Connectome Smoothing and a Law of Large Graphs

Runze Tang[1], Michael Ketcha[2], Joshua T. Vogelstein[2], Carey E. Priebe[1], Daniel L. Sussman[3*]

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 Mathematics & Statistics, Boston University, Boston, MA
* sussman@bu.edu

Abstract

Estimating the mean of a population of graphs based on a sample is a core problem in connectomics and network science at large. Often, this problem is especially difficult because the sample or cohort size is relatively small as compared to the number of parameters to estimate. This problem will only be exacerbated as new technologies enable the collection of higher resolution connectomes. 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 together with tools for dimension selection and diagonal augmentation to smooth the estimates and improve performance over the naive methodology for small sample sizes. Theoretical results for the stochastic blockmodel show that this method will offer major improvements when there are many vertices. Similarly, in analyzing human connectome data, we demonstrate that the low-rank methods outperform the standard sample mean for many settings. These results indicate that low-rank methods should be an important part of the tool box for researchers studying populations of graphs.

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 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, we take averages almost everywhere, from data in Euclidean space to more complex objects like images, shapes, and documents. In this manuscript we consider the challenges of estimating a population mean based on a sample of brains, 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 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 often leads 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 can be realized [Trunk, 1979].

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 doesn't mean that other estimators should not be considered. In many situations where other structural information is hypothesized, other estimators may be preferable.

In complex data settings such as shape data, language data, or graph data, we also must take care in how we define 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 in which edges appear. As we will describe below, our definition of the mean graph is the expectation of the adjacency matrix.

This 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, we take the CoRR brain graphs with Desikan atlases as an example. The dataset contains 454 brain scans with 70 vertices. Each vertex represents a region defined by the Desikan atlases, while an edge exists between two vertices if there is at least one white-matter tract connecting the corresponding regions of the brain. More details about this dataset are given in Section 4.3 and Section 6.3. By observing M graphs sampled from the 454 graphs, our goal is to estimate the mean graph of the population P, defined as

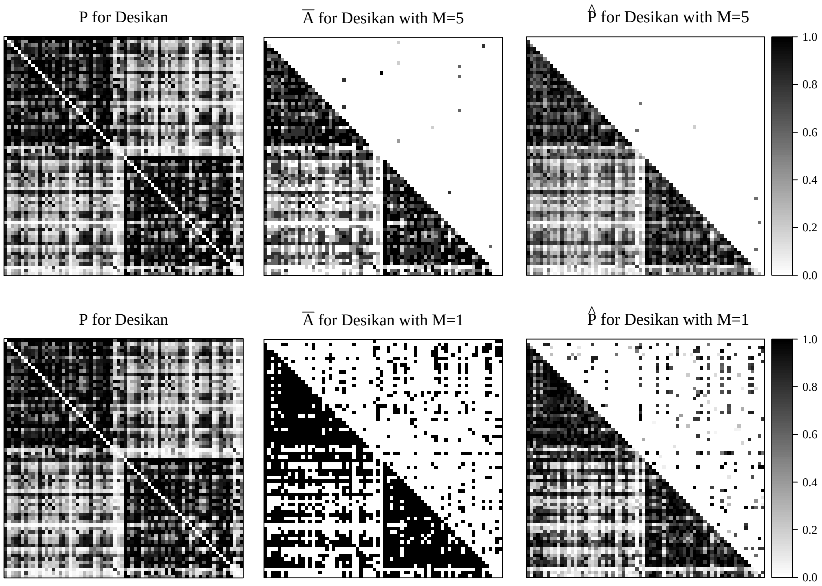


Fig 1: **Heat maps of the population mean P, the sample mean $\bar{A}$, and the estimator $\widehat{P}$ based on sample sizes $M = 1$ and $M = 5$.** The three heat maps in the upper level indicate the population mean for the 454 graphs (left), sample mean for the 5 sampled graphs (center), and $\widehat{P}$ for the same 5 sampled graphs with dimension $d = 11$ selected using the Zhu and Ghodsi method (right). Details about how to construct $\widehat{P}$ are discussed in Section 3.2. Darker pixels indicate a higher probability of an edge between the given vertices. By calculating the mean squared error based on this sample, we can see that $\widehat{P}$ (with mean squared error equals 0.015) outperforms $\bar{A}$ (with mean squared error equals 0.016), with a 3% relative improvement. In order to see where the improvements are clearly, in the upper triangular of the heat maps for $\bar{A}$ and $\widehat{P}$, we highlight the edges (18 edges highlighted for $\bar{A}$ and 6 for $\widehat{P}$) which have absolute estimation error larger than 0.4. In the lower level, we plot three heat maps in the similar way based on sample size $M = 1$. For this specific sampled graph, $\widehat{P}$ is calculated with dimension $d = 12$. A clearly smoothing effect can be seen in the heat map of $\widehat{P}$ (with mean squared error equals 0.049), which leads to a 53% relative improvement compared to $\bar{A}$ (with mean squared error equals 0.104). Similarly, we use the same absolute estimation error threshold 0.4 and highlight 504 edges for $\bar{A}$ and 234 edges for $\widehat{P}$.

the entry-wise mean of all the 454 graphs. We plot the population mean graph P on the left panel in Fig. 1.

The element-wise sample mean is a reasonable estimator if we consider the general independent edge model (IEM) [Bollobás et al., 2007] without taking any additional structure into account. However, with only a small sample size, such as when the sample size is much less than the number of vertices, it does not perform very well. Now take a sample of size $M = 5$ in the CoRR dataset example with Desikan atlases mentioned above, then we calculate the entry-wise sample mean $\bar{A}$ and plot it in the top middle panel of Fig. 1. Darker pixels indicate a higher probability of an edge between the given vertices. We can see $\bar{A}$ gives a fair estimate of P. However, there are still an amount of edges being estimated very inaccurately. In order to see these inaccurate estimates clearly, in the upper triangular of the heat maps for $\bar{A}$, we highlight the 18 edges which have absolute estimation error larger than 0.4 based on the same color scale. When the sample size is small, the performance of $\bar{A}$ degrades due to its high variances. Such phenomenon is most obvious when we decrease the sample size from $M = 5$ to $M = 1$. In the bottome middle panel of Fig. 1, we plot 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, we use the same absolute estimation error threshold 0.4 and highlight 504 edges in the upper triangular. Intuitively, an estimator incorporating structure in the distribution of graphs, assuming the estimator is computationally tractable, is preferable to the entry-wise sample mean. In general, we don't have any knowledge about this structure so it can be hard to take advantage of in practice.

One of the most important structures in graphs is the community structure in which vertices are clustered into groups that share similar connectivity structure. The stochastic blockmodel (SBM) [Holland et al., 1983] is one model that captures this structural property and is widely used in modeling networks. From population mean P plotted in Fig. 1, we can see the brain is a 2-block model at the highest level, representing the two hemispheres. More generally, the latent positions model (LPM) [Hoff et al., 2002], provides a way to parameterize the graph structure by latent positions associated with each vertex. Latent position models can capture strong community structure like the stochastic blockmodel, but may also allow for more variance within communities and other structures. One example of an LPM which captures this middle ground is the random dot product graph (RDPG) [Young and Scheinerman, 2007, Nickel, 2007] which motivates our estimator. It generalizes the positive semi-definite SBM by allowing for mixed membership and degree corrections.

Using estimates of the latent positions based on a truncated eigen-decomposition of the adjacency matrix, we propose an estimator which captures the low-rank structure of the mean graph for the RDPG model. Details about this estimator are discussed in Section 3.2. These estimates will improve performance since they will be biased towards the low-rank structure of the RDPG model and will have much lower overall variance than naive element-wise sample means. Here we consider the same random sample of size $M = 5$ based on the Desikan atlas in Fig. 1 and plot the estimate $\widehat{P}$ in the top right panel. Note that compared to the sample mean $\bar{A}$ (with mean squared error equals 0.016), $\widehat{P}$ (with mean squared error equals 0.015) has a finer gradient of values which in this case leads to a 3% relative improvement in estimation of the true probability matrix P. In order to see where the improvements are clearly, in the upper triangular of the heat map for $\widehat{P}$, we also highlight the 6 edges which have absolute estimation error larger than 0.4, where 18 edges are highlighted for $\bar{A}$ based on the same threshold. The smoothing effect is much more obvious when we decrease the sample size from $M = 5$ to $M = 1$. In the lower level of Fig.1, we plot the heat map of $\widehat{P}$ based on sample size $M = 1$. From the figure, we can see that $\widehat{P}$ 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, $\widehat{P}$ (with mean squared error equals 0.049) outperforms $\bar{A}$ (with mean squared error equals 0.104), with a 53% relative improvement. Similarly, we use the same absolute estimation error threshold 0.4 and highlight the 234 edges for $\widehat{P}$.

