Connectome Smoothing via Low-rank Approximations

Runze Tang[1], Michael Ketcha[2], Alexandra Badea[3], Evan D. Calabrese[3], Daniel S. Margulies[4], Joshua T. Vogelstein[2,5], Carey E. Priebe[1], Daniel L. Sussman[6*]

1 Department of Applied Math & Statistics, The Johns Hopkins University, Baltimore, MD

2 Department of Biomedical Engineering, The Johns Hopkins University, Baltimore, MD

3 Department of Radiology, and Department of Biomedical Engineering, Duke University, Durham, NC

4 Max Planck Research Group for Neuroanatomy & Connectivity, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany;

4 Child Mind Institute, New York, NY

5 Department of Mathematics & Statistics, Boston University, Boston, MA

* sussman@bu.edu

Abstract

In statistical connectomics, the quantitative study of brain networks, estimating the mean of a population of graphs based on a sample is a core problem. Often, this problem is especially difficult because the sample or cohort size is relatively small, sometimes even a single subject. While using the element-wise sample mean of the adjacency matrices is a common approach, this method does not exploit any underlying structural properties of the graphs. We propose using a low-rank method which incorporates tools for dimension selection and diagonal augmentation to smooth the estimates and improve performance over the naïve methodology for small sample sizes. Theoretical results for the stochastic blockmodel show that this method offers major improvements when there are many vertices. Similarly, we demonstrate that the low-rank methods outperform the standard sample mean for a variety of independent edge distributions as well as human connectome data derived from magnetic resonance imaging, especially when sample sizes are small. Moreover, the low-rank methods yield "eigen-connectomes", which correlate with the lobe-structure of the human brain and superstructures of the mouse brain. These results indicate that low-rank methods are an important part of the tool box for researchers studying populations of graphs in general, and statistical connectomics in particular.

Keywords: networks, connectome, low-rank, estimation

1 Introduction

Estimation of the mean of a population based on samples is at the core of statistics. The sample mean, motivated by its intuitive appeal, the law of large numbers, and the central limit theorem, has its place as one of the most important statistics for this task. In modern settings, averages of many different kinds of data representations are taken, including scalars and vectors, but also objects like images, shapes, and documents. In this manuscript we consider the challenges of estimating a population mean based on a sample of brain networks, as represented by their structural connectomes.

The mean of a population of graphs is a high-dimensional object, consisting of $O(N^2)$ parameters for graphs with N nodes (vertices). When the number of samples M is much smaller than N^2, or even N, estimating such a high-dimensional estimands using naive unbiased methods can lead to inaccurate estimates with very high variance. Furthermore, using these estimates for subsequent inference tasks such as testing can lead to low power and accuracy. By exploiting a bias-variance trade-off, it is often fruitful to develop estimators which have some bias but greatly reduced variance. When these estimators are biased towards low-dimensional structures which well approximate the full dimensional population mean, major improvements in estimation can be realized (Trunk, 1979). Furthermore, a more parsimonious representation of the mean improves interpretability and allows for further exploratory analysis.

In complex data settings such as shape data, language data, or graph data, caution is necessary even in properly defining the mean. For a population, we define the mean graph as the weighted adjacency matrix with weights given by the proportion of times the corresponding edge appears in the population. This definition naturally extends the definition of the mean for standard Euclidean data. As with real valued data, one may want to define the mean of a population of graphs to be a graph. This is captured in the notion of the median graph (Jiang et al., 2001). However, this may be too restrictive for populations of graphs where there is high variation. As described below, our definition of the mean graph is the expectation of the adjacency matrix.

The population mean is becoming an important object for statistical connectomics applications. For example, Ginestet et al. (2014) proposed a way to test if there is a difference between the distributions for two groups of networks. While hypothesis testing is the end goal of their work, estimation is a key intermediate step which may be improved by accounting for underlying structure in the mean matrix. Thus, improving the estimation procedures for the mean graph is not only important by itself, but also can be applied to help improve other statistical inference procedures.

To better illustrate the idea, consider brain graphs (connectomes) available through the Consortium for Reliability and Reproducibility (CoRR) (Zuo et al., 2014) after processing with ndmg pipeline (Kiar et al., 2017, 2016). In particular, the SWU4 dataset contains 454 brain scans (Qiu et al., 2017). Each vertex in the estimated graphs represents a region defined by the Desikan atlases, while an edge exists between two vertices if there is at least one streamline connecting

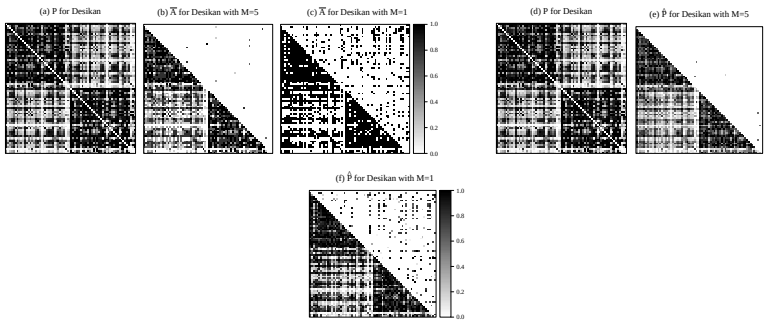


Fig 1: **Heat maps depicting the weighted adjacency matrices of the population mean P, the sample mean $\bar{A}$, and the estimator $\hat{P}$.** The heat map of the population mean for the 454 graphs are depicted in Panel (a) (also in Panel (d)). The two heat maps in the second column indicate the sample mean for 5 sampled graphs (Panel (b)), and the low-rank estimate $\hat{P}$ for the same 5 sampled graphs with rank $d = 11$ selected using the Zhu and Ghodsi method (Panel (e)). Section 3.3 discusses details about how to construct $\hat{P}$. Darker pixels indicate a higher probability of an edge between the given vertices. By calculating the mean squared error based on this sample, $\hat{P}$ (with mean squared error equal to 0.015) outperforms $\bar{A}$ (with mean squared error equal to 0.016), with a 3% relative improvement. To highlight the improvements, the upper triangular area of the heat maps for $\bar{A}$ and $\hat{P}$ shows the edges (18 edges for $\bar{A}$ and 6 for $\hat{P}$) which have absolute estimation error larger than 0.4. In the third column, two heat maps using a sample size $M = 1$ ($\hat{P}$ is calculated with a dimension $d = 12$) show a smoothing effect in the heat map of $\hat{P}$ (with mean squared error equal to 0.049), which leads to a 53% relative improvement compared to $\bar{A}$ (with mean squared error equal to 0.104). Similarly, the same absolute estimation error threshold of 0.4 highlights 504 edges for $\bar{A}$ and 234 edges for $\hat{P}$.

the corresponding regions of the brain. More details about this dataset are given in Section 4.2 and Section A.4. Our goal is to estimate the mean graph of the population P, defined as the entry-wise mean of all the 454 graphs. Fig. 1 shows the population mean graph P on the left panel. Darker pixels indicate a higher proportion of graphs having an edge between the given vertices.

The entry-wise sample mean is a reasonable estimator if it is assumed that each edge is present independently of all other edges without taking any additional structure into account. However, with only a small sample size, the entry-wise sample mean does not perform very well. Fig. 1(b) depicts the entry-wise sample mean $\bar{A}$ when the sample size is $M = 5$ in the SWU4 dataset example. While $\bar{A}$ gives a reasonable estimate of P, there are still some node pairs with very inaccurate estimates. The upper triangular area of the heat map for $\bar{A}$ depicts the 18 vertex-pairs which have an absolute estimation error larger than 0.4. When the sample size is small, the performance of $\bar{A}$ degrades due

to its high variance. Such phenomena are most obvious when the sample size decreases from $M = 5$ to $M = 1$. Fig. 1(c) shows the heat map of $\bar{A}$ based on sample size $M = 1$. Since there is only one observed graph, $\bar{A}$ is binary and thus very bumpy. Similarly, when the same absolute estimation error threshold is 0.4, 504 (out of 2415) edges in the upper triangular area are highlighted.

Intuitively, an estimator incorporating known structure in the true distribution of the graphs, assuming the estimator is computationally tractable, is preferable to the entry-wise sample mean. We propose an estimator based on a low-rank structure of a family of random graphs. Section 3.3 discusses details about this estimator. These estimates improve the performance since they have much lower overall variance than naive entry-wise sample means, thereby offsetting the bias towards the low-rank structure in terms of overall error.

When we use the same random sample size of $M = 5$ is in Fig. 1, the plot of low-rank estimate $\hat{P}$ in Panel (e) shows a finer gradient of values which leads, in this case, to a 3% relative improvement in estimation of the true probability matrix, P ($\bar{A}$ with mean squared error equal to 0.016 and $\hat{P}$ with mean squared error equal to 0.015). The upper triangular area of the heat map for $\hat{P}$ depicts the 6 edges which have absolute estimation error larger than 0.4, whereas 18 edges are highlighted for $\bar{A}$ based on the same threshold.

The smoothing effect is much more obvious when sample size is decreased from $M = 5$ to $M = 1$, as in Fig.1(f). $\hat{P}$ smooths the estimate, especially for edges across the two hemispheres, in the lower left and corresponding upper right block (which is not shown in the heat map). Based on the calculations, $\hat{P}$ (with mean squared error equals 0.049) outperforms $\bar{A}$ (with mean squared error equals 0.104), with a 53% relative improvement in estimation. Similarly, the same absolute estimation error threshold of 0.4 highlights 234 edges for $\hat{P}$, less than 50% as many as $\bar{A}$.

In addition to potentially improving the overall accuracy of the estimate of the mean graph, low-rank methods also produce parsimonious and interpretable representations of the data as represented by the random dot product graph model parameters. As demonstrated later, the interpretations of these model parameters illustrate hypotheses which relate the structure of the connectome to well established anatomical structures (see Fig. 2 and Section 4.5) and suggest a basis for mapping structural hierarchies in brain organization.

2 Statistical Connectome Models

This work considers the scenario of having M graphs, represented as adjacency matrices, $A^{(1)}, A^{(2)}, \ldots, A^{(M)}$, each having N vertices with $A^{(m)} \in \{0,1\}^{N \times N}$ for each m. We assume there is a known correspondence for vertices in different graphs, so that vertex i in graph m corresponds to vertex i in graph m' for any i, m, m'. The graphs we consider are undirected and unweighted with no self-loops, so each $A^{(m)}$ is a binary symmetric matrix with zeros along the diagonal.

In connectomics, the brain imaging data for each subject can be processed

to output a graph, where each vertex represents a well-defined anatomical region present in each subject. For structural brain imaging, such as diffusion tensor MRI, an edge may represent the presence of anatomical connections between the two regions as estimated based on tractography for diffusion tensor magnetic resonance imaging (Gray et al., 2012). Similarly, for functional brain imaging, such as fMRI, an edge between two regions may represent the presence of correlated brain activity between the two regions.

For the purpose of this paper, we assume that the graphs are sampled independently and identically from some distribution. To this end, the mean graph is the expectation of each adjacency matrix.

Definition 2.1 (Mean Graph)**.** *Suppose that* $A^{(1)}, \ldots, A^{(M)} \overset{iid}{\sim} \mathcal{G}$ *for some random graph distribution* $\mathcal{G}$*, with* $A^{(m)} \in \{0,1\}^{N \times N}$ *for each* m*. The* mean graph *is defined as* $\mathbb{E}[A^{(m)}]$*, where the graphs are identically distributed* $\mathbb{E}[A^{(m)}] = \mathbb{E}[A^{(m')}]$ *for any* m, m'*.*

2.1 Independent Edge Model

The most general model we consider is the independent edge model (IEM) with parameter $P \in [0,1]^{N \times N}$ (Bollobás et al., 2007). An edge exists between vertex i and vertex j with probability P_{ij}, and each edge is present independently of all other edges. Note that the IEM is a generalization of the Erdős-Rényi random graphs, where each edge is present with probability p independently of all other edges (Gilbert, 1959; Erdős and Rényi, 1959). For this case, the mean graph is $P = \mathbb{E}[A^{(m)}]$.

2.2 Random Dot Product Graph

In graphs, the adjacencies between vertices generally depend on unobserved properties of the corresponding vertices. For example, in a connectomics setting, two brain regions with similar properties might have similar connectivity patterns. The latent positions model (LPM) proposed by (Hoff et al., 2002) captures such structure, where each vertex is associated with a latent position that influences the adjacencies for that vertex. In this model, each vertex i has an associated latent vector $X_i \in \mathbb{R}^d$. Conditioned on the latent positions, the existence of each edge is independent and the probability the edge is present only depends on the latent vectors of the incident vertices through a link function. If d is much smaller than the number of vertices N and the link function is known, LPMs are more parsimonious models compared to IEM, requiring only dN parameters rather than $\binom{N}{2} = N(N-1)/2$.

A specific instance of an LPM examined in this work is the random dot product graph model (RDPG) (Young and Scheinerman, 2007; Nickel, 2008) where the link function is the dot product, so the probability of an edge being present between two nodes is the dot product of their latent vectors. The latent position is determined by properties of that vertex, and vertices whose latent

positions are close are more likely to have an edge between them than vertices whose latent positions are far from each other.

Formally, let $\mathcal{X} \subset \mathbb{R}^d$ be a set such that $x, y \in \mathcal{X}$ implies $x^\top y = \sum_{i=1}^d x_i y_i \in [0,1]$. Let $X_1, \ldots, X_N \in \mathcal{X}$ be column vectors representing the N latent positions and let $X = [X_1 | \cdots | X_N]^\top \in \mathbb{R}^{N \times d}$. A random graph G with random adjacency matrix A is said to be an RDPG if for each adjacency matrix $a \in \{0,1\}^{n \times n}$,

$$\Pr[A = a | X] = \prod_{i<j} (X_i^\top X_j)^{a_{ij}} (1 - X_i^\top X_j)^{1 - a_{ij}}.$$

In the RDPG model, each vertex i is associated with latent position X_i, and conditioned on the latent positions X, the entries A_{ij} are distributed independently as Bernoulli$(X_i^\top X_j)$ for $i < j$. Note that the probability matrix is the outer product of the latent position matrix with itself, $P = XX^\top$. This imposes two properties on P, namely that P is positive-semidefinite and $\text{rank}(P) = \text{rank}(X) \leq d$. These properties lead us to our proposed estimator. Importantly, even if P is positive semi-definite and low-rank A will always be indefinite, since it has a zero diagonal, and will with high probability have full rank.

In the RDPG model, by representing a low-rank matrix in terms of the latent positions, where each vertex is represented as a vector in $\mathbb{R}^d$ and the entries of the matrix are given by the inner products of these vectors, we can analyze and visualize the geometry of these vectors to interpret how each vertex is behaving in the context of the larger graph. By using the SWU4 brain graphs again as an example (see Section 4.2 and A.4) and embedding the mean graph P, the average of all 454 graphs, we determined the estimated latent positions $\hat{X} \in \mathbb{R}^{N \times d}$, where $N = 70$ is the number of vertices and $d = 8$ is the dimension selected by the Zhu and Ghodsi's method (Zhu and Ghodsi, 2006).

Fig. 2 shows the first 4 dimensions of $\hat{X}$ in the brain. The value of $\hat{X}_{ij}$ determines the color of the i-th brain region for the j-th dimension; i.e. the j-th element of the estimated latent vector for the i-th region. Red represents a positive value while blue represents the negative one, and the darker the color, the smaller the magnitude of the $\hat{X}_{ij}$.

