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, USA
2 Department of Biomedical Engineering, The Johns Hopkins University, Baltimore, MD, USA
3 Department of Mathematics & Statistics, Boston University, Boston, MA, USA
* sussman@bu.edu

Abstract

Estimating the mean of a population of graphs based on a sample is a core problem in network science. While using the element-wise sample mean of the adjacency matrices is a common approach, this method does not exploit any underlying structure of graphs. We propose using a low-rank method together with tools for dimension selection and diagonal augmentation to improve performance over the naive methodology. 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 the standard sample mean. These results indicate that low-rank methods should be a key part of the tool box for researchers studying populations of graphs.

Author Summary

Sampling from populations of networks is common task in areas such as connectomics and gene expression analysis. When confronted with a sample from a population of graphs estimating the mean graph is a non-trivial task since the number of parameters in the mean graph grows quadratically with the number of vertices in the graph. We propose using low-rank estimates of the mean graph in order to reduce the variance of the estimate in exchange for a slight increase in bias. This results in an estimator which is often superior to using the naive element-wise mean, especially when there are very few graphs in the sample. We demonstrate the improved performance via theoretical results for the stochastic blockmodel as well as in an extensive analysis of connectomic data derived from diffusion tensor magnetic resonance imaging.

1 Introduction

Estimation of the mean of a population based on a sample 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. Nowadays we take averages almost everywhere, from data in Euclidean space to more complex objects, like shapes, documents and graphs.

The mean of a population of graphs is a high dimensional object, consisting of $O(N^2)$ parameters for a graph with N vertices. Estimating high dimensional estimands with a small sample size using naive unbiased methods often leads to inaccurate estimates with very high variance. By exploiting the bias-variance trade-off, it is often fruitful to develop estimators which have some bias with 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.

Additionally, in a striking result, Stein (1956) and James and Stein (1961) showed the arithmetic average should not always be the first choice and is inadmissable in even simple settings by todays standards. Twenty-seven years later, Gutmann (1982) proved that this phenomenon cannot occur when the sample spaces are finite. But even when the sample mean is admissible, it doesn't close the door to other estimators in all cases. In many situations where other structural information is hypothesized, for instance a collection of graphs as considered in this paper, 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. As with real valued data, one may want to define the mean of a graph to itself be a graph such as for 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. Instead, we define the mean graph as the weigted adjacency matrix with weights given by the proportion of times the corresponding edge appears in the population. This population mean is becoming more and more important both in statistical inference and in various applications like connectomics, social networks, and computational biology.

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.

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, it does not perform very well especially when only a small sample size is available. Intuitively, an estimator incorporating the mean-graph structure is preferable to the entry-wise MLE. 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. 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 a LPM which captures this middle ground is the random dot product graph (RDPG) Nickel (2007); Young and Scheinerman (2007) which motivates our estimator.

Using the estimates of the latent positions based on a truncated eigen-decomposition of the adjacency matrix, in the RDPG setting we consider an estimator for the mean of the collection of graphs which captures the low-rank structure of the RDPG model. These estimates will improve performance since they will be biased towards the low-rank structure of the RDPG model and will have significantly lower overall variance than naive element-wise sample means. In this study, we show via theory, simulations and real data analysis that it frequently outperforms the element-wise MLE, especially

in small sample sizes. 59

2 Models and Estimators 60

This work considers the scenario of having M graphs represented as adjacency matrices, 61
$\{A^{(m)}\}$ $(m = 1, \cdots, M)$, each having N vertices. We assume there is a known 62
correspondence for vertices in different graphs, so that vertex i in graph m corresponds 63
to vertex i in graph m' for any i, m, m'. The graphs we consider are undirected and 64
unweighted with no self-loops, so each $A^{(m)}$ is a binary symmetric matrix with zeros 65
along the diagonal. An example application of this arises in the field of connectomics, 66
where functional brain imaging data for each subject can be represented as a graph. 67
Each vertex represent a well defined anatomical region present in each subject, and an 68
edge between two regions is defined to exist if correlation in activity between the regions 69
surpasses a certain threshold. Similarly, for structural brain imaging an edge may 70
represent the presence of anatomical connections between the two regions. 71

For the purpose of this paper, we will also assume that the graphs are sampled 72
independently and identically from some distribution. We consider three nested models 73
for these distributions, the independent edge model, the random dot product model, 74
and the stochastic blockmodel, and two estimators motivated by these models. 75

