n = 319$ vertices in either the left or right hemisphere are partitioned according to hemisphere. Again, the weighted *Drosophila* mushroom body is directed, so assuming symmetry would not be sensible. We investigate whether the *Drosophila* mushroom body is ER, planted partition, symmetric heterogeneous, or asymmetric heterogeneous SBM, using Kruskal-Wallis (KW), Distance Correlation (DCorr), and Analysis of Variance (ANOVA). Each method identifies the planted partition SBM as the most appropriate block model. This has the interpretation that the best-fit SBM includes a shared distribution for the on-diagonal (Left,Left) and

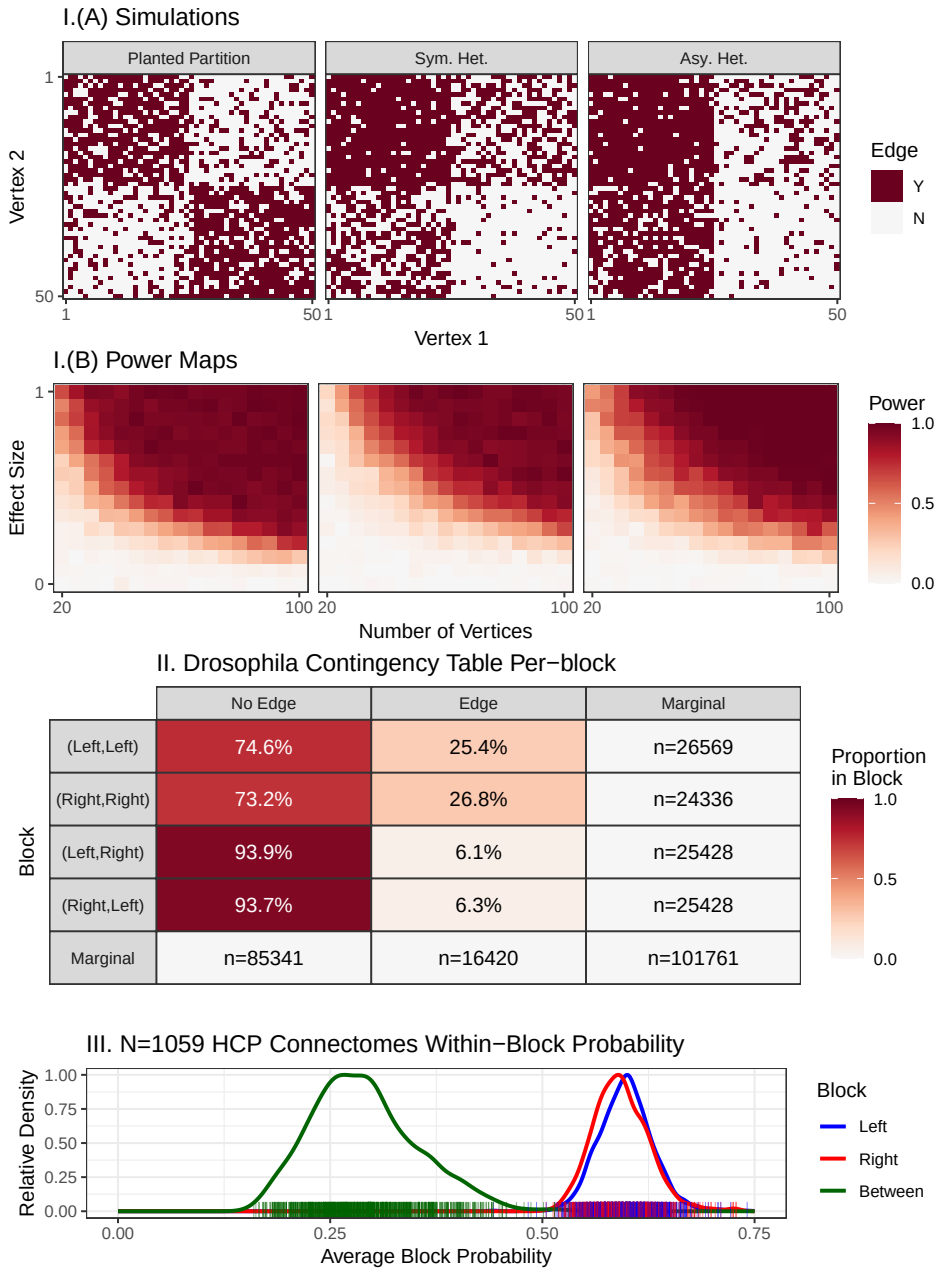


Figure 9: **Estimating optimal block structure**. **I.(B)** χ^2 test is effective for identifying the ideal block structure across disparate candidate block structures from **I.(A)**, as power improves as both effect size and graph size increase. **II.** The *Drosophila* mushroom body displays a planted partition structure (χ^2 test, p-value=0.0), where *(Left, Left)* and *(Right, Right)* blocks share a different probability from the *(Left, Right)* and *(Right, Left)* blocks. **III.** Similarly, all $N = 1059$ HCP diffusion connectomes show planted partition structure, with a similar interpretation to the *Drosophila* result.

(Right,Right) blocks, and a different shared distribution for the off-diagonal (Left,Right) and (Right,Left) blocks. An important considerations is that while the best-fit SBM is symmetric, the graph itself is directed. This has the implication that while the best-fit SBM would posit that edges in the (Left,Right) and (Right,Left) blocks have the same distribution, realizations of the (Left,Right) and (Right,Left) block will not necessarily be identical.

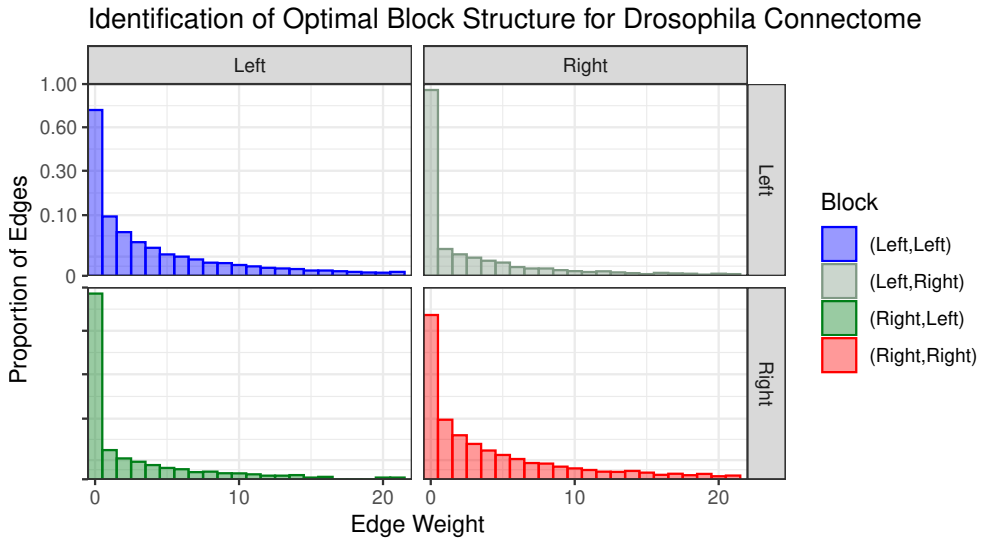


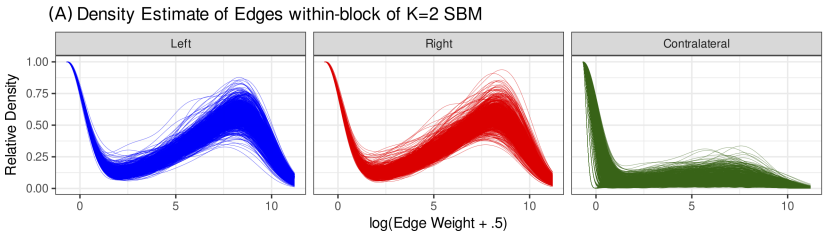
Figure 10: **Identifying the appropriate block structure of the *Drosophila* mushroom body**. We investigate the appropriate block structure in the *Drosophila* mushroom body, with $n = 319$ vertices in the left or right hemisphere. Each block corresponds to the proportion of edges with the listed edge weight. As in Figure 7, the proportion of edges are shown on a scale in which bar height corresponds to the square root of the proportion, due to the presence of a large number of zero-weight edges. We investigate whether the SBM is ER, planted partition, symmetric heterogeneous, or asymmetric heterogeneous SBM, using Kruskal-Wallis (KW), Distance Correlation (DCorr), and Analysis of Variance (ANOVA) for model selection. All approaches identify planted partition SBM as the best-fit block structure.

In Figure 11, we investigate the appropriate block structure for diffusion connectomes, analogous to the single graph investigations using the *Drosophila* mushroom body in Figure 10. Panel **(A)** demonstrates the distribution of edges associated with each block of $\mathbf{B}$. Panel **(B)** shows the fraction of diffusion connectomes that accept each of the candidate hypotheses, using 3 different approaches for weighted graph model selection: Kruskal-Wallace [57], Distance Correlation (Dcorr) [96], and Ananysis of Variance (ANOVA) [36, 86]. Diffusion connectomes tend to display planted partition structure across all model selection approaches.

8.5 Same Network, Different Communities In the case of 2-block SBMs with positive semi-definite block probability matrix $\mathbf{B} = [a, b; b, c]$, there are two structures of interest: affinity and core-periphery. In affinity structure, $a, c \gg b$, that is the within-block connectivity is relatively higher than that of between-block connectivity. In the core-periphery structure, $a \gg b, c$, that is one block has relatively higher within-block connectivity than those of other block's within-block probability and between-block connectivity.

