ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Statistical Connectomics

Jaewon Chung,[1] Eric Bridgeford,[2] Jesús Arroyo,[3] Benjamin D. Pedigo,[1] Ali Saad-Eldin,[1] Vivek Gopalakrishnan,[1] Liang Xiang,[1] Carey E. Priebe,[4] and Joshua T. Vogelstein[5]

[1]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA; email: j1c@jhu.edu

[2]Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21218, USA

[3]Center for Imaging Science, Johns Hopkins University, Baltimore, Maryland 21218, USA

[4]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, Maryland 21218, USA

[5]Department of Biomedical Engineering, Institute for Computational Medicine, Kavli Neuroscience Discovery Institute, Johns Hopkins University, Baltimore, Maryland 21218, USA; email: jovo@jhu.edu

Annu. Rev. Stat. Appl. 2021. 8:463–92

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-042720-023234

Keywords

connectomics, networks, graphs, statistical models

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.

1. INTRODUCTION

The idea of the brain as a network of interconnected neuronal elements has existed since the late nineteenth 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 (Shepherd 1991, Rieke 1997, Russell & Norvig 2016). Recent advances in neuroimaging (Hagmann 2005, Biswal et al. 2010, Chung & Deisseroth 2013), along with large-scale projects, opened new frameworks for studying the brain by modeling brain connectivity as networks, or connectomes (Van Essen et al. 2013, Zuo et al. 2014, Alexander et al. 2017). 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" (Vogelstein et al. 2019).

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 (Hagmann 2005, Sporns et al. 2005, Vogelstein et al. 2019). Connectomes can have additional structures. For example, edges can have weights that describe the strength of connection, and they can 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*,*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 nonzero 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, 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 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 (Bullmore & Bassett 2011, Mhembere et al. 2013). One major drawback to this method is that features are not independent of one another, making results from subsequent inference using these features difficult

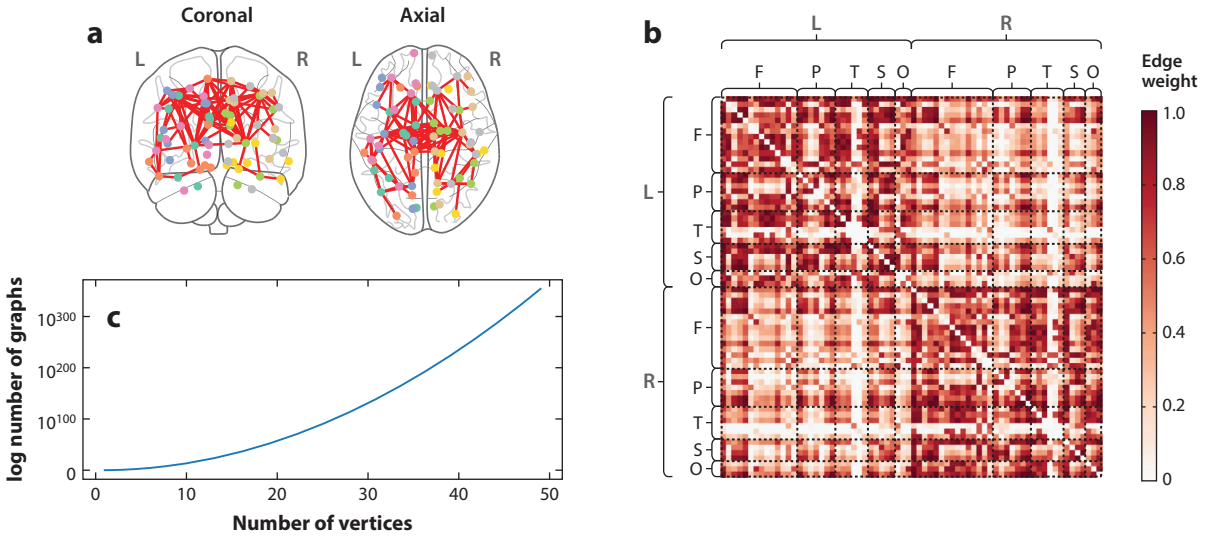


Figure 1

Different representations of a connectome: human structural connectome estimated from averaging 1,059 human connectomes from the Human Connectome Project (HCP) (Van Essen et al. 2013). 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), temporal (T), and subcortical (S) structures. (*a*) Connectivity shown in the coronal and axial views. Dots correspond to the center of mass of a region, lines correspond to connections, and line thickness corresponds to the magnitude of the connection. Only the largest 5% of edges are shown for visualization purposes. Note that infinitely many spatial arrangements 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 substructures. However, given any permutation matrix $\mathbf{P}$, the permuted adjacency matrix $\mathbf{PAP}^T$ is still a valid matrix of the 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.

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 (Varoquaux et al. 2010, Craddock et al. 2013). 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 provides 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 (Zheng et al. 2009, Athreya et al. 2017, Zhang et al. 2018a, Arroyo et al. 2019). 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 under consideration. We vary parameters for the data,

Table 1 Notations and symbols used in this article

Symbol	Description	Symbol	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
$\mathbf{A}_i$	ith 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
$\mathbf{A}^{(l)}$	lth element in sequence of $\mathbf{A}$		

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. **Table 1** provides an overview of the notation used throughout the article.

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 one-to-one 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 weights in human structural connectomes are nonnegative 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, 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 article, graphs are considered undirected and unweighted and with no self-loops, that is, $\text{diag}(\mathbf{A}) = \vec{0}$, unless specified otherwise.

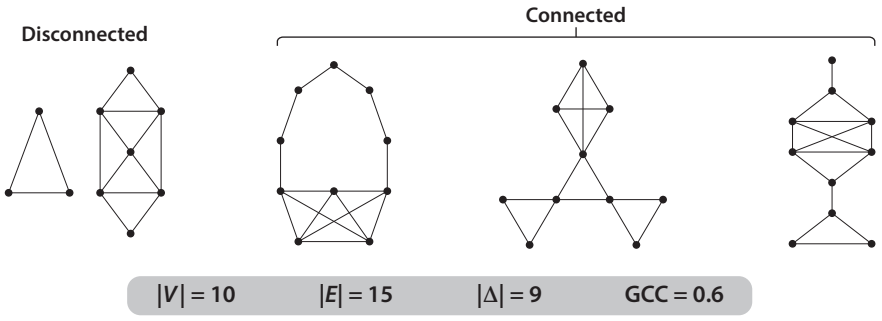


Figure 2

Four networks with the same network statistics. Each network 10 vertices ($|V|$), 15 edges ($|E|$), and 9 triangles ($|\Delta|$), and the global clustering coefficient (GCC) is 0.6. However, these graphs have distinctive topologies. For example, the left-most network is disconnected, while the others are connected. This suggests that, given a small set of network statistics, one cannot identify from which network the features are computed.

2.2. Bag of Features

Network statistics, or features, are abstract representations that capture either global or local structures of a network (Priebe et al. 2010, Mhembere et al. 2013). 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 individuals and individuals with depression, the difference in global clustering coefficient, which measures how likely vertices tend to cluster together, can be computed (Bullmore & Sporns 2009). These network statistics have enjoyed applications in many connectomics studies that compare different populations of networks (Bullmore & Bassett 2011, Ghoshdastidar et al. 2017). However, there are an infinite number of such statistics, and we lack general guidance on which statistics to compute. Furthermore, no set of network statistics can adequately characterize a network (Matejka & Fitzmaurice 2017, Chen et al. 2018). These considerable shortcomings further motivate the use of other representations of networks, and the examples below demonstrate the shortcomings of studying bags of features.

2.2.1. Nonidentifiability of graph features. Summary statistics, such as the mean, variance, and correlation, are often used to describe real valued data sets, which can provide insight in understanding the data. However, Anscombe's quartet illustrates four drastically different distributions of eleven points that have the same summary statistics (Anscombe 1973). This suggests that any small number of summary statistics can fail to meaningfully characterize the data.

In network analysis, a variety of network level statistics can be computed to summarize networks. Similar to 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, and triangles and the same global clustering coefficient but have different properties, such as connectedness and symmetry. Other works have also explored the distributions of network statistics (Matejka & Fitzmaurice 2017, Chen et al. 2018).

2.2.2. Network features are correlated and relatively uninformative. We consider all nonisomorphic, 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 nonisomorphic networks are considered since isomorphic networks have identical network features.