The 1st dimension, depicted in Panel (a) of the figure, shows the average level of connectivity for each vertex. Panel (b) shows a distinction between the left and right hemisphere, as conveyed in the 2nd dimension. Similarly, the other dimensions qualitatively correspond to different lobes. For example, the red color corresponds to the Frontal and Temporal lobes in the 3rd dimension, while the light blue roughly matches the Occipital lobe in the 4th dimension.

These "eigen-connectomes" demonstrate noteworthy similarity to the structural connectome harmonics of Atasoy et al. (2016). While prior work has established the utility of eigen-connectomes as a basis set for modeling of functional dynamics, the current study demonstrates the correspondence of these dimensions with lobular divisions. Future work integrating the connections between structural features with the modeling of dynamics may provide insight into the

spatial distribution of large-scale cortical hierarchies (Margulies et al., 2016). Additionally, such representations allow the use of techniques from multivariate analysis to further study the estimated population mean.

The connection between eigenvectors corresponding to each dimension and lobes will be explored more rigorously in Section 4.5.

2.3 Stochastic Blockmodel as an RDPG

One of the most common structures for graphs is that vertices tend to cluster into communities such that vertices of the same community behave similarly, connecting to similar sets of nodes. This structural property is captured by the stochastic blockmodel (SBM) (Holland et al., 1983), where each vertex is assigned to a block and the probability that an edge exists between two vertices depends only on their respective block memberships.

The SBM is parameterized by the number of blocks K (generally much less than the number of vertices N), the block probability matrix $B \in [0,1]^{K\times K}$, and the vector of block memberships $\tau \in \{1,\dots,K\}^N$, where for each $i \in [N]$, $\tau_i = k$ means vertex i is a member of block k.

Conditioned on τ, each entry of the adjacency matrix A_{ij} $(i < j)$ is independently sampled from the Bernoulli distribution with parameter B_{τ_i,τ_j}. To ensure that the SBM can be considered as an RDPG, we impose that the B matrix for the SBM is positive semidefinite. For notational convenience the sub-model of the SBM with positive semidefinite B matrix is referred to as simply the SBM.

To analyze the estimator $\hat{P}$ motivated by RDPG (Section 3.3 discusses how to construct $\hat{P}$), the SBM can be viewed as an RDPG by decomposing B as $\nu\nu^\top$, where $\nu \in \mathbb{R}^{K\times d}$ with rows given by $\nu_1^\top,\dots,\nu_K^\top$ and each row $\nu_k^\top$ is the shared latent position for all vertices assigned to block k. Let $X \in \mathbb{R}^{N\times d}$ have rows given by $X_1^\top = \nu_{\tau_1}^\top, X_2^\top = \nu_{\tau_2}^\top,\dots,X_N^\top = \nu_{\tau_N}^\top$. Since τ_i and τ_j represent the blocks that vertex i and vertex j belong to respectively:

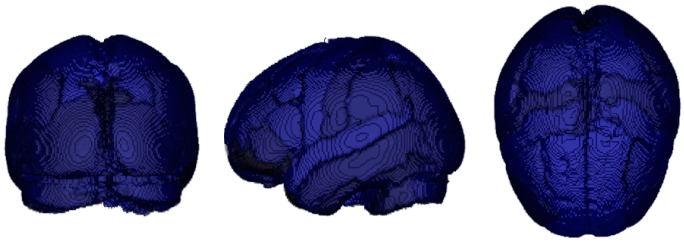
$$\Pr[A_{ij} = 1|\tau] = B_{\tau_i,\tau_j} = \nu_{\tau_i}^\top \nu_{\tau_j} \in [0,1].$$

The SBM is therefore a special case of an RDPG where all vertices in the same block have identical latent positions.

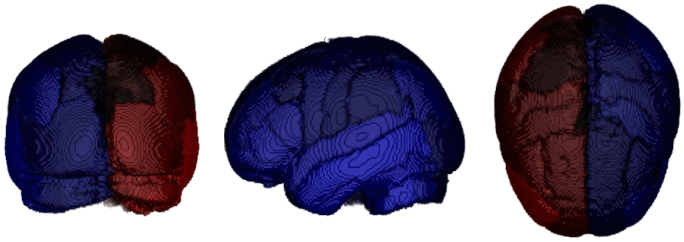
An example SBM is illustrated in Fig. 3. We consider a 5-block SBM and plot the corresponding probability matrix and one adjacency matrix generated from it with 200 vertices. Panel (a) shows the block-constant structure of the probability matrix and panel (b) shows how this structure is reflected in the random binary adjacency matrix.

Remark 2.2. *Rather than allowing vertices to differ from each other as in the RDPG, the SBM presumes all nodes within the same block have the same expected degree. Vertices in the same block are all stochastically equivalent. To better describe complex network structures, many generalizations of the SBM have been explored to incorporate the local variation of vertices to the block structure. Airoldi et al. (2008) proposed mixed membership stochastic blockmodels, which associates each vertex with multiple blocks with a probability vector rather*

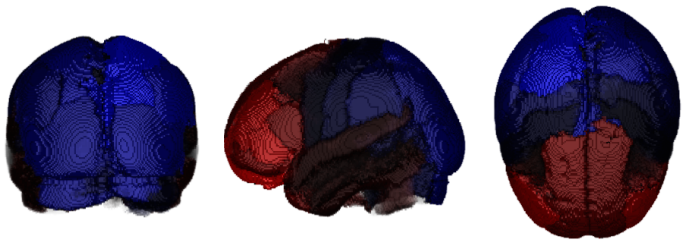
(a) 1st dimension



(b) 2nd dimension



(c) 3rd dimension



(d) 4th dimension

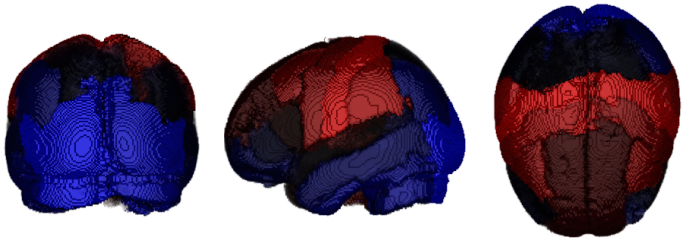


Fig 2: **Brain plots colored separately for each of the first 4 dimensions of $\hat{X}$ under the Desikan atlas.** The color of the i-th brain region for the j-th dimension is determined by the value of $\hat{X}_{ij}$, i.e. the j-th element of the estimated latent vector for the i-th region. Red represents a positive value while blue represents the negative one. And the darker the color is, the smaller the magnitude of the value is. The 1st dimension depicted in Panel (a) of the figure provides an average level of the entire brain. Then in Panel (b) particularly, there is a distinction of the left and right hemisphere as conveyed in the 2nd dimension. Similarly, the other dimensions appear to correlate with the anatomical lobe structures of the brain (see Section 4.5 for more details).

than a single block as SBM requires. To model variation of the expected degrees of different vertices within the same block, Karrer and Newman (2011) proposed degree-corrected SBM, which assigns additional parameters to each vertex to adjust the expected degree relatively. These generalizations aim to capture variations among vertices while maintaining aspects of the original community structure. The RDPG is useful in this regard since any positive semidefinite SBM with degree-correction and mixed membership can be represented as an RDPG and visa versa.

3 Estimators

3.1 Decision Theory

Estimating the mean of a population of samples requires a procedure. In statistical decision theory, the risk of a procedure δ for a given parameter value $\theta \in \Theta$, $R(\delta, \theta)$, is defined as the expected loss. The loss function can be squared Euclidean distance, absolute distance, or other distances (Bickel and Doksum, 2007) and the expectation is taken with respect to the given parameter. A procedure δ' dominates another procedure δ if the risk of δ' is no bigger than the risk of δ regardless of the parameter, and strictly less than the risk of δ in some situations.

In a striking result, Stein (1956) and James and Stein (1961) showed that even the arithmetic mean can be dominated by another procedure. In particular, James and Stein showed that the sample mean for a multivariate normal distribution with at least three dimensions has strictly higher risk than a procedure that introduces shrinkage, and can be strictly improved by carefully biasing the estimate towards any given fixed point. Twenty-seven years later, Gutmann (1982) proved that this phenomenon cannot occur when the sample spaces are finite, as is the case for graphs. However, while there must be some cases where the sample mean is preferred, this does not imply that other estimators should not be considered. In many situations where other structural information is hypothesized, other estimators may be preferable.

3.2 Element-wise sample mean $\bar{A}$

The most natural estimator to consider is the element-wise sample mean. This estimator, defined as $\bar{A} = \frac{1}{M}\sum_{m=1}^{M} A^{(m)}$, is the maximum likelihood estimator (MLE) for the mean graph P if the graphs are sampled from an IEM distribution. It is unbiased so $\mathbb{E}[\bar{A}] = P$ with entry-wise variance $\mathrm{Var}(\bar{A}_{ij}) = P_{ij}(1 - P_{ij})/M$. Moreover, for the independent edge model, $\bar{A}$ is the uniformly minimum-variance unbiased estimator, so it has the smallest variance among all unbiased estimators. Similarly, it enjoys the many asymptotic properties of the MLE as $M \to \infty$ for fixed N. However, if graphs with a large number of vertices are of interest, $\bar{A}$ is not consistent for P as the number of vertices N becomes large for fixed M, while our estimator $\hat{P}$ discussed in Section 3.3 is

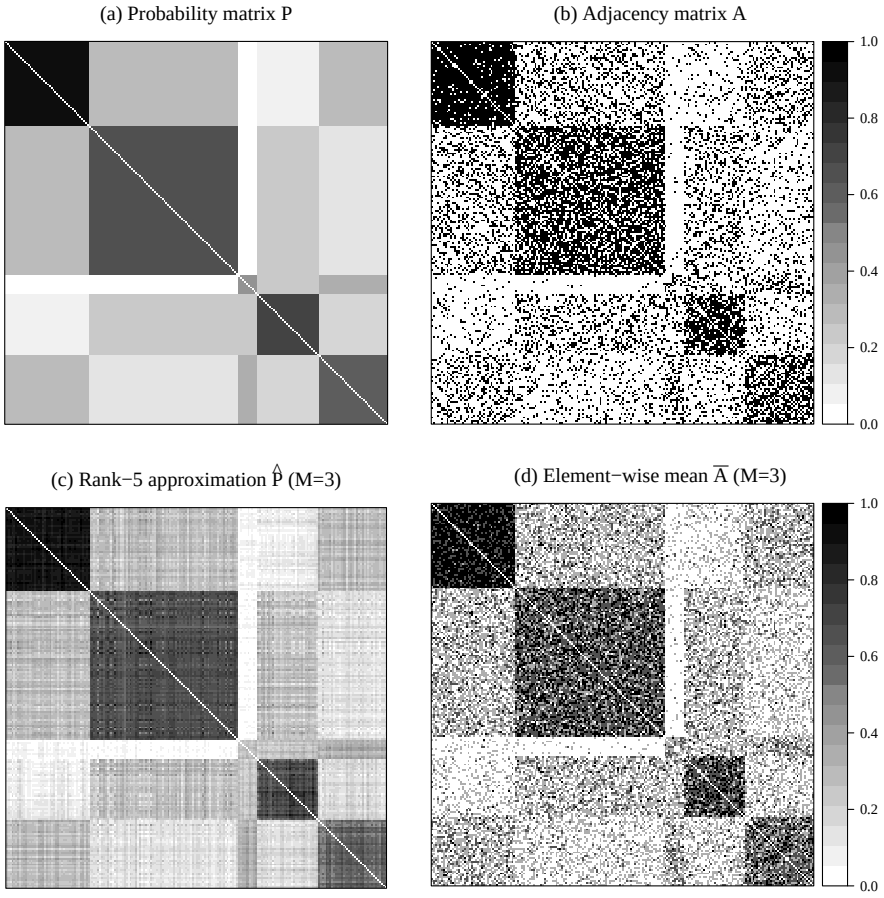


Fig 3: **Example illustrating the stochastic blockmodel.** Appendix D gives detailed parameters. Panel (a) shows the mean graph P with $K = 5$ blocks and $N = 200$ vertices. Panel (b) shows an adjacency matrix A sampled according to the probabilities from P. While A is a noisy version of P, much of the structure of P is preserved in A, a property exploited in our estimation procedure. Panel (d) shows the element-wise mean $\bar{A}$ based on three graphs sampled independently and identically according to the probability matrix P (see Section 3.2). Finally, Panel (c) shows our proposed estimate $\hat{P}$, a rank-5 approximation of $\bar{A}$, thresholding the values to be between 0 and 1 (see Section 3.3). Visual inspection shows that the low-rank estimate $\hat{P}$ more closely approximates the probability matrix P as compared to $\bar{A}$.

consistent for low-rank P.

Additionally, $\bar{A}$ does not exploit any underlying graph structure. If the graphs are distributed according to an RDPG or SBM, then $\bar{A}$ is no longer the maximum likelihood estimator since it is not guaranteed to satisfy the properties of the mean graph for that model. The performance can be especially poor when the sample size M is small, such as when $M \ll N$. For example, when $M = 1$, $\bar{A}$ is simply the binary adjacency matrix $A^{(1)}$, which is an inaccurate estimate for an arbitrary P compared to estimates which exploit underlying structure, such as the low-rank structure of the RDPG model.

3.3 Low-Rank Estimator $\hat{P}$

Motivated by the low-rank structure of the RDPG mean matrix, we propose the estimator $\hat{P}$ based on the spectral decomposition of $\bar{A}$ which yields a low rank approximation of $\bar{A}$. This estimator does not estimate the parameters of a stochastic blockmodel, as we wish to allow for the increased flexibility of the RDPG. However this can be implemented as a step towards estimating SBM parameters.

This estimator is similar to the estimator proposed by Chatterjee (2015) but incorporates additional adjustments which serve to improve the performance for the specific task of estimating the mean graph. Additionally, we consider an alternative dimension selection technique. Details of the dimension selection procedures and the diagonal augmentation are in Section A.2 and A.3, respectively. To summarize, the overall strategy to compute $\hat{P}$ is described in Algorithm 1. A key component of this algorithm is the low-rank approximation (see Algorithm 2 for details).

The first step is to calculate the sample mean $\bar{A}$. In Step 2 to Step 4, the algorithm augments the diagonal of $\bar{A}$ based on Marchette et al. (2011), selects the dimension $\hat{d}$ to embed (see Section A.2), and computes the low-rank approximation $\tilde{P}^{(0)}$ based on the embedding. Then in Step 5 and Step 6, the algorithm augments the diagonal again based on (Scheinerman and Tucker, 2010) (see Section A.3) which yields an improved low-rank estimate $\tilde{P}^{(1)}$. Finally, Step 7 thresholds the matrix entries to ensure all elements are between 0 and 1.