In this section, we examine the two spectral embedding clustering approaches described in Section 5.1.1 which produce different clusterings depending on the SBM model [20, 77]. In short, ASE clustering tends to favor core-periphery structure while LSE clustering tends to favor affinity structure.

We consider graphs generated from 4-block SBM with $n = 4000$ vertices, membership vector $\vec{\pi} =$



(B) Identified Block Structure

Method	ER	Sym. Hom.	Sym. Het.
Kruskal-Wallis	0	1059	0
DCorr	0	1059	0
ANOVA	0	1059	0

Figure 11: **Identification of appropriate block structure in diffusion connectomes**. We investigate the appropriate block structure in the diffusion connectomes from the HCP Dataset, with $n = 70$ vertices, and $N = 1059$ graphs. **(A)** the empirical distribution of edges for each of the 4 blocks of edges for each between and within-hemisphere pair for the left and right hemispheres respectively. As the diffision connectomes are inherently symmetric, the off-diagonal blocks are inherently symmetric. The hypothesized models are that the graph is ER, planted partition SBM (Plant Part.), or the symmetric heterogeneous SBM (Sym. Het.). **(B)** The number of connectomes from the dataset for which the specified candidate model is selected. All methods for selection of optimal block structure identify diffusion connectomes as planted partition SBM, which has the interpretation that the optimal structure is to assume that the on-diagonal left and right blocks share a common distribution that differs from the off-diagonal contralateral blocks. This conclusion holds across all diffusion connectomes within the dataset.

$[0.25, 0.25, 0.25, 0.25]$, and the block probability matrix

$$\mathbf{B} = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0.01 & 0.02 & 0.01 & 0.002 \\ 0.02 & 0.1 & 0.002 & 0.015 \\ 0.01 & 0.002 & 0.01 & 0.02 \\ 0.002 & 0.015 & 0.02 & 0.01 \end{bmatrix} \end{array}$$

The above 4-block SBM exhibit both affinity and core-periphery structures when projected down to 2-blocks, which are shown below:

$$\mathbf{B}_{affinity} \approx \begin{array}{c} \\ AB \\ CD \end{array} \begin{array}{c} \begin{array}{cc} AB & CD \end{array} \\ \begin{bmatrix} 0.04 & 0.007 \\ 0.007 & 0.04 \end{bmatrix} \end{array}, \quad \mathbf{B}_{core} \approx \begin{array}{c} \\ AC \\ BD \end{array} \begin{array}{c} \begin{array}{cc} AC & BD \end{array} \\ \begin{bmatrix} 0.01 & 0.01 \\ 0.01 & 0.06 \end{bmatrix} \end{array}$$

Blocks AB and CD form $B_{affinity}$, which exhibit the affinity structure, while blocks AC and BD form B_{core}, which exhibit the core-periphery structure. A network is sampled from the 4-block SBM model, and spectral clustering is performed (see Section 5.1.4) with embedding dimension $\hat{d} = 2$ and $K = 2$ number of clusters. Figure 12 shows the spectral clustering results. On the left Figure, clustering with `LSE` shows the blocks forming affinity structures are grouped together, and, on the right Figure, clustering with `ASE` shows the blocks forming core-periphery structures grouped together. Thus, the two different spectral clustering methods provide two different groups that are both meaningful.

8.6 Detecting Communities with Spectral Clustering Many of the techniques described above rely on knowing an *a priori* grouping of nodes or edges, but in many real-world examples this information is not available. Additionally, one may seek to discover communities in the network, either for modeling the network as a block-model or to reveal groups of similar nodes.

As described in Section 5.1.4, one can embed a graph via `ASE` or `LSE` and then use `GMM` to reveal communities of nodes. Here, we separately embed both the left and right hemisphere induced subgraphs of the *Drosophila* larva connectome using `ASE` (see [76] for an extensive investigation) with

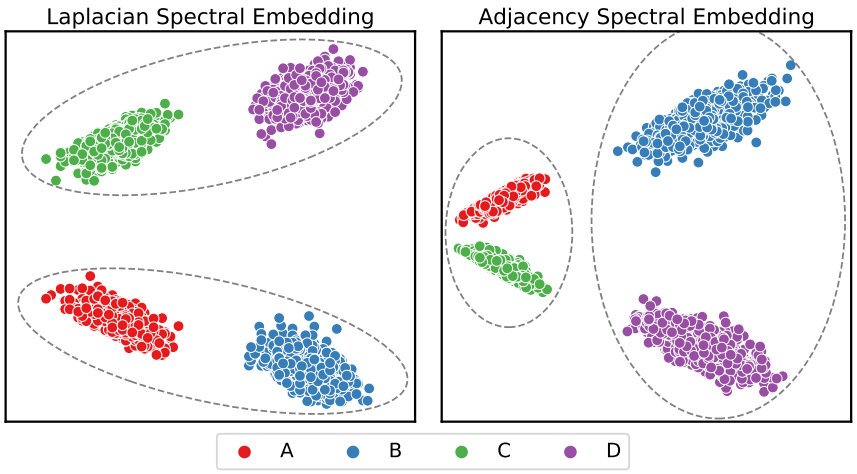


Figure 12: **Different clustering results from ASE and LSE.** For both ASE and LSE, the network was embedded into $d = 2$ dimensions, and GMM with $K = 2$ clusters were fit. The dots represent vertices in the embedded space and the colors correspond to block memberships. The dashed black ellipses define the vertices that were clustered into same group. *(Left)* clustering the embeddings from LSE results in affinity clustering. *(Right)* clustering the embeddings from ASE results in core-periphery clustering.

$\hat{d} = 3$. GMM was performed independently on both hemispheres, with the clustering assignments and embeddings shown in Figure 13. Note that while the embedding and clustering of both hemispheres were performed separately, similar structures emerge for the left and right. In particular, each cluster is mostly comprised of a single cell type. Thus, spectral clustering can provide neuroscientists to find meaningful communities when the assignment is not known.

9 Applications for Multi-Graph Data In this section, we explore the applications of the multiple graph models in Section 3.2 and the algorithms in Section 5.2 using simulated, *Drosophila* mushroom body, and HCP connectomes.

9.1 Matching Vertices between Subgraphs For many statistical approaches on graphs, knowing an alignment or matching between the vertices of one graph and another can be useful. For instance, if each neuron on the left hemisphere of the brain has a corresponding neuron in the right hemisphere, then both hemispheres could be jointly embedded and compared using techniques such as OMNI or MASE. In the case of the mushroom body network, hemilateral neuron pairs were identified for 198 of the neurons considered in Figure 13, yielding 99 neuron pairs.

Here, we test the ability of graph matching techniques to identify this structure in an unsupervised manner, based only on the network topology (note that the neuron pairs considered here were based on both topology and morphology). We perform unseeded graph matching between the subset of left and right hemisphere neurons for which pairs are known. We restart the algorithm 256 times, and choose the run with the best objective function value (not matching accuracy). Results are shown in Figure 14. This matching correctly identified 78.8% (78/99) of neuron pairs, and all incorrectly matched neurons were matched to a neuron of the correct cell type.

Given a new connectome, where the correspondence between neurons is not known, this method can provide neuroscientists with a faster and statistically-grounded estimate of neuron pairing.

9.2 Testing for Significant Edges We consider two populations of networks generated from an ER model and a 2-block Kidney – Egg SBM model. All networks have $n = 20$ vertices and $\pi = [0.25, 0.75]$. The block probability matrices for each population is given by $\mathbf{B}^{(1)} = [p, p; p, p]$ and $\mathbf{B}^{(2)} = [p+\delta, p; p, p]$ where $p = 0.5$. The difference between the two population is in the first block, $\mathbf{B}_{11}$, and δ is the