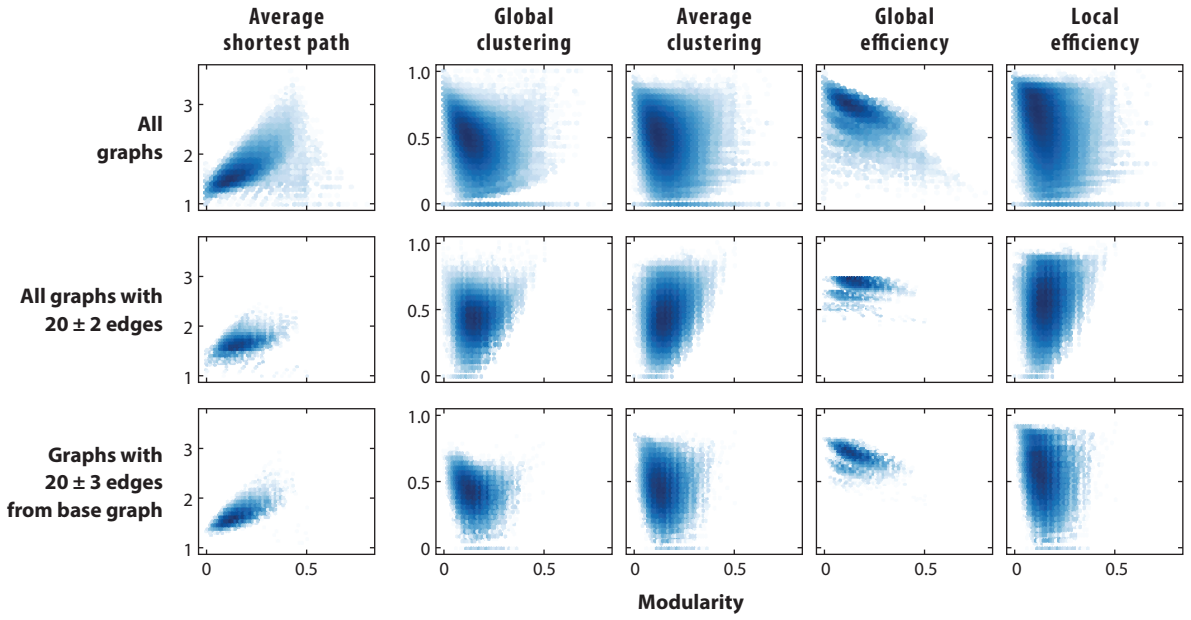


Figure 3

Density plots of network statistics. (*Top row*) The distributions of network 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 sets 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.

For each network, the following six graph network statistics are computed: (*a*) average path length, (*b*) global clustering coefficient, (*c*) average clustering coefficient, (*d*) global efficiency, (*e*) local efficiency, and (*f*) modularity. These are some of the most commonly computed statistics (Sporns et al. 2005, Bullmore & Sporns 2009). The distributions 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 is 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. A similar pattern is shown in the analysis of HCP data, as shown in the **Supplemental Appendix (Section A.1)**. Changing only a few edges on a network can yield a network with almost any possible configuration according to these statistics, and therefore they 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.

Supplemental Material >

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 (Craddock et al. 2013, Varoquaux &

Craddock 2013, Zhang et al. 2018b). 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 (Genovese et al. 2002, Efron 2008). 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 (Simes 1986, Benjamini & Hochberg 1995, Zalesky et al. 2010). On the other hand, Bonferroni corrections are considered too conservative and, therefore, lack the sensitivity for connectomics (Simes 1986).

More intricate methods represent each connectome as a long vector containing all of its edges (Richiardi et al. 2011, Amico et al. 2017). Vector representations can allow for correlation of edges and direct application of common machine learning algorithms, but still discard 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 (Grover & Leskovec 2016, Athreya et al. 2017, Arroyo et al. 2019). Algorithms that operate on Euclidean data [e.g., Gaussian mixture modeling (GMM) for clustering vertices, random forests for classifying vertices, multivariate hypothesis tests for testing for differences between vertices] can be employed for subsequent analysis (Priebe et al. 2017, Tang et al. 2018).

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 that comprise a network, called community structure, can describe both the local and global patterns of the network. At a local scale, we can examine the properties of vertices that are within the same community. At a global scale, we can measure associations between connectivity patterns of communities across groups or other covariates (Faskowitz et al. 2018, Kim & Levina 2019, Arroyo & Levina 2020). Furthermore, the community structure in spatial resolution connectomes from human magnetic resonance imaging can be used to delineate regions of the brain called parcellations (Thirion et al. 2014).

Community detection in networks has been studied extensively (Newman 2013, Fortunato & Hric 2016). Typically, the community structure is identified by modularity optimization methods (Clauset et al. 2004, Blondel et al. 2008). In this article, we present spectral methods that rely on statistical models for community detection, which have strong statistical guarantees for recovering true communities (Sussman et al. 2012, Lyzinski et al. 2017, Athreya et al. 2017, Arroyo et al. 2019). It is important to note that analysis of communities depends on the performance of the community detection algorithms.

2.6. Bag of Networks

In the bag of networks 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

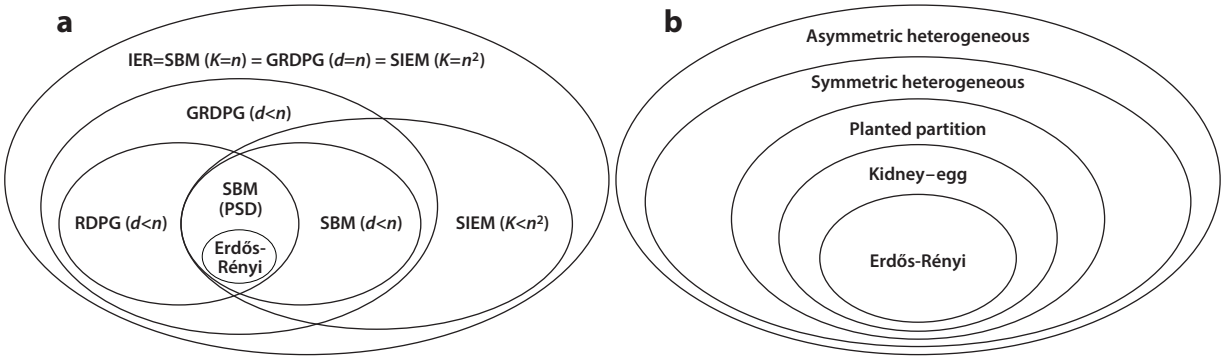


Figure 4

Hierarchical relationships of statistical models. (*a*) Relationships among all the single graph statistical models. The Erdős-Rényi (ER) model is a stochastic block model (SBM) with one community. An SBM with a positive semidefinite (PSD) block probability matrix **B** is also a random dot product graph (RDPG). Any SBM with K blocks fewer than the number of vertices n, or RDPG, and some structured independent edge models (SIEMs) with K groups fewer than n^2 can be represented as a d-dimensional generalized random dot product graph (GRDPG) with d fewer n. The inhomogeneous Erdős-Rényi (IER) model is equivalent to an n-block SBM, n-dimensional GRDPG, and n^2-group SIEM. (*b*) Relationships among the two-block SBMs. The most complex model is the asymmetric heterogeneous SBM, and the simplest model is the ER, which is a degenerate case of two-block SBM.

networks. For example, bag of vertices representation can be used to test whether two networks are different (Tang et al. 2017a,b). For studying more than two networks, geometry in the space of the networks is defined and the networks are represented in that geometry, which is then used for finding differences across groups (Ginestet et al. 2017, Arroyo et al. 2019, Xia & Li 2019).

Another group of methods finds subsets of vertices, or subgraphs, that contain the most information about certain covariates (Vogelstein et al. 2012, Wang et al. 2018, Arroyo Relión et al. 2019, L. Wang et al. 2019b, Guha & Rodriguez 2020). Estimating signal subgraphs is useful since networks can be extremely large (i.e., millions of vertices), which presents computational challenges and can potentially improve the performance of subsequent inference tasks, such as classification. Different approaches for finding subgraphs have been proposed, but all approaches leverage the network topologies inherent in connectomics data.

3. STATISTICAL MODELS

Connectomes can be modeled using statistical models designed for network data (Goldenberg et al. 2010, Kolaczyk & Csárdi 2014). 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 4** shows the relationship between all the single graph models presented in this article. Section 3.2 provides an overview of some models for multiple networks. While other statistical models for multiple network data exist (Durante et al. 2017, Nielsen & Witten 2018, Zhang et al. 2018a, S. Wang et al. 2019c), we focus on some recent models that are used in spectral inference for connectomics data. In **Supplemental Appendix Section B**, we describe some extensions to these models.

Supplemental Material >

3.1. Single Graph Models

3.1.1. Erdős-Rényi random graphs. The simplest random graph model is the Erdős-Rényi (ER) model (Erdős & Rényi 1959). For a given set of n vertices, each distinct pair of vertices is connected independently with probability $p \in [0, 1]$. Specifically, $\mathbf{A} \sim \mathrm{ER}_n(p)$ if $\mathbf{A}$ has entries $\mathbf{A}_{ij} \sim \mathrm{Bernoulli}(p)$ 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 (Newman et al. 2002, Rukhin & Priebe 2010).

3.1.2. Stochastic block model. First introduced by Holland et al. (1983), the SBM is a model that can produce graphs with vertices grouped into K communities (Rohe et al. 2011, Sussman et al. 2012, Wasserman & Anderson 1987). There are two simple variations of the SBM in which the vertex assignment vector $\vec{\tau} \in \{1, \ldots, K\}^n$ is known a priori, and in which $\vec{\tau}$ 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 $\vec{\tau} \in \{1, \ldots, K\}^n$ is known a priori, the a priori SBM is parameterized only by the block connectivity matrix $\mathbf{B}$, and the model is $\mathbf{A} \sim \mathrm{SBM}_n(\vec{\tau}, \mathbf{B})$ if $\mathbf{A}$ has entries $\mathbf{A}_{ij} \sim \mathrm{Bernoulli}(\mathbf{B}_{kl})$ 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 $\mathbf{A} \sim \mathrm{SBM}_n(\vec{\pi}, \mathbf{B})$ if $\mathbf{A}$ has entries $\mathbf{A}_{ij} \,|k = \tau_i, l = \tau_j \sim \mathrm{Bernoulli}(\mathbf{B}_{kl})$, where $\tau_i \sim \mathrm{Multinomial}(\vec{\pi})$ for $i = 1, \ldots, n$.

Throughout this article, we focus particularly on a few variations of the two-block SBM ($K = 2$) with block connectivity matrix $\mathbf{B} = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$, 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 a 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 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. Erdős-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 4*b* summarizes the relationships of SBMs.