For a given dimension d we consider the estimator $\text{lowrank}_d(\bar{A})$ defined as the best rank-d positive-semidefinite approximation of $\bar{A}$. Let $\hat{S}$ be a diagonal matrix with non-increasing entries along the diagonal corresponding to the largest d eigenvalues of $\bar{A}$ and let $\hat{U}$ have columns given by the corresponding eigenvectors. Similarly, let $\tilde{S}$ be the diagonal matrix with non-increasing entries along the diagonal corresponding to the rest $N-d$ eigenvalues of $\bar{A}$ and let $\tilde{U}$ have columns given by the corresponding eigenvectors. Since the graphs are symmetric, the eigen-decomposition can be computed as $\bar{A}$ as $\hat{U}\hat{S}\hat{U}^\top + \tilde{U}\tilde{S}\tilde{U}^\top = [\hat{U}|\tilde{U}](\hat{S} \oplus \tilde{S})[\hat{U}|\tilde{U}]^T$. The d-dimensional adjacency spectral embedding (ASE) of $\bar{A}$ is given by $\hat{X} = \hat{U}\hat{S}^{1/2} \in \mathbb{R}^{N \times d}$. For an RDPG, the rows of $\hat{X}$ are estimates of the latent vectors for each vertex (Sussman et al., 2014). Using the adjacency spectral embedding, the low-rank approximation is $\bar{A}$ to be

Algorithm 1 Algorithm to compute $\hat{P}$

Input: Adjacency matrices $A^{(1)}, A^{(2)}, \cdots, A^{(M)}$, with each $A^{(m)} \in \{0,1\}^{N \times N}$
Output: Estimate $\hat{P} \in [0,1]^{N \times N}$

1: $\bar{A} \leftarrow \left(\sum_{m=1}^{M} A^{(m)} \right) / M$;
2: $D^{(0)} \leftarrow \text{diag}(\bar{A}\mathbf{1})/(N-1)$;
3: $\hat{d} \leftarrow \text{dimselect}(\bar{A} + D^{(0)})$; (see Section A.2)
4: $\tilde{P}^{(0)} \leftarrow \text{lowrank}_{\hat{d}}(\bar{A} + D^{(0)})$; (see Algorithm 2)
5: $D^{(1)} \leftarrow \text{diag}(\tilde{P}^{(0)})$;
6: $\tilde{P}^{(1)} \leftarrow \text{lowrank}_{\hat{d}}(\bar{A} + D^{(1)})$; (see Algorithm 2)
7: $\hat{P} \leftarrow \min(\max(\tilde{P}^{(1)}, 0), 1)$.

$\hat{X}\hat{X}^\top = \hat{U}\hat{S}\hat{U}^\top$. Algorithm 2 in the appendix gives the steps to compute this low-rank approximation for a general symmetric matrix A.

To compute the estimator $\hat{P}$, the rank d must be specified; there are various ways of dealing with dimension selection. In this paper, we explore an elbow selection method proposed in Zhu and Ghodsi (2006) and the universal singular value thresholding (USVT) method (Chatterjee, 2015). Section A.2 discusses details of these methods.

Moreover, since the adjacency matrices are hollow, with zeros along the diagonal, there is a missing data problem that leads to inaccuracies if $\hat{P}$ is computed based only on $\bar{A}$. To compensate for this issue, we use an iterative method developed in Scheinerman and Tucker (2010). Section A.3 discusses details of the iterative method.

Algorithm 1 gives the steps involved to compute the low-rank estimate $\hat{P}$. The bottom panels of Fig. 3 demonstrate the two estimators $\hat{P}$ and $\bar{A}$ for the stochastic blockmodel given by Panel (a). The estimates are based on a sample of size $M = 3$ and in this instance visual inspection demonstrates that $\hat{P}$ performs much better than $\bar{A}$. As in the succeeding sections, this procedure will frequently yield improvements in estimation as compared to using the sample mean $\bar{A}$. While this is not surprising for random dot product graphs, where we are able to show theoretical results to this effect, we also see this effect for connectome data and more general independent edge graphs. In the following sections, this estimator is explored in the context of the stochastic blockmodel.

4 Statistical Efficiency under Low Rank Models

4.1 Theoretical Results

To estimate the mean of a collection of graphs, we consider the two estimators from Section 3: the entry-wise sample mean $\bar{A}$ and the low-rank $\hat{P}$ motivated by the RDPG. We first evaluated the mean squared error for our estimators, $\text{MSE}(\hat{P}_{ij}) = \mathbb{E}[\hat{P}_{ij} - P]^2$ and $\text{MSE}(\bar{A}) = \mathbb{E}[\bar{A}_{ij} - P]^2$. While one can directly

compare the difference in mean squared errors between the two estimators, it is frequently useful to consider the relative efficiency between two estimators. In our case, this is $\mathrm{RE}(\bar{A}_{ij}, \hat{P}_{ij}) = \frac{\mathrm{MSE}(\hat{P}_{ij})}{\mathrm{MSE}(\bar{A}_{ij})}$, with values above 1 indicating $\bar{A}$ should be preferred while values below 1 indicate $\hat{P}$ should be preferred. Relative efficiency is a useful metric for comparing estimators because it will frequently be invariant to the scale of the noise in the problem and hence is more easily comparable across different settings.

In this section, entry-wise relative efficiency is computed to analyze the performance of these two estimators under the SBM. The asymptotic relative efficiency is also evaluated. The asymptotic relative efficiency is the limit of the relative efficiency as the number of vertices $N \to \infty$ but with the number of graphs M fixed, and the scaled relative efficiency, $N \cdot \mathrm{RE}(\bar{A}_{ij}, \hat{P}_{ij})$ which normalizes the relative efficiency so that the asymptotic scaled relative efficiency is non-zero and finite. Somewhat surprisingly, the data show that the asymptotic relative efficiency will not depend on this fixed sample size M.

For this asymptotic framework, we assume the block memberships τ_i are drawn iid from a categorical distribution with block membership probabilities given by $\rho \in [0,1]^K$. In particular, this implies that for each block k, the proportion of vertices in block k, $|\{i : \tau_i = k\}|/N$, will converge to ρ_k as $N \to \infty$ by the law of large numbers. We will also assume that for a given N, the block membership probabilities are fixed for all graphs. Denote the block probability matrix as $B = \nu\nu^\top \in [0,1]^{K \times K}$. By definition, the mean of the collection of graphs generated from this SBM is $P \in [0,1]^{N \times N}$, where $P_{ij} = B_{\tau_i, \tau_j}$. After observing M graphs on N vertices $A^{(1)}, \cdots, A^{(M)}$ sampled independently from the SBM conditioned on τ, the two estimators can be calculated, $\bar{A}$ and $\hat{P}$.

Lemma 4.1. *For the above setting, for any $i \neq j$, if $\mathrm{rank}(B) = K = d$; for large enough N,*

$$\mathbb{E}[(\hat{P}_{ij} - P_{ij})^2] \approx \frac{1/\rho_{\tau_i} + 1/\rho_{\tau_j}}{MN} P_{ij}(1 - P_{ij}),$$

and

$$\lim_{N \to \infty} N \cdot \mathrm{Var}(\hat{P}_{ij}) = \frac{1/\rho_{\tau_i} + 1/\rho_{\tau_j}}{M} P_{ij}(1 - P_{ij}).$$

The first part of this lemma ensures that the estimator is asymptotically unbiased for P, and the second part gives the form of the asymptotic variance of $\hat{P}$. Note that ρ_{τ_i} represents the probability that a vertex is assigned to the same block as vertex i, i.e. τ_i-th block. The proof of this lemma is outlined in Section A.5 and is based on results for the variance of the adjacency spectral embedding from Athreya et al. (2016). Lemma 4.1 shows that the MSE of $\hat{P}_{ij}$ has order $O(M^{-1}N^{-1})$ in terms of M and N, regardless of the SBM parameters. Similar to $\bar{A}$, the estimate will get better as the number of observations M increases. However, it also benefits from more vertices since the number of parameters for a low-rank P grows slowly compared to the number of potential edges as the number of vertices increases. That is, $\hat{P}$ will perform better as the number of vertices N increases.

Moreover, since $\bar{A}_{ij}$ is the sample mean of M independent Bernoulli random variables with parameter P_{ij}, then

$$\mathbb{E}[(\bar{A}_{ij} - P_{ij})^2] = \frac{P_{ij}(1 - P_{ij})}{M}.$$

Based on this MSE result of $\bar{A}_{ij}$ and the MSE result of $\hat{P}_{ij}$ given by Lemma 4.1, the next theorem follows.

Theorem 4.2. *In the same setting as in Lemma 4.1, for any $i \neq j$, if* $\operatorname{rank}(B) = K = d$*, then for large enough N:*

$$\operatorname{RE}(\bar{A}_{ij}, \hat{P}_{ij}) \approx \frac{1/\rho_{\tau_i} + 1/\rho_{\tau_j}}{N}. \tag{1}$$

And the asymptotic relative efficiency (ARE) is

$$\operatorname{ARE}(\bar{A}_{ij}, \hat{P}_{ij}) = \lim_{N \to \infty} \operatorname{RE}(\bar{A}_{ij}, \hat{P}_{ij}) = 0.$$

Proof. Combining the MSE result of $\bar{A}_{ij}$

$$\mathbb{E}[(\bar{A}_{ij} - P_{ij})^2] = \frac{P_{ij}(1 - P_{ij})}{M},$$

and Lemma 4.1, i.e. for large enough N,

$$\mathbb{E}[(\hat{P}_{ij} - P_{ij})^2] \approx \frac{1/\rho_{\tau_i} + 1/\rho_{\tau_j}}{MN} P_{ij}(1 - P_{ij}),$$

and therefore there is a large enough N,

$$\operatorname{RE}(\bar{A}_{ij}, \hat{P}_{ij}) = \frac{\mathbb{E}[(\hat{P}_{ij} - P_{ij})^2]}{\mathbb{E}[(\bar{A}_{ij} - P_{ij})^2]} \approx \frac{1/\rho_{\tau_i} + 1/\rho_{\tau_j}}{N}.$$

And the ARE result follows directly by taking the limit of RE as $N \to \infty$. □

This theorem indicates that under the SBM, $\hat{P}$ is a much better estimate of the mean of the collection of graphs P than $\bar{A}$, especially when N is large. Note that a relative efficiency less than 1 indicates that $\hat{P}$ should be preferred over $\bar{A}$, so under the above assumptions, as $N \to \infty$, $\hat{P}$ performs far better than $\bar{A}$. Note that even though the RE could be greater than 1 for some N, eventually the RE will go to 0 as N increases. The result shows that the relative efficiency is of order $O(N^{-1})$ and $N \cdot \operatorname{RE}(\bar{A}_{ij}, \hat{P}_{ij})$ (which is denoted as scaled RE) converges to $1/\rho_{\tau_i} + 1/\rho_{\tau_j}$ when $N \to \infty$. An important aspect of Theorem 4.2 is that the ARE does not depend on the number of graphs M, so the larger the graphs are, the better $\hat{P}$ is relative to $\bar{A}$, regardless of M.

Note that the asymptotic result here is for a number of vertices going to infinity with a fixed number of graphs. This setting will be very useful in connectomics analysis, especially as the collection of larger and larger brain

networks grow from initially small sample sizes as the technology to scale these connectome collection techniques is developed. For example, Calabrese et al. (2015) recently reported a high resolution magnetic resonance microscopy based estimate of the mouse brain using a single mouse.

The approximate formula Eq. 1 indicates that the sizes of the blocks can greatly impact the relative efficiency. As an example, consider a 2-block SBM. If each of the blocks contain half the vertices, then for each pair of vertices, the relative efficiency is approximately $4/N$. If the first block gets larger, with $\rho_1 \to 1$, then the RE for estimating P_{ij} with $\tau_i = \tau_j = 1$ will tend to its minimum of $2/N$. On the other hand as $\rho_1 \to 1$, if $\tau_i = 1$ and $\tau_j = 2$, then, since $\rho_2 = 1 - \rho_1$, the relative efficiency for estimating such an edge pair will be approximately 1 and the same will hold if $\tau_i = \tau_j = 2$. Note that the maximum value for the relative efficiency of two vertices from different blocks in a two-block model is achieved when $\rho_1 = 1/N$ and $\rho_2 = (N-1)/N$ in which case the relative efficiency is $N/(N-1) \approx 1$. (Note values of ρ_s below $1/N$ correspond to graphs where no vertices are typically in that block, so the effective minimum that can be considered for ρ_s is $1/N$.)

Fig. 4 shows curves for $2/\rho_1$ and $1/\rho_1 + 1/\rho_2$ i, the scaled asymptotic RE for pairs of vertices both in block one and pairs of vertices in different blocks, respectively, in terms of ρ_1. We vary ρ_1 between 0 and 1 to demonstrate how the number of pairs of vertices with the corresponding block memberships impacts the overall relative efficiency. Note, $N \cdot \mathrm{RE}(\bar{A}_{ij}, \hat{P}_{ij})$ achieves for i and j from different blocks when $\rho_k = 1/K$ for all k, which corresponds to $\rho_1 = \rho_2 = 1/2$ in the 2-block case. When the vertices are from the same block the scaled relative efficiency

If instead of assuming that the graphs follow an SBM distribution, we assume that the graphs are distributed according to an RDPG distribution, similar gains in relative efficiency can be realized. While there is no compact analytical formula for the relative efficiency of $\hat{P}$ versus $\bar{A}$ in the general RDPG case, using the same ideas as in Theorem 4.2, we can show that $\mathrm{RE}(\bar{A}_{ij}, \hat{P}_{ij}) = O(1/N)$.

Proposition 4.3. *Suppose that $A^{(1)}, A^{(2)}, \ldots, A^{(M)}$ are independently and identically distributed from an RDPG distribution with common latent positions $X_1, \ldots, X_n$, which are independently and identically distributed from some distribution. As the number of vertices $N \to \infty$, it holds for any $i \neq j$ that $\mathrm{RE}(\bar{A}_{ij}, \hat{P}_{ij}) = O(1/N)$, where again the asymptotic relative efficiency in N does not depend on M.*

The proof of this proposition closely follows the proofs of Lemma 4.1 and Theorem 4.2, and hence it is omitted here.