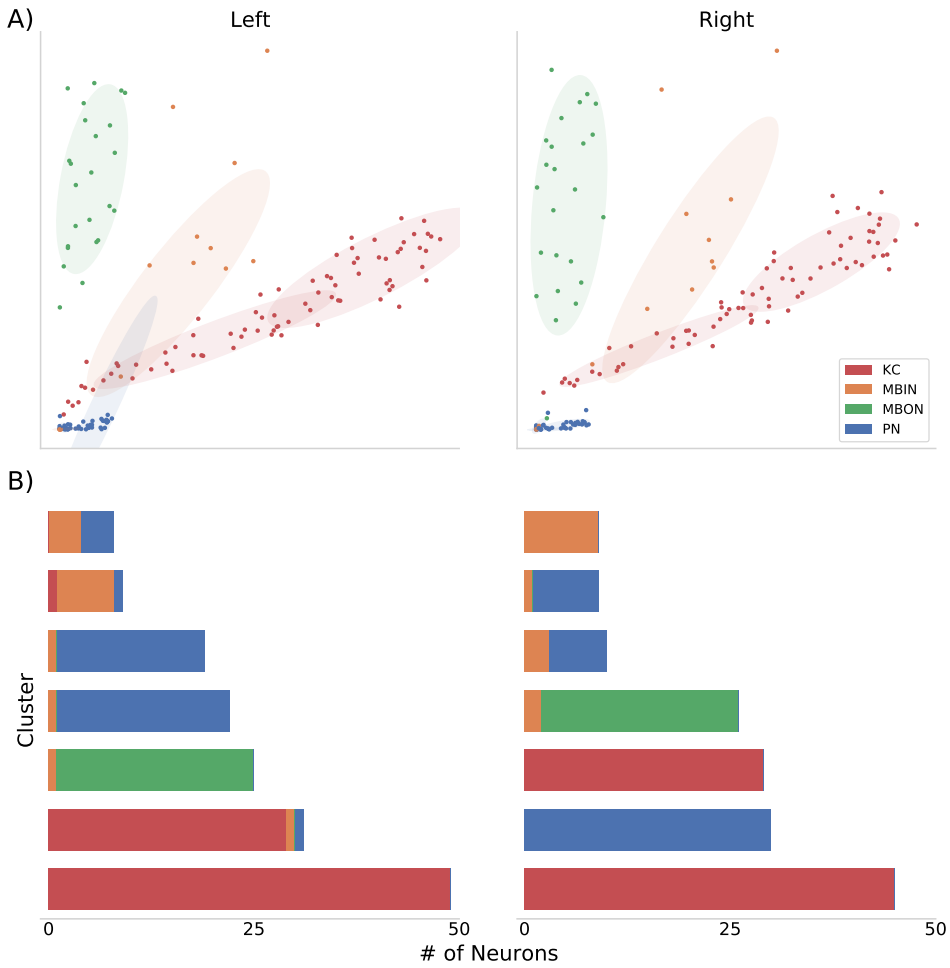


Figure 13: **Spectral clustering of the *Drosophila* mushroom body network. (A)** First "in" embedding dimension is plotted against the first "out" embedding dimension for both the left and right hemisphere networks (note that the clustering was performed in six dimensions, but only two are shown here for visualization). Each point represents a neuron, colored by its corresponding cell type. Ellipses show the clusters predicted by Gaussian mixture modeling, colored according to the cell type with the most neurons in that cluster. **(B)** Stacked barplots showing each cluster's composition in terms of neuron cell type, for both the left and the right hemisphere clusterings. Each cluster is mostly comprised of a single cell type for both left and right hemisphere networks, meaning that spectral clustering can recover true communities.

magnitude of the difference which ranges from 0 to $(1-p)$. In other words, δ is the effect size. Total of m networks are sampled ($\frac{m}{2}$ networks per population). For each edge, the t-test test statistics is computed between the two populations, which are then ranked from largest to smallest in magnitude. Ranking of the test statistics and a cutoff is utilized rather than p-value corrections (e.g. Bonferroni) to control for false positive rate. In this case, the ten edges with largest magnitudes are considered since we expect ten edges to be different. Non-parametric tests are not considered since many of them are based on ranking the underlying data, which is not sensible for binary data. The performance is evaluated with recall@10, which quantifies the fraction of the top ten ranked edges are indeed the truly different edges, averaged over 100 repeated trials.

Figure 15(A) shows that when the effect size is small ($\delta \leq 0.05$), significant edges cannot be detected even at largest sample sizes ($m = 1000$). On the other hand, when effect size is large ($\delta \geq 0.45$),

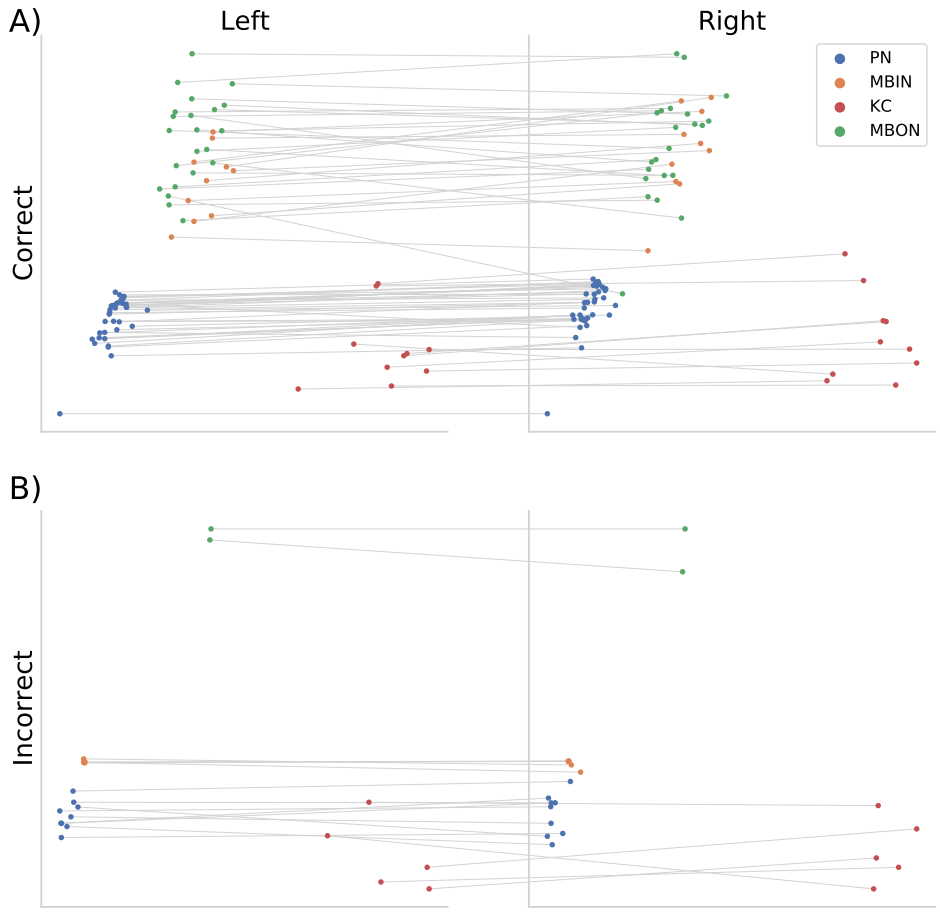


Figure 14: **Graph matching on the *Drosophila* mushroom body network.** All panels show the first two dimensions of PCA on the ASE embedding of the mushroom body network (for visualization purposes). Each point represents a neuron in the network which has a manually identified pair in the opposite hemisphere, and colors represent the cell type of a given neuron. Lines show the neuron pair that was predicted by graph matching. **(A)** All of the correctly matched neuron pairs. 78.8% of neuron pairs (78/99) were correctly matched. **(B)** All of the incorrectly matched neuron pairs. Note that all of the incorrectly matched neurons are matched to neurons of the same cell type.

significant edges can be perfectly detected at relatively small sample sizes ($m \geq 30$).

Connectivity in human brains was analyzed using the structural connectomes (Section 7.2). For each edge, the class conditional mean, which is the estimated connectivity probability, is computed for females ($m = 572$) and males ($m = 488$). The sample sizes and difference in conditional means, which is the estimated effect size, are used to find the closest recall@10 values from the simulated experiment, denoted empirical trustworthiness shown in Figure 15(B). Thus, empirical trustworthiness is the confidence in which one can trust that a significant edge is truly significant. There are 2380 possible total edges in connectomes with 70 vertices, but only 49 edges have trustworthiness ≥ 0.9, meaning one can only trust significance for small set of edges.

9.3 Testing for Significant Edges in Weighted Networks We consider two populations of networks generated from a 2 block SBM, except edges are now sampled from truncated normal distribution to emulate correlation matrices. All networks have $n = 20$ vertices and $\vec{\pi} = [0.25, 0.75]$. The block edge

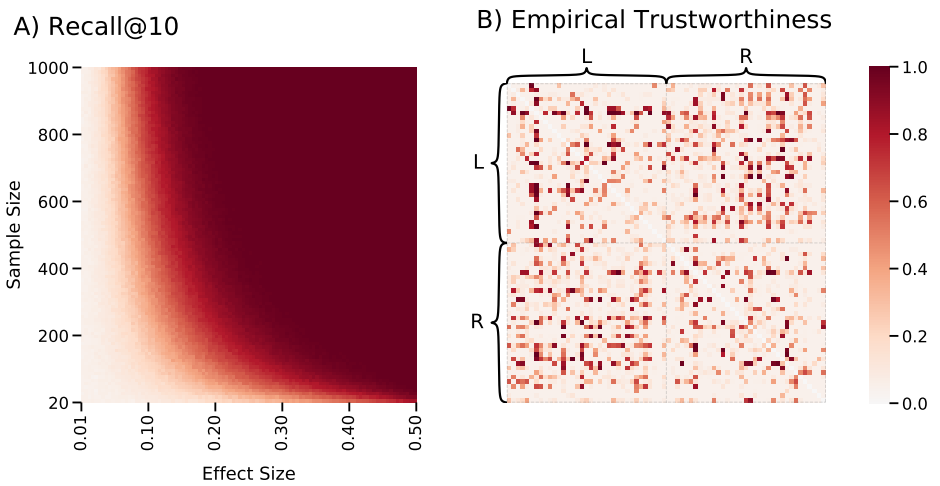


Figure 15: **Performance of finding significant edges in two different populations of networks. (A)** Recall for varying sample size and effect size when comparing two populations of binary networks using t-test. The color bar represents recall@10 averaged over 100 trials. When effect size is small, significant edges cannot be detected even at large sample size. When effect size is large, significant edges can be detected at small sample sizes ($m = 1000$). **(B)** Analysis of structural connectomes from the HCP data, and the vertices are organized by left (L) and right (R) hemispheres. Edge weights are binarized to parallel the simulations. Heatmap shows the empirical trustworthiness of significant edges when comparing each edge between females and males.

distribution matrices for each population is given by

$$\mathbf{B}^{(1)} = \begin{bmatrix} \mathsf{TN}(0, 0.25, -1, 1) & \mathsf{TN}(0, 0.25, -1, 1) \\ \mathsf{TN}(0, 0.25, -1, 1) & \mathsf{TN}(0, 0.25, -1, 1) \end{bmatrix}$$

$$\mathbf{B}^{(2)} = \begin{bmatrix} \mathsf{TN}(0 + \delta, 0.25 + \phi, -1, 1) & \mathsf{TN}(0, 0.25, -1, 1) \\ \mathsf{TN}(0, 0.25, -1, 1) & \mathsf{TN}(0, 0.25, -1, 1) \end{bmatrix}$$

where $\mathsf{TN}(\mu, \sigma^2, a, b)$ denotes a truncated normal distribution with mean μ and variance σ^2 such that all values are in $[a, b]$. Total of m networks are sampled ($m/2$ networks per population). One population has the same edge weight distribution for all edges, and the second population's first block edges has either a different mean, δ, or variance, $0.25 + \phi$. For each edge, test statistics are computed with three different tests: 1) t-test, 2) Mann-Whitney (MW) U test, which is a non-parametric test of medians, and 3) two-sample Kolomogrov-Smirnov (KS) test, which is test of two distributions. Similar to experiment 1, the test statistics are sorted to find the ten most significant edges, and the performance is evaluated with recall.