3.1.3. Structured independent edge model. The structured independent edge model (SIEM) is a generalization of the 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 $\mathbf{M} \in \{1, \ldots, K\}^{n \times n}$, which is known a priori. Given the community assignment matrix $\mathbf{M}$, the SIEM is $\mathbf{A} \sim \mathrm{SIEM}_n(\mathbf{M}, \vec{p})$ if $\mathbf{A}_{ij} \sim \mathrm{Bernoulli}(p_k)$ where $\mathbf{M}_{ij} = k$, for

$i, j \in [n]$ and $k \in [K]$. $\vec{p} = [p_1, \ldots, p_K]^\top \in [0,1]^K$ is the edge probability vector that governs the probability of an edge between vertices.

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

3.1.4. Random dot product graphs. Random dot product graphs (RDPGs) belong to the class of latent position random graphs (Hoff et al. 2002). 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 is given by a link function. In RDPGs, the space $\mathcal{X}$ is a constrained subspace of Euclidean space $\mathbb{R}^d$ and the link function is the dot product (Young & Scheinerman 2007, Scheinerman & Tucker 2010, Sussman et al. 2014). 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 (Sussman et al. 2012), vertex classification (Tang et al. 2013), or two-sample hypothesis testing for graphs with matched and nonmatched vertices for a pair of graphs (Tang et al. 2017a,b; Priebe et al. 2019).

The RDPG is a flexible model, and other models of interest can be seen as special cases of the RDPG. An SBM whose block connectivity matrix $\mathbf{B}$ is positive semidefinite is an RDPG with K distinct latent positions. Thus, an 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 (Airoldi et al. 2008) or hierarchical communities (Lyzinski et al. 2017).

3.1.5. Generalized random dot product graphs. Unlike the RDPG model, the generalized random dot product graph (GRDPG) does not assume that $\mathbf{P}$ is a positive semidefinite probability matrix (Rubin-Delanchy et al. 2017). 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 an RDPG model. To represent any SBM as a 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. Inhomogeneous Erdős-Rényi random graphs. The inhomogeneous Erdős-Rényi (IER) is a model where each pair of nodes has a unique probability of an edge existing between the two, and it is therefore the most general independent edge model. For a given set of n vertices, the IER model is parameterized 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 $i, j \in [n]$. An IER model 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 the IER. Additionally, an SBM with $K = n$, a SIEM with $K = n^2$, and a 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 contains 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 6.

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. In the joint random dot product graph (JRDPG) model, we consider a collection of m RDPGs, all with the same generating latent positions. Similar to an RDPG, given an appropriately constrained Euclidean subspace $\mathbb{R}^d$, this model is parameterized by a latent positions matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ where $d \ll n$. The model is $(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(m)}) \sim \text{JRDPG}(\mathbf{X})$ where $\mathbf{A}^{(l)}_{ij} \sim \text{Bernoulli}(\mathbf{X}_i \mathbf{X}_j^\top)$ 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 $\mathbf{X}$ (Athreya et al. 2017, Levin et al. 2017). 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 the model still remains very useful, as shown in Section 6.

3.2.2. Common subspace independent-edge model. In the common subspace independent-edge (COSIE) model, the heterogeneous networks are described via a shared latent structure on the vertices, but it also permits sufficient heterogeneity via individual matrices for each graph (Arroyo et al. 2019). 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 $(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}) \sim \text{COSIE}(\mathbf{V}; \mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)})$, 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 (Crainiceanu et al. 2011, Afshin-Pour et al. 2012, Lock et al. 2013, L. Wang et al. 2019a).

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 nonzero correlation. Correlated graph models have numerous applications, such as when a graph is estimated repeatedly for the same source at different points in time.

The $\mathbf{R}$-correlated $(\mathbf{P}, \mathbf{Q})$ model (Lyzinski & Sussman 2017) 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 IER 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 $\mathbf{R}_{ij} \leq \sqrt{\frac{\mathbf{Q}_{ij}(1-\mathbf{Q}_{ij})}{\mathbf{P}_{ij}(1-\mathbf{P}_{ij})}}$ (Lyzinski & Sussman 2017).

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

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 $\mathbf{R} = \rho\mathbf{1}_{n\times n}$ (that is, the matrix of edge correlations $\mathbf{R}$ 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 $\mathbf{R} = \rho\mathbf{1}_{n\times n}$ (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. ALGORITHMS

In this section, we introduce algorithms used for statistical analysis of networks.

4.1. Single Graph Algorithms

In this section, we introduce methods for embedding a network as a way of learning representations that can be utilized for subsequent inference tasks.

4.1.1. Adjacency and Laplacian spectral embedding. 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 $|\mathbf{S}_{11}| \geq |\mathbf{S}_{22}| \geq \cdots \geq |\mathbf{S}_{nn}|$. 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 the 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 nonnegative 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 as the 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 many definitions exist for the 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 (Rohe et al. 2016). The LSE of a directed graph is processed similarly to that of a directed ASE.

Spectral embedding is the first step in many subsequent inference tasks. For example, spectral clustering for community detection (Section 4.1.4) can be achieved via GMM on $\hat{\mathbf{X}}$ from either ASE or LSE. The resulting cluster assignments can further be used to estimate the parameters for a posteriori SBMs.

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 (Jackson 2005, Chatterjee et al. 2015), we consider the method of Zhu & Ghodsi (2006) when applying any spectral embeddings in real data. Given $\mathbf{A} = \mathbf{USU}^\top$ 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 the 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 the largest eigenvalues as input to the profile likelihood function and take the second elbow as the estimate of $\hat{d}$.

4.1.2. Diagonal augmentation. Many connectomes have no self-loops, resulting in all zeros in the diagonal entries of the adjacency matrices. When computing spectral embeddings of graphs, the zero diagonal results in increased errors in estimation (Tang et al. 2018). Furthermore, the sum of eigenvalues of the adjacency matrices is zero, leading to an indefinite matrix, which violates assumptions of the statistical models such as the RDPG.

Diagonal augmentation (diag-aug) is a method for imputing the diagonals of adjacency matrices from graphs with no self-loops (Scheinerman & Tucker 2010, Marchette et al. 2011, Tang et al. 2018). The diagonals are imputed with the average of the nondiagonal 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 $\mathbf{A} \in \mathbb{R}^{n\times n}$, and $\tilde{\mathbf{D}} \in \mathbb{R}^{n\times n}$ 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 the 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).

4.1.3. Pass-to-ranks. Connectomes often have 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 (Kiar et al. 2018). 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 kth 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 nonzero 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).

4.1.4. 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

Supplemental Material >

spectral embeddings of many statistical models (e.g., SBM, RDPG) suggest GMM for clustering (see the **Supplemental Appendix, Section C.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, and minimum description length (Akaike 1974, Rissanen 1978, Schwarz et al. 1978). By default, we use penalized likelihood via BIC to estimate K (Priebe et al. 2019). In practice, various covariance types and initialization methods for GMM and number of clusters are swept over to compute the best estimated number of clusters, $\hat{K}$ (Scrucca et al. 2016, Athey & Vogelstein 2019).

4.2. Multiple Graph Algorithms

In this section, we introduce embedding methods for populations of networks and various algorithms applicable for multiple networks, such as seeded graph matching and community detection.

4.2.1. 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, omnibus embedding (OMNI) is a consistent method (see the **Supplemental Appendix, Section C.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:

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

The embeddings gives the matrix

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

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

4.2.2. Multiple adjacency spectral embedding. Multiple adjacency spectral embedding (MASE) is a consistent method for estimation (see the **Supplemental Appendix, Section C.2.1**) of underlying parameters for each graph under the COSIE model (Arroyo et al. 2019). MASE is a three-step process:

1. Each adjacency matrix, $\mathbf{A}^{(i)}$, is embedded into d dimensions via ASE, and the matrix $\hat{\mathbf{U}} = [\mathrm{ASE}(\mathbf{A}^{(1)}), \mathrm{ASE}(\mathbf{A}^{(2)}), \ldots, \mathrm{ASE}(\mathbf{A}^{(m)})] \in \mathbb{R}^{n \times dm}$ 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 $\hat{\mathbf{R}}^{(i)} \in \mathbb{R}^{d \times d}$.

4.2.3. Spectral clustering for community detection. Similar to the procedure described in Section 4.1.4, spectral clustering in the multi-graph setting can also be performed. Clustering is performed on the average latent position matrix, $\bar{\mathbf{X}} := \frac{1}{m}\sum_{i=1}^{m}\hat{\mathbf{X}}^{(i)}$, in the JRDPG model and the vertex subspace matrix, $\hat{\mathbf{V}}$, in the COSIE model. The clustering procedure proceeds identically to the one described in Section 4.1.4.

4.2.4. Seeded graph matching. 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:

$$\min \|\mathbf{A}\,\mathbf{P} - \mathbf{P}\,\mathbf{B}\|_F^2 \qquad 1.$$

$$\text{s.t. } \mathbf{P} \in \mathcal{P},$$

where $\mathcal{P}$ is the set of permutation matrices in $\mathbb{R}^{n\times n}$. Seeded graph matching is a modification of the graph matching algorithm that allows for the specification of seed sets W_1, W_2 with seeding ψ: $W_1 \to W_2$ and is solved via fast approximate quadratic assignment (Vogelstein et al. 2015). As the seeded graph matching problem is computationally intractable, seeded graph matching 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:

Algorithm 1.