Remark 4.4. *As noted above, if the graphs are distributed according to an SBM or an RDPG, the relative efficiency is approximately invariant to the number of graphs M when N is large. If on the other hand, the graphs are generated according to a full-rank independent edge model, then the relative efficiency can change more dramatically as M changes. The reason for this is because for larger M, more of the eigenvectors of $\bar{A}$ will begin to concentrate around the*

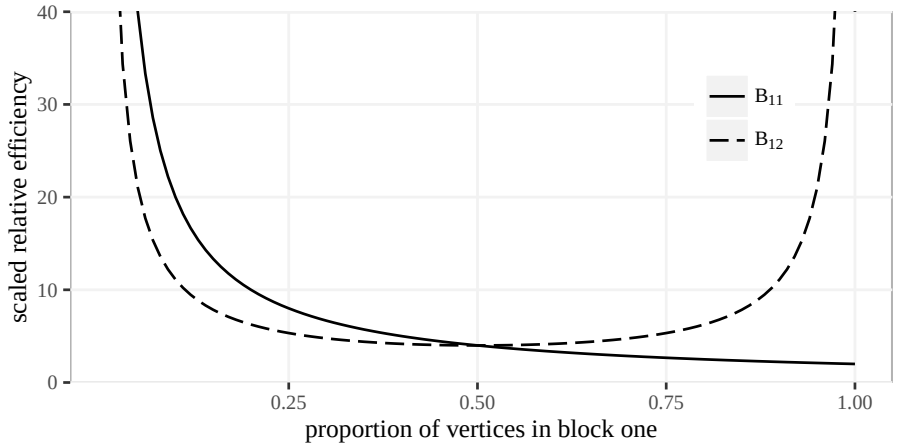


Fig 4: **Asymptotic scaled relative efficiency $N \cdot \mathrm{RE}(\bar{A}, \hat{P})$ in a 2-block SBM.** For each two distinct pairs of edge probabilities in a 2-block SBM, the scaled relative efficiency only depends on the proportion of vertices in each block. The scaled asymptotic relative efficiency is shown as ρ_1 changes from $0, 1$ for pairs of vertices where either both are in block one, or one is in block one and one is in block two. These curves all intersect at a scaled relative efficiency of 4 when $\rho_1 = 1/2 = \rho_2$. Improvements using low-rank methods are greater for larger blocks, such as for B_{11} when ρ_1 is close to 1, while the improvements are smaller for block pairs with relatively few vertex pairs such as B_{11} when ρ_1 is small and B_{12} when ρ_1 is near 0 or 1 . Note that the curve for B_{22} would be the same as that for B_{11} but reflected around the vertical line when $\rho_1 = 1/2$. Overall, $\hat{P}$ performs best for large blocks while the improvements may be very minor for blocks with only a few vertices.

eigenvectors of the mean graph. This leads to the fact that the optimal embedding dimension for estimating the mean will increase, making $\bar{A}$ and the low-rank approximation at the optimal dimension closer together. As a result, $\mathrm{RE}(\bar{A}, \hat{P})$ *will increase as M increases for full-rank models. Indeed, for large M it is possible that* $\mathrm{RE}(\bar{A}, \hat{P}) \geq 1$ *since it is not guaranteeed that $\hat{P}$ will choose the optimal dimension. The lack of gaps in the eigenvalues of the mean graph makes dimension reduction quite dangerous. In an extreme case, the low-rank assumption will be most violated when all eigenvalues of the mean graph are almost equal. This leads to a certain type of structure, which is close to a constant times the identity matrix. However, such structure is not seen in connectomics. This will be discussed further in Section 4.2 when applying the estimator to the SWU4 dataset.*

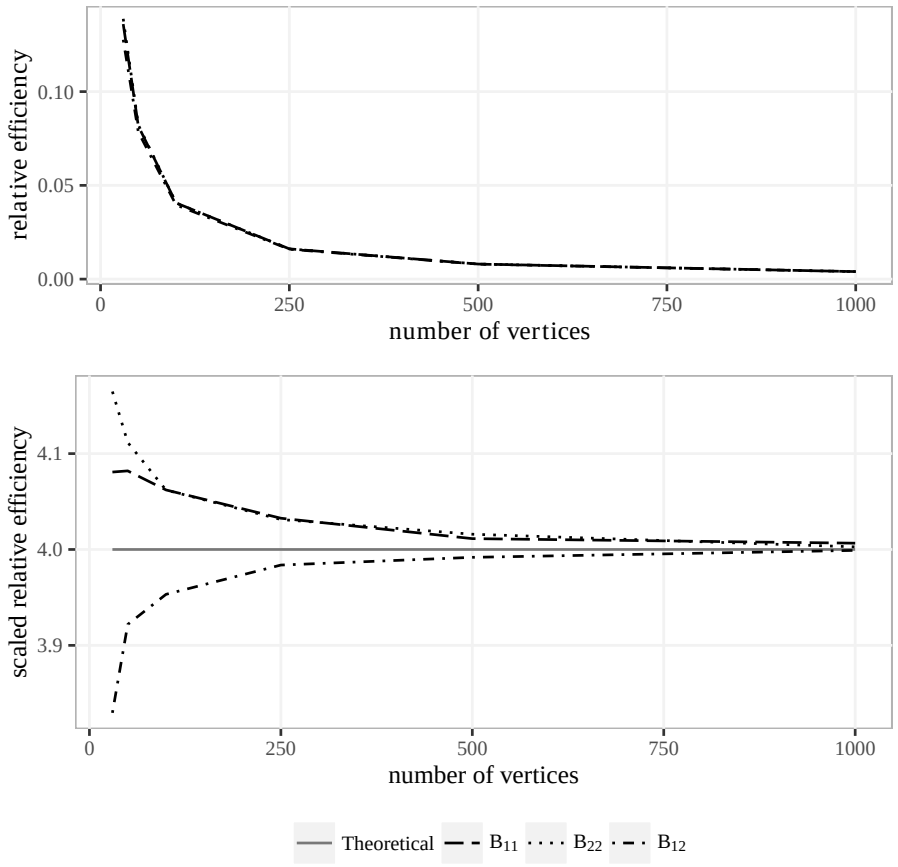


Fig 5: **Finite sample relative efficiency based on simulations.** The top panel shows the estimated relative efficiency $\hat{\text{RE}}(\bar{A}, \hat{P})$ as a function of N for fixed $M = 100$ based on simulations of an SBM. For each value of N, 1000 Monte Carlo replicates of the SBM from Section 4.2 estimated the RE. Each curve corresponds to an average across vertex pairs corresponding to the three distinct block probabilities B_{11}, B_{12}, and B_{22} in the two-block SBM. Recall that values below 1 indicate that $\hat{P}$ is performing better than $\bar{A}$. The relative efficiencies are all very close so the lines are indistinguishable. To distinguish the three curves, the bottom panel shows the corresponding scaled relative efficiencies, $N \cdot \hat{\text{RE}}(\bar{A}, \hat{P})$. The solid horizontal line indicates the theoretical asymptotic scaled relative, which is $1/\rho_s + 1/\rho_t = 4$, since $\rho_1 = \rho_2 = 4$. All the curves converge quickly to this theoretical limit.

4.2 Finite Sample Simulations

In this section, the theoretical results from Section 4.1 regarding the relative efficiency between $\bar{A}$ and $\hat{P}$ *via* Monte Carlo simulation experiments in an idealized setting will be illustrated. These numerical simulations will also allow us to investigate the finite sample performance of the two estimators. Note that in Section 4.6, the model assumptions will be broken slightly to run experiment in a more realistic setting.

Here, we consider the following 2-block SBM with parameters

$$B = \begin{bmatrix} 0.42 & 0.2 \\ 0.2 & 0.7 \end{bmatrix}, \qquad \rho = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}.$$

When calculating $\hat{P}$, the dimension selection step from Algorithm 1 is omitted and replaced with the true dimension $d = \text{rank}(B) = 2$. Note that for large N, many dimension selection methods will often correctly select the true dimension (Chatterjee, 2015; Fishkind et al., 2012).

$M = 100$ graphs from 1000 Monte Carlo replicates using the above SBM distribution with different numbers of vertices $N \in \{30, 50, 100, 250, 500, 1000\}$ show the relative efficiency $\text{RE}(\bar{A}_{ij}, \hat{P}_{ij})$ can be estimated since P is known for this simulation. Since the relative efficiency only depends on the block memberships of the pair i, j, $D_{st} = \{(i,j) : \tau_i = s, \tau_j = t, 1 \leq i < j \leq n\}$ as defined as the set of ordered pairs of vertices which belong to s-th block and t-th block respectively. The relative efficiency for each block pair can be estimated using

$$\hat{\text{RE}}_{st}(\bar{A}, \hat{P}) = \frac{\sum_{(i,j)\in D_{st}} \hat{\text{MSE}}(\hat{P}_{ij})}{\sum_{(i,j)\in D_{st}} \hat{\text{MSE}}(\bar{A}_{ij})}$$

for $s, t \in \{1, 2\}$, where $\hat{\text{MSE}}$ denotes the estimated mean squared error based on the Monte Carlo replicates. For the remaining simulations and real data analysis, we will always be considering estimated relative efficiency and estimated mean squared error rather than analytic results, and hence we will frequently omit that these are estimated values.

In Fig. 5, we plot the (estimated) relative efficiency (top panel) and the scaled (estimated) relative efficiency (bottom panel), $N \cdot \hat{\text{RE}}_{st}(\bar{A}, \hat{P})$. The different dashed lines denote the RE and scaled RE associated with different block pairs, either B_{11}, B_{12}, or B_{22}. As expected from Theorem 4.2, the top panel indicates that the relative efficiencies are all very close together and much less than 1, decreasing at the rate of $1/N$, indicating that $\hat{P}$ is performing better than $\bar{A}$.

Based on Theorem 4.2, the scaled RE converges to $1/\rho_{\tau_i} + 1/\rho_{\tau_j} = 4$ as $N \to \infty$ for all pairs i, j. This is plotted as a solid line in the bottom panel. The figure shows that $N \cdot \hat{\text{RE}}_{st}(\bar{A}, \hat{P})$ converges to scaled asymptotic RE quite rapidly. Error bars were omitted, as the standard errors are very small for these estimates.

Remark 4.5. *An intriguing aspect of these finite sample results is that the scaled relative efficiencies behave differently for small graphs with fewer vertices.*

The estimates of the edge probabilities for pairs of vertices in different blocks are much better than the estimates for edges within each block. The reason for this is unclear and could be due to the actual values of the true probability, but it may also be due to the fact that there are approximately twice as many pairs of vertices in different blocks, $N^2/4$, than there are in the same block, $N^2/8 - N/4$. This could lead to an increase in effective sample size which may cause the larger differences displayed in the left parts of Fig. 5. However, these differences are nearly indistinguishable for unscaled relative efficiency overall.

Low-Rank Estimator for Human Connectomes In practice, observed graphs do not follow the independent edge model, let alone an RDPG or SBM, but the mean of a collection of graphs is still of interest for these cases. To demonstrate that the estimator $\hat{P}$ is still useful in such cases, its performance on structural connectomic data was tested. The graphs are based on diffusion tensor MR images of SWU4 dataset collected and available at the Consortium for Reliability and Reproducibility (CoRR) (Zuo et al., 2014) (see Section A.4.1). The dataset contains 454 different brain scans, each of which was processed to yield an undirected, unweighted graph with no self-loops, using the pipeline described in (Kiar et al., 2017, 2016). The vertices of the graphs represent different regions in the brain defined according to an atlas. Here, three atlases were used: the JHU atlas with 48 vertices (Oishi et al., 2010), the Desikan atlas with 70 vertices (Desikan et al., 2006), and the CPAC200 atlas with 200 vertices (Sikka et al., 2014). An edge exists between two vertices whenever there is at least one white-matter tract connecting the corresponding two regions of the brain. Further details of the dataset are provided in Section A.4.

4.3 Estimating Mean Connectomes from Diffusion MRI Data

A cross validation on the 454 graphs of each size evaluated the performance of the two estimators. Specifically, for a given atlas, each Monte Carlo replicate corresponds to sampling M graphs out of the 454 and computing the low-rank estimator $\hat{P}$ and the sample mean $\bar{A}$ using the M selected graphs. These estimates were compared to the sample mean for the remaining $454 - M$ adjacency matrices. While we cannot interpret this mean graph as the probability matrix for an IEM distribution (see Section 4.6), the sample mean for the remaining graphs does give the proportion of times each pair of vertices are adjacent in the population from the sampled graphs.

While in previous sections the mean squared error for either an individual entry or for an entire block in the SBM were evaluated, in this section and the next section we will focus on the overall error for estimating the mean graph. In particular, the average of the mean squared error across all pairs of vertices, in conjunction with $\text{MSE}(\bar{A})$ and $\text{MSE}(\hat{P})$ will determine the relative efficiency,

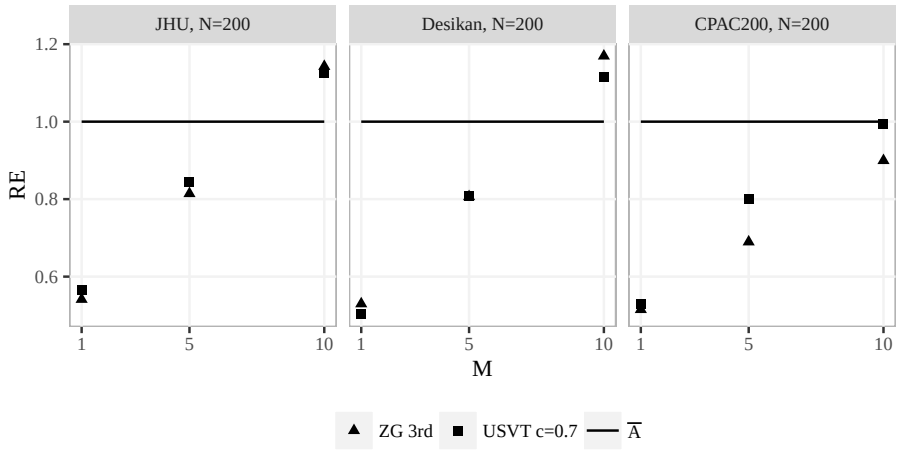


Fig 6: **Relative efficiencies of $\bar{A}$ versus $\hat{P}$ for the SWU4 data set.** A sampling of graphs from each atlas: JHU, Desikan, and CPAC 200, had $\bar{A}$ and $\hat{P}$ computed using different sample sizes M and different dimension selection procedures, ZG and USVT. For each of the two methods for computing $\hat{P}$, relative efficiencies were estimated with respect to the sample mean $\bar{A}$. Confidence intervals all had lengths less than 0.015, and hence they were omitted for clarity. The largest improvements using $\hat{P}$ occur when M is small and N is large, where the RE are smaller than 1. On the other hand, once $M = 10$, $\bar{A}$ tends to do nearly as well or better than $\hat{P}$. Overall, the relative efficiencies are greater for smaller sample sizes M and larger number of vertices N.

where $\text{MSE}(\bar{A})$ is defined as

$$\text{MSE}(\bar{A}) = \binom{N}{2}^{-1} \sum_{i<j} \mathbb{E}[(\bar{A}_{ij} - P_{ij})^2]$$

and $\text{MSE}(\hat{P})$ is similarly defined. As in the previous section, analytical evaluations of the MSE will not be evaluated, and instead estimates of the MSE and relative efficiencies *via* from Monte Carlo simulations will be employed.

1000 cross-validation simulations on each of the three atlases were run for sample sizes of $M = 1, 5, 10$. For $M = 1$, there are only 454 distinct samples, so $\bar{A}$ and $\hat{P}$ is compared for each of these 454 possible samples of size 1. As long as we determine which dimension to embed, the two estimates $\bar{A}$ and $\hat{P}$ can be calculated based on the sample. In practice, we employed algorithms like Zhu and Ghodsi's method or USVT to select the dimension d. These methods are relatively easy to compute. Section A.2 discusses details of the methods. Fig. 6 shows plots of the estimated relative efficiencies between $\bar{A}$ and $\hat{P}$.

For each atlas and each sample size, the Zhu and Ghodsi method (Zhu and

Ghodsi, 2006) was compared with the USVT method (Chatterjee, 2015). Both have similar overall performance. Confidence intervals were omitted (calculated by assuming a normal distribution) for the estimated relative efficiencies since all confidence intervals have lengths less than 0.015. This confidence interval indicates that all relative efficiencies are significantly different from 1, aside from the relative efficiency for the CPAC200 atlas at $M = 10$.

Again, the largest improvements using $\hat{P}$ occur when M is small and N is large, where the RE are smaller than 1. On the other hand, once $M = 10$, $\bar{A}$ tends to do nearly as well or better than $\hat{P}$. In addition, $\hat{P}$ offers certain advantages, especially since low-rank estimates can often be more easily interpretable by considering the latent position representation described in Section 2.3.