Figure 16 shows the results by varying the sample size, mean, and variance. Figure 16 top row shows that all three tests can identify edges that are different in means, and that no particular test is superior than another. Figure 16 bottom row shows that only KS test can detect changes in variance when the means are kept the same. This is because t-test and MW test ultimately test for differences in centrality (e.g. mean or median), where as KS tests for any differences between a pair of observed distributions.

Functional connectivity in human brains was analyzed using functional connectomes estimated using fMRI data from the HCP dataset. In functional connectomes, the edges represent correlations of changes in blood flow between a pair of ROIs, which is a proxy for correlations of brain activity. For each edge, the class-conditional mean and the variance of truncated normal distribution are computed for males ($m = 330$) and females ($m = 407$). Networks are then simulated as above using the 2-block weighted SBM, but the parameters for first block, $\mathbf{B}_{11}$, is substituted with class-conditional means

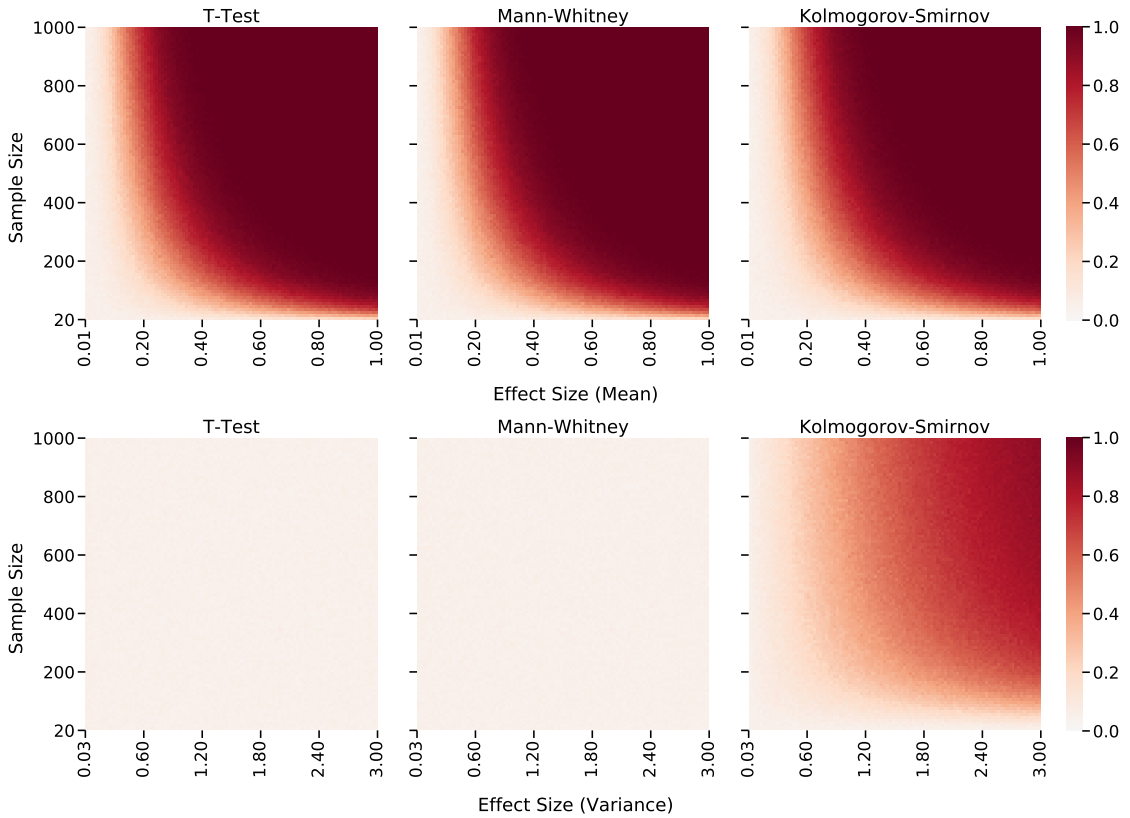


Figure 16: **Performance of finding significant edges that have different weight distributions.** Recall@10 for each edge when comparing two populations of weighted networks using t-test, Mann-Whitney, and Kolmogorov-Smirnov tests. The color bar represents recall averaged over 100 trials. *(Top row)* Results for varying the mean δ and sample size wile keeping the variance is same ($\phi = 0$). All three tests perform equally, and can detect significant edges when edge distributions differ in means. *(Bottom row)* Results for varying the variance ϕ and sample size wile keeping the mean same ($\delta = 0$). T-test and Mann-Whitney test cannot detect changes in variance regardless of the sample and effect size. KS test is the only test that can detect changes in variance.

and variances. The performance is measured with recall@10, denoted empirical trustworthiness in Figure 17, is measured using KS test. Again, the empirical trustworthiness shows how one can trust that the edge is truly different. There are 70 vertices with 2380 total edges, but only 256 edges have trustworthiness ≥ 0.9.

9.4 Testing for Significant Edges Using Communities in Binary Networks In previous Section 9.2, the community structure was ignored even though the generative process produced two communities. In the following experiment, the community assignments are used to test whether all edges within a community or across communities are significantly different. Formally, the following hypothesis test is considered:

$$H_0 : \mathbb{P}[\mathbf{B}_{ij} | Y = 0] = \mathbb{P}[\mathbf{B}_{ij} | Y = 1]$$
$$H_A : \mathbb{P}[\mathbf{B}_{ij} | Y = 0] \neq \mathbb{P}[\mathbf{B}_{ij} | Y = 1]$$

where $\mathbb{P}[\mathbf{B}_{ij} | Y = y]$ denotes class-conditional distribution of edges that belong to community i and j, and $i, j \in [K]$ where K denotes the number of communities. When $i = j$, the edges are incident to vertices that belong to the same community. When $i \neq j$, the edges are incident to vertices that do not belong in the same community. In this setting, a total of $\frac{(K)(K+1)}{2}$ null hypothesis are tested.

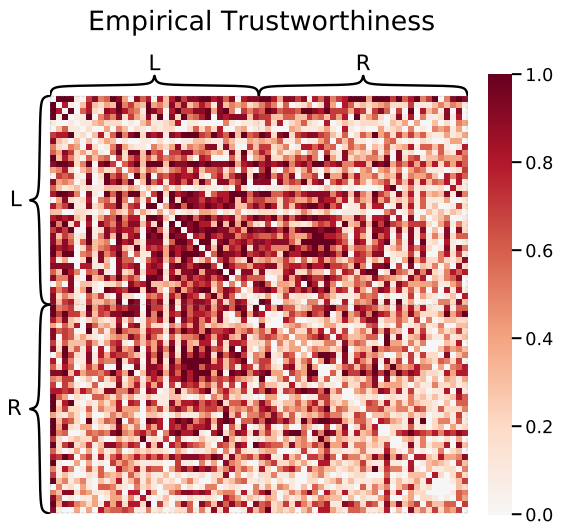


Figure 17: Functional connectomes are derived from the HCP data. Vertices are defined by Desikan parcellation into 70 ROIs, and are organized by hemisphere, denoted left hemisphere (L) and right hemisphere (R). Edge weights are correlations represent correlation of brain activity between a pair of ROIs. For each edge, the class-conditional means and the variances for females ($m = 407$) and males ($m = 330$) are computed, which are used to simulate weighted 2-block SBM. Recall@10, denoted empirical trustworthiness, is measured from test statistics using KS test. Out of the 2380 total edges, only 256 edges have trustworthiness ≥ 0.9.

We consider two populations of networks with the connectivity probability matrices as below,

$$\mathbf{B}^{(1)} = \begin{bmatrix} p & p \\ p & p \end{bmatrix}, \ \mathbf{B}^{(2)} = \begin{bmatrix} p+\delta & p \\ p & p \end{bmatrix}$$

with $n = 50$ vertices, and membership vector, $\vec{\pi} = [0.5, 0.5]$. Total of m networks are sampled ($m/2$ networks per population). Since $K = 2$, community assignment results in three sets of edges, two within communities and one across communities. The t-test statistic was computed for each set of edges, and significant edges are identified by the hypothesis test that resulted in largest test-statistic. The performance is measured by precision, which measures false positive rate, and recall, which measures true positive rate.