In this study, we show via theory, simulations, and real data analysis that the low-rank estimator frequently outperforms the element-wise sample mean, especially in small sample sizes.

In Section 2, we outline a nested collection of models which we consider for our theorems and simulations, and in Section 3 we describe the entry-wise sample mean and introduce our specific low-rank estimator, which accounts for the unknown dimension and attempts to correct for other issues found in real world problems. Our main results are presented in Section 4 where we present theorems and simulations for the stochastic blockmodel, an investigation of a connectome dataset, and a synthetic data analysis. We conclude with a discussion of these results in Section 5 and with details on our proposed method and the data which is analyzed in Section 6.

2 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. 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 also 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 since the graphs are identically distributed* $\mathbb{E}[A^{(m)}] = \mathbb{E}[A^{(m')}]$ *for any* m, m'*.*

We present here three nested models for these distributions, the independent edge model, the random dot product model, and the (positive semi-definite) stochastic blockmodel. In Section 3, we present two estimators motivated by these models.

2.1 Independent Edge Model

The first 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. For this case, we aim to estimate the mean graph $P = \mathbb{E}[A^{(m)}]$ based on the observed adjacency matrices $A^{(1)}, \dots, A^{(M)}$.

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, the two brain regions with similar properties will have similar connectivity patterns to other regions of the brain. 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 $\binom{N}{2}$.

A specific instance of an LPM that we examine in this work is the random dot product graph model (RDPG) [Young and Scheinerman, 2007, Nickel, 2007] 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 direction of the latent position is determined by properties of that vertex, and vertices whose latent positions point in similar directions are more likely to have an edge between them than vertices whose latent positions point in orthogonal directions. Similarly, the magnitudes of the latent position encode the vertices' overall tendency to form edges, with a larger magnitude leading to more edges incident with the vertex.

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 $X_1, \dots, X_N \in \mathcal{X}$ be the column vectors representing the N latent positions and write $X = [X_1 | \cdots | X_N]^\top \in \mathbb{R}^{N \times d}$. A random graph G with adjacency matrix A is said to be an RDPG if for any adjacency matrix a of the same size,

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

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.

Low-rank methods can often be more easily interpreted in a vertexy way. In particular, 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, one can analyze and visualize the geometry of these vectors in order to interpret how each vertex is behaving in the context of the larger graph. Now we take the CoRR brain graphs with Desikan atlases as an example. More details about this dataset are given in Section 4.3 and Section 6.3. By embedding the mean graph P which is the average of all 454 graphs, we get the estimated latent positions $\widehat{X} \in \mathbb{R}^{N \times d}$, where $N = 70$ is the number of vertices and $d = 8$ is the dimension selected by the Zhu and Ghodsi's method. We color the brain using the first 5 dimensions of $\widehat{X}$ as in Fig. 2 respectively. By the figure of the second dimension, we can see a clear distinction of the left and right hemisphere as conveyed in the second dimension. Additionally, such a representation allows the use of techniques from multivariate analysis to further study the estimated population mean.

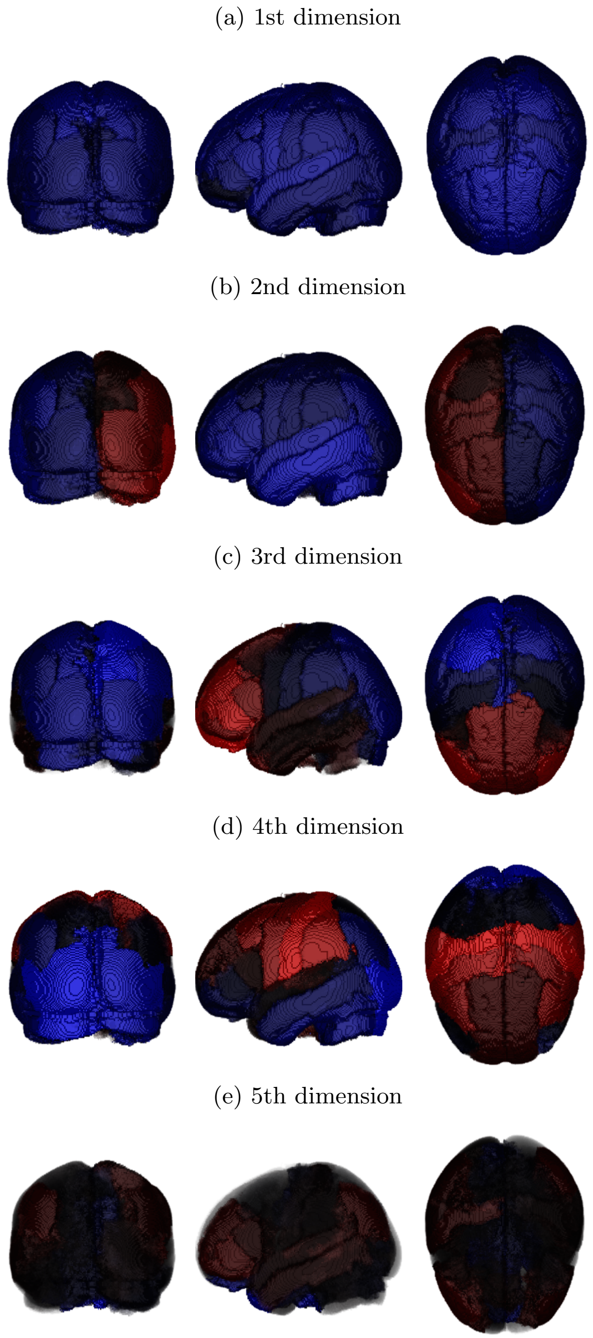
2.3 Stochastic Blockmodel as a Random Dot Product Graph

One of the most important structures for graphs is the community structure in which vertices are clustered into different communities such that vertices of the same community behave similarly. 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 $\tau \in \{1, \ldots, K\}^N$, 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 we will refer to the sub-model of the SBM with positive semidefinite B matrix as the SBM.

In order to analyze the estimator $\widehat{P}$ motivated by RDPG (details about how to construct $\widehat{P}$ are discussed in Section 3.2), we will represent the SBM as an RDPG by decomposing B as $\nu\nu^\top$, where $\nu \in \mathbb{R}^{K \times d}$ with rows given by $\nu_1^\top, \ldots, \nu_K^\top$ and each row $\nu_k^\top$ is the shared latent position for all vertices assigned to block k. For $X \in \mathbb{R}^{N \times d}$ with rows given by $X_1^\top = \nu_{\tau_1}^\top, \ldots, X_N^\top = \nu_{\tau_N}^\top$, since



(a) 1st dimension

(b) 2nd dimension

(c) 3rd dimension

(d) 4th dimension

(e) 5th dimension

Fig 2: **Brain plots colored by the first 5 dimensions of $\widehat{X}$ for the Desikan atlas respectively.** We plot the brain using the first 5 dimensions of $\widehat{X}$. From the figure, we can see the embeddings have their own neuro-meaning, for example there is a clear distinction of the left and right hemisphere as conveyed in the second dimension. Also, the first dimension provides an average level of the entire brain.

τ_i and τ_j represent the blocks that vertex i and vertex j belong to respectively, we have

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

In this way, the SBM can be seen as 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. We omit the details of the parameter setting of this example for conciseness. From the top two panels of the figure, we can clearly see the structure of 25 blocks in both the probability matrix and the adjacency matrix as a result of 5 different blocks among vertices.

Remark 2.2. *As we mentioned, under the SBM, all vertices in the same block have identical latent positions. Rather than allowing vertices differ from each other as RDPG, SBM presumes all nodes within the same block have the same expected degree. To better describe complex networks in some situations, a bunch of generalizations of the SBM have been explored in order 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 than a single block as SBM requires. Also, in order 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. All these generalizations are trying to drag SBM towards RDPG a little bit so that the model can capture variations among vertices while keeping the community structure.*

3 Estimators

In this section, we present two estimators, the standard element-wise sample mean $\bar{A}$, and a low-rank estimator $\widehat{P}$. We describe the low-rank aspects of this estimator in this section and present further important details regarding diagonal augmentation and dimension estimation in Section 6.