2.1 Independent Edge Model 76

The first model we consider is the independent edge model (IEM) with parameter 77
$P \in [0,1]^{N \times N}$ (Bollobás et al., 2007). An edge exists between vertex i and vertex j with 78
probability P_{ij} and each edge is present independently of all other edges. For this case, 79
the mean graph is the For this case, we aim to estimate the mean matrix $P = \mathbb{E}[A^{(m)}]$ 80
base on the observed adjacency matrices $A^{(1)}, \ldots, A^{(M)}$. 81

2.2 Estimator $\bar{A}$ 82

Under the IEM, the MLE for the mean graph P is the element-wise sample mean of the 83
adjacency matrices $\bar{A} = \frac{1}{M} \sum_{m=1}^{M} A^{(m)}$. It is unbiased with entry-wise variance 84
$\mathrm{Var}(\bar{A}_{ij}) = P_{ij}(1 - P_{ij})/M$. Moreover, $\bar{A}$ is the uniformly minimum-variance unbiased 85
estimator, so it has the smallest variance among all unbiased estimators and enjoys the 86
many asymptotic properties of the MLE as $M \to \infty$. 87

However, $\bar{A}$ doesn't exploit any graph structure, and sometimes the performance is 88
not very good especially when M is small. For example, when $M = 1$, $\bar{A}$ is exactly the 89
binary graph we observe, which is an inaccurate estimate for an arbitrary P compared 90
to estimates which exploit underlying structure. Additionally, there are no useful 91
asymptotic properties for $\bar{A}$ as the number of vertices N becomes large. 92

2.3 Random Dot Product Graph 93

In graphs, the adjacencies between vertices generally depend on unobserved properties 94
of the corresponding vertices. For example, in a connectomics setting, the two brain 95
regions with similar properties will have similar connectivity patterns to other regions of 96
the brain. The latent positions graph (LPG) model proposed by Hoff et. al. (2002) Hoff 97
et al. (2002) captures such structure, where each vertex is associated with a latent 98
position that influences the adjacencies for that vertex. In this model, each vertex i has 99
an associated latent vector $X_i \in \mathbb{R}^d$. Based on the latent positions, the existence of 100
edges are conditionally independent with probability that only depends on the latent 101
vectors of the incident vertices through a link function. If d is much smaller than the 102

number of vertices N and the link function is known, LPMs are more parsimonious models compared to IEM, requiring only dN parameters rather then $\binom{N}{2}$.

A specific instance of a LPM that we examine in this work is the random dot product graph model (RDPG) Nickel (2007); Young and Scheinerman (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.

Formally, let $\mathcal{X} \subset \mathbb{R}^d$ be a set such that $x, y \in \mathcal{X}$ implies $\langle x, y \rangle = \sum_i x_i y_i \in [0, 1]$. Let $X_1, \ldots, X_n \in \mathcal{X}$ and write $X = [X_1 | \cdots | X_N]^T \in \mathbb{R}^{N \times d}$. A random graph G with adjacency matrix A is said to be an RDPG if

$$\Pr[A|X] = \prod_{i<j} \langle X_i, X_j \rangle^{A_{ij}} \left(1 - \langle X_i, X_j \rangle\right)^{1 - A_{ij}}.$$

In the RDPG model, each vertex i is associated with latent position X_i and conditioned on the latent positions X, the edges $A_{ij} \overset{iid}{\sim} \mathrm{Bern}(\langle X_i, X_j \rangle)$. Note that the probability matrix is the outter product of the latent position matrix with itself, $P = XX^T$. This imposes two properties on P, namely that P is positive-semidefinite and $\mathrm{rank}(P) = \mathrm{rank}(X) \leq d$. These properties lead us to our proposed estimator.

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

Motivated by the low-rank structure of the RDPG mean matrix, we consider the estimator $\hat{P}$ 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 et al. (2015) but we analyze its performance more specifically in the random graph setting.

For a given dimension d we consider the estimator $\hat{P}$ defined as the best rank-d positive-semidefinite approximation of $\bar{A}$. Since the graphs are symmetric, we can compute the eigendecomposition of $\bar{A}$ as $\hat{U}\hat{S}\hat{U}^T + \tilde{U}\tilde{S}\tilde{U}^T$, where $\hat{S}$ is a diagonal matrix with non-increasing entries along the diagonal corresponding to the largest d eigenvalues of A and $\hat{U}$ has columns given by the corresponding eigenvectors. The d-dimensional adjacency spectral embedding (ASE) of $\bar{A}$ is given by $\hat{X} = \hat{U}\hat{S}^{1/2} \in \mathbb{R}^{N \times d}$. For an RDPG, the rows of $\hat{X}$ are estimates of the latent vectors for each vertex (Sussman et