Figure 18 shows the results of using t-test as the effect size is changed using known and estimated community assignments. When the community assignment is known *a priori*, significant edges can be perfectly detected with no false positives at low sample sizes ($m = 10$) and effect size ($\delta \geq 0.05$). However, estimating community assignments results in large number of false positives edges as shown in precision plots for both JRDPG and COSIE models since recovery of community assignments is correlated with magnitude of the effect size. When effect size is small, communities cannot be reliably recovered for both JRDPG and COSIE models, which results in false positive tests. As effect size increases, community recovery improves and the number of false positive edges decrease at effect size ($\delta \geq 0.2$).

9.5 Testing for Significant Edges Using Communities in Weighted Networks We consider weighted 2-block SBM similar to that of Section 9.3, but with $n = 50$ vertices, membership vector, $\vec{\pi} = [0.5, 0.5]$,

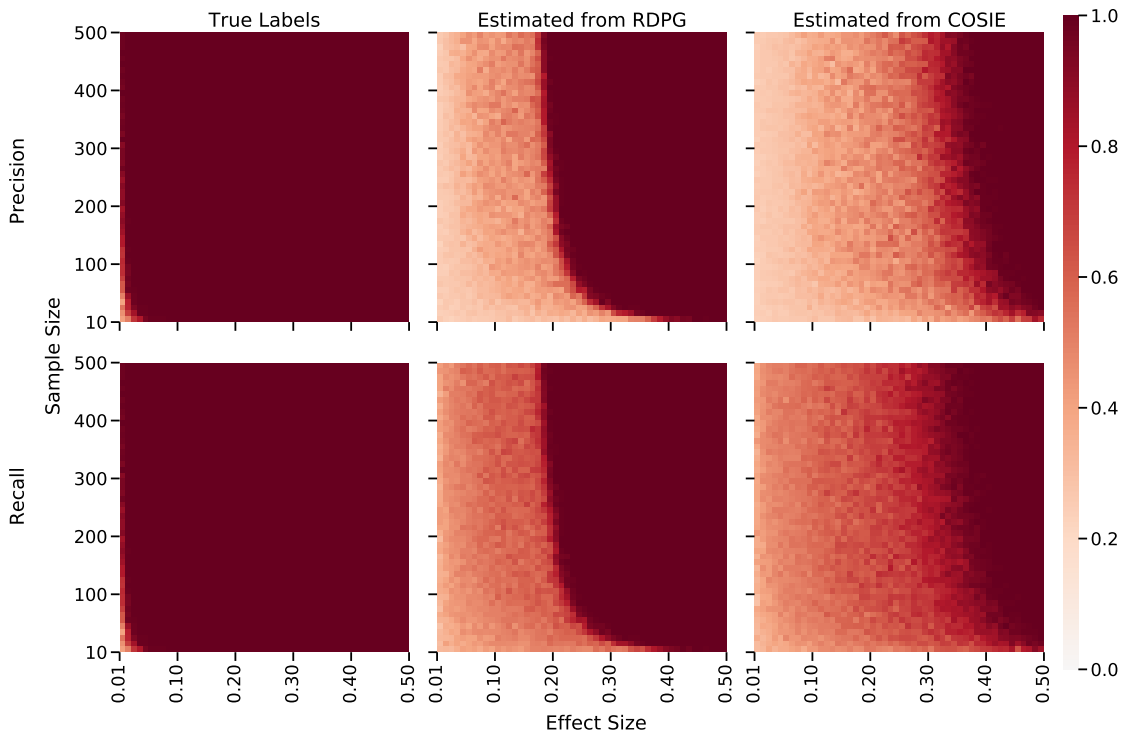


Figure 18: **Performance of finding significant edges using either known or estimated community structure.** Precision *(top row)* and recall *(bottom row)* for significant edges using t-test on sets of edges from within community or across communities averaged over 50 trials. *(Left column)* shows the precision and recall when using true community assignments. At low sample sizes ($m = 10$) and low effect size ($\delta \geq 0.05$), community wise testing results in perfect precision and recall. *(Middle column)* shows the results for using community assignments estimated under the JRDPG model. *(Right column)* shows the results for using community assignments estimated under the COSIE model. Since recovery of community assignment is related to the effect size, spectral clustering results in misclassified vertices. As a result, precision is low at effect sizes ≤ 0.2. As effect size increases, the communities become more identifiable, and results in increased precision for JRDPG and COSIE models. However, COSIE model requires larger effect size to reach precision ≥ 0.95.

and block edge distribution is as below:

$$\mathbf{B}^{(1)} = \begin{bmatrix} \mathsf{TN}(0, 0.25, -1, 1) & \mathsf{TN}(0, 0.25, -1, 1) \\ \mathsf{TN}(0, 0.25, -1, 1) & \mathsf{TN}(0, 0.25, -1, 1) \end{bmatrix}$$

$$\mathbf{B}^{(2)} = \begin{bmatrix} \mathsf{TN}(0+\delta, 0.25+\phi, -1, 1) & \mathsf{TN}(0, 0.25, -1, 1) \\ \mathsf{TN}(0, 0.25, -1, 1) & \mathsf{TN}(0, 0.25, -1, 1) \end{bmatrix}$$

We proceed with the same experiment as that of Section 9.4, while changing the means (δ) or the variances (ϕ). The community assignment is estimated using `OMNI` under JRDPG model and `MASE` under COSIE model. The KS test statistic was computed for each set of edges, and significant edges are identified by the hypothesis test that resulted in largest test-statistic. The performance is measured with precision and recall.

Figure 19 shows the results when varying the mean (δ) and Figure 20 shows the results when varying the variance (ϕ). When the community assignment is known *a priori*, significant edges can be perfectly detected with no false positives at low sample sizes ($m = 10$) and effect size ($\delta \geq 0.1, \phi \geq 0.12$). When means are changed, communities can be perfectly recovered under JRDPG model, but communities cannot be reliably recovered under COSIE model. When the edge distributions are different by

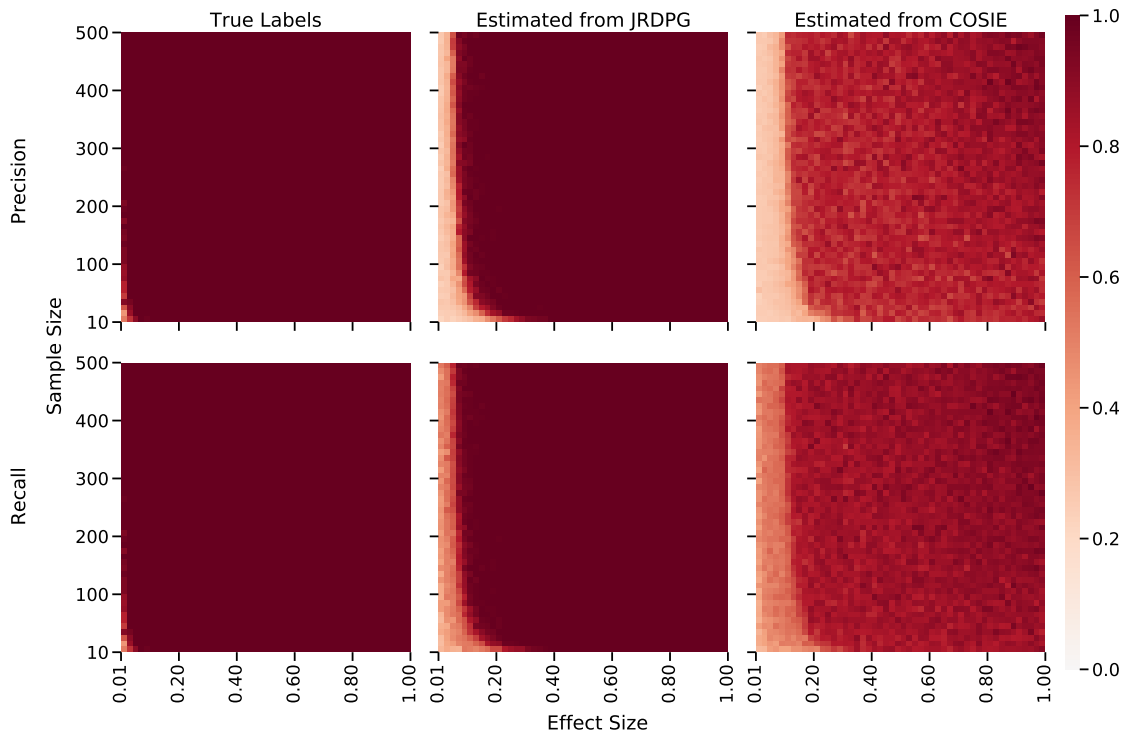


Figure 19: Precision *(top row)* and recall *(bottom row)* for significant edges using K−S test averaged over 50 trials. Effect size (x-axis) is the difference in means of the truncated normal distribution for $\mathbf{B}_{1,1}$. *(Left column)* shows the precision and recall when using known community assignments. At low sample sizes ($m = 10$) and low effect size ($\delta \geq 0.1$), community wise testing results in perfect precision and recall. *(Middle column)* shows the results for using community assignments estimated under the JRDPG model. Even at low effect size ($\delta \geq 0.15$), communities can be perfectly recovered. All significant edges can be detected without false positives. *(Right column)* shows the results for using community assignments estimated under the COSIE model. Under this model, communities cannot be perfectly recovered, resulting in false positive edges and false negative edges.

variance, recovering communities is impossible regardless of the statistical model. This suggest that both JRDPG and COSIE models are not appropriate when studying differences in variances.