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

The most natural estimator to consider is to take the average of the observed adjacency matrices which yields 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, $\bar{A}$ is the uniformly minimum-variance unbiased estimator, so it has the smallest variance among all unbiased estimators and enjoys the many asymptotic properties of the MLE as $M \to \infty$ for fixed N. However, if graphs with a

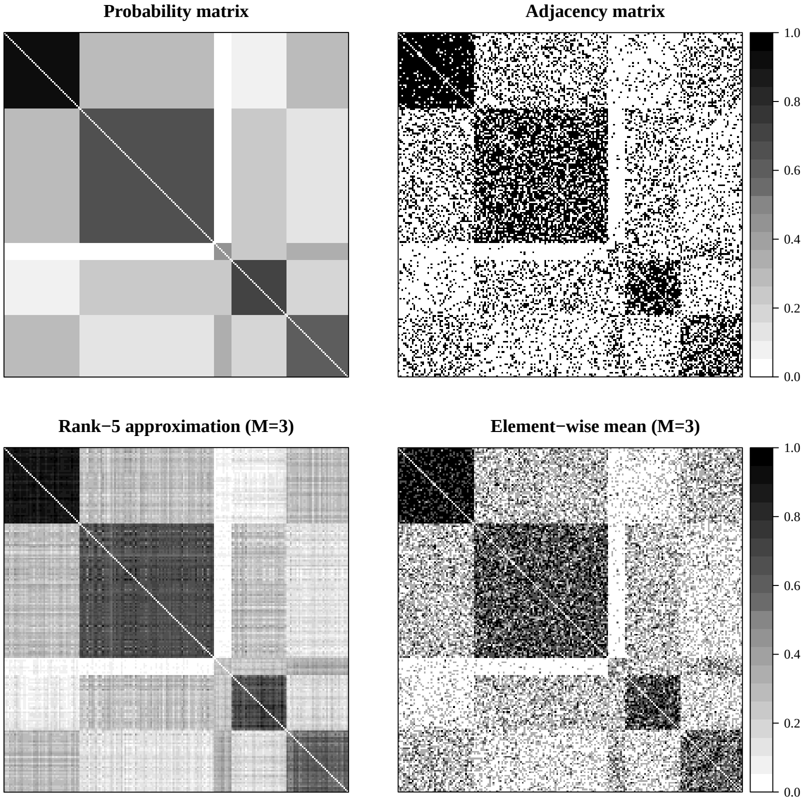


Fig 3: **Example illustrating the stochastic blockmodel.** The top left figure shows the mean graph P with $K = 5$ blocks and $N = 200$ vertices and the top right figure shows an adjacency matrix A sampled according to the probabilities from P. While A is a noisy version of P, much of the structure of P is preserved in A, a property we will exploit in our estimation procedure. Based on three graphs sampled independently and identically according to the probability matrix P, we construct the element-wise mean $\bar{A}$, shown in the lower right panel (see Section 3.1). Finally, by taking a rank-5 approximation of $\bar{A}$ and thresholding the values to be between 0 and 1, we construct our proposed estimate $\widehat{P}$, shown in the lower left panel (see Section 3.2). By visual inspection, it is clear that the low-rank estimate $\widehat{P}$ more closely approximates the probability matrix P as compared to $\bar{A}$.

large number of vertices are of interest, there are no useful asymptotic properties for $\bar{A}$ as the number of vertices N becomes large for fixed M.

Additionally, $\bar{A}$ doesn't exploit any 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 occurs for the RDPG.

3.2 Low-Rank Estimator $\widehat{P}$

Motivated by the low-rank structure of the RDPG mean matrix, we propose the estimator $\widehat{P}$ based on the spectral decomposition of $\bar{A}$ which yields a low rank approximation of $\bar{A}$. This estimator is similar to the estimator proposed by Chatterjee [2015] but additionally we propose adjustments to canonical low-rank methods which serve to improve the performance for the specific task of estimating the mean graph. Additionally, we consider an alternative dimension selection technique as discussed in Section 6.1. To summarize, our overall strategy to computer $\widehat{P}$ is described in Algorithm 1. The key component of this algorithm is the low-rank estimator. Details of this vital step to compute the actual low-rank approximation is in Algorithm 2.

```
**Algorithm 1** Algorithm to compute $\widehat{P}$
**Input:** Adjacency matrices $A^{(1)}, A^{(2)}, \cdots, A^{(M)}$, with each $A^{(m)} \in \{0,1\}^{N \times N}$
**Output:** Estimate $\widehat{P} \in [0,1]^{N \times N}$
Calculate the sample mean $\bar{A} = \frac{1}{M} \sum_{m=1}^{M} A^{(m)}$;
Calculate the scaled degree matrix $D^{(0)} = \mathrm{diag}(\bar{A}\mathbf{1})/(N-1)$;
Select the dimension $d$ based on the eigenvalues of $\bar{A}+D^{(0)}$; (see Section 6.1)
Set $\tilde{P}^{(0)}$ to $\mathrm{lowrank}_d(\bar{A} + D^{(0)})$; (see Algorithm 2)
Set $D^{(1)}$ to $\mathrm{diag}(\tilde{P}^{(0)})$, the diagonal matrix with diagonal matching $\tilde{P}^{(0)}$;
Set $\tilde{P}^{(1)}$ to $\mathrm{lowrank}_d(\bar{A} + D^{(1)})$; (see Algorithm 2)
Set $\widehat{P}$ to $\tilde{P}^{(1)}$ with values $< 0$ set to 0 and values $> 1$ set to 1.
```

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

Algorithm 2 Algorithm to compute the rank-d approximation of a matrix.

```
Input: Symmetric matrix $A \in \mathbb{R}^{N \times N}$ and dimension $d \leq N$.
Output: $\mathrm{lowrank}_d(A) \in \mathbb{R}^{N \times N}$
Compute the algebraically largest $d$ eigenvalues of $A$, $s_1 \geq s_2 \geq \ldots \geq s_d$ and corresponding unit-norm eigenvectors $u_1, u_2, \ldots, u_d \in \mathbb{R}^N$;
Set $\widehat{S}$ to the $d \times d$ diagonal matrix $\mathrm{diag}(s_1, \ldots, s_d)$;
Set $\widehat{U} = [u_1, \ldots, u_d] \in \mathbb{R}^{N \times d}$;
Set $\mathrm{lowrank}_d(A)$ to $\widehat{U}\widehat{S}\widehat{U}^\top$;
```

To compute our estimator $\widehat{P}$, we also need to specify what rank d to use and there are various ways of dealing with dimension selection. In this paper, we use an elbow selection method proposed in Zhu and Ghodsi [2006] and the universal singular value thresholding (USVT) method [Chatterjee, 2015]. Details for these methods are discussed in Section 6.1.

Moreover, since the adjacency matrices are hollow, with zeros along the diagonal, there is a missing data problem that leads to inaccuracies if we compute $\widehat{P}$ based only on $\bar{A}$. To compensate for this issue, we use an iterative method developed in Scheinerman and Tucker [2010]. Details are discussed in Section 6.2.

Algorithm 1 gives the steps involved to compute the low-rank estimate $\widehat{P}$. The bottom panels of Fig. 3 demonstrate the two estimators $\widehat{P}$ and $\bar{A}$ for the stochastic blockmodel given by the upper left panel. The estimates are based on a sample of size $M = 3$ and in this instance visual inspection demonstrates that $\widehat{P}$ performs much better than $\bar{A}$. As we will see in the succeeding sections, this procedure will frequently yield improvements in estimation as compared to using the sample mean $\bar{A}$. While this is unsurprising 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, we explore this estimator in the context of the stochastic blockmodel.

4 Results

4.1 Asymptotic Theory

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 $\widehat{P}$ motivated by the RDPG. We evaluate our estimators in terms of mean squared error, either $\mathrm{MSE}(\widehat{P}_{ij}) = \mathbb{E}[\widehat{P}_{ij} - P]^2$ or $\mathrm{MSE}(\bar{A}) = \mathbb{E}[\bar{A}_{ij} - P]^2$. While we 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 $\mathrm{RE}(\bar{A}_{ij}, \widehat{P}_{ij}) = \frac{\mathrm{MSE}(\widehat{P}_{ij})}{\mathrm{MSE}(\bar{A}_{ij})}$, with values above 1 indicating $\bar{A}$ should be preferred while values below 1 indicate $\widehat{P}$ 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, we analyze the performance of these two estimators under the SBM by computing the entry-wise relative efficiency (RE). We also consider the asymptotic relative efficiency, which 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, $N \cdot \mathrm{RE}(\bar{A}_{ij}, \widehat{P}_{ij})$ which in our case normalizes the relative efficiency so that the asymptotic scaled relative efficiency is non-zero and finite. Somewhat surprisingly, we will see 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 $|\{i : \tau_i = k\}|/N$ of vertices in block k 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 block probability matrix $B = \nu\nu^\top \in [0,1]^{K \times K}$. 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, we can calculate the two estimators $\bar{A}$ and $\widehat{P}$.