As discussed in Section 1, Fig. 1 further illustrates the differences between the two estimators for a samples with size $M = 5$ and $M = 1$. For the sample with $M = 5$, $\hat{P}$ was calculated using Zhu and Ghodsi's 3rd elbow to select $d = 11$. As discussed above, $\hat{P}$ has a finer gradient of estimated probabilities and has fewer outliers than $\bar{A}$, as shown in the upper triangular part of panels (b) and (e). Moreover, for the sample with size $M = 5$ discussed above, Fig. 7 shows the values for the absolute estimation error $|\bar{A} - P|$ and $|\hat{P} - P|$. In addition, the absolute difference is included $|\bar{A} - \hat{P}|$ to show the overall difference between the two estimates. The lower triangular sections show the actual absolute difference while the upper triangular matrix highlights the vertex pairs with absolute differences larger than 0.4. There are 18 edges for $\bar{A}$ and only 6 edges for $\hat{P}$ being highlighted in the figure. Note that approximately 13% of all pairs of vertices are adjacent in all 454 graphs and hence $\bar{A}$ will always have zero error for those pairs of vertices. Nonetheless, $\hat{P}$ typically outperforms $\bar{A}$.

To investigate the difference in performance with respect to the geometry of the brain, in Fig. 8 we plot the 50 edges with the largest differences $|\bar{A}_{ij} - P_{ij}| - |\hat{P}_{ij} - P_{ij}|$ according to the location of the corresponding regions in the brain. Red edges indicate that $\hat{P}$ overestimates P, while blue means that $\hat{P}$ underestimates P. The edge width is determined by the estimation error for $\hat{P}$, where pairs with larger estimation error are represented by thicker lines. The five regions corresponding to vertices that contribute most to the difference are also highlighted, which are the vertices i with the largest value of $\sum_j (|\bar{A}_{ij} - P_{ij}| - |\hat{P}_{ij} - P_{ij}|)$.

Notably, three of these top five regions form a contiguous group of regions. The top five regions are the inferior temporal, middle temporal, and transverse temporal regions in the left hemisphere and the parahippocampal and pars opercularis regions in the right hemisphere of the Desikan atlas. Furthermore, the regions that differ most appear to be predominantly localized to the temporal lobe, where increased regionally-specific noise in the diffusion weighted imagining data may increase the estimation error.

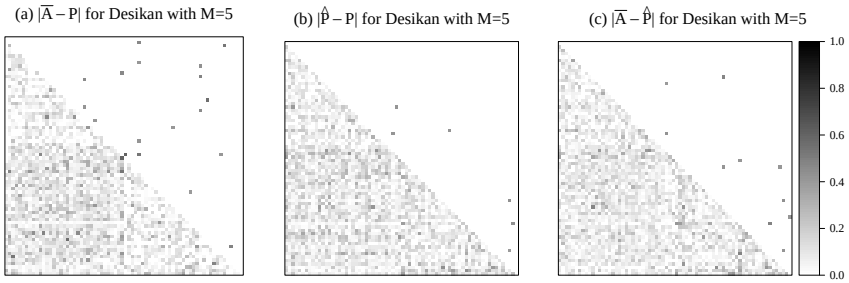


Fig 7: **Heat plots of absolute estimation error for $\bar{A}$ and $\hat{P}$ (lower triangle) and absolute errors above 0.4 (upper triangle).** These heat plots show the absolute estimation error $|\bar{A} - P|$, $|\hat{P} - P|$ and $|\bar{A} - \hat{P}|$ for a sample of size $M = 5$ from the Desikan dataset. The embedding dimension for $\hat{P}$ is $d = 11$ selected by the 3rd elbow of the ZG method. The lower triangular matrix shows the actual absolute difference, while the upper triangular matrix only highlights the edges with absolute differences larger than 0.4. The fact that 18 edges from $\bar{A}$ are highlighted and only 6 edges from $\hat{P}$ are highlighted indicates that $\hat{P}$ has fewer large outliers compared to $\bar{A}$.

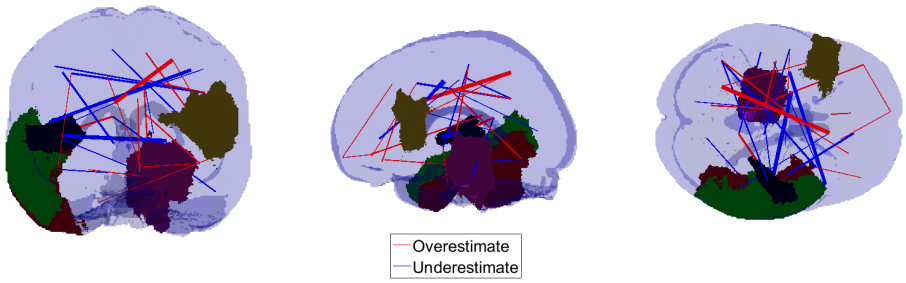


Fig 8: **Top 5 regions of the brain (vertices in graphs) and top 50 connections between regions (edges in graphs) with the largest differences $|\bar{\boldsymbol{A}}_{\boldsymbol{ij}} - \boldsymbol{P}_{\boldsymbol{ij}}| - |\hat{\boldsymbol{P}}_{\boldsymbol{ij}} - \boldsymbol{P}_{\boldsymbol{ij}}|$.** Red edges indicate that $\hat{P}$ overestimates P while blue means that $\hat{P}$ underestimates P. The edge width is determined by the estimation error. Connections with larger estimation error are represented by thicker lines. This figure shows the regions and connections of the brain where $\hat{P}$ outperforms $\bar{A}$ the most for estimating P.

4.4 Challenges of the SWU4 Dataset

While our estimator $\hat{P}$ performs well when the sample size M is small and the number of vertices N is large, the CoRR dataset itself does not strictly adhere to the low-rank assumptions of our theory. As discussed in Remark 4.4, whether the dataset has low-rank structure was first investigated. In Fig. 9, the relative error $\|\text{lowrank}_d(P) - P\|_F^2 / \|P\|_F^2$ of using a rank-d approximation of P (see Algorithm 2) is plotted as solid curves. The rate at which this curve tends to zero provides an indication of the relative performance of using $\hat{P}_d$ as compared to $\bar{A}$ when M is large. For all three atlases, while these error rates do tend to zero relatively quickly, substantial errors remain for any low-rank approximation. This can be compared to the dashed lines which show how these errors would behave if P was truly low-rank where the ranks are selected by Zhu and Ghodsi's method. In particular, the selected dimensions are 13 for JHU, 8 for Desikan, and 37 for CPAC200. As it turns out, the SWU4 dataset is not even approximately low-rank.

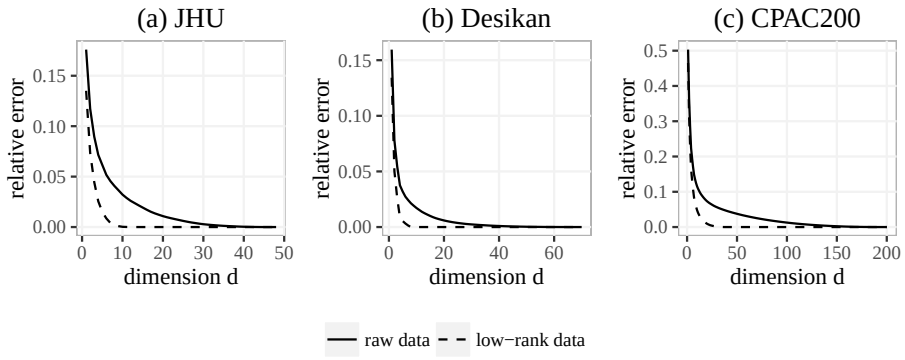


Fig 9: **Relative error of the rank-d approximation of the population mean.** The solid curves show the relative error $\|\text{lowrank}_d(P) - P\|_F^2 / \|P\|_F^2$ of using a rank-d approximation of P (see Algorithm 2) for three different atlases. The relative error decays relatively slowly when d is close to N, which indicates that P is not low-rank. Also, if P actually has the low-rank property, the relative error plot will look like the dashed curves, where we revised P to be low-rank by only keeping a few large eigenvalues.

Two other aspects of this dataset provide challenges for these low-rank methods. First, there are a large number of negative eigenvalues which $\hat{P}$ will not capture. We can adapt our low-rank methods by including large negative eigenvalues as well. However, for low sample sizes excluding negative eigenvalues improved performance. Second, approximately 12.8% of the entries of P are exactly equal to 1. For these edges, $\bar{A}$ will always have exactly zero error, while $\hat{P}$ will necessarily give a less accurate estimate.

Despite these challenges, our results show that when the sample size is rela-

tively small, such as $M = 1$ or $M = 5$, and for the atlases with a larger number of vertices, $\hat{P}$ gives a better estimate than $\bar{A}$ for the CoRR dataset. Importantly, this improvement is robust to the embedding dimension provided the dimension is not underestimated.

4.5 Lobe Structure behind the Low-rank Methods

In previous sections, we have shown how low-rank methods help us improve the accuracy of estimation while providing convenient interpretations simultaneously. In this section, the focus is on how the lobe structure is reflected in the low-rank estimate and the estimated latent positions in particular.

Classically, the brain can be divided into lobes, originally based purely on anatomical considerations, now widely recognized to also play a functional role Vanderah and Gould (2015). While different anatomists partition brain regions differently, there is general agreement on four cortical lobes per hemisphere: frontal, parietal, occipital, and temporal (Fischl, 2012; Desikan et al., 2006; Salat et al., 2004). For the Desikan atlas, there are 70 different regions (35 regions for each hemisphere), with each region belonging to a single lobe. However, three regions of the Desikan atlas per hemisphere (Banks of Superior Temporal Sulcus, Corpus Callosum, and the "Unknown" region) do not have obvious lobe assignment. These three regions were clustered into a new lobe category named "other" to resolve this issue.

In general, one might expect that properties of regions within a lobe are more similar to those across lobes, even upon conditioning on anatomical proximity, as regions within lobes would be expected to share more functional roles. To see whether the embedded latent positions X preserve this property or not, we propose a test statistic T to be the average differences between vertices within the same lobe minus the average differences between vertices across different lobes, i.e.

$$T(X, l) = \sum_{i \neq j} \frac{\mathbf{1}_{l(i)=l(j)} \|X_i - X_j\|_2}{\sum_{k \neq l} \mathbf{1}_{l(k)=l(l)}} - \frac{\mathbf{1}_{l(i)\neq l(j)} \|X_i - X_j\|_2}{\sum_{k \neq l} \mathbf{1}_{l(k)\neq l(l)}}, \tag{2}$$

where $l(i)$ represents the lobe assignment for vertex i. If the latent positions X and the lobe assignment l are independent, then $T(X, l)$ will be close to zero. A small test statistic $T(X, l)$ indicates that latent positions of the regions within the same lobe are closer compared to the ones across the lobes.

To test this hypothesis the lobe labels of each region were randomly permutated and the test statistics calculated $T(X, l')$ based on the permuted lobe labels l'. After performing 1000 Monte Carlo replicates of unconditional permutations, the p-value is less than 10^{-12}.

However, the anatomical geometry might contribute to the dependence between X and l, with spatially proximal vertices having similar connectivity patterns. Hence, a small test statistic $T(X, l)$ is evidence that the low-rank methods preserve the lobe structure only if we also condition on anatomy geometry: $H_0 : X$ and l are conditionally independent given anatomical geometry,

$H_A : X$ and l are conditionally dependent given anatomical geometry.

A test of these hypotheses will focus on how much of the lobe structure is really captured by the latent positions associated with the low-rank methods without being affected by the inherent spatial relationship. Note that the test of conditional independence, performed without any anatomical geometry constraints, has less power compared to the test of unconditional independence which is performed with a random permutation of lobe labels as discussed above.

To test under the anatomical geometry conditions, the lobe assignment l was permutated in a way such that the number of regions in each lobe remain the same and the regions within the same randomly defined lobe are still spatially connected. A flip was defined to be a swap of two pairs of vertices which preserves the number of regions in each lobe and also maintains the constraint that lobes are spatially contiguous. The lobes are flipped a limited number of times in order to study how the number of flips impacts the test statistic. Appendix C discusses the flipping procedure.

1000 simulations with the test statistics of $T(X, l')$ are shown for a fixed number of flips after the permutation. The number of flips was varied from 1 to 10 (Fig. 10). In the violin plot, the dashed line indicates the value of $T(X, l)$ based on the true lobe assignment. $T(X, l)$ for the true lobe assignment move away from the the 2.5%-quantile based for the random permutations. The figure shows corresponding p-values. When the number of flips is larger than 7, the p-value is less than 0.05, which suggests that latent positions in the same lobe are more similar to each other, even after accounting for the fact that geometrically proximal regions may also have similar latent positions. Note that when the number of flips is small, this test will have very little power since the null distribution is just a small deviation from the original lobe structure.

4.6 Synthetic Data Analysis for Full Rank IEM

While the theory we have developed is based on the assumption that the mean graph is low rank, as we have seen in Section 4.2, $\hat{P}$ can perform well even when this assumption is false. To further illuminate this point, a synthetic data analysis under a more realistic full-rank independent edge model was performed. As discussed in Section 4.4, the sample mean of the 454 graphs in the Desikan dataset is actually of full rank. For this simulation, we will use the sample mean as the probability matrix P. A sampling of independent graphs from the full rank IEM with the probability matrix P show that for the synthetic data sets of size $M = 1, 5$, and 10, $\hat{P}$ performs even better than $\bar{A}$ in the real data experiments. Fig. 11 shows the resulting estimated MSE for $\bar{A}$ (solid line) and $\hat{P}$ (dashed line), as a function of the embedding dimension for simulated data based on the full rank probability matrix P shown in the left panel of Fig. 1. These results are similar to those presented in Section 4.2, though overall $\hat{P}$ performs even better than in the real data experiments. When M is small, $\hat{P}$ outperforms $\bar{A}$ with a flexible range of embedding dimensions including those selected by the Zhu and Ghodsi method. On the other hand, when M is large enough, both estimators perform well with the decision between the two being

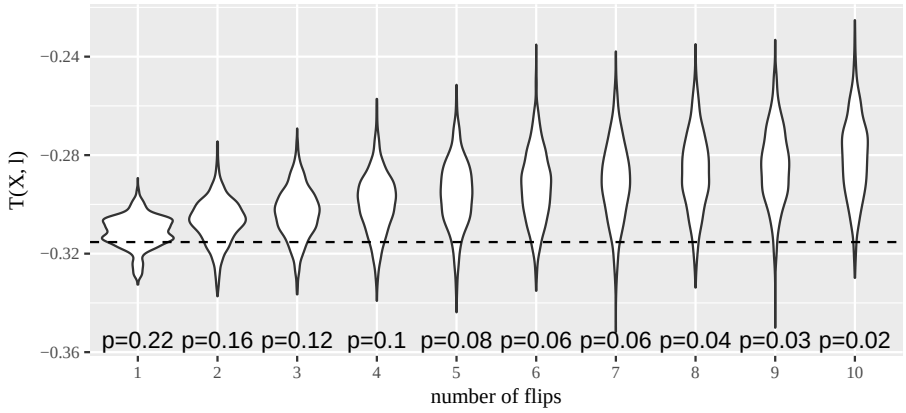


Fig 10: **Violin plot of the permutation test.** 1000 simulations were run for each number of flips. Dashed line represents the statistic $T(x, l)$ based on true lobe assignment. As the null hypotheses move further from the original lobe assignment, the 95% central region of the null distributions shifts away from $T(x, l)$ under the true lobe assignment.

less conclusive. This simulation again shows the robustness of $\hat{P}$ to deviations from the RDPG model, specifically if the probability matrix is full-rank.