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 $\Delta f(\mathbf{P}^{(i)})$.
b. Compute the search direction $\mathbf{Q}^{(i)} \in \text{argmax}\,(\text{tr}(\mathbf{Q}^T\,\Delta f(\mathbf{P}^{(i)})))$ via the Hungarian algorithm.
c. Compute step size $\alpha^{(i)} \in \text{argmax}(f(\alpha^{(i)}\,\mathbf{P}^{(i)} + (1-\alpha^{(i)})\,\mathbf{Q}^{(i)}))$.
d. Update $\mathbf{P}^{(i+1)} := \alpha^{(i)}\,\mathbf{P}^{(i)} + (1-\alpha^{(i)})\,\mathbf{Q}^{(i)}$.

3. Compute $\hat{\mathbf{P}} \in \text{argmax}(\text{tr}(\mathbf{P}^{\top}\,\mathbf{P}^{(final)}))$ via the Hungarian algorithm.

5. APPLICATIONS FOR SINGLE GRAPH DATA

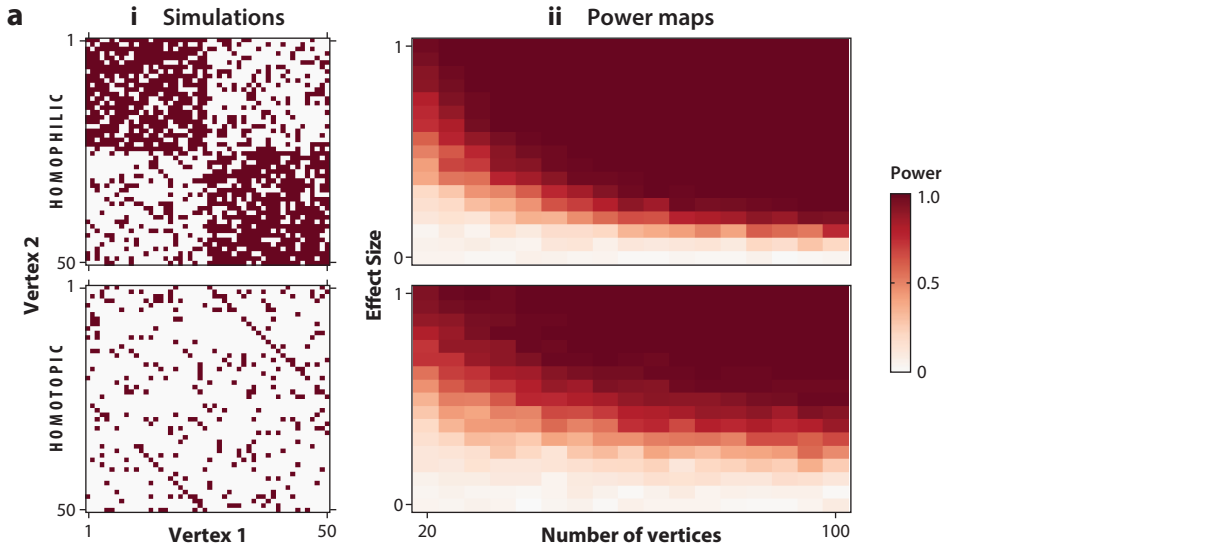
In this section, we explore the applications of the single graph models in Section 3.1 and the algorithms in Section 4.1. The *Drosophila* mushroom body connectome and HCP data are analyzed (see the **Supplemental Appendix, Sections D.1 and D.2**, for a description) along with simulated examples, and the **Supplemental Appendix (Section E)** contains additional analysis in weighted connectomes.

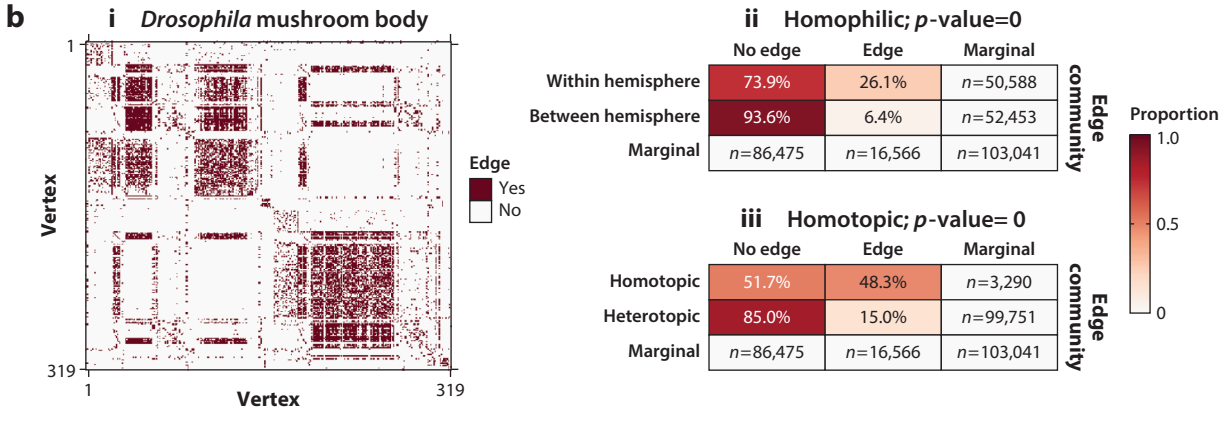
Supplemental Material >

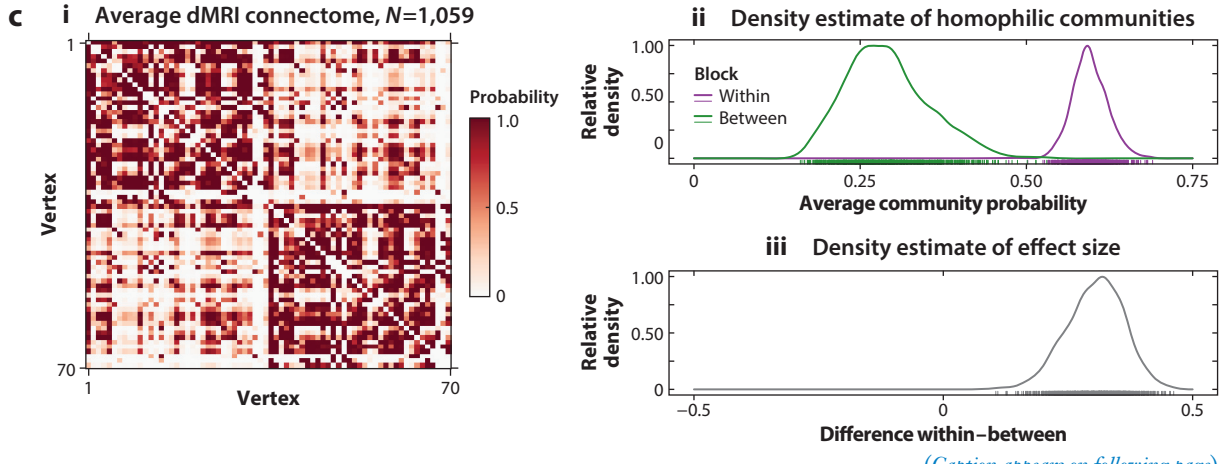
5.1. Testing for Differences Between Communities of Edges

In **Figure 5**, we compare a number of different strategies using Fisher's exact test (Fisher 1925) 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:

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







(*Caption appears on following page*)

Figure 5 (*Figure appears on preceding page*)

Comparing communities of edges in graphs. (*a*, *i–ii*) Fisher's exact test shows reasonable statistical power across both homophilic and homotopic block structures (*a*, *i*) with power converging to 1 as effect size and number of vertices grow (*a*, *ii*). (*b*) The *Drosophila* mushroom body in subpanel *b*, *i* shows both homophilic planted partition and homotopic structure (*b*, *ii–iii*) (Fisher's exact test, p-values = 0). (*c*) All $N = 1{,}059$ Human Connectome Project diffusion connectomes (*c*, *i*) show homophilic planted partition structure (*c*, *ii*), with within-hemisphere connectivity exceeding between-hemisphere connectivity (*c*, *iii*) (Fisher's exact test, $N = 1{,}059$, Bonferroni corrected p-values $<10^{-21}$).

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 5*a***, subpanel *i*. 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 5*a***, subpanel *ii*. 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 5*b***, 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 displays 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 nonbilateral vertices). **Figure 5*b***, subpanel *i* 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 5*b***, subpanel *ii*. 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 nonbilateral (heterotopic) connectivity, in **Figure 5*b***, subpanel *iii*. 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 data set. 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 5*c***, subpanel *i* shows the average unweighted diffusion connectome over all participants in the study. **Figure 5*c***, subpanel *ii* 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 $N = 1{,}059$ connectomes, 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 $N = 1{,}059$ connectomes, shown in **Figure 5*c***, subpanel *iii*. All 1,059 diffusion connectomes have significantly higher within-hemisphere connectivity than between-hemisphere connectivity at $\alpha = 0.05$ after Bonferroni correction (Fisher's exact test, $N = 1{,}059$, maximum p-value $<10^{-20}$).

Supplemental Material >

The **Supplemental Appendix (Section E.1)** investigates the fit of independent edge models for weighted graphs from the *Drosophila* mushroom body and the HCP diffusion connectomes. We leverage the Mann–Whitney–Wilcoxon test, a nonparametric test of whether there exists a difference in medians between the two edge clusters. We again find that the weighted *Drosophila* mushroom body shows both homotopic planted partition structure and homophily, and that the weighted HCP connectomes all show homotopic planted partition structure ($N = 1{,}059$), a conclusion consistent with our results on the unweighted graphs.