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

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

And

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

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 $\widehat{P}$. The proof of this lemma is outlined in Section 6.4 and is based on results for the variance of the adjacency spectral embedding from Athreya et al. [2013]. From Lemma 4.1, we can see that the MSE of $\widehat{P}_{ij}$ is of order $O(M^{-1}N^{-1})$ approximately. Similar to $\bar{A}$, the estimate will get better as the number of observations M increases. Furthermore, it also benefits from a larger graph because of the use of low-rank structure. That is, $\widehat{P}$ will perform better as the number of vertices of the graph N increases.

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

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

Based on this MSE result of $\bar{A}_{ij}$ and the MSE result of $\widehat{P}_{ij}$ given by Lemma 4.1, we can conclude the following theorem naturally.

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

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

And the asymptotic relative efficiency (ARE) is

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

Proof. Combine 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,

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

we have for large enough N,

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

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

This theorem indicates that under the SBM, $\widehat{P}$ is a much better estimate of the mean of the collection of graphs P than $\bar{A}$. Note that a relative efficiency less than 1 indicates that $\widehat{P}$ should be preferred over $\bar{A}$, so under the above assumptions, as $N \to \infty$, $\widehat{P}$ performs far better than $\bar{A}$. From the result, we see that the relative efficiency is of order $O(N^{-1})$ and $N \cdot \mathrm{RE}(\bar{A}_{ij}, \widehat{P}_{ij})$ 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 $\widehat{P}$ is relative to $\bar{A}$, regardless of M.

Note that the asymptotic result here is for number of vertices going to infinity with a fixed number of graphs. Such setting is very useful in connectomics analysis since we anticipate the collection of larger and larger brain network which will also likely initially correspond to smaller sample sizes as the technology to scale these connectome collection techniques is developed.

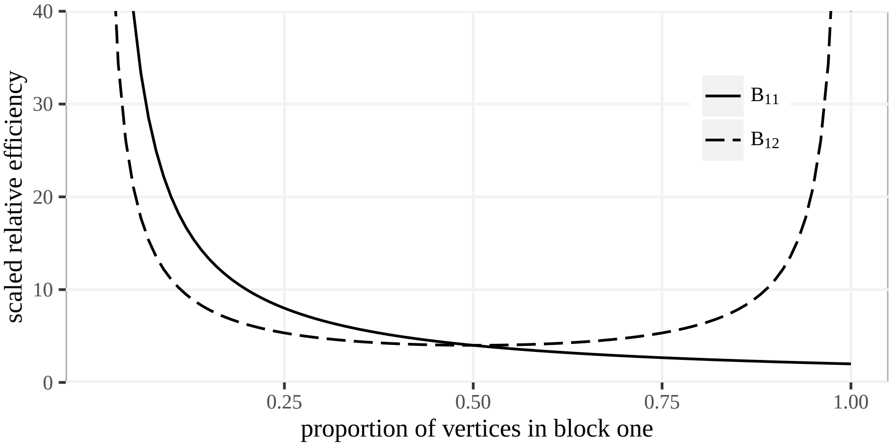


Fig 4: **Asymptotic scaled relative efficiency $N \cdot \mathrm{RE}(\bar{A}, \widehat{P})$ in a 2-block SBM.** For each distinct pair of edge probabilities in a 2-block SBM specified in Eq. 2, the scaled relative efficiency only depends on the proportion of vertices in each block. We show the scaled asymptotic relative efficiency 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, $\widehat{P}$ performs best for large blocks while the improvements may be very minor for blocks with only a few vertices.

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 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 typically no vertices are in that block, so the effective minimum we can consider for ρ_s is $1/N$.)

To illustrate Eq. 1 of Theorem 4.2, we consider a 2-block SBM with parameters

$$B = \begin{bmatrix} 0.42 & 0.2 \\ 0.2 & 0.7 \end{bmatrix}, \qquad \rho = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}, \tag{2}$$

so that $|\{i : \tau_i = 1\}| \approx |\{i : \tau_i = 2\}|$, especially for large N. Note that this simulation only focuses on the rank-2 setting primarily for the interpretability. When calculating $\widehat{P}$, we omit the dimension selection step from Algorithm 1 and instead use the true dimension $d = \text{rank}(B) = 2$. Fig. 4 shows $2/\rho_1$ and $1/\rho_1 + 1/\rho_2$, the scaled asymptotic RE for pairs of vertices both in block one and pairs of vertices in different blocks, respectively, in the 2-block SBM we specified earlier. 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. For $N = 500$ and $M = 100$, estimates of the scaled RE based on simulations agree very closely with their corresponding theoretical values displayed in the figure. Note that when $\rho_1 = 0.5$, the scaled RE has value 4.0, which agrees with the result in Fig. 5 for simulated data.

If instead of assuming that the graphs follow an SBM distribution, we assume 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 $\widehat{P}$ versus $\bar{A}$ in the general RDPG case, using the same ideas as in Theorem 4.2, we can show that $\text{RE}(\bar{A}_{ij}, \widehat{P}_{ij}) = O(1/N)$.

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*

$$\text{RE}(\bar{A}_{ij}, \widehat{P}_{ij}) = O(1/N).$$

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 we omit it here.

Remark 4.4. *As we 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 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, $\text{RE}(\bar{A}, \widehat{P})$ will increase as M increases for full-rank models. Indeed, for large M we could have $\text{RE}(\bar{A}, \widehat{P}) \geq 1$ since we cannot guarantee that $\widehat{P}$ 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 we don't see such structure in connectomics. We will discuss this further in Section 4.3 when applying our estimator to the CoRR dataset.*

4.2 Finite Sample Toy Model Simulations 467

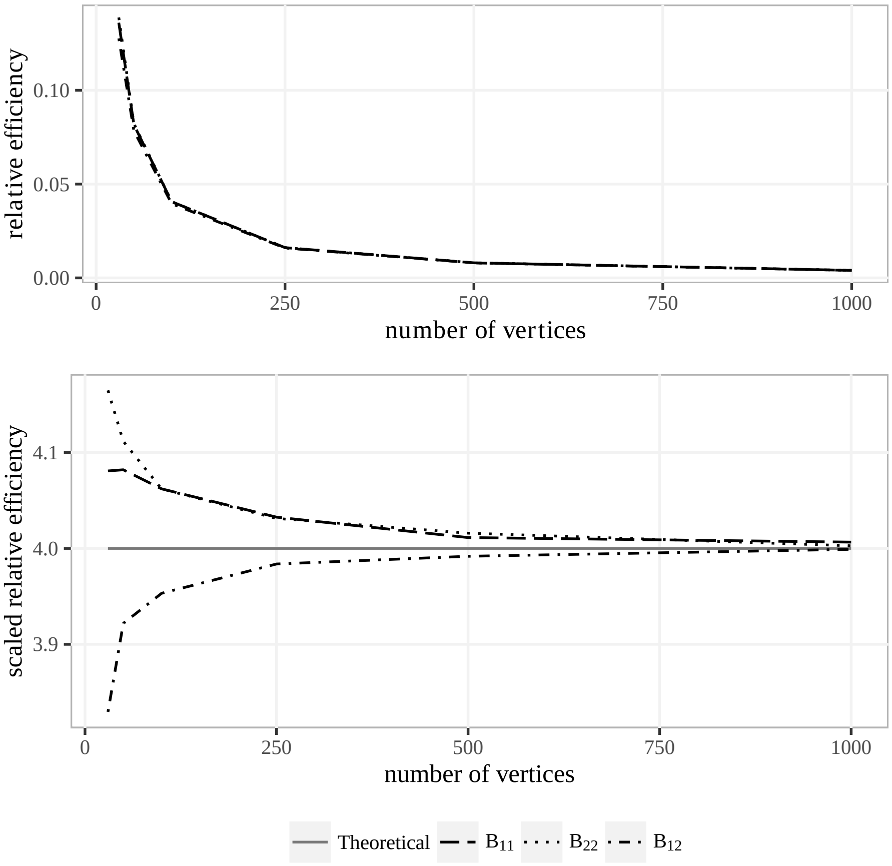


Fig 5: **Finite sample relative efficiency based on simulations.** The top panel shows the estimated relative efficiency $\widehat{\mathrm{RE}}(\bar{A}, \widehat{P})$ as a function of N for fixed $M = 100$ based on simulations of an SBM. For each value of N, we used 1000 Monte Carlo replicates of the SBM from Section 4.2 to estimate 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 $\widehat{P}$ 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, $N \cdot \widehat{\mathrm{RE}}(\bar{A}, \widehat{P})$. 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.

In this section, we will illustrate the theoretical results from Section 3.1 468 regarding the relative efficiency between $\bar{A}$ and $\widehat{P}$ via Monte Carlo simulation experiments in an idealized setting. These numerical simulations will also allow us to investigate the finite sample performance of the two estimators. Note that in Section 4.4, we will break the model assumptions slightly and run experiment in a more realistic setting.