9.6 Testing for Significant Vertices In this section, we test for significant vertices using different representations of vertices. Simplest representation is a set of edges, where the corresponding row (or column) of a vertex in the adjacency matrices are collected and tested for difference. Another is the low-dimensional latent-space representation using the JRDPG and COSIE models, and the latent positions of vertices are tested for difference. Since all representations are multivariate, hypothesis are tested using Hotelling's test, which is a multivariate generalization of t-test.

We consider a population of planted partition SBM and a symmetric heterogeneous SBM in two different settings. In both settings, the planted partition SBM has $\mathbf{B}^{(1)} = [0.125, 0.0625; 0.0625, 0.125]$ block probability matrix. In setting 1, the symmetric heterogeneous SBM has $\mathbf{B}^{(2)} = [0.125, 0.088; 0.088, 0.25]$ block probability matrix, and in setting 2, $\mathbf{B}^{(2)} = [0.125, 0.0625; 0.0625, 0.25]$. The vertices that belong to the second block, which has the different within-block probability, are considered significant vertices, and we vary the number of vertices that belong to the second block. Total of $m = 100$ networks are sampled per population, and the p-values are computed using Hotelling's on each of the three vertex representations for each vertex. Vertices with p-values less than $\alpha = 0.05$ after Bonferroni correction are considered significant. The performance is measured via true positive rate (TPR), false positive rate (FPR), and recall@K, where K is the number of significant vertices.

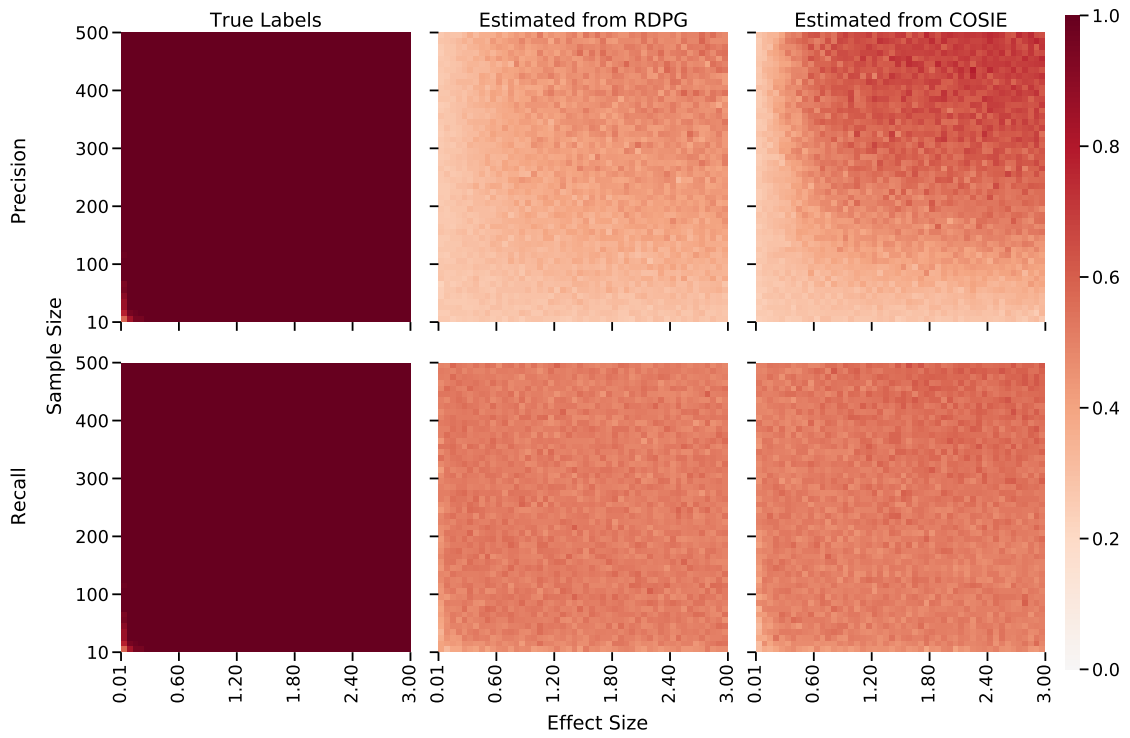


Figure 20: Precision *(top row)* and recall *(bottom row)* for significant edges using K-S test sets on of edges from within community or across communities averaged over 50 trials. Effect size (x-axis) is the difference in variances of the truncated normal distribution for $B_{1,1}$. *(Left column)* shows the precision and recall when using known community assignments. At low sample sizes ($m = 10$) and low effect size ($\phi \geq 0.12$), community wise testing results in perfect precision and recall. *(Middle column)* shows the results for using community assignments estimated under the JRDPG model. *(Right column)* shows the results for using community assignments estimated under the COSIE model. Communities cannot be recovered under both JRDPG and COSIE model regardless of effect size and sample size. As a result, community-wise testing result in large number of false positive edges.

Figure 21 shows that the p-values cannot necessarily be trusted. That is, in some settings, the significant vertices cannot be trusted due to uncontrolled FPR. However, the sorting of p-values can be trusted in both settings. Thus, in situations when the underlying model is not known (i.e. in real data), one should trust the sorting of the p-values (or test statistic), but not the magnitudes.

10 Summary

1. Don't rely on network statistics to characterize populations of connectomes. In general, network statistics don't characterize the data that well, and are correlated with one another. Thus, any claim that a specific statistic explains a phenotypic property of a person is based on spurious reasoning.
2. Do use statistical models developed for networks. Statistical models allow for testing a variety of hypotheses, such as testing for appropriate models and finding significant vertices or communities.
3. Do use spectral clustering methods for determining community structure. Theoretical and empirical results show that spectral clustering methods can estimate meaningful and trustworthy community structures. However, note that different methods can provide different, but complementary results.
4. Do use appropriate hypothesis tests. For example, t-test is appropriate for binary connectomes, but typically invalid and/or under-powered for weighted connectomes.

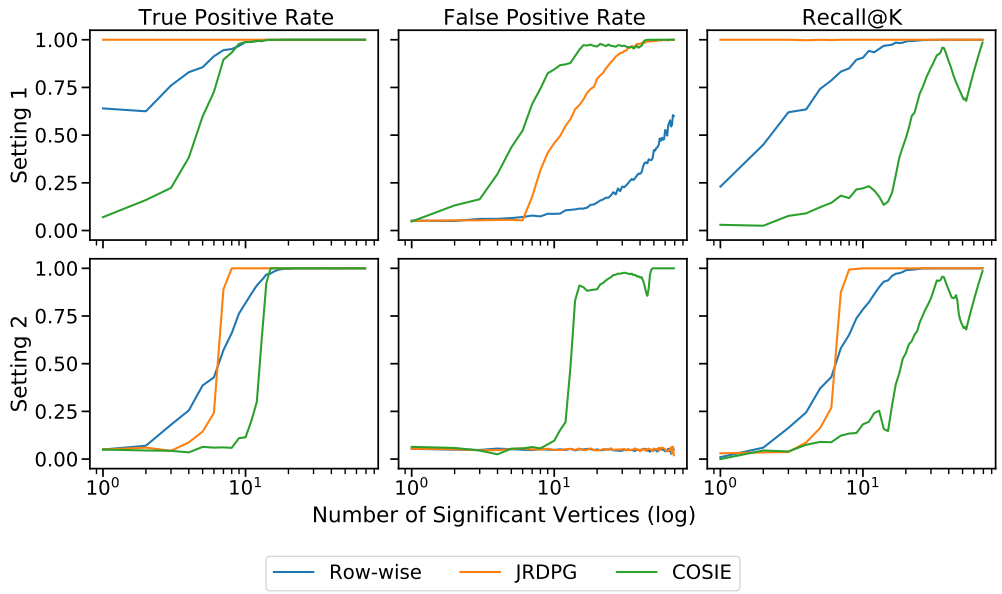


Figure 21: **Performance for finding significant vertices using various representations of vertices.** We compare a population of graphs from a planted partition SBM and another from a symmetric heterogeneous SBM in two different settings. The number of vertices for each graph is kept constant ($n = 70$), but the number of significantly different vertices is varied (x-axis). *(Top row)* In this setting, all three representations are not valid as the false positive rate increases with the number of significant vertices. *(Bottom row)* In this setting, row-wise and JRDPG representations are valid while COSIE representation is not. In both settings, the sorting of the p-values can be trusted as recall@K increases as number of significant vertices increase.

5. Don't trust the p-values when performing multiple hypothesis tests. Multiple testing requires corrections to control the false positive rate, all of which are inappropriate for connectomics data.
6. Do trust the sorting of the p-values when performing multiple hypothesis tests. That is, consider the tests with smallest p-values to reject the null hypothesis as the sorting can be trusted, but not necessarily the magnitudes of p-values.

Connectomics is an exciting area and is full of interesting ideas, which has led to the emergence of a variety of analysis frameworks. However, the use of statistical modeling in connectomics is still relatively sparse, especially compared to other areas of science. The key conceptual hurdle in statistical modeling of connectomes is to model the entire connectome rather than just edges or features while taking into account the structures and interactions within a connectome. This article provides an overview of current analysis frameworks of connectomics data, and how statistical models can be incorporated to improve current analysis methods.

Code All graph related simulations and analysis were performed using `GraSPy` (https://neurodata.io/graspy/) and all multivariate hypothesis testing was done using `hyppo` (https://neurodata.io/hyppo) [23, 73].