We also note that the finite-sample relative efficiency for this synthetic data is even more favorable to $\hat{P}$ than for the real data, with relative efficiency lower than $1/3$ for $M = 1$ in the synthetic data analysis as compared to relative efficiency which were at best around $1/2$ for $M = 1$ in the original data. From this observation, we can postulate that the degradation in the performance of $\hat{P}$ in real data can at least partially be attributed to the fact that the independent edge assumption does not hold for real data. It also suggests that more elaborate models of connectomes will be valuable for various inferential tasks.

5 Application to a Mouse Connectome

As a further application of low-rank methods, an MRI-DTI mouse brain connectome (Calabrese et al., 2015) with N=1 specimen was evaluated. The data acquisition protocol is described in the appendix, Section A.4.2, and resulted in a 296 weighted, directed graphs with vertices again corresponding to regions in the brain. The 296 regions were organized into a multilevel, hierarchical structure. Analysis of the fine-grained and the first level of the hierarchy partitioned the label set into eight superstructures, with four in each hemisphere: forebrain, midbrain, hindbrain, and white matter.

The original matrix $W \in (\mathbb{R}^+)^{296\times 296}$ is a weighed adjacency matrix with W_{ij} denoting the number of tracts passing through ROIs i and j. As the original

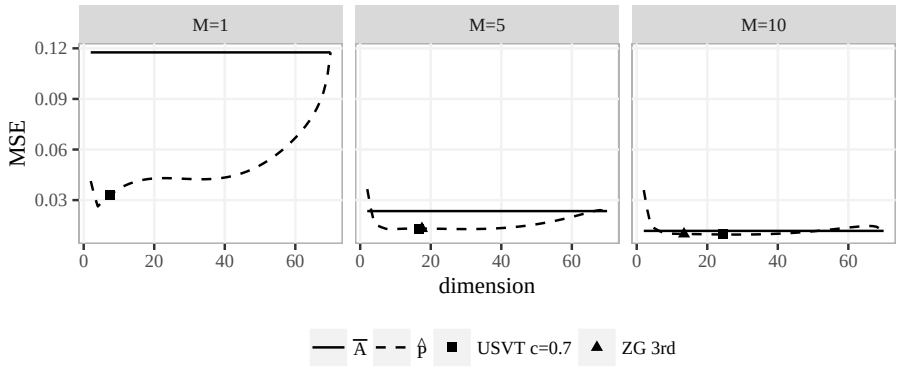


Fig 11: **Comparison of $\hat{P}$ and $\bar{A}$ for synthetic data analysis.** As in Fig. 5, this figure shows $\hat{\text{MSE}}$ for $\bar{A}$ (solid line) and $\hat{P}$ (dashed line) for simulated data with different sample sizes M based on the sample mean for the Desikan dataset. Again, the average of dimensions selected by the USVT method (square) and the ZG method (triangle) tend to nearly approximate the optimal dimension. Overall, the structure of these plots well approximates the structure for the real data indicating that performance for the independent edge model will tend to translate in structure to non-independent edge scenarios. On the other hand, the relative efficiency $\hat{\text{RE}}(\bar{A}, \hat{P})$ is lower for this synthetic data analysis than for the CoRR data.

weights had very heavy tails, these weights were transformed by settingA_{ij} was set to $A_{ij} = \log(W_{ij} + 1)$ to transform the original weights, which had very heavy tails. This resulted in the weighted adjacency matrix in Figure 12 (a).

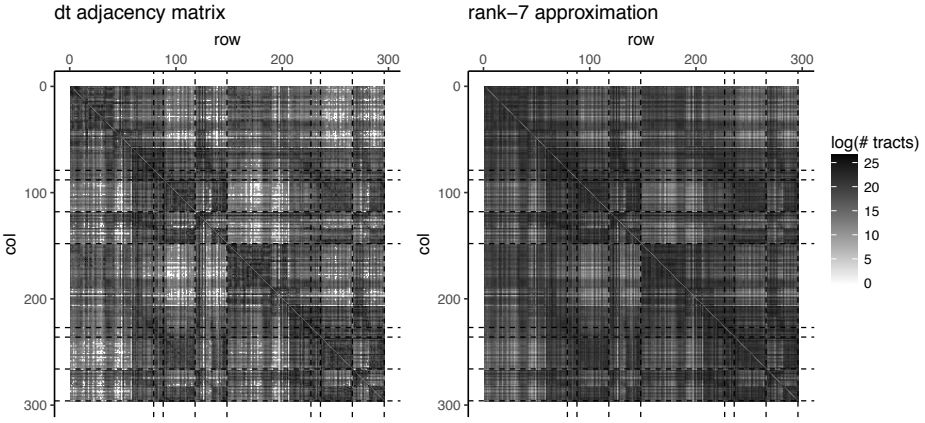


Fig 12: The left panel shows the original weighted adjacency with weights transformed by the transformation $w \mapsto \log(w+1)$. The right panel shows the rank-7 approximation of this matrix where the rank was chosen using the Zhu and Ghodsi (2006) method. In both panels, dashed lines show the boundaries between the eight different superstructures.

Using the procedure described in Algorithm 1, we found a low-rank approximation to the re-weighted matrix A. In this instance, the Zhu and Ghodsi (2006) procedure resulted in a rank-7 approximation which is shown in the right panel of Figure 12. Since the sample size is only one, cross-validation cannot be employed to analyze the prediction performance of the low-rank methods, but visually it appears that the rank-7 approximation captures many of the features in the original adjacency matrix.

As with the human connectome, we sought to understand the relationship between the structure of the graph, as conveyed through the adjacency spectral embedding, and the eight superstructures which partition the regions of the mouse brain. Panel (a) of Figure 13 shows entries of the four scaled singular vectors of the weighted adjacency matrix A corresponding to the largest singular values. The points are colored according to the four superstructures and the shapes are determined by the hemisphere. The ordering of the points groups together nodes in the same superstructure and hemisphere. The second and fourth vectors have structure which correlates closely with the four superstructures and the two hemispheres, respectively. Additionally, the first vector appears to separate the midbrain from the other three superstructures.

The right panel of Figure 13 shows a scatter plot of the entries of the fourth and second vectors along with the class boundaries for the eight-class quadratic discriminant analysis classifier, which is based on Gaussian mixture models with no constraints on the covariance structure Hastie et al. (2001). This classifier

achieves a training error rate of $87/296 \approx 0.29$. The error rate is particularly high for the white matter, with 58/60 vertices being classified incorrectly, meaning that ignoring the white matter, $29/236 \approx 0.12$ vertices were misclassified. Panel (c) shows the normalized confusion matrix for the eight classes, indicating that the forebrain and hindbrain classes are well separated while the white matter and midbrain have more substantial overlap. This matches with the general structure of the white matter which is not defined at the first hierarchical level of the atlas according to spatial structure, while the fore-, mid-, and hind-brain superstructures are.

Finally, as for the eight human superstructures, the same permutation analysis was also performed for the mouse connectome. Figure 14 shows the violin plots for the metric defined in Eq. (2) permuted in the same way as described in the Appendix. This test again indicates that the eight superstructures are significant even after accounting for the spatial structure of the regions.

6 Discussion

Motivated by the RDPG model, our methodology takes advantage of the low-rank structure of graphs by applying a low-rank approximation to the entry-wise MLE. We give a closed form for the asymptotic relative efficiency between the entry-wise MLE $\bar{A}$ and our estimator $\hat{P}$ in the case of a stochastic blockmodel, demonstrating that when the number of vertices N is sufficiently large, low-rank methods provide a substantial improvement. In particular, for a stochastic blockmodel with fixed number of blocks K, block size proportion ρ, and number of graphs M, the low-rank estimator $\hat{P}$ has MSE which is on the order of N times lower than the MSE for $\bar{A}$.

Moreover, our estimator outperforms the entry-wise MLE in a cross validation analysis of the SWU4 brain graphs and in low- and full-rank simulation settings when M is small. These results illustrate that $\hat{P}$ performs well even when the low-rank assumption is violated and that $\hat{P}$ is robust and can be applied in practice.

One of the key observations from our real data analysis was that the largest improvements using the low-rank method occurred when the number of graphs M was small, and that it provided only minor improvements, or even slightly degraded performance, when M was large. However, even in large scale studies the low-rank methods will be useful for estimating graph means for subpopulations, e.g. the population of females over 60 with some college education. Using the element-wise sample mean for such small strata, which may have fewer than ten subjects, will frequently result in a degradation of performance. Similarly, Durante et al. (2016) proposed a Bayesian nonparametric approach for modeling the population distribution of network-valued data which reduces dimensionality via a mixture model and our methods could be easily adapted to those ideas.

While the low-rank methods considered in this paper will often offer substantial improvements, further refinements of these methods which account for

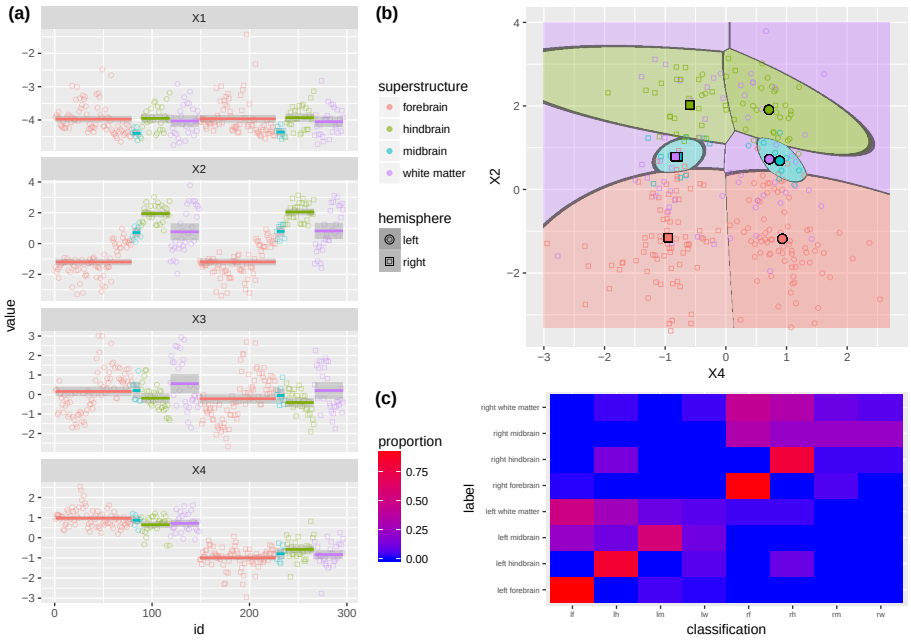


Fig 13: **Latent positions analysis of the mouse connectome** Panel (a) shows the entries of the the four scaled singular vectors of the weighted adjacency matrix A corresponding to largest singular values. The ordering of the entries is the same as the ordering in Figure 12, where vertices in the same superstructures and hemispheres are grouped together. The colors of points denote the superstructures and the shapes denote the hemispheres. Panel (b) shows a scatter plot of the the fourth and second singular vectors. The background coloring corresponds to fitting a classifier, based on a mixture of Gaussians, to classify the eight different superstructures. Panel (c) shows the normalized confusion matrix for the mixture of Gaussians learned in panel (b). Each entry corresponds to the proportion of nodes with a given label that were predicted to be each other label, with the true label given by the row and the predicted label given by the column.

the particular traits of connectomics data would be useful to improve estimation further. For example, we assume that the adjacency matrix is observed without contamination, however in practice there will be noise in the observed graph and one may seek to account for this noise with more robust methods. This may be especially fruitful when each graph has weighted edges and the weights themselves have noisy or heavy-tailed distributions. Rank-based methods and robust likelihood methods could be very useful in that case (Huber and Ronchetti, 2009; Qin and Priebe, 2013).

Another issue that arose in the analysis of the connectome dataset was the presence of structural ones in the mean graph for the population. These struc-

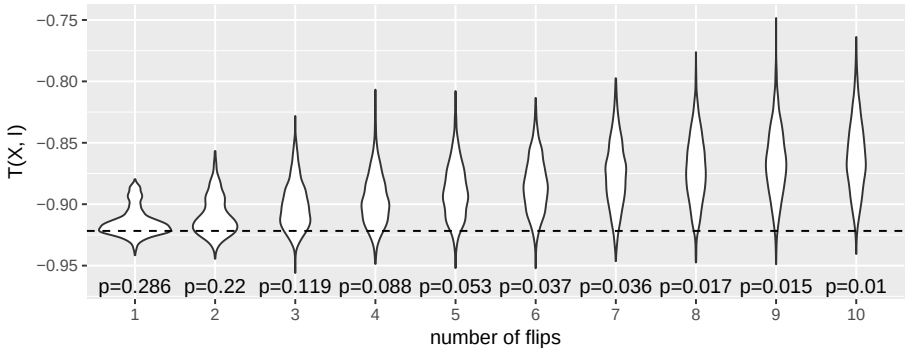


Fig 14: **Violin plot for the mouse connectome permutation test.** As in Fig. 10, 1000 simulations for each number of flips show that as the number of flips increases, the distribution under the null moves further from the dashed line. This indicates that the latent positions correlate with the superstructures even after conditioning on the spatial structure. Dashed line represents the situation based on true superstructure assignment.

tural ones appear since edges between certain regions of the brain are present in all members of the healthy population. For these always-present edges, the low-rank methods will have non-zero error while the sample mean will always have zero error. Detecting and incorporating structural ones and zeros could yield methods that share the best elements of both methods considered here.

For the SWU4 dataset, we performed a cross-validation framework and compared the estimates based on a subsample to the mean for the held-out set. Another option would be to compare the estimates $\bar{A}$ and $\hat{P}$ to the mean for the entire population including the subsample. Both of these analyses lead to very similar results in the cases presented above, but for various reasons one may prefer one analysis over another. The cross-validation method is most reasonable from a prediction perspective where prediction about new samples is of interest. If instead one is interested in learning directly about the mean of a population, especially a finite population, the sub-sampling approach may be the most logical choice.

While in this paper the focus on the estimation of the probability matrix P based on the low-rank structure, many future directions are quite interesting, such as fitting the SBM or clustering the vertices to detect different brain regions. For example, Abraham et al. (2013) introduced a region-extraction approach based on a sparse penalty with dictionary learning; Calhoun et al. (2001) performed independet component analysis of fMRI data to draw group inferences; Wang et al. (2017) consider a joint embedding model for feature extractions. In future work, it would be interesting to explore further statistical inferences based on our method.

Acknowledgments

This work is graciously supported by the XDATA program of the Defense Advanced Research Projects Agency (DARPA) administered through Air Force Research Laboratory contract FA8750-12-2-0303; the DARPA SIMPLEX program through SPAWAR contract N66001-15-C-4041; and DARPA GRAPHS contract N66001-14-1-4028. We would also like to acknowledge support from NIH through K01 AG041211, and P41 EB015897.

References

Alexandre Abraham, Elvis Dohmatob, Bertrand Thirion, Dimitris Samaras, and Gael Varoquaux. Extracting brain regions from rest fmri with total-variation constrained dictionary learning. In *MICCAI-16th International Conference on Medical Image Computing and Computer Assisted Intervention-2013*. Springer, 2013.

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014, 2008.

Robert J Anderson, James J Cook, Natalie A Delpratt, John C Nouls, Bin Gu, James O McNamara, Brian B Avants, G Allan Johnson, and Alexandra Badea. An HPC pipeline with validation framework for small animal multivariate brain analysis (SAMBA). September 2017.

Selen Atasoy, Isaac Donnelly, and Joel Pearson. Human brain networks function in connectome-specific harmonic waves. *Nature communications*, 7:10340, January 2016.