Here, we consider the same 2-block SBM as in Eq. 2. To be clear, we restate the parameters here:

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

Similarly, when calculating $\widehat{P}$, we omit the dimension selection step from Algorithm 1 and instead use the true dimension $d = \operatorname{rank}(B) = 2$.

To investigate the finite sample relative efficiency, we first sample 1000 Monte Carlo replicates from the above SBM distribution with different numbers of vertices $N \in \{30, 50, 100, 250, 500, 1000\}$ and a fixed number of graphs $M = 100$. The relative efficiency $\mathrm{RE}(\bar{A}_{ij}, \widehat{P}_{ij})$ 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, we estimate the relative efficiency for each block pair using

$$\widehat{\mathrm{RE}}_{st}(\bar{A}, \widehat{P}) = \frac{\sum_{\tau_i = s, \tau_j = t, i \neq j} \widehat{\mathrm{MSE}}(\widehat{P}_{ij})}{\sum_{\tau_i = s, \tau_j = t, i \neq j} \widehat{\mathrm{MSE}}(\bar{A}_{ij})}$$

for $s, t \in \{1, 2\}$, where $\widehat{\mathrm{MSE}}$ 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 when it is clear from context.

In Fig. 5, we plot the (estimated) relative efficiency (top panel) and the scaled (estimated) relative efficiency (bottom panel), $N \cdot \widehat{\mathrm{RE}}_{st}(\bar{A}, \widehat{P})$. 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 $\widehat{P}$ is performing better than $\bar{A}$.

Based on Theorem 4.2, we also have that 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. From the figure, we see that $N \cdot \widehat{\mathrm{RE}}_{st}(\bar{A}, \widehat{P})$ converges to scaled asymptotic RE quite rapidly. We omit error bars 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, overall these differences are nearly indistinguishable for unscaled relative efficiency.*

4.3 CoRR Brain Graphs: Cross-Validation

In practice, 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 $\widehat{P}$ is still useful in such cases, we tested its performance on structural connectomic data. The graphs are based on diffusion tensor MR images collected and available at the Consortium for Reliability and Reproducibility (CoRR) [Zuo et al., 2014, Gorgolewski et al., 2015] (see Section 6.3).

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 Roncal et al. [2013] and Kiar et al. [In Preparation]. The vertices of the graphs represent different regions in the brain defined according to an atlas. We used three atlases, the JHU atlas with 48 vertices, the Desikan atlas with 70 vertices, and the CPAC200 atlas with 200 vertices. 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 6.3.

4.3.1 Experiment Results

In order to evaluate the performance of the two estimators, we used a cross validation on the 454 graphs of each size. Specifically, for a given atlas, each Monte Carlo replicate corresponds to sampling M graphs out of the 454 and computing the low-rank estimator $\widehat{P}$ and the sample mean $\bar{A}$ using the M selected graphs. We then compared these estimates 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 4.4), the sample mean for the remaining graphs does give the proportion of times each pair of vertices are adjacent in the population from which the graphs were sampled.

While in previous sections we evaluated the mean squared error for either an individual entry or for an entire block in the SBM, in this section and the next section we will focus on the overall error for estimating the mean graph. In particular we will use the average of the mean squared error across all pairs of vertices and we define $\mathrm{MSE}(\bar{A}) = \binom{N}{2}^{-1} \sum_{i<j} \mathbb{E}[\bar{A}_{ij} - P_{ij}]$ and similarly for $\mathrm{MSE}(\widehat{P})$, which are also used for the relative efficiency. As in the previous section, we will not use analytical evaluations of the MSE and instead estimate the MSE and relative efficiencies via Monte Carlo simulations.

We ran 1000 simulations on each of the three atlases for sample size $M = 5, 10$. For $M = 1$, we only have 454 different possibilities. So instead of running

1000 simulations, we looked through all 454 possible sample with size 1. As long as we determine which dimension to embed, the two estimates $\bar{A}$ and $\widehat{P}$ can be calculated based on the sample. In practice, we use algorithms like Zhu and Ghodsi's method or USVT to select the dimension d. These methods are neither computationally advanced nor requiring sophisticated algorithms. Details are discussed in Section 6.1. We plot the estimated relative efficiencies between $\bar{A}$ and $\widehat{P}$ in Fig. 6.

For each atlas and each sample size, we compare the Zhu and Ghodsi method [Zhu and Ghodsi, 2006] with the USVT method [Chatterjee, 2015] and note that both perform reasonably well relative to the full-dimensional $\bar{A}$. We omit confidence intervals for the estimated relative efficiencies since all confidence intervals have lengths less than 0.015, indicating that all relative efficiencies, aside from the relative efficiency for the CPAC200 atlas at $M = 10$, are very different from 1.

Again we can see that the largest improvements using $\widehat{P}$ 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 $\widehat{P}$. Nonetheless, when applied to subgroups inference, such as all females between the age of 21 and 25, $\widehat{P}$ can be really helpful for better exploring differences between groups compared to $\bar{A}$ due to a small sample size of each subgroup. In addition, $\widehat{P}$ 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.

To further illustrate the differences between the two estimators, we considered a single random sample of size $M = 5$ based on the Desikan atlas. We calculated $\bar{A}$ and $\widehat{P}$, using Zhu and Ghodsi's 3rd elbow to select $d = 11$. In Fig. 1, the estimates $\bar{A}$ and $\widehat{P}$ as well as the sample mean of 454 graphs (as a close estimate of P) are plotted in the upper level. Since the sample size is small, there are a lot of pairs of vertices with no edges or 5 edges in the 5 observations. This leads to the white and black pixels in the image corresponding to $\bar{A}$. On the other hand, $\widehat{P}$ has a finer gradient of values which in this case leads to a more accurate estimate. By calculating the mean squared error based on this sample, we can see that $\widehat{P}$ (with mean squared error equals 0.015) outperforms $\bar{A}$ (with mean squared error equals 0.016), with a 3% relative improvement. In order to see where the improvements are clearly, in the upper triangular of the heat maps for $\bar{A}$ and $\widehat{P}$, we highlight the edges (18 edges highlighted for $\bar{A}$ and 6 for $\widehat{P}$) which have absolute estimation error larger than 0.4.

Moreover, for the same sample discussed above, Fig. 7 shows the values for the absolute estimation error $|\bar{A} - P|$ and $|\widehat{P} - P|$. In addition, we include the absolute difference $|\bar{A} - \widehat{P}|$ 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 from $\bar{A}$ and 6 edges from $\widehat{P}$ being highlighted in the figure, further indicating the superior performance of $\widehat{P}$. Note that approximately 13% of all pairs of vertices are adjacent in all 454 graphs

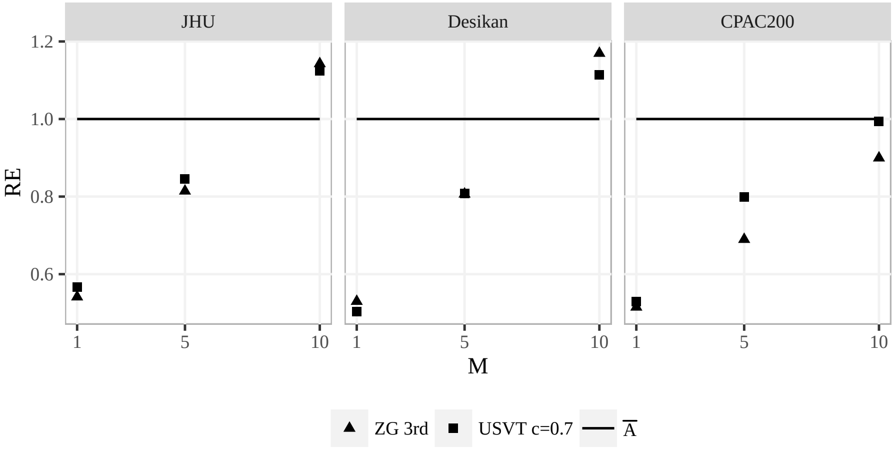


Fig 6: **Relative efficiencies of $\bar{A}$ versus $\widehat{P}$ for the CoRR data set.** For each atlas, JHU, Desikan, and CPAC 200, we sampled graphs which we used to compute $\bar{A}$ and $\widehat{P}$. We compared different sample sizes M and different dimension selection procedures, ZG and USVT. For each of the two methods for computing $\widehat{P}$, we estimated their relative efficiencies with respect to the sample mean $\bar{A}$. Confidence intervals all had lengths less than 0.015, and hence we omitted them for clarity. Overall, the relative efficiencies are greater for smaller sample sizes M and larger number of vertices N.