DISCLOSURE STATEMENT The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS This work is graciously supported by the DARPA, under agreement numbers FA8650-18-2-7834 and FA8750-17-2-0112 and Microsoft Research.

References

[1] Afshin-Pour B, Hossein-Zadeh GA, Strother SC, Soltanian-Zadeh H. 2012. Enhancing reproducibil-

ity of fmri statistical maps using generalized canonical correlation analysis in npairs framework. *NeuroImage* 60:1970–1981

[2] Airoldi EM, Blei DM, Fienberg SE, Xing EP. 2008. Mixed Membership Stochastic Blockmodels. *J. Mach. Learn. Res.* 9:1981–2014

[3] Akaike H. 1974. A new look at the statistical model identification. *IEEE transactions on automatic control* 19:716–723

[4] Alexander LM, Escalera J, Ai L, Andreotti C, Febre K, et al. 2017. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific data* 4:170181

[5] Amico E, Marinazzo D, Di Perri C, Heine L, Annen J, et al. 2017. Mapping the functional connectome traits of levels of consciousness. *Neuroimage* 148:201–211

[6] Anscombe FJ. 1973. Graphs in statistical analysis. *The american statistician* 27:17–21

[7] Arroyo J, Athreya A, Cape J, Chen G, Priebe CE, Vogelstein JT. 2019. Inference for multiple heterogeneous networks with a common invariant subspace. *arXiv preprint arXiv:1906.10026*

[8] Arroyo J, Levina E. 2020. Simultaneous prediction and community detection for networks with application to neuroimaging. *arXiv preprint arXiv:2002.01645*

[9] Arroyo Relión JD, Kessler D, Levina E, Taylor SF, et al. 2019. Network classification with applications to brain connectomics. *The Annals of Applied Statistics* 13:1648–1677

[10] Athey TL, Vogelstein JT. 2019. Autogmm: Automatic gaussian mixture modeling in python. *arXiv preprint arXiv:1909.02688*

[11] Athreya A, Fishkind DE, Tang M, Priebe CE, Park Y, et al. 2017. Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research* 18:8393–8484

[12] Athreya A, Lyzinski V, Marchette DJ, Priebe CE, Sussman DL, Tang M. 2016. A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A* 78:1–18

[13] Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B Methodological* 57:289–300

[14] Biswal BB, Mennes M, Zuo XNN, Gohel S, Kelly AMC, et al. 2010. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences* 107:4734–4739

[15] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008:P10008

[16] Bonferroni CE. 1936. Teoria statistica delle classi e calcolo delle probabilità. Libreria internazionale Seeber

[17] Bullmore ET, Bassett DS. 2010. Brain Graphs: Graphical Models of the Human Brain Connectome. *Annu. Rev. Clin. Psychol.* 7:113–140

[18] Bullmore ET, Bassett DS. 2011. Brain graphs: graphical models of the human brain connectome. *Annual review of clinical psychology* 7:113–140

[19] Bullmore ET, Sporns O. 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10:186–198

[20] Cape J, Tang M, Priebe CE. 2019. On spectral embedding performance and elucidating network structure in stochastic blockmodel graphs. *Network Science* 7:269–291

[21] Chatterjee S, et al. 2015. Matrix estimation by universal singular value thresholding. *The Annals of Statistics* 43:177–214

[22] Chen H, Soni U, Lu Y, Maciejewski R, Kobourov S. 2018. Same stats, different graphs, In *International Symposium on Graph Drawing and Network Visualization*, pp. 463–477, Springer

[23] Chung J, Pedigo BD, Bridgeford EW, Varjavand BK, Vogelstein JT. 2019. GraSPy: Graph Statistics in Python

[24] Chung K, Deisseroth K. 2013. CLARITY for mapping the nervous system. *Nat. Methods* 10:508–513

[25] Clauset A, Newman ME, Moore C. 2004. Finding community structure in very large networks.

Physical review E 70:066111
[26] Craddock RC, Jbabdi S, Yan CG, Vogelstein JT, Castellanos FX, et al. 2013. Imaging human connectomes at the macroscale. *Nat. Methods* 10:524–539
[27] Crainiceanu CM, Caffo BS, Luo S, Zipunnikov VM, Punjabi NM. 2011. Population value decomposition, a framework for the analysis of image populations. *Journal of the American Statistical Association* 106:775–790
[28] Cullina D, Kiyavash N. 2016. Improved achievability and converse bounds for erdos-rényi graph matching. *ACM SIGMETRICS Performance Evaluation Review* 44:63–72
[29] Draves B, Sussman DL. 2020. Bias-variance tradeoffs in joint spectral embeddings. *arXiv preprint arXiv:2005.02511*
[30] Durante D, Dunson DB, Vogelstein JT. 2017. Rejoinder: Nonparametric Bayes Modeling of Populations of Networks. *J. Am. Stat. Assoc.* 112:1547–1552
[31] Efron B. 2008. Simultaneous inference: When should hypothesis testing problems be combined? *Ann. Appl. Stat.* 2:197–223
[32] Eichler K, Li F, Litwin-Kumar A, Park Y, Andrade I, et al. 2017. The complete connectome of a learning and memory centre in an insect brain. *Nature* 548:175–182
[33] Erdős P, Rényi A. 1959. On random graphs, I. *Publ. Math. Debrecen* 6:290–297
[34] Faskowitz J, Yan X, Zuo XN, Sporns O. 2018. Weighted stochastic block models of the human connectome across the life span. *Scientific reports* 8:1–16
[35] Fisher R. 1925. Statistical methods for research workers. Edinburgh Oliver & Boyd
[36] Fisher RA. 1925. Theory of Statistical Estimation. *Math. Proc. Cambridge Philos. Soc.* 22:700–725
[37] Fortunato S, Hric D. 2016. Community detection in networks: A user guide. *Physics reports* 659:1–44
[38] Garyfallidis E, Brett M, Amirbekian B, Rokem A, Van Der Walt S, et al. 2014. Dipy, a library for the analysis of diffusion mri data. *Frontiers in neuroinformatics* 8:8
[39] Garyfallidis E, Brett M, Correia M, Williams G, Nimmo-Smith I. 2012. Quickbundles, a method for tractography simplification. *Frontiers in Neuroscience* 6:175
[40] Genovese CR, Lazar NA, Nichols T. 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15:870–878
[41] Ghoshdastidar D, Gutzeit M, Carpentier A, von Luxburg U. 2017. Two-sample tests for large random graphs using network statistics. *arXiv preprint arXiv:1705.06168*
[42] Ginestet CE, Li J, Balachandran P, Rosenberg S, Kolaczyk ED, et al. 2017. Hypothesis testing for network data in functional neuroimaging. *The Annals of Applied Statistics* 11:725–750
[43] Goldenberg A, Zheng AX, Fienberg SE, Airoldi EM, et al. 2010. A survey of statistical network models. *Foundations and Trends® in Machine Learning* 2:129–233
[44] Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13:723–773
[45] Grover A, Leskovec J. 2016. node2vec: Scalable feature learning for networks, In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864
[46] Guha S, Rodriguez A. 2020. Bayesian regression with undirected network predictors with an application to brain connectome data. *Journal of the American Statistical Association* :1–34
[47] Hagmann P. 2005. From diffusion mri to brain connectomics :141
[48] Hoff PD, Raftery AE, Handcock MS. 2002. Latent space approaches to social network analysis. *Journal of the american Statistical association* 97:1090–1098
[49] Holland PW, Laskey KB, Leinhardt S. 1983. Stochastic blockmodels: First steps. *Social networks* 5:109–137
[50] Jackson JE. 2005. A user's guide to principal components, vol. 587. John Wiley & Sons
[51] Jenkinson M, et al. 2012. FSL. *NeuroImage* 62:782–90

[52] Karrer B, Newman ME. 2011. Stochastic blockmodels and community structure in networks. *Physical review E* 83:016107

[53] Kiar G, Bridgeford EW, Roncal WRG, Chandrashekhar V, Mhembere D, et al. 2018. A high-throughput pipeline identifies robust connectomes but troublesome variability. *bioRxiv* :188706

[54] Kim Y, Levina E. 2019. Graph-aware modeling of brain connectivity networks. *arXiv preprint arXiv:1903.02129*

[55] Klein A, Tourville J. 2012. 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in Neuroscience* 6:171

[56] Kolaczyk ED, Csárdi G. 2014. Statistical analysis of network data with r, vol. 65. Springer

[57] Kruskal WH, Wallis WA. 1952. Use of ranks in One-Criterion variance analysis. *J. Am. Stat. Assoc.* 47:583–621

[58] Levin K, Athreya A, Tang M, Lyzinski V, Park Y, Priebe CE. 2017. A central limit theorem for an omnibus embedding of multiple random graphs and implications for multiscale network inference. *arXiv preprint arXiv:1705.09355*

[59] Lock EF, Hoadley KA, Marron JS, Nobel AB. 2013. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The Annals of Applied Statistics* 7:523

[60] Lyzinski V, Fishkind DE, Fiori M, Vogelstein JT, Priebe CE, Sapiro G. 2015. Graph matching: Relax at your own risk. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38:60–73

[61] Lyzinski V, Sussman DL. 2017. Matchability of heterogeneous networks pairs. *Information and Inference: A Journal of the IMA*

[62] Lyzinski V, Sussman DL, Fishkind DE, Pao H, Chen L, et al. 2013. Spectral clustering for divide-and-conquer graph matching. *Parallel Comput.* :1310.1297