5.2. 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 Section 3.1.2, 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 6*a***, we perform simulations where the true graph is either planted partition, symmetric heterogeneous, or asymmetric heterogeneous, as shown in **Figure 6*a***, subpanel *i*. 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 increase. **Figure 6*a***, subpanel *ii* shows that the test features both empirical validity and empirical consistency, as in **Figure 5**.

In **Figure 6*b***, we investigate the appropriate block structure for the unweighted *Drosophila* mushroom body, which 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 distinct 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 consideration 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 6*c* investigates the optimal block structure for the $N = 1{,}059$ diffusion connectomes from the HCP data set. The figure shows the average connectivity for the three 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 three possible block structures for

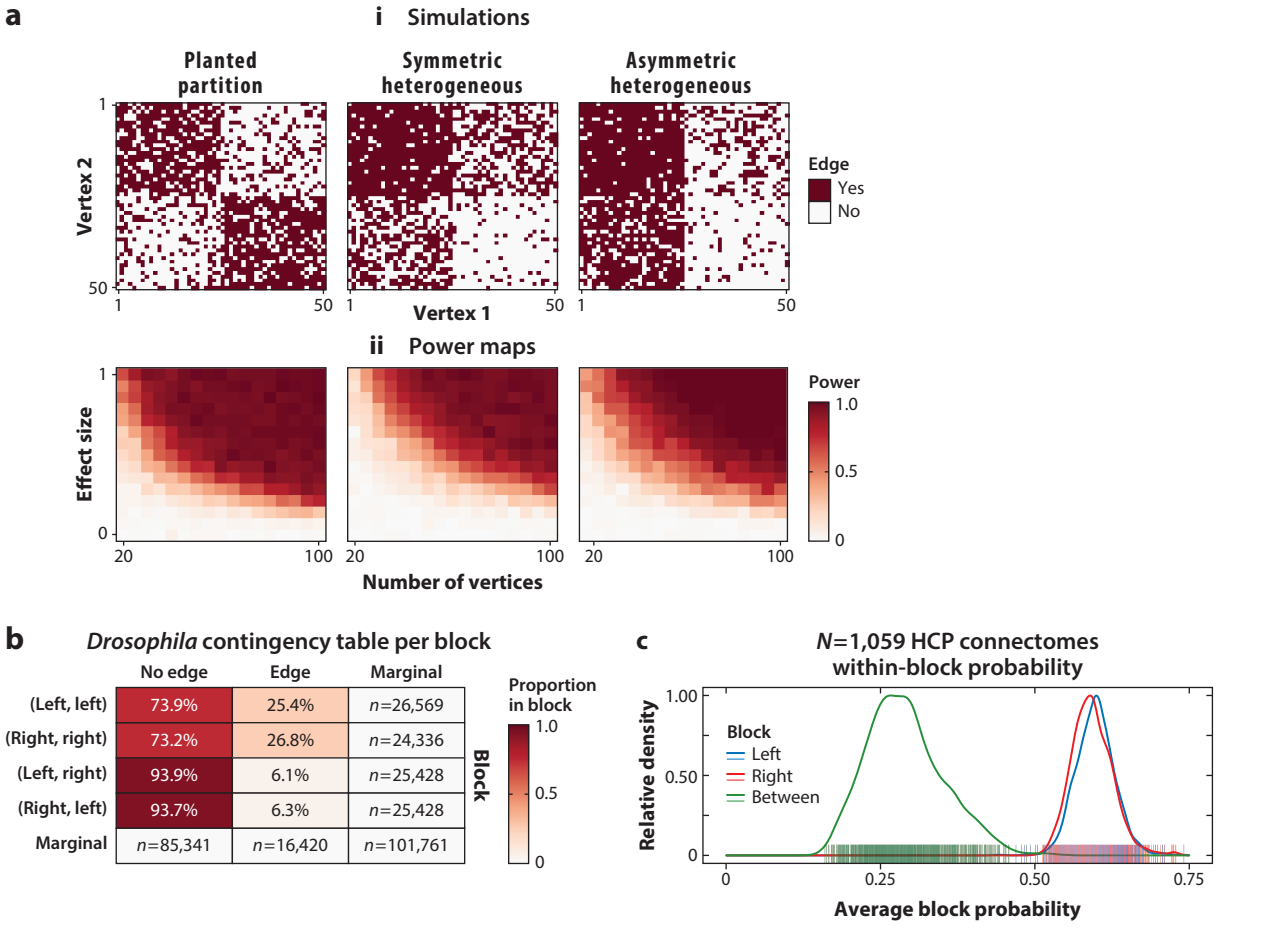


Figure 6

Estimating optimal block structure. (*a*) The χ^2 test is effective for identifying the ideal block structure across disparate candidate block structures (*a*, *i*), as power improves as both effect size and graph size increase (*a*,*ii*). (*b*) 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. (*c*) Similarly, all $N = 1{,}059$ Human Connectome Project (HCP) diffusion connectomes show planted partition structure, with a similar interpretation to the *Drosophila* result.

the diffusion connectome: ER, planted partition, and symmetric heterogeneous. On all $N = 1{,}059$ connectomes, the optimal block structure is planted partition, using the χ^2 test.

Supplemental Material >

The **Supplemental Appendix (Section E.2)** proposes the use of several testing variants (analysis of variance, or ANOVA; Kruskal–Wallis; and distance correlation) for the weighted *Drosophila* mushroom body and the weighted HCP diffusion connectomes for investigating the optimal block structure. All tests yield the same conclusion (and in the case of the HCP data set, again for all $N = 1{,}059$ connectomes) that the mushroom body and diffusion connectomes display planted partition structure. An important consideration is that the implication for weighted graphs is, rather than the on-diagonal and off-diagonal blocks sharing the same probability (as is the case for the unweighted graphs), the two on-diagonal blocks, (left, left) and (right, right), share a common distribution. The implication is similar for the two off-diagonal blocks, (left, right) and (right, left).

5.3. Same Network, Different Communities

In the case of two-block SBMs with positive semidefinite 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 the between-block connectivity. In the core-periphery structure, $a \gg b, c$; that is, one block has relatively higher within-block connectivity than other blocks' within-block probability and between-block connectivity.

In this section, we examine the two spectral embedding clustering approaches described in Section 4.1.1, which produce different clusterings depending on the SBM (Priebe et al. 2019, Cape et al. 2019). In short, ASE clustering tends to favor core-periphery structure, while LSE clustering tends to favor affinity structure.

We consider graphs generated from a four-block SBM with $n = 4{,}000$ vertices, membership vector $\vec{\pi} = [0.25, 0.25, 0.25, 0.25]$, 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 four-block SBM exhibits both affinity and core-periphery structures when projected down to two blocks, which are shown below:

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

Blocks AB and CD form B_{affinity}, which exhibits the affinity structure, while blocks AC and BD form B_{core}, which exhibits the core-periphery structure. A network is sampled from the four-block SBM, and spectral clustering is performed (see Section 4.1.4) with embedding dimension $\hat{d} = 2$ and $K = 2$ the number of clusters. **Figure 7** shows the spectral clustering results. In **Figure 7*a***, clustering with LSE shows that the blocks forming affinity structures are grouped together, and in **Figure 7*b***, 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.

5.4. 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 4.1.4, 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 Priebe et al. 2017 for an extensive investigation) with $\hat{d} = 3$. GMM was performed independently on both hemispheres, with the clustering assignments and embeddings shown in **Figure 8**. 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 mostly comprises a single cell type. Thus, spectral clustering can provide neuroscientists a way to find meaningful communities when the assignment is not known.

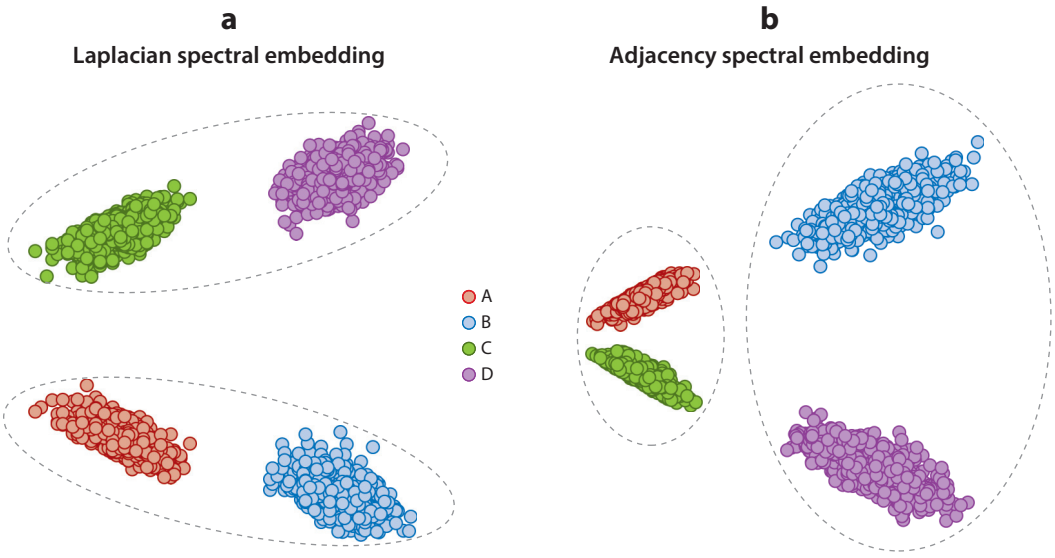


Figure 7

Different clustering results from adjacency spectral embedding (ASE) and Laplacian spectral embedding (LSE). For both ASE and LSE, the network was embedded into $d = 2$ dimensions, and GMM with $K = 2$ clusters was 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 the same group. (*a*) Clustering the embeddings from LSE results in affinity clustering. (*b*) Clustering the embeddings from ASE results in core-periphery clustering.