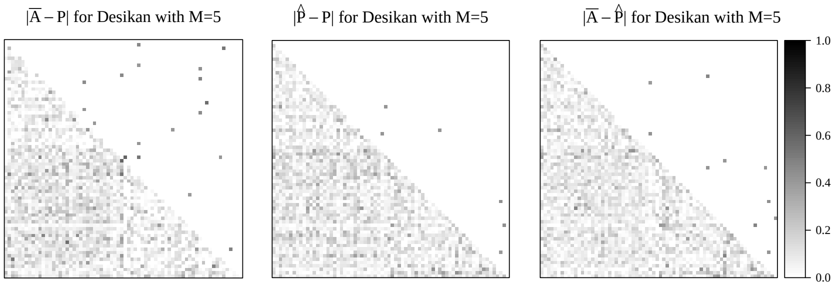


Fig 7: **Heat plots of absolute estimation error for $\bar{\boldsymbol{A}}$ and $\widehat{\boldsymbol{P}}$ (lower triangle) and absolute errors above 0.4 (upper triangle).** These heat plots show the absolute estimation error $|\bar{A}-P|$, $|\widehat{P}-P|$ and $|\bar{A}-\widehat{P}|$ for a sample of size $M=5$ from the Desikan dataset. The embedding dimension for $\widehat{P}$ 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 $\widehat{P}$ are highlighted indicates that $\widehat{P}$ has fewer large outliers compared to $\bar{A}$.

and hence $\bar{A}$ will always have zero error for those pairs of vertices. Nonetheless, $\widehat{P}$ 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 $|\bar{A}_{ij}-P_{ij}|-|\widehat{P}_{ij}-P_{ij}|$ according to the location of the corresponding regions in the brain. Red edges indicate that $\widehat{P}$ overestimates P, while blue means that $\widehat{P}$ underestimates P. The edge width is determined by the estimation error for $\widehat{P}$, where pairs with larger estimation error are represented by thicker lines. We also highlight the five regions corresponding to vertices that contribute most to the difference, meaning the vertices i with the largest value of $\sum_j(|\bar{A}_{ij}-P_{ij}|-|\widehat{P}_{ij}-P_{ij}|)$. 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 parsopercularis regions in the right hemisphere of the Desikan atlas.

4.3.2 Exploration of Dimension Selection Procedures

To further investigate the impact of the dimension selection procedures, we also considered all possible dimensions for $\widehat{P}$ by ranging d from 1 to N in order to investigate the impact of the dimension selection procedures. We plot $\widehat{\text{MSE}}$ of $\bar{A}$ and $\widehat{P}$ in Fig. 9. The horizontal axis gives dimension d, which only impacts $\widehat{P}$, which is why estimated MSE of $\bar{A}$ is shown as flat.

When d is small, $\widehat{P}$ underestimates the dimension and throws away impor-

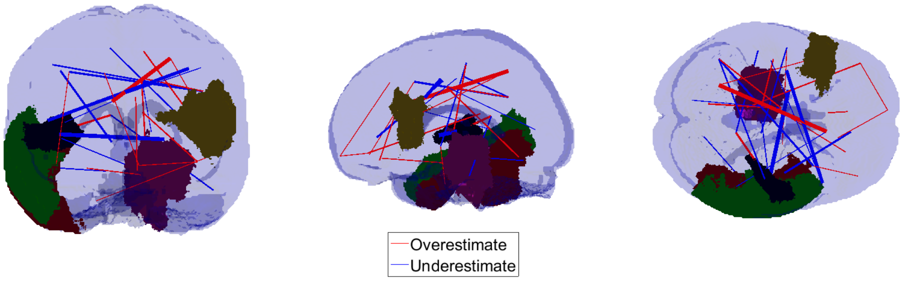


Fig 8: **Top 5 regions of the brain (vertices in graphs) and top 50 connections between regions (edges in graphs) with the largest differences $|\bar{A}_{ij} - P_{ij}| - |\widehat{P}_{ij} - P_{ij}|$.** Red edges indicate that $\widehat{P}$ overestimate P while blue means that $\widehat{P}$ underestimates 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 $\widehat{P}$ outperforms $\bar{A}$ the most for estimating P.

tant information, which leads to relatively poor performance. When $d = N$, $\widehat{P}$ is equal to $\bar{A}$, so that the curve for $\widehat{\text{MSE}}$ for $\widehat{P}$ ends at $\widehat{\text{MSE}}(\bar{A})$.

In the figure, we denote the 3rd elbow found by the Zhu and Ghodsi method by a triangle, and denote the dimension selected by USVT with threshold 0.7 by a square. 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 $\widehat{\text{MSE}}$. Meanwhile, $\widehat{P}$ reduces the variance by taking advantages of inherent low-rank structure of the mean graph. Such smoothing effect is especially obvious while we only have 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, $\widehat{P}$ smooths the connectomes estimate and improves the performance. Additionally, we see that there is a large range of dimensions where the performance for $\widehat{P}$ 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 $\widehat{P}$.

4.3.3 Challenges of the CoRR Dataset

While our estimator $\widehat{P}$ 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, we first check whether the dataset has the low-rank property or not. In Fig. 10, we plot the relative error $\|\text{lowrank}_d(P) - P\|_F^2/\|P\|_F^2$ of using a rank-d approximation of P (see Algorithm 2) as solid curves. The rate at which this curve tends to zero provides an indication of the relative performance of using $\widehat{P}_d$ as compared to $\bar{A}$ when M is large. For all three atlases, while these error rates do tend to zero relatively quickly, substantial

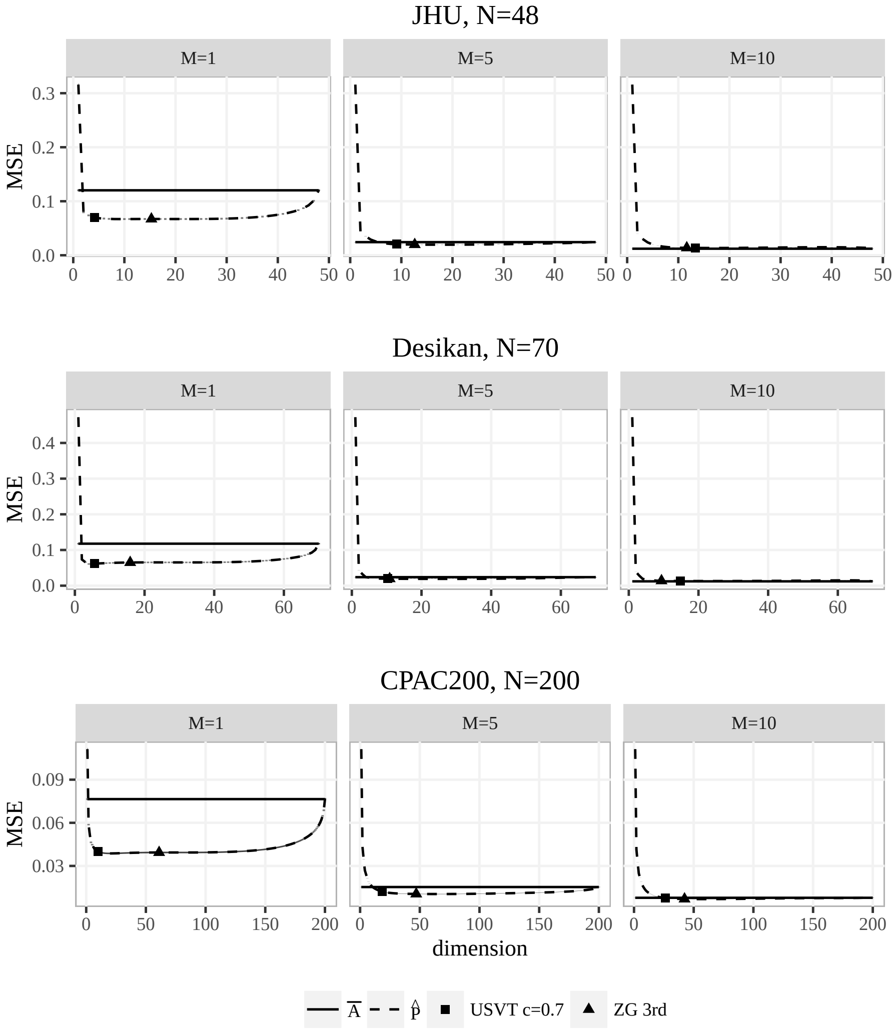


Fig 9: **Comparison of $\widehat{\text{MSE}}$ of $\widehat{P}$ and $\bar{A}$ for three atlases at four sample sizes for the CoRR data.** These plots show the mean squared error for $\bar{A}$ (solid line) and $\widehat{P}$ (dashed line) for three dataset (JHU, Desikan, and CPAC200) while embedding the 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, $\widehat{P}$ outperforms $\bar{A}$ with a flexible range of the embedding dimension including the average of the dimensions selected by Zhu and Ghodsi and USVT.