[63] Lyzinski V, Sussman DL, Tang M, Athreya A, Priebe CE. 2014. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electron. J. Stat.* 8:2905–2922

[64] Lyzinski V, Tang M, Athreya A, Park Y, Priebe CE. 2016. Community detection and classification in hierarchical stochastic blockmodels. *IEEE Transactions on Network Science and Engineering* 4:13–26

[65] Lyzinski V, Tang M, Athreya A, Park Y, Priebe CE. 2017. Community Detection and Classification in Hierarchical Stochastic Blockmodels. *IEEE Transactions on Network Science and Engineering* 4:13–26

[66] Marchette D, Priebe C, Coppersmith G. 2011. Vertex nomination via attributed random dot product graphs, In *Proceedings of the 57th ISI World Statistics Congress*, vol. 6, p. 16

[67] Matejka J, Fitzmaurice G. 2017. Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing, In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1290–1294

[68] Mazziotta J, et al. 2001. A four-dimensional probabilistic atlas of the human brain. *Journal of the American Medical Informatics Association* 8:401–430

[69] Mhembere D, Roncal WG, Sussman D, Priebe CE, Jung R, et al. 2013. Computing scalable multivariate glocal invariants of large (brain-) graphs, In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 297–300, IEEE

[70] Newman ME. 2013. Spectral methods for community detection and graph partitioning. *Physical Review E* 88:042822

[71] Newman ME, et al. 2003. Random graphs as models of networks. *Handbook of graphs and networks* 1:35–68

[72] Nielsen AM, Witten D. 2018. The multiple random dot product graph model. *arXiv preprint arXiv:1811.12172*

[73] Panda S, Palaniappan S, Xiong J, Swaminathan A, Ramachandran S, et al. 2019. mgcpy: A Comprehensive High Dimensional Independence Testing Python Package

[74] Priebe C, Coppersmith G, Rukhin A. 2010. You say graph invariant, i say test statistic. *ASA Sec-*

tions on Statistical Computing Statistical Graphics SCGN Newsletter 21:11–14
[75] Priebe CE, Coppersmith G, Rukhin A. 2010. You say “graph invariant,” i say “test statistic”. *Statistical Computing and Statistical Graphics*
[76] Priebe CE, Park Y, Tang M, Athreya A, Lyzinski V, et al. 2017. Semiparametric spectral modeling of the drosophila connectome. *arXiv preprint arXiv:1705.03297*
[77] Priebe CE, Park Y, Vogelstein JT, Conroy JM, Lyzinski V, et al. 2019. On a two-truths phenomenon in spectral graph clustering. *Proceedings of the National Academy of Sciences* 116:5995–6000
[78] Richiardi J, Eryilmaz H, Schwartz S, Vuilleumier P, Van De Ville D. 2011. Decoding brain states from fmri connectivity graphs. *Neuroimage* 56:616–626
[79] Rieke F. 1997. Spikes: Exploring the Neural Code, vol. 20. Cambridge: MIT Press
[80] Rissanen J. 1978. Modeling by shortest data description. *Automatica* 14:465–471
[81] Rohe K, Chatterjee S, Yu B. 2011. Spectral Clustering and the High-Dimensional Stochastic Blockmodel. *Ann. Stat.* 39:1878–1915
[82] Rohe K, Qin T, Yu B. 2016. Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences* 113:12679–12684
[83] Rubin-Delanchy P, Priebe CE, Tang M, Cape J. 2017. A statistical interpretation of spectral embedding: the generalised random dot product graph. *arXiv preprint arXiv:1709.05506*
[84] Rukhin A, Priebe CE. 2010. A comparative power analysis of the maximum degree and size invariants for random graph inference. *Journal of Statistical Planning and Inference* 141:1041–1046
[85] Russell SJ, Norvig P. 2016. Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited,
[86] Scheffé H. 1999. The analysis of variance. John Wiley & Sons
[87] Scheinerman ER, Tucker K. 2010. Modeling graphs using dot product representations. *Computational statistics* 25:1–16
[88] Schwarz G, et al. 1978. Estimating the dimension of a model. *The annals of statistics* 6:461–464
[89] Scrucca L, Fop M, Murphy TB, Raftery AE. 2016. mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal* 8:289
[90] Shepherd GM. 1991. Foundations of the Neuron Doctrine. Oxford University Press, 1st ed.
[91] Simes RJ. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73:751–754
[92] Smith SM, et al. 2004. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* 23 Suppl 1:S208–19
[93] Sporns O, Tononi G, Kötter R. 2005. The human connectome: a structural description of the human brain. *PLoS computational biology* 1:e42
[94] Sussman DL, Tang M, Fishkind DE, Priebe CE. 2012. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association* 107:1119–1128
[95] Sussman DL, Tang M, Priebe CE. 2014. Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* 36:48–57
[96] Székely GJ, Rizzo ML, Bakirov NK. 2007. Measuring and testing dependence by correlation of distances. *Ann. Stat.* 35:2769–2794
[97] Tang M, Athreya A, Sussman DL, Lyzinski V, Park Y, Priebe CE. 2017. A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics* 26:344–354
[98] Tang M, Athreya A, Sussman DL, Lyzinski V, Priebe CE, et al. 2017. A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli* 23:1599–1630
[99] Tang M, Priebe CE, et al. 2018. Limit theorems for eigenvectors of the normalized laplacian for random graphs. *The Annals of Statistics* 46:2360–2415
[100] Tang M, Sussman DL, Priebe CE. 2013. Universally consistent vertex classification for latent positions graphs. *Ann. Statist.* 41:1406–1430

[101] Tang R, Ketcha M, Badea A, Calabrese ED, Margulies DS, et al. 2018. Connectome smoothing via low-rank approximations. *IEEE transactions on medical imaging* 38:1446–1456
[102] Thirion B, Varoquaux G, Dohmatob E, Poline JB. 2014. Which fmri clustering gives good brain parcellations? *Frontiers in neuroscience* 8:167
[103] Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, et al. 2013. The wu-minn human connectome project: an overview. *Neuroimage* 80:62–79
[104] Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens T, et al. 2012. The human connectome project: a data acquisition perspective. *Neuroimage* 62:2222–2231
[105] Varoquaux G, Craddock RC. 2013. Learning and comparing functional connectomes across subjects. *Neuroimage* 80:405–415
[106] Varoquaux G, Gramfort A, Poline JB. 2010. Brain covariance selection: better individual functional connectivity models using population prior. *arXiv preprint* :1–9
[107] Vogelstein JT, Bridgeford EW, Pedigo BD, Chung J, Levin K, et al. 2019. Connectal coding: discovering the structures linking cognitive phenotypes to individual histories. *Current opinion in neurobiology* 55:199–212
[108] Vogelstein JT, Conroy JM, Lyzinski V, Podrazik LJ, Kratzer SG, et al. 2015. Fast approximate quadratic programming for graph matching. *PLOS one* 10:e0121002
[109] Vogelstein JT, Roncal WG, Vogelstein RJ, Priebe CE. 2012. Graph classification using signal-subgraphs: Applications in statistical connectomics. *IEEE transactions on pattern analysis and machine intelligence* 35:1539–1551
[110] Wang L, Zhang Z, Dunson D. 2019. Symmetric bilinear regression for signal subgraph estimation. *IEEE Transactions on Signal Processing* 67:1929–1940
[111] Wang L, Zhang Z, Dunson D, et al. 2019. Common and individual structure of brain networks. *The Annals of Applied Statistics* 13:85–112
[112] Wang S, Arroyo J, Vogelstein JT, Priebe CE. 2019. Joint embedding of graphs. *IEEE transactions on pattern analysis and machine intelligence*
[113] Wang S, Shen C, Badea A, Priebe CE, Vogelstein JT. 2018. Signal subgraph estimation via vertex screening. *arXiv preprint arXiv:1801.07683*
[114] Wasserman S, Anderson C. 1987. Stochastic a posteriori blockmodels: Construction and assessment. *Soc. Networks* 9:1–36
[115] Woolrich MW, et al. 2009. Bayesian analysis of neuroimaging data in FSL. *NeuroImage* 45:S173–86
[116] Xia Y, Li L. 2019. Matrix graph hypothesis testing and application in brain connectivity alternation detection. *Statistica Sinica* 29:303–328
[117] Young SJ, Scheinerman ER. 2007. Random Dot Product Graph Models for Social Networks, In *Algorithms and Models for the Web-Graph*, pp. 138–149, Springer Berlin Heidelberg
[118] Zalesky A, Fornito A, Harding IH, Cocchi L, Yücel M, et al. 2010. Whole-brain anatomical networks: does the choice of nodes matter? *Neuroimage* 50:970–983
[119] Zhang J, Sun WW, Li L. 2018. Network response regression for modeling population of networks with covariates. *arXiv preprint arXiv:1810.03192*
[120] Zhang Z, Descoteaux M, Zhang J, Girard G, Chamberland M, et al. 2018. Mapping population-based structural connectomes. *NeuroImage* 172:130–145
[121] Zheng AX, Fienberg SE, Airoldi EM, Goldenberg A. 2009. A Survey of Statistical Network Models. *Foundations and Trends in Machine Learning* 2:129–233
[122] Zhu M, Ghodsi A. 2006. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis* 51:918–930
[123] Zuo XN, Anderson JS, Bellec P, Birn RM, Biswal BB, et al. 2014. An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific data* 1:140049