6. 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 4.2 using simulated and HCP data. The **Supplemental Appendix (Section F)** contains additional exploration of weighted connectomes.

Supplemental Material >

6.1. 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 in 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 8**, 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 9**. This matching correctly identified 78.8% (78 of 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.

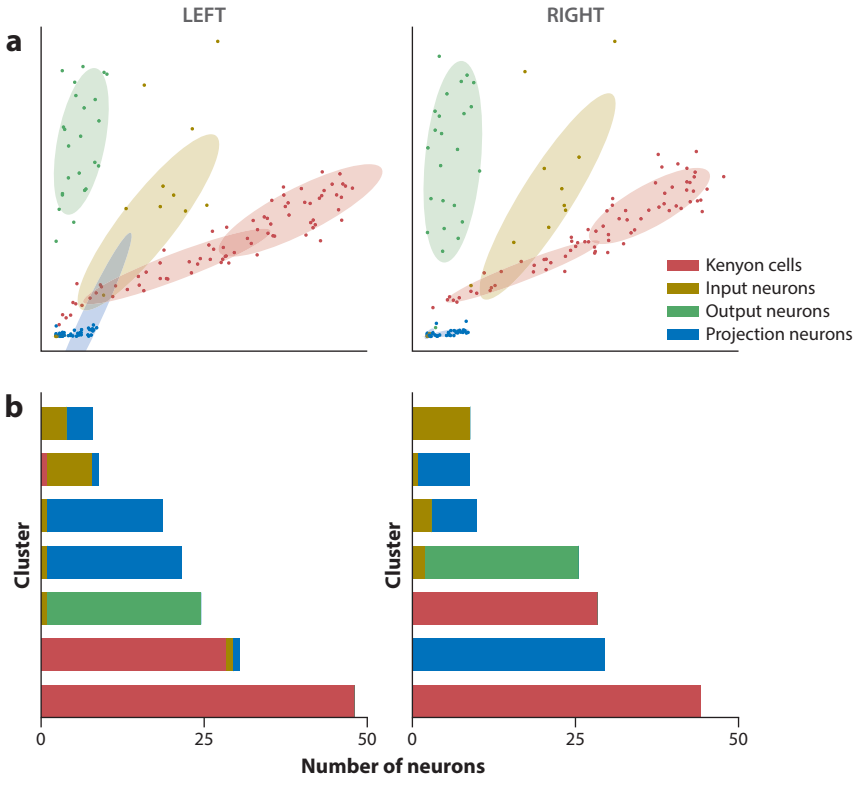


Figure 8

Spectral clustering of the *Drosophila* mushroom body network. (*a*) The 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. Each color corresponds to one of Kenyon cells, input neurons, output neurons, or projections neurons. (*b*) Stacked bar graphs showing each cluster's composition in terms of neuron cell type, for both the left- and the right-hemisphere clusterings. Each cluster mostly comprises a single cell type for both left- and right-hemisphere networks, meaning that spectral clustering can recover true communities.

6.2. Testing for Significant Edges

We consider two populations of networks generated from an ER model and a two-block kidney-egg SBM model. All networks have $n = 20$ vertices and $\pi = [0.25, 0.75]$. The block probability matrices for each population are 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 populations is in the first block, $\mathbf{B}_{11}$, and δ is the magnitude of the difference, which ranges from 0 to $(1 - p)$. In other words, δ is the effect size. In total, m networks are sampled ($\frac{m}{2}$ networks per population). For each edge, the t-test statistic is computed between the two populations, and these are then ranked from largest to smallest in magnitude. Ranking of the test statistics and a cutoff are utilized, rather than p-value corrections (e.g., Bonferroni correction), to control for false positive rate. In this case, the ten edges with the largest magnitudes are considered since we expect ten edges to be different. Nonparametric tests are not considered since many of them are based on ranking the underlying data, which is not

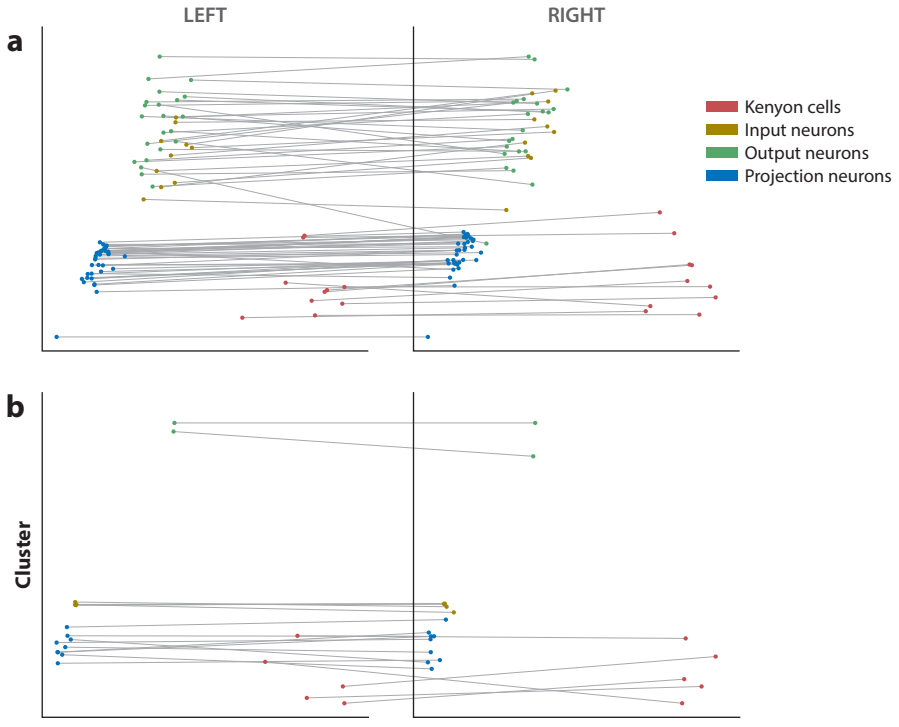


Figure 9

Graph matching on the *Drosophila* mushroom body network. All panels show the first two dimensions of principal component analysis on the adjacency spectral embedding of the mushroom body network (for visualization purposes). Each point represents a neuron in the network that 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 of 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.

sensible for binary data. The performance is evaluated with recall@10, which quantifies the fraction of the top ten ranked edges that are indeed truly different edges, averaged over 100 repeated trials.

Figure 10*a* shows that when the effect size is small ($\delta \leq 0.05$), significant edges cannot be detected even at the largest sample sizes ($m = 1{,}000$). However, when effect size is large ($\delta \geq 0.45$), significant edges can be perfectly detected at relatively small sample sizes ($m \geq 30$).

Connectivity in human brains was analyzed using the structural connectomes (see the **Supplemental Appendix, Section D.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 as empirical trustworthiness in **Figure 10*b***. Thus, empirical trustworthiness is the confidence with which one can trust that a significant edge is truly significant. There are 2,380 possible total edges in connectomes with 70 vertices, but only 49 edges have trustworthiness ≥0.9, meaning one can only trust significance for a small set of edges.

Supplemental Material >

The **Supplemental Appendix (Section F.2)** investigates the edge-wise testing in weighted connectomes. We leverage the *t*-test, Mann–Whitney–Wilcoxon test, and Kolmogorov-Smirnov

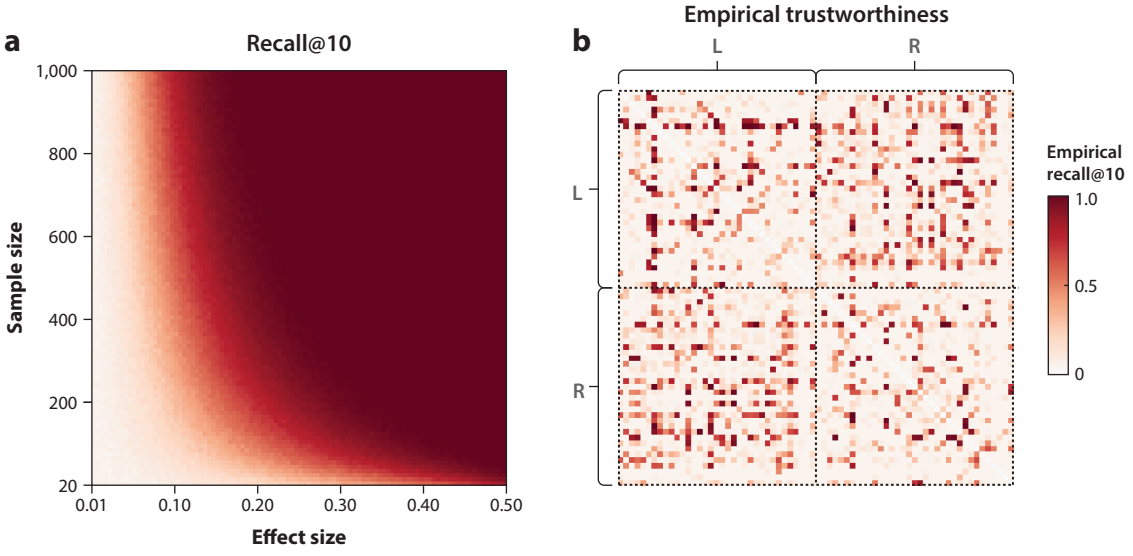


Figure 10

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 a *t*-test. The color bar represents recall@10 averaged over 100 trials. When the effect size is small, significant edges cannot be detected even at large sample sizes. When effect size is large, significant edges can be detected at small sample sizes ($m = 1{,}000$). (*b*) Analysis of structural connectomes from the Human Connectome Project data, with the vertices organized by left (L) and right (R) hemispheres. Edge weights are binarized to parallel the simulations. The heat map shows the empirical trustworthiness of significant edges when comparing each edge between females and males.