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. As can be expected, the CoRR dataset is actually high-rank, which violates the low-rank assumption.

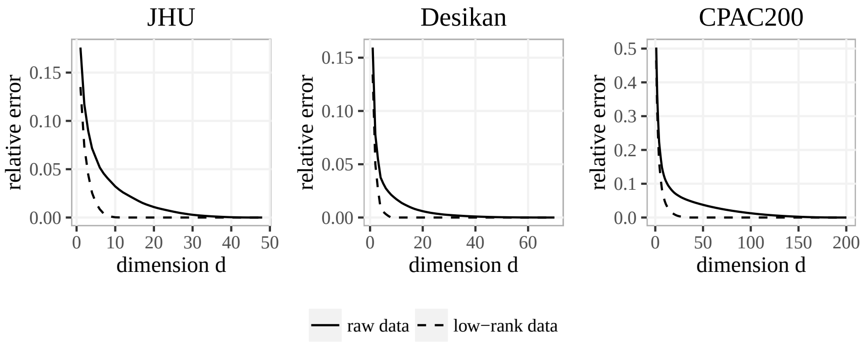


Fig 10: **Relative error of the rank-d approximation of the population mean.** The solid curves show the relative error $\|\text{lowrank}_d(P) - P\|_F^2 / \|P\|_F^2$ of using a rank-d approximation of P (see Algorithm 2) for three different atlases. The relative error decays relatively slowly when d 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 parts of this dataset provide challenges for our low-rank methods. First, there are a large number of negative eigenvalues which $\widehat{P}$ will not capture. We can adapt our low-rank methods by including large negative eigenvalues as well however we found that 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 zero error, while $\widehat{P}$ will necessarily give a less accurate estimate.

Despite all these challenges, our results show that when the sample size is relatively small, such as $M = 1$ or $M = 5$, and for the atlases with a larger number of vertices, $\widehat{P}$ still 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.

4.3.4 Lobe Structure behind the Low-rank Methods

In previous sections, we have shown that how the low-rank methods help us improve the accuracy of estimation while providing convenient interpretations simultaneously. Certainly there is more behind it. In this section, we focus on the lobe structure in particular.

The lobes of the brain is an anatomical classification, which has been shown to be related to different brain functions. Basically, there are 4 different lobes for each hemisphere, i.e. Frontal, Parietal, Occipital, and Temporal. For the Desikan atlas, there are 70 different regions (35 regions for each hemisphere). Each region belongs to one lobe. However, 8 regions of the Desikan atlas (Unknown, Banks of Superior Temporal Sulcus, and Corpus Callosum, for both hemispheres) do not have obvious lobe assignment. So we cluster them into a new lobe category named “other” to resolve this issue.

Generally, the regions within the same lobe should be more similar compared to the regions across the lobes. In order to see whether the embedded latent positions X preserve this property or not, we propose a test statistics T to be the average differences between vertices within the same lobe minus the average differences between vertices across different lobes, i.e.

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

where $l(i)$ represents the lobe assignment for vertex i. If the latent positions X and the lobe assignment l are independent, then we expect $T(X, l)$ to 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, which is evidence that the low-rank methods preserves the lobe structure.

However, the anatomical geometry might contribute to the dependence between X and l and we do not want to be distracted by this factor. So instead of testing the dependence between X and l, we are more interested in the following hypothesis test:

H_0: X and l are conditionally independent given anatomical geometry.

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

In this situation, we can focus on how much of the lobe structure is really captured by the low-rank methods without affected by the inherent spatial relationship. Note that this test is significantly underpowered compared to the test on a unconditional independence. To test under the anatomical geometry conditions, we permute the lobe assignment l in a way that the number of regions in each lobe remain the same and the regions within the same lobe are still spatially connected. In order to permute under such constraints, we define a flip to be a swap of two pairs of vertices which keeps the number of regions in each lobe, and do it several times. Details of the flipping procedure is discussed in Appendix B. For a fixed number of flips, we ran 1000 simulations and calculate the test statistics $T(X, l)$ after the permutation. We vary the number of flips from 1 to 10 and the results are shown in Fig. 11. In this violin plot, we use dashed line to represent the $T(X, l)$ based on the true lobe assignment. As the number of flips increases, $T(X, l)$ converges to the expected value under the assumption that X and l are independent in a constrained way mentioned above. And we can see that the $T(X, l)$ according to the true lobe assignment move away from the 95% confidence interval. We also calculate the corresponding p-values and labeled them in the figure. We can see that when the number of flips is larger than 7, the p-value is less than 0.05, which suggests that the latent positions based on the low-rank methods preserves the lobe structure regardless of the anatomical geometry.

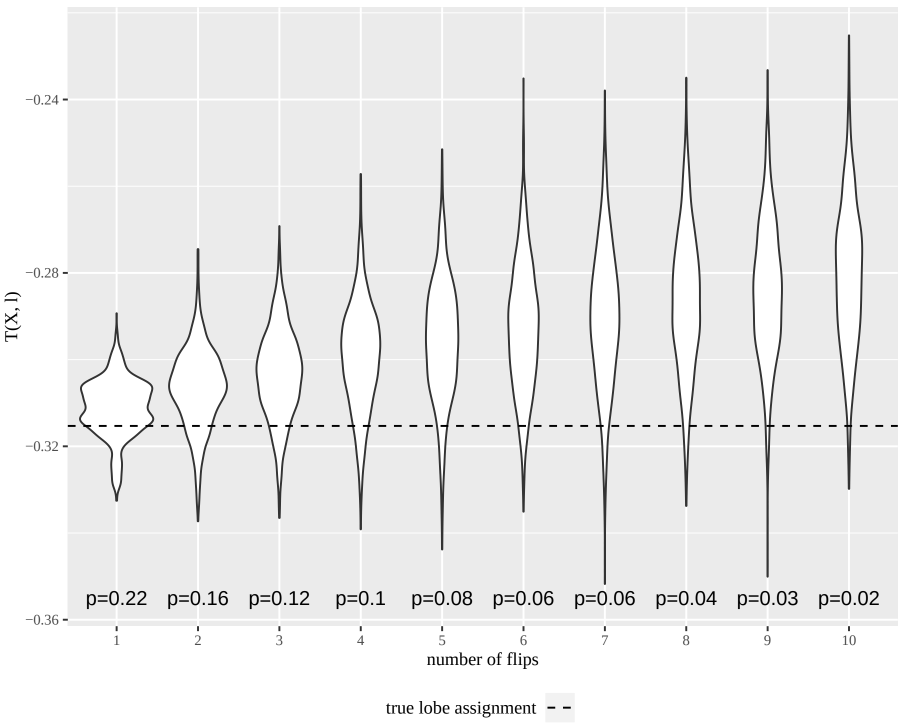


Fig 11: **Violin plot of the permutation test.** We run 1000 simulations for each number of flips. Dashed line represents the situation based on true lobe assignment. As the number of flips increases, $T(X, l)$ converges to the expected value under the assumption that X and l are independent in a constrained way mentioned above. And we can see that the $T(X, l)$ according to the true lobe assignment move away from the 95% confidence interval.

4.4 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 4.3, $\widehat{P}$ often performs well even when this assumption is false. To further illuminate this point, we performed a synthetic data analysis under a full-rank independent edge model where we used the sample mean of the 454 graphs in the Desikan dataset as the probability matrix P. As in the previous section, we simulated datasets from the full rank IEM distribution with probability matrix P of size $M = 1, 5$, and 10 and used $\bar{A}$ and $\widehat{P}$, where we varied the rank of $\widehat{P}$ from 1 to 70.

Fig. 12 shows the resulting estimated MSE for $\bar{A}$ (solid line) and $\widehat{P}$ (dashed line) for simulated data based on the full rank probability matrix P shown in the left panel of Fig. 1. We see that the results are very similar to those presented in Section 4.3, though overall $\widehat{P}$ performs even better than in the real data experiments. When M is small, $\widehat{P}$ outperforms $\bar{A}$ with a flexible range of the embedding dimension 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 less conclusive. This simulation again shows the robustness of $\widehat{P}$ to deviations from the RDPG model, specifically if the probability matrix is full-rank.

We also note that the finite-sample relative efficiency in these cases shows is even more favorable to $\widehat{P}$, with relative efficiencies lower than $1/3$ for $M = 1$, than for the real data, where relative efficiencies were at best around $1/2$ for $M = 1$. From this observation, we can postulate that the degradation in the performance of $\widehat{P}$ in real data can at least partially be attributed to the fact that the independent edge assumption does not hold for real data.