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

T E J Behrens, H Johansen Berg, S Jbabdi, M F S Rushworth, and M W Woolrich. Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *NeuroImage*, 34(1):144–155, January 2007.

Peter J Bickel and Kjell A Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*, volume i. Pearson Prentice Hall, second edition, 2007.

Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

Gerald G Brown and Herbert C Rutemiller. Means and variances of stochastic vector products with applications to random linear models. *Management Science*, 24(2):210–216, 1977.

Evan Calabrese, Alexandra Badea, Gary Cofer, Yi Qi, and G Allan Johnson. A diffusion MRI tractography connectome of the mouse brain and comparison with neuronal tracer data. *Cereb. Cortex*, 25(11):4628–4637, November 2015.

Vince D Calhoun, Tulay Adali, Godfrey D Pearlson, and JJ Pekar. A method for making group inferences from functional mri data using independent component analysis. *Human brain mapping*, 14(3):140–151, 2001.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

Daniele Durante, David B Dunson, and Joshua T Vogelstein. Nonparametric bayes modeling of populations of networks. *Journal of the American Statistical Association*, (just-accepted), 2016.

P Erdős and A Rényi. On random graphs, I. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

Donniell E Fishkind, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. 2012.

Eleftherios Garyfallidis, Matthew Brett, Bagrat Amirbekian, Ariel Rokem, Stefan van der Walt, Maxime Descoteaux, Ian Nimmo-Smith, and Dipy Contributors. Dipy, a library for the analysis of diffusion MRI data. *Frontiers in neuroinformatics*, 8:8, February 2014.

E N Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

Cedric E Ginestet, Prakash Balanchandran, Steven Rosenberg, and Eric D Kolaczyk. Hypothesis testing for network data in functional neuroimaging. *arXiv preprint arXiv:1407.5525*, 2014.

Krzysztof J Gorgolewski, Natacha Mendes, Domenica Wilfling, Elisabeth Wladimirow, Claudine J Gauthier, Tyler Bonnen, Florence JM Ruby, Robert Trampel, Pierre-Louis Bazin, Roberto Cozatl, et al. A high resolution 7-tesla resting-state fmri test-retest dataset with cognitive and physiological measures. *Scientific data*, 2, 2015.

William Gray, John Bogovic, Joshua Vogelstein, Bennett Landman, Jerry Prince, and R Vogelstein. Magnetic resonance connectome automated pipeline: An overview. *IEEE Pulse*, 2(3):42–48, 2012.

Sam Gutmann. Stein's paradox is impossible in problems with finite sample space. *The Annals of Statistics*, pages 1017–1020, 1982.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. New York, NY, USA, 2001.

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

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

Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2009. ISBN 978-0-470-12990-6.

W. James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I*, pages 361–379. Univ. California Press, Berkeley, Calif., 1961.

Xiaoyi Jiang, Andreas Müunger, and Horst Bunke. On median graphs: Properties, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1144–1151, 2001.

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

Gregory Kiar, William Gray Roncal, Disa Mhembere, Eric Bridgeford, Randal Burns, and Joshua T. Vogelstein. ndmg: Neurodata's mri graphs pipeline. `http://m2g.io`, August 2016.

Gregory Kiar, Krzysztof J Gorgolewski, Dean Kleissas, William Gray Roncal, Brian Litt, Brian Wandell, Russel A Poldrack, Martin Wiener, R Jacob Vogelstein, Randal Burns, et al. Science in the cloud (sic): A use case in mri connectomics. *Giga Science*, 6(5):1–10, 2017.

David Marchette, Carey Priebe, and Glen Coppersmith. Vertex nomination via attributed random dot product graphs. In *Proceedings of the 57th ISI World Statistics Congress*, volume 6, page 16, 2011.

Daniel S Margulies, Satrajit S Ghosh, Alexandros Goulas, Marcel Falkiewicz, Julia M Huntenburg, Georg Langs, Gleb Bezgin, Simon B Eickhoff, F Xavier Castellanos, Michael Petrides, Elizabeth Jefferies, and Jonathan Smallwood. Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences of the United States of America*, 113(44):12574–12579, November 2016.

Christine Leigh Myers Nickel. *Random Dot Product Graphs A Model For Social Networks*. PhD thesis, Johns Hopkins University, 2008.

Kenichi Oishi, Andreia V Faria, Peter CM van Zijl, and Susumu Mori. *MRI atlas of human white matter*. Academic Press, 2010.

Yichen Qin and Carey E Priebe. Maximum l q-likelihood estimation via the expectation-maximization algorithm: A robust estimation of mixture models. *Journal of the American Statistical Association*, 108(503):914–928, 2013.

Jiang Qiu, Qinglin Zhang, and Dongtao Wei. Swu 4 - southwest university connectome data. `http://fcon_1000.projects.nitrc.org/indi/CoRR/html/swu_4.html`, 2017. Accessed: 2017-12-01.

David H Salat, Randy L Buckner, Abraham Z Snyder, Douglas N Greve, Rahul SR Desikan, Evelina Busa, John C Morris, Anders M Dale, and Bruce Fischl. Thinning of the cerebral cortex in aging. *Cerebral cortex*, 14(7):721–730, 2004.

Edward R Scheinerman and Kimberly Tucker. Modeling graphs using dot product representations. *Computational Statistics*, 25(1):1–16, 2010.

S Sikka, B Cheung, R Khanuja, S Ghosh, C Yan, Q Li, J Vogelstein, R Burns, S Colcombe, C Craddock, et al. Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac). In *5th INCF Congress of Neuroinformatics, Munich, Germany*, volume 10, 2014.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, vol. I*, pages 197–206. University of California Press, Berkeley and Los Angeles, 1956.

Daniel L Sussman, Minh Tang, and Carey E Priebe. Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):48–57, 2014.

G. V. Trunk. A problem of dimensionality: A simple example. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (3):306–307, 1979.

Todd Vanderah and Douglas J Gould. *Nolte's The Human Brain: An Introduction to its Functional Anatomy, 7e*. Elsevier, 7 edition edition, 1 June 2015.

Shangsi Wang, Joshua T Vogelstein, and Carey E Priebe. Joint embedding of graphs. *arXiv preprint arXiv:1703.03862*, 2017.

Stephen J Young and Edward R Scheinerman. Random dot product graph models for social networks. In *Algorithms and models for the web-graph*, pages 138–149. Springer, 2007.

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

Xi-Nian Zuo, Jeffrey S Anderson, Pierre Bellec, Rasmus M Birn, Bharat B Biswal, Janusch Blautzik, John CS Breitner, Randy L Buckner, Vince D Calhoun, F Xavier Castellanos, et al. An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific data*, 1, 2014.

A Methods

A.1 Algorithm for low-rank approximation

Algorithm 2 Rank-d approximation of a matrix.

Input: Symmetric matrix $A \in \mathbb{R}^{N \times N}$ and dimension $d \leq N$.
Output: $\text{lowrank}_d(A) \in \mathbb{R}^{N \times N}$

1: Compute the algebraically largest d eigenvalues of A, $s_1 \geq s_2 \geq \ldots \geq s_d$ and corresponding orthonormal eigenvectors $u_1, u_2, \ldots, u_d \in \mathbb{R}^N$;
2: $\hat{S} \leftarrow \text{diag}(s_1, \ldots, s_d)$ and $\hat{U} \leftarrow [u_1, \ldots, u_d]$;
3: Return $\hat{U}\hat{S}\hat{U}^\top$;

A.2 Choosing Dimension

Often in dimensionality reduction techniques, the choice for dimension d, relies on analyzing the set of the ordered eigenvalues, looking for a "gap" or "elbow" in the scree-plot. Zhu and Ghodsi (2006) present an automated method for finding this gap in the scree-plot that takes only the ordered eigenvalues as an input and uses Gaussian mixture modeling to find these gaps. The mixture modeling results in multiple candidate dimensions or elbows, and our analysis indicated that underestimating the dimension is much more harmful than overestimating the dimension. For this reason, the 3rd elbow was employed in the experiments performed for this work. While Zhu and Ghodsi (2006) only defines the 1st elbow, we define the s-th elbow as in Algorithm 3.

Algorithm 3 Algorithm to compute the Zhu and Ghodsi's elbow

Input: The number of Zhu and Ghodsi's elbow s, with eigenvalues $\lambda_1, \ldots, \lambda_N$
Output: The s-th Zhu and Ghodsi's elbow

1: Calculate the 1st elbow d_1 based on $\lambda_1, \ldots, \lambda_N$ according to (Zhu and Ghodsi, 2006)
2: **for** $i = 2$ to s **do**
3: Calculate the i-th elbow d_i based on $\lambda_{d_{i-1}+1}, \ldots, \lambda_N$ according to (Zhu and Ghodsi, 2006)
4: **end for**

Universal Singular Value Thresholding (USVT) is a simple estimation procedure proposed in Chatterjee (2015) that can work for any matrix that has "a

little bit of structure". In the current setting, it selects the dimension d as the number of singular values that are greater than a constant c times $\sqrt{N/M}$. The specific constant c must be selected carefully based on the mean and variance of the entries, and since overestimating the dimension was not overly harmful, we chose a relatively small value of $c = 0.7$.

Overall, selecting the appropriate dimension is a challenging task and numerous methods could be applied successfully depending on the setting. On the other hand, in our setting, many dimensions will yield nearly optimal mean squared errors and the two methods did not pick drastically different dimensions. Thus efforts to ensure the selected dimension is in the appropriate range are more important than finding the best dimension.

A.2.1 Exploration of Dimension Selection Procedures

To further investigate the impact of the dimension selection procedures, we also considered all possible dimensions for $\hat{P}$ by ranging d from 1 to N. $\hat{\text{MSE}}$ of $\bar{A}$ and $\hat{P}$ was plotted in Fig. 15. The horizontal axis gives dimension d, which only impacts $\hat{P}$, which is why estimated MSE of $\bar{A}$ is shown as flat. When d is small, $\hat{P}$ underestimates the dimension and throws away important information, which leads to relatively poor performance. When $d = N$, $\hat{P}$ is equal to $\bar{A}$, so that the curve for $\hat{\text{MSE}}$ for $\hat{P}$ ends at $\hat{\text{MSE}}(\bar{A})$. In the figure, a triangle denotes the 3rd elbow found by the Zhu and Ghodsi method, and a square denotes the dimension selected by USVT with threshold 0.7. Both dimension selection algorithms tend to select dimensions which nearly minimize the mean squared error.

When M is 1 or 5, $\bar{A}$ has large variance which leads to large $\hat{\text{MSE}}$. Meanwhile, $\hat{P}$ reduces the variance by taking advantages of inherent low-rank structure of the mean graph. Such smoothing effect is especially obvious while there is only 1 observation. When $M = 1$, all weights of the graph are either 0 or 1, leading to a very bumpy estimate $\bar{A}$. In this case, $\hat{P}$ smooths the connectomes estimate and improves the performance. Additionally, there is a large range of dimensions where the performance for $\hat{P}$ is superior to $\bar{A}$. With a larger M, the performance of $\bar{A}$ improves so that its performance is frequently superior but nearly identical to $\hat{P}$.

A.3 Graph Diagonal Augmentation

The graphs examined in this work have no self-loops and thus the diagonal entries of the adjacency matrix and the mean graph are all zero. However, when computing the low-rank approximation, these structural zeros lead to increased errors in the estimation of the mean graph. While this problem has been investigated in the single graph setting, with multiple graphs, the problem is exacerbated since the variance of the other entries is lower, so the relative impact of the bias in the diagonal entries is higher. Moreover, the sum of eigenvalues of the hollow matrix will be zero, leading to an indefinite matrix, which

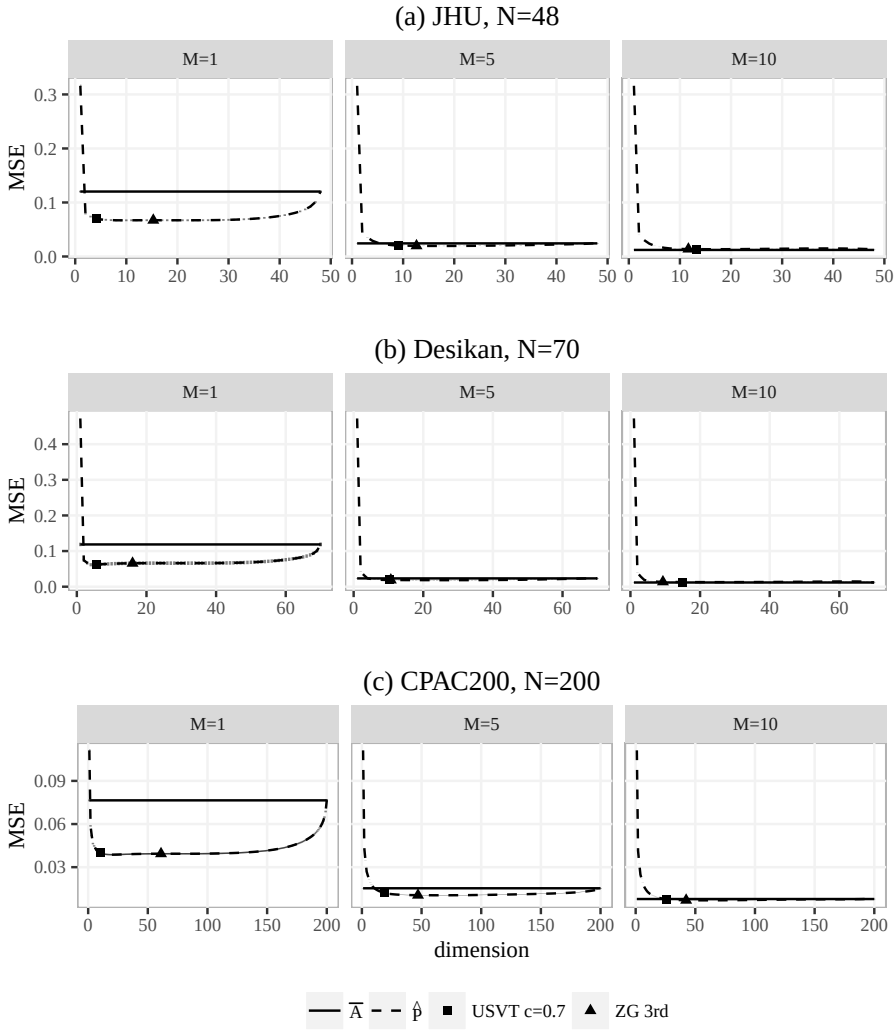


Fig 15: **Comparison of $\hat{\text{MSE}}$ of $\hat{P}$ and $\bar{A}$ for three atlases at three sample sizes for the CoRR data.** These plots show the mean squared error for $\bar{A}$ (solid line) and $\hat{P}$ (dashed line) for three datasets (JHU, Desikan, and CPAC200) while embedding the graphs into different dimensions and with different sample sizes M. The average dimensions chosen by the 3rd elbow of Zhu and Ghodsi is denoted by a triangle and those chosen by USVT with threshold equaling 0.7 is denoted by a square. Vertical intervals, visible mainly in the $N = 48, 70$ and $M = 1$ plots, represent the 95% confidence interval for the mean squared errors. When M is small, $\hat{P}$ outperforms $\bar{A}$ with a flexible range of the embedding dimension including the average of the dimensions selected by Zhu and Ghodsi and USVT.

violates the positive semi-definite assumption. So it is important to remedy the situation that we don't observe the diagonal entries.