(KS) test, which is a nonparametric test of whether there exists a difference in empirical cumulative distributions between edge weights. We find that the KS test is the only test that is appropriate for weighted edges since the KS test can detect changes not only in the means but also in the variance. In weighted HCP connectomes, we find that 256 edges have trustworthiness ≥ 0.9 and that a very small fraction of edges can be trusted to be significant.

6.3. Testing for Significant Vertices

In this section, we test for significant vertices using different representations of vertices. The simplest representation is a set of edges, where the corresponding rows (or columns) 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, where the latent positions of vertices are tested for difference. Since all representations are multivariate, hypotheses are tested using Hotelling's *t*-squared test, which is a multivariate generalization of the *t*-test.

We consider a population of planted partition SBMs 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. In total, $m = 100$ networks are sampled per population, and the *p*-values are computed using Hotelling's *t*-squared test on each of the three vertex

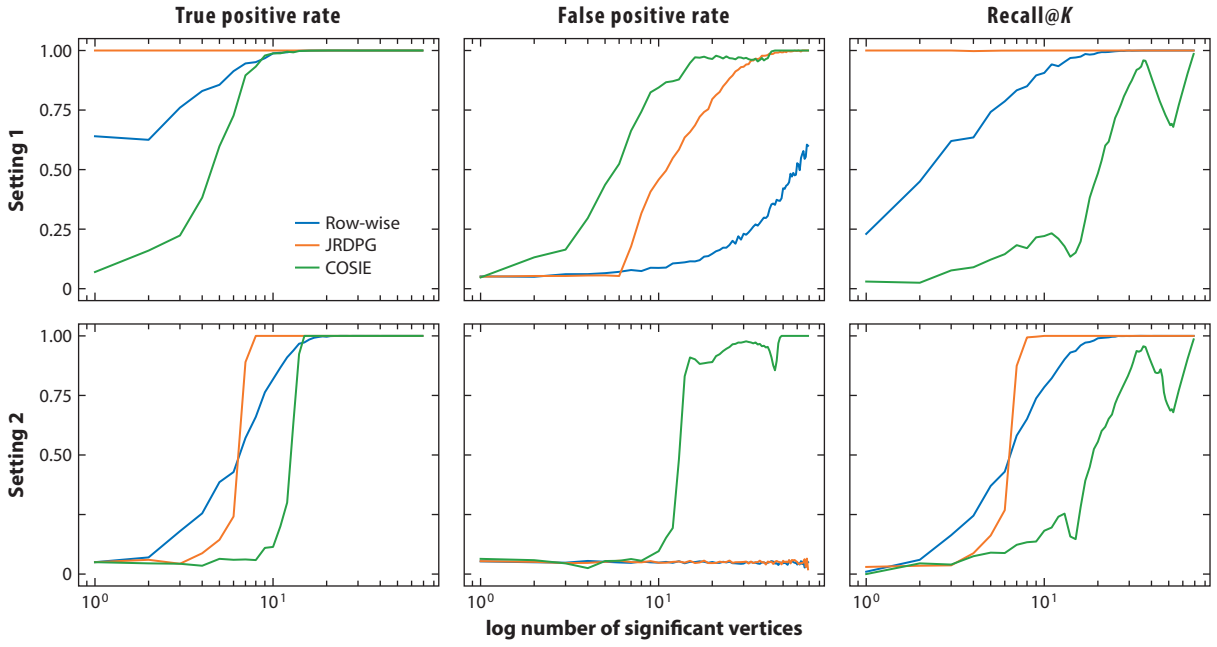


Figure 11

Performance for finding significant vertices using various representations of vertices. We compare a population of graphs from a planted partition stochastic block model (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 joint random dot product graph (JRDPG) representations are valid, while common subspace independent-edge model (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.

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, false positive rate, and recall@K, where K is the number of significant vertices.

Figure 11 shows that the p-values cannot necessarily be trusted. That is, in some settings, the significant vertices cannot be trusted due to an uncontrolled false positive rate. 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.

7. CONCLUSION

Connectomics is an exciting area and is full of interesting ideas; consequently, a variety of analysis frameworks have emerged. However, the use of statistical modeling in connectomics is still relatively sparse, especially compared with 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.

SUMMARY POINTS

1. Do not rely on network statistics to characterize populations of connectomes. In general, network statistics do not 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, the *t*-test is appropriate for binary connectomes but typically invalid and/or underpowered for weighted connectomes.
5. Do not 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 the smallest *p*-values to reject the null hypothesis, as the sorting can be trusted but not necessarily the magnitudes of *p*-values.

DISCLOSURE STATEMENT

Joshua T. Vogelstein received funding from Microsoft Research within the past three years. The other 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 Defense Advanced Research Projects Agency (DARPA) under agreement numbers FA8650-18-2-7834 and FA8750-17-2-0112, and Microsoft Research. All graph-related simulations and analysis were performed using `Graspologic` (**https://neurodata.io/graspy/**), and all multivariate hypothesis testing was done using `hyppo` (**https://neurodata.io/hyppo**) (Chung et al. 2019, Panda et al. 2019).

LITERATURE CITED

Afshin-Pour B, Hossein-Zadeh GA, Strother SC, Soltanian-Zadeh H. 2012. Enhancing reproducibility of fMRI statistical maps using generalized canonical correlation analysis in NPAIRS framework. *NeuroImage* 60:1970–81

Airoldi EM, Blei DM, Fienberg SE, Xing EP. 2008. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* 9:1981–2014

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Control* 19:716–23

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. *Sci. Data* 4:170181

Amico E, Marinazzo D, Di Perri C, Heine L, Annen J, et al. 2017. Mapping the functional connectome traits of levels of consciousness. *Neuroimage* 148:201–11
Anscombe FJ. 1973. Graphs in statistical analysis. *Am. Stat.* 27:17–21
Arroyo J, Athreya A, Cape J, Chen G, Priebe CE, Vogelstein JT. 2019. Inference for multiple heterogeneous networks with a common invariant subspace. arXiv:1906.10026 [stat.ME]
Arroyo J, Levina E. 2020. Simultaneous prediction and community detection for networks with application to neuroimaging. arXiv:2002.01645 [stat.ME]
Arroyo Relión JD, Kessler D, Levina E, Taylor SF. 2019. Network classification with applications to brain connectomics. *Ann. Appl. Stat.* 13:1648–77
Athey TL, Vogelstein JT. 2019. AutoGMM: Automatic Gaussian mixture modeling in Python. arXiv:1909.02688 [cs.LG]
Athreya A, Fishkind DE, Tang M, Priebe CE, Park Y, et al. 2017. Statistical inference on random dot product graphs: a survey. *J. Mach. Learn. Res.* 18:8393–484
Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57:289–300
Biswal BB, Mennes M, Zuo XNN, Gohel S, Kelly AMC, et al. 2010. Toward discovery science of human brain function. *PNAS* 107:4734–39
Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* 2008:P10008
Bullmore ET, Bassett DS. 2011. Brain graphs: graphical models of the human brain connectome. *Annu. Rev. Clin. Psychol.* 7:113–40
Bullmore ET, Sporns O. 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10:186–98
Cape J, Tang M, Priebe CE. 2019. On spectral embedding performance and elucidating network structure in stochastic blockmodel graphs. *Netw. Sci.* 7:269–91
Chatterjee S. 2015. Matrix estimation by universal singular value thresholding. *Ann. Stat.* 43:177–214
Chen H, Soni U, Lu Y, Maciejewski R, Kobourov S. 2018. Same stats, different graphs. In *International Symposium on Graph Drawing and Network Visualization*, ed. T Biedl, A Kerren, pp. 463–77. New York: Springer
Chung J, Pedigo BD, Bridgeford EW, Varjavand BK, Helm HS, Vogelstein JT. 2019. GraSPy: graph statistics in Python. *J. Mach. Learn. Res.* 20:1–7
Chung K, Deisseroth K. 2013. CLARITY for mapping the nervous system. *Nat. Methods* 10:508–13
Clauset A, Newman ME, Moore C. 2004. Finding community structure in very large networks. *Phys. Rev. E* 70:066111
Craddock RC, Jbabdi S, Yan CG, Vogelstein JT, Castellanos FX, et al. 2013. Imaging human connectomes at the macroscale. *Nat. Methods* 10:524–39
Crainiceanu CM, Caffo BS, Luo S, Zipunnikov VM, Punjabi NM. 2011. Population value decomposition, a framework for the analysis of image populations. *J. Am. Stat. Assoc.* 106:775–90
Durante D, Dunson DB, Vogelstein JT. 2017. Rejoinder: nonparametric Bayes modeling of populations of networks. *J. Am. Stat. Assoc.* 112:1547–52
Efron B. 2008. Simultaneous inference: When should hypothesis testing problems be combined? *Ann. Appl. Stat.* 2:197–223
Erdős P, Rényi A. 1959. On random graphs, I. *Publ. Math. Debrecen* 6:290–97
Faskowitz J, Yan X, Zuo XN, Sporns O. 2018. Weighted stochastic block models of the human connectome across the life span. *Sci. Rep.* 8:1–16
Fisher R. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd
Fortunato S, Hric D. 2016. Community detection in networks: a user guide. *Phys. Rep.* 659:1–44
Genovese CR, Lazar NA, Nichols T. 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15:870–78
Ghoshdastidar D, Gutzeit M, Carpentier A, von Luxburg U. 2017. Two-sample tests for large random graphs using network statistics. arXiv:1705.06168 [stat.ME]
Ginestet CE, Li J, Balachandran P, Rosenberg S, Kolaczyk ED. 2017. Hypothesis testing for network data in functional neuroimaging. *Ann. Appl. Stat.* 11:725–50

Goldenberg A, Zheng AX, Fienberg SE, Airoldi EM. 2010. A survey of statistical network models. *Found. Trends Mach. Learn.* 2:129–233

Grover A, Leskovec J. 2016. node2vec: Scalable feature learning for networks. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–64. New York: ACM

Guha S, Rodriguez A. 2020. Bayesian regression with undirected network predictors with an application to brain connectome data. *J. Am. Stat. Assoc.* **https://doi.org/10.1080/01621459.2020.1772079**

Hagmann P. 2005. *From diffusion MRI to brain connectomics*. PhD Thesis, École Polytechnique Fédérale de Lausanne, Lausanne, Switz.