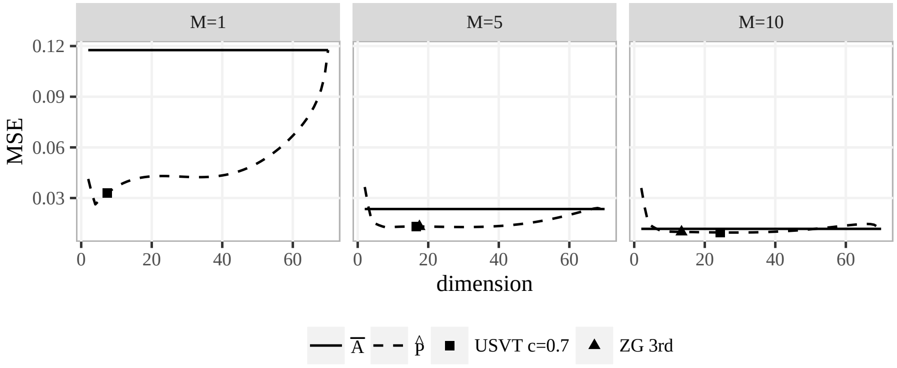


Fig 12: **Comparison of $\widehat{P}$ and $\bar{A}$ for synthetic data analysis.** As in Fig. 5, this figure shows $\widehat{\text{MSE}}$ for $\bar{A}$ (solid line) and $\widehat{P}$ (dashed line) for simulated data with different sample sizes M based on the sample mean for the Desikan dataset. Again, the average of dimensions selected by the USVT method (square) and the ZG method (triangle) tend to nearly approximate the optimal dimension. Overall, we see that 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 $\widehat{\text{RE}}(\bar{A}, \widehat{P})$ is lower for this synthetic data analysis than for the CoRR data.

5 Discussion

Motivated by the RDPG model, our methodology takes advantage of the low-rank structure of the graphs by applying 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 $\widehat{P}$ 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, we show that for a stochastic blockmodel with fixed number of blocks K, block size proportion ρ, and number of graphs M, the low-rank estimator $\widehat{P}$ 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 CoRR brain graphs and in low- and full-rank simulation settings. These results illustrate that $\widehat{P}$ performs well even when the low-rank assumption is violated and that $\widehat{P}$ 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 degraded performance slightly 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. [2014] used low-rank deviations from a full rank population mean to model collections of graphs 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 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 our analysis of the connectome dataset was the presence of structural ones in the mean graph for the population. These structural ones appear since edges between certain regions of the brain are present in all or nearly all members of the healthy population. The low-rank methods tend to miss these always-present edges while the sample mean will always capture them. Detecting and incorporating structural ones and zeros could yield methods that share the best elements of both methods considered here.

For the CoRR dataset, we used a cross-validation framework where we 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 $\widehat{P}$ 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.

6 Methods

6.1 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, we used the 3rd elbow in the experiments performed for this work.

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 our 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 again we found that 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, we have observed that in our setting, many dimensions will yield nearly optimal mean squared errors. Thus efforts to ensure the selected dimension is in the appropriate range are more important than finding the best dimension.

6.2 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 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. Earlier, Scheinerman and Tucker [2010] proposed using an iterative method to impute the diagonal entries. In this work, we combine these two ideas 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, we omit the diagonal entries since these are known to be zero.

6.3 Dataset Description

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 comprised 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 [neu, Kiar, 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. The resulting graphs are undirected, unweighted, and have no self-loops.

6.4 Outline for the Proof of the Theorems

Here we provide an outline of the proof of Lemma 4.1 which provides the approximate MSE of $\widehat{P}$ in the stochastic blockmodel case. The result depends on using the asymptotic results (see Theorem A.1) for the distribution of eigenvectors from Athreya et al. [2013] 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 $\mathrm{Var}(A_{ij}) = \frac{1}{M}P_{ij}(1-P_{ij})$. The results of Athreya et al. [2013] provide a central limit theorem for estimates of the latent position in an RDPG model for a single graph. Details of this theorem are in Theorem A.1. 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 $\widehat{P}_{ij} = \widehat{X}_i^\top \widehat{X}_j$ is a noisy version of the dot product of $\nu_s^\top \nu_t$ from Section 2.3 and each $\widehat{X}_i$ 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 $\widehat{P}_{ij}$ 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 $\widehat{P}_{ij}$ and $\bar{A}_{ij}$ is approximately $(\rho_{\tau_i}^{-1} + \rho_{\tau_j}^{-1})/N$ when N is sufficiently large.

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.

References

Neurodata's mri to graphs pipeline. `http://m2g.io`. Accessed: 2016-05-23.

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.

Avanti Athreya, CE Priebe, M Tang, V Lyzinski, DJ Marchette, and DL Sussman. A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A*, pages 1–18, 2013.

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.

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

Daniele Durante, David B Dunson, and Joshua T Vogelstein. Nonparametric bayes modeling of populations of networks. *arXiv preprint arXiv:1406.7851*, 2014.

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.

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

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. Gremlin: Graph estimation from mr images leading to inference in neuroscience. 2016.

Gregory Kiar, William Gray Roncal, Disa Mhembere, Eric Bridgeford, Randal Burns, Carey Priebe, and Joshua T. Vogelstein. m2g: A reliable and robust open-source one-click pipeline for mri to graph connectome estimation. In Preparation.

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.

Christine Leigh Myers Nickel. *Random dot product graphs: A model for social networks*, volume 68. 2007.

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.

W. Gray Roncal, Z. H. Koterba, D. Mhembere, D. M. Kleissas, J. T. Vogelstein, R. Burns, A. R. Bowles, D. K. Donavos, S. Ryman, R. E. Jung, L. Wu, V. Calhoun, and R. J. Vogelstein. Migraine: Mri graph reliability analysis and inference for connectomics. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 313–316, Dec 2013.

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

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.

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 Proofs for Theory Results

Here we present the proofs of the results in Section 4.1. To keep the ideas clear and concise, we leave out some details 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.

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 $\widehat{U}\widehat{S}\widehat{U}^\top$ be the best rank-d positive semidefinite approximation of $\bar{A}$, then we define $\widehat{P} = \widehat{X}\widehat{X}^\top$, where $\widehat{X} = \widehat{U}\widehat{S}^{1/2}$.

The proofs presented here will rely on a central limit theorem developed in Athreya et al. [2013]. We modify the theorem slightly to account for the multiple graph setting and present it in the special case of the stochastic blockmodel.

Theorem A.1 (Corrolary of Theorem 1 in Athreya et al. [2013])**.** *In the setting above, let $X = [X_1, \ldots, X_N]^\top \in \mathbb{R}^{K \times d}$ 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 $\tau_j = t$,*

$$\Pr\left\{\sqrt{n}(W\widehat{X}_i - \nu_s) \le z, \sqrt{n}(W\widehat{X}_j - \nu_t) \le z'\right\} = \Phi(z, \Sigma(\nu_s)/M)\Phi(z', \Sigma(\nu_t)/M) + o(1) \tag{3}$$

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 A.2. *In the same setting as Theorem 4.2, for any $1 \le s, t \le K$, we have*

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

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, we could assume that $\nu = US$ where $U = [u_1, \cdots, u_K]^\top$ is orthonormal in columns and S is a diagonal matrix. Here we can conclude that $\nu_s^\top = u_s^\top S$. Defining $R = \mathrm{diag}(\rho_1, \cdots, \rho_K)$, we have

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

Thus

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

□

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

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

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

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

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

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

and

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

Also, Corollary 4.11 in Athreya et al. [2013] says $\widehat{X}_i$ and $\widehat{X}_j$ are asymptotically independent. Thus, conditioning on $X_i = \nu_s$ and $X_j = \nu_t$, we have $\lim_{n\to\infty} \mathbb{E}[\widehat{X}_i^\top \widehat{X}_j] = \lim_{n\to\infty} \mathbb{E}[(W_n\widehat{X}_i)^\top]\mathbb{E}[W_n\widehat{X}_j] = \nu_s^\top \nu_t = P_{ij}$.

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

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

and

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

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

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

□

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 $\widehat{P}$.

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

Here we describe the details of the flipping procedure for the permutation test mentioned in Section 4.3.4.

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. We denote such spatial adjacency by an adjacency matrix S for the 70 regions, where $S_{ij} = 1$ means region i and region j are contain a pair of voxels, v_i and v_j, which are spatially adjacent. If this is true, then we say region j is a neighbor of region i. We denote the lobe i.d. for region i by l_i.

Now we define a uniform 1-flip to be:

1. Select 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 select 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. Reassign region j_1 to lobe l_{i_1} and reassign region i_2 to lobe l_{j_2}.

By the 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 keeps the same.

In the permutation test, we apply a uniform k-flip and calculate the test statistic $T(X, l)$ 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.