Marchette et al. (2011) proposed the simple method of imputing the diagonals to be equal to the average of the non-diagonal entries for the corresponding row, or in equivalently the degree of the vertex divided by $n - 1$. Earlier, Scheinerman and Tucker (2010) proposed using an iterative method to impute the diagonal entries. In this work, these two ideas are combined by first using the row-average method (see Step 3 of Algorithm 1) and then using one step of the iterative method (see Step 6 of Algorithm 1). Note that when computing errors, the diagonal entries are omitted since these are known to be zero.

A.4 Dataset Description

A.4.1 Human Connectomes

The original dataset is from the Emotion and Creativity One Year Retest Dataset provided by Qiu, Zhang and Wei from Southwest University available at the Consortium for Reliability and Reproducibility (CoRR) (Zuo et al., 2014; Gorgolewski et al., 2015). It is composed of 235 subjects, all of whom were college students. Each subject underwent two sessions of anatomical, resting state DTI scans, spaced one year apart. Due to incomplete data, only 454 scans are available.

When deriving MR connectomes, the NeuroData team parcellates the brain into groups of voxels as defined by anatomical atlases (Kiar et al., 2016). The atlases are defined either physiologically by neuroanatomists (Desikan and JHU), or are generated using an automated segmentation algorithm (CPAC200). Once the voxels in the original image space are grouped into regions, an edge is placed between two regions when there is at least one white-matter tract, derived using a tractography algorithm, connecting the corresponding two parts of the brain (Garyfallidis et al., 2014). The resulting graphs are undirected, unweighted, and have no self-loops.

A.4.2 Mouse Connectome

Images of the fixed specimen were acquired on a 9.4 T small animal magnet using a 3D diffusion weighted imaging sequence. 120 unique diffusion directions were acquired using a b value of 4000 s/mm2, interleaved with 11 non-diffusion weighted scans. Images were acquired in 235 hours, and reconstructed at 43 micron resolution. The mouse brain was labeled with 296 regions of interest, 148 per hemisphere (Anderson et al., 2017).

To construct a structural connectome of the mouse brain fiber data was reconstructed (max 4 fiber orientations/voxel), then probabilistic tractography was performed using FSL (Behrens et al., 2007), (5000 samples per voxel, 21 μm step size, 45 degrees curvature threshold). The 296 seed regions had connectivity estimates produced by counting the number of fibers that originate from one region and fall onto all other regions. This was normalized by the volume of the seed region and resulted in a 296x296 weighted, directed graph.

A.5 Outline for the Proof of the Theorems

Here the proof of Lemma 4.1 is outlined, which provides the approximate MSE of $\hat{P}$ in the stochastic blockmodel case. The result depends on using the asymptotic results (see Theorem B.1) for the distribution of eigenvectors from Athreya et al. (2016) which extend to the multiple graph setting in a straightforward way.

The first key observation is that since $\bar{A}$ is computed from iid observations each with expectation P, $\bar{A}$ is unbiased for P and $\text{Var}(A_{ij}) = \frac{1}{M}P_{ij}(1-P_{ij})$. The results of Athreya et al. (2016) provide a central limit theorem for estimates of the latent position in an RDPG model for a single graph. Theorem B.1 describes important details. Since the variance of each entry is scaled by $1/M$ in $\bar{A}$, the analogous result for $\bar{A}$ is that the estimated latent positions will follow an approximately normal distribution with variance scaled by $1/M$ compared to the variance for a single graph.

Since $\hat{P}_{ij} = \hat{X}_i^\top \hat{X}_j$ is a noisy version of the dot product of $\nu_s^\top \nu_t$ from Section 2.3 and each $\hat{X}_i$ is approximately independent and normal, we can use common results for the variance of the inner product of two independent multivariate normals (Brown and Rutemiller, 1977). After simplifications that occur in the stochastic blockmodel setting, we can derive that the variance of $\hat{P}_{ij}$ converges to $\left(1/\rho_{\tau_i} + 1/\rho_{\tau_j}\right) P_{ij}(1-P_{ij})/(N \cdot M)$ as $N \to \infty$. Since the variance of $\bar{A}_{ij}$ is $P_{ij}(1-P_{ij})/M$, the relative efficiency between $\hat{P}_{ij}$ and $\bar{A}_{ij}$ is approximately $(\rho_{\tau_i}^{-1} + \rho_{\tau_j}^{-1})/N$ when N is sufficiently large.

B Proofs for Theory Results

Here the proofs are presented of the results in Section 4.1. To keep the ideas clear and concise, some details are removed, which are only slight changes to previous works. We assume the block memberships τ_i are drawn iid from a categorical distribution with block membership probabilities given by $\rho \in [0,1]^K$ where $\sum_i \rho_i = 1$. We will also assume that for a given N, the block memberships are fixed for all graphs.

We denote matrix of between-block edge probabilities by $B = \nu\nu^\top \in [0,1]^{K \times K}$ which we assume has rank K and is positive definite. By definition, the mean of the collection of graphs generated from this SBM is P, where $P_{ij} = B_{\tau_i, \tau_j}$.

We observe M graphs on N vertices $A^{(1)}, \cdots, A^{(M)}$ sampled independently from the SBM conditioned on τ. Define $\bar{A} = \frac{1}{M}\sum_{t=1}^{M} A^{(t)}$. Let $\hat{U}\hat{S}\hat{U}^\top$ be the best rank-d positive semidefinite approximation of $\bar{A}$, then we define $\hat{P} = \hat{X}\hat{X}^\top$, where $\hat{X} = \hat{U}\hat{S}^{1/2}$.

The proofs presented here will rely on a central limit theorem developed in Athreya et al. (2016). The theorem was modified slightly to account for the multiple graph setting and is presented in the special case of the stochastic blockmodel.

Theorem B.1 (Corollary of Theorem 1 in Athreya et al. (2016))**.** *In the setting above, let $X = [X_1, \ldots, X_N]^\top \in \mathbb{R}^{N \times d}$ have row i equal to $X_i = \nu_{\tau_i}$ (recall that*

τ_i are drawn from $[K]$ according to the probabilities ρ). Then there exists an orthogonal matrix W such that for each row i and j and any $z \in \mathbb{R}^d$, conditioned on $\tau_i = s$ and $\tau_j = t$,

$$\begin{aligned} &\Pr\left\{\sqrt{N}(W\hat{X}_i - \nu_s) \leq z, \sqrt{N}(W\hat{X}_j - \nu_t) \leq z'\right\} \\ =&\Phi(z, \Sigma(\nu_s)/M)\Phi(z', \Sigma(\nu_t)/M) + o(1) \end{aligned} \tag{3}$$

where $\Sigma(x) = \Delta^{-1}\mathbb{E}[X_j X_j^\top (x^\top X_j - (x^\top X_j)^2)]\Delta^{-1}$ and $\Delta = \mathbb{E}[X_1 X_1^T]$ is the second moment matrix, with all expectations taken unconditionally. The function Φ is the cumulative distribution function for a multivariate normal with mean zero and the specified covariance, and $o(1)$ denotes a function that tends to zero as $N \to \infty$.

The proof of this result follows very closely the proof of the result in the original paper with only slight modifications for the multiple graph setting.

We now prove a technical lemma which yields the simplified form for the variance under the stochastic blockmodel.

Lemma B.2. *In the same setting as Theorem 4.2, for any $1 \leq s, t \leq K$:*

$$\nu_s^\top \Sigma(\nu_t)\nu_s \ = \frac{1}{\rho_s}\nu_s^\top \nu_t \ (1 - \nu_s^\top \nu_t).$$

Proof. Under the stochastic blockmodel with parameters (B, ρ), we have $X_i \overset{iid}{\sim} \sum_{k=1}^K \rho_k \delta_{\nu_k}$, where $\nu = [\nu_1, \cdots, \nu_K]^\top$ satisfies $B = \nu\nu^\top$. Without loss of generality, it can be assumed that $\nu = US$ where $U = [u_1, \cdots, u_K]^\top$ is orthonormal in columns and S is a diagonal matrix. Here it can be concluded that $\nu_s^\top = u_s^\top S$. Defining $R = \mathrm{diag}(\rho_1, \cdots, \rho_K)$, allows

$$\Delta = \mathbb{E}[X_1 X_1^\top] = \sum_{k=1}^K \rho_k \nu_k \nu_k^\top = \nu^\top R \nu = SU^\top RUS.$$

Thus

$$\begin{aligned} \nu_s^\top \Sigma(\nu_t)\nu_s =& \sum_{k=1}^K \nu_s^\top \Delta^{-1}\rho_k \nu_k \nu_k^\top \Delta^{-1}\nu_s(\nu_t^\top \nu_k)(1 - \nu_t^\top \nu_k) \\ =& \sum_{k=1}^K \rho_k (u_s^\top U^\top R^{-1} U u_k)^2 (\nu_t^\top \nu_k)(1 - \nu_t^\top \nu_k) \\ =& \sum_{k=1}^K \rho_k (e_s^\top R^{-1} e_k)^2 (\nu_t^\top \nu_k)(1 - \nu_t^\top \nu_k) \\ =& \sum_{k=1}^K \rho_k \delta_{sk} \rho_s^{-2} (\nu_t^\top \nu_k)(1 - \nu_t^\top \nu_k) \\ =& \frac{1}{\rho_s}\nu_t^\top \nu_s (1 - \nu_t^\top \nu_s) \end{aligned}$$

□

Lemma B.3 (Lemma 4.1)**.** *In the same setting as above, for any i, j, conditioning on $X_i = \nu_{\tau_i}$ and $X_j = \nu_{\tau_j}$:*

$$\lim_{N\to\infty} N \cdot \mathrm{Var}(\hat{P}_{ij}) = \frac{1/\rho_{\tau_i} + 1/\rho_{\tau_j}}{M} P_{ij}(1 - P_{ij}).$$

And for N large enough, conditioning on $X_i = \nu_{\tau_i}$ and $X_j = \nu_{\tau_j}$:

$$\mathbb{E}[(\hat{P}_{ij} - P_{ij})^2] \approx \frac{1/\rho_{\tau_i} + 1/\rho_{\tau_j}}{MN} P_{ij}(1 - P_{ij}).$$

Proof. Conditioned on $X_i = \nu_k$, we have by Theorem B.1,

$$\mathbb{E}[W\hat{X}_i] = \nu_k + o(1)$$

and

$$N \cdot \mathrm{Cov}(W\hat{X}_i, W_n\hat{X}_i) = \Sigma(\nu_k)/M.$$

Also, Corollary 3 in Athreya et al. (2016) says $\hat{X}_i$ and $\hat{X}_j$ are asymptotically independent. Thus, conditioning on $X_i = \nu_s$ and $X_j = \nu_t$, we have $\lim_{N\to\infty} \mathbb{E}[\hat{X}_i^\top \hat{X}_j] = \lim_{N\to\infty} \mathbb{E}[(W_N\hat{X}_i)^\top]\mathbb{E}[W_N\hat{X}_j] = \nu_s^\top \nu_t = P_{ij}$.

Since $\hat{P}_{ij} = \hat{X}_i^\top \hat{X}_j$ is a noisy version of the dot product of $\nu_s^\top \nu_t$, combined with Lemma B.2 and the results above, by Equation 5 in (Brown and Rutemiller, 1977), conditioning on $X_i = \nu_s$ and $X_j = \nu_t$:

$$\mathbb{E}[\hat{X}_i^\top \hat{X}_j] = \mathbb{E}[(W_N\hat{X}_i)^\top]\mathbb{E}[W_N\hat{X}_j] = \nu_s^\top \nu_t + o(1) = P_{ij} + o(1)$$

and

$$\begin{aligned}
&N \cdot \mathrm{Var}(\hat{P}_{ij}) \\
=&\frac{1}{M}\left(\nu_s^\top \Sigma(\nu_t)\nu_s + \nu_t^\top \Sigma(\nu_s)\nu_t^\top\right) + \frac{1}{M^2 N}\left(tr(\Sigma(\nu_s)\Sigma(\nu_t))\right) + o(1) \\
=&\frac{1}{M}\left(\nu_s^\top \Sigma(\nu_t)\nu_s + \nu_t^\top \Sigma(\nu_s)\nu_t^\top\right) + o(1) \\
=&\frac{1/\rho_s + 1/\rho_t}{M} P_{ij}(1 - P_{ij}) + o(1).
\end{aligned}$$

Since $\hat{P}_{ij} = \hat{X}_i^\top \hat{X}_j$ is asymptotically unbiased for P_{ij}, when n is large enough:

$$\mathbb{E}[(\hat{P}_{ij} - P_{ij})^2] = \mathrm{Var}(\hat{P}_{ij}) \approx \frac{1/\rho_s + 1/\rho_t}{MN} P_{ij}(1 - P_{ij}) + o(1).$$

□

The proof for Theorem 4.2 is now a simple application of the above lemmas to the ratio of the mean squared errors for $\bar{A}$ and $\hat{P}$.

C Flipping Procedure in the Hypothesis Testing for Lobe Structure behind the Low-rank Methods

Here the details of the flipping procedure are described for the permutation test mentioned in Section 4.5. As mentioned before, there are 10 lobes and 70 regions based on the Desikan atlas. We say two regions are adjacent if they share a common boundary. Such spatial adjacency is denoted by an adjacency matrix S for the 70 regions, where $S_{ij} = 1$ means region i and region j contain a pair of voxels, v_i and v_j, which are spatially adjacent. If this is true, then region j is defined as a neighbor of region i. The lobe i.d. for region i is denoted by l_i.

Now a uniform 1-flip can be defined by:

1. Selecting a pair of adjacent regions (region i_1 and region j_1) across the boundary of lobes uniformly, i.e. $S_{i_1 j_1} = 1$ and $l(i_1) \neq l(j_1)$;
2. Uniformly selecting another pair of adjacent regions (region i_2 and region j_2 where $i_1 \neq i_2$ and $j_1 \neq j_2$) across the same boundary of lobes uniformly, i.e. $S_{i_2 j_2} = 1$ and $l(i_1) = l(i_2)$ and $l(j_1) = l(j_2)$;
3. Reassigning region j_1 to lobe l_{i_1} and reassign region i_2 to lobe l_{j_2}.

By this definition, after a uniform 1-flip, the number of regions in each lobe stays the same, where only two regions are changed to a different lobe.

Then we can define a uniform k-flip naturally as sequentially performing uniform 1-flip k times. Note that after a uniform k-flip, the number of regions in each lobe still stays the same.

In the permutation test, a uniform k-flip was applied and the test statistic $T(X, l)$ was calculated based on the lobe assignment after flipping. The p-value is computed as the proportion of uniform k-flips with a T value smaller than the T value for the true lobe assignments.

D Stochastic Blockmodel Parameter Setting

Here the parameters in the stochastic blockmodel example depicted in Figure 3 are given. It is a 5-block SBM with

$$B = \begin{bmatrix} 0.90 & 0.27 & 0.05 & 0.10 & 0.30 \\ 0.27 & 0.67 & 0.02 & 0.26 & 0.14 \\ 0.05 & 0.02 & 0.44 & 0.25 & 0.33 \\ 0.10 & 0.26 & 0.25 & 0.70 & 0.18 \\ 0.30 & 0.14 & 0.33 & 0.18 & 0.58 \end{bmatrix},$$

$$\rho = \begin{bmatrix} 0.22 & 0.39 & 0.05 & 0.16 & 0.18 \end{bmatrix}.$$