Hoff PD, Raftery AE, Handcock MS. 2002. Latent space approaches to social network analysis. *J. Am. Stat. Assoc.* 97:1090–98

Holland PW, Laskey KB, Leinhardt S. 1983. Stochastic blockmodels: first steps. *Soc. Netw.* 5:109–37

Jackson JE. 2005. *A User's Guide to Principal Components*. New York: Wiley

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. **https://doi.org/10.1101/188706**

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

Kolaczyk ED, Csárdi G. 2014. *Statistical Analysis of Network Data with R*. New York: Springer

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:1705.09355 [stat.ME]

Lock EF, Hoadley KA, Marron JS, Nobel AB. 2013. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *Ann. Appl. Stat.* 7:523

Lyzinski V, Sussman DL. 2017. Matchability of heterogeneous networks pairs. *Inform. Inference*. **https://doi.org/10.1093/imaiai/iaz031**

Lyzinski V, Tang M, Athreya A, Park Y, Priebe CE. 2017. Community detection and classification in hierarchical stochastic blockmodels. *IEEE Trans. Netw. Sci. Eng.* 4:13–26

Marchette D, Priebe CE, Coppersmith G. 2011. Vertex nomination via attributed random dot product graphs. In *Bulletin of the International Statistical Institute Proceedings of the 58th World Statistics Congress 2011, Dublin*, pp. 5047–52. The Hague, Neth.: Int. Stat. Inst. **https://2011.isiproceedings.org/papers/950095.pdf**

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. 1290–94. New York: ACM

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. Piscataway, NJ: IEEE

Newman ME. 2002. Random graphs as models of networks. In *Handbook of Graphs and Networks: From the Genome to the Internet*, ed. S Bornholdt, HG Schuster, pp. 35–68. New York: Wiley

Newman ME. 2013. Spectral methods for community detection and graph partitioning. *Phys. Rev. E* 88:042822

Nielsen AM, Witten D. 2018. The multiple random dot product graph model. arXiv:1811.12172 [stat.ME]

Panda S, Palaniappan S, Xiong J, Bridgeford EW, Mehta R, et al. 2019. hyppo: a comprehensive multivariate hypothesis testing Python package. arXiv:1907.02088 [stat.CO]

Priebe CE, Coppersmith G, Rukhin A. 2010. You say "graph invariant," I say "test statistic". *Stat. Comput. Stat. Graph.* 21:11–14

Priebe CE, Park Y, Tang M, Athreya A, Lyzinski V, et al. 2017. Semiparametric spectral modeling of the *Drosophila* connectome. arXiv:1705.03297 [stat.ML]

Priebe CE, Park Y, Vogelstein JT, Conroy JM, Lyzinski V, et al. 2019. On a two-truths phenomenon in spectral graph clustering. *PNAS* 116:5995–6000

Richiardi J, Eryilmaz H, Schwartz S, Vuilleumier P, Van De Ville D. 2011. Decoding brain states from fMRI connectivity graphs. *Neuroimage* 56:616–26

Rieke F. 1997. *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press

Rissanen J. 1978. Modeling by shortest data description. *Automatica* 14:465–71
Rohe K, Chatterjee S, Yu B. 2011. Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Stat.* 39:1878–915
Rohe K, Qin T, Yu B. 2016. Co-clustering directed graphs to discover asymmetries and directional communities. *PNAS* 113:12679–84
Rubin-Delanchy P, Priebe CE, Tang M, Cape J. 2017. A statistical interpretation of spectral embedding: the generalised random dot product graph. arXiv:1709.05506 [stat.ML]
Rukhin A, Priebe CE. 2010. A comparative power analysis of the maximum degree and size invariants for random graph inference. *J. Stat. Plann. Inference* 141:1041–46
Russell SJ, Norvig P. 2016. *Artificial Intelligence: A Modern Approach*. Essex, UK: Pearson
Scheinerman ER, Tucker K. 2010. Modeling graphs using dot product representations. *Comput. Stat.* 25:1–16
Schwarz G. 1978. Estimating the dimension of a model. *Ann. Stat.* 6:461–64
Scrucca L, Fop M, Murphy TB, Raftery AE. 2016. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* 8:289
Shepherd GM. 1991. *Foundations of the Neuron Doctrine*. Oxford, UK: Oxford Univ. Press. 1st ed.
Simes RJ. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73:751–54
Sporns O, Tononi G, Kötter R. 2005. The human connectome: a structural description of the human brain. *PLOS Comput. Biol.* 1:e42
Sussman DL, Tang M, Fishkind DE, Priebe CE. 2012. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Am. Stat. Assoc.* 107:1119–28
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
Tang M, Athreya A, Sussman DL, Lyzinski V, Park Y, Priebe CE. 2017a. A semiparametric two-sample hypothesis testing problem for random graphs. *J. Comput. Graph. Stat.* 26:344–54
Tang M, Athreya A, Sussman DL, Lyzinski V, Priebe CE, et al. 2017b. A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli* 23:1599–630
Tang M, Sussman DL, Priebe CE. 2013. Universally consistent vertex classification for latent positions graphs. *Ann. Stat.* 41:1406–30
Tang R, Ketcha M, Badea A, Calabrese ED, Margulies DS, et al. 2018. Connectome smoothing via low-rank approximations. *IEEE Trans. Med. Imaging* 38:1446–56
Thirion B, Varoquaux G, Dohmatob E, Poline JB. 2014. Which fMRI clustering gives good brain parcellations? *Front. Neurosci.* 8:167
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
Varoquaux G, Craddock RC. 2013. Learning and comparing functional connectomes across subjects. *Neuroimage* 80:405–15
Varoquaux G, Gramfort A, Poline JB. 2010. Brain covariance selection: better individual functional connectivity models using population prior. arXiv:1008.5071 [stat.ML]
Vogelstein JT, Bridgeford EW, Pedigo BD, Chung J, Levin K, et al. 2019. Connectal coding: discovering the structures linking cognitive phenotypes to individual histories. *Curr. Opin. Neurobiol.* 55:199–212
Vogelstein JT, Conroy JM, Lyzinski V, Podrazik LJ, Kratzer SG, et al. 2015. Fast approximate quadratic programming for graph matching. *PLOS ONE* 10:e0121002
Vogelstein JT, Roncal WG, Vogelstein RJ, Priebe CE. 2012. Graph classification using signal-subgraphs: applications in statistical connectomics. *IEEE Trans. Pattern Anal. Mach. Intel.* 35:1539–51
Wang L, Zhang Z, Dunson D. 2019a. Common and individual structure of brain networks. *Ann. Appl. Stat.* 13:85–112
Wang L, Zhang Z, Dunson D. 2019b. Symmetric bilinear regression for signal subgraph estimation. *IEEE Trans. Signal Proc.* 67:1929–40
Wang S, Arroyo J, Vogelstein JT, Priebe CE. 2019c. Joint embedding of graphs. *IEEE Trans. Pattern Anal. Mach. Intel.* **https://doi.org/10.1109/TPAMI.2019.2948619**
Wang S, Shen C, Badea A, Priebe CE, Vogelstein JT. 2018. Signal subgraph estimation via vertex screening. arXiv:1801.07683 [stat.ME]

Wasserman S, Anderson C. 1987. Stochastic a posteriori blockmodels: construction and assessment. *Soc. Netw.* 9:1–36

Xia Y, Li L. 2019. Matrix graph hypothesis testing and application in brain connectivity alternation detection. *Stat. Sin.* 29:303–28

Young SJ, Scheinerman ER. 2007. Random dot product graph models for social networks. In *Algorithms and Models for the Web Graph*, ed. B Kamiński, P Prałat, P Szufel, pp. 138–49. New York: Springer

Zalesky A, Fornito A, Harding IH, Cocchi L, Yücel M, et al. 2010. Whole-brain anatomical networks: does the choice of nodes matter? *Neuroimage* 50:970–83

Zhang J, Sun WW, Li L. 2018a. Network response regression for modeling population of networks with covariates. arXiv:1810.03192 [stat.ME]

Zhang Z, Descoteaux M, Zhang J, Girard G, Chamberland M, et al. 2018b. Mapping population-based structural connectomes. *NeuroImage* 172:130–45

Zheng AX, Fienberg SE, Airoldi EM, Goldenberg A. 2009. A survey of statistical network models. *Found. Trends Mach. Learn.* 2:129–233

Zhu M, Ghodsi A. 2006. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Comput. Stat. Data Anal.* 51:918–30

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. *Sci. Data* 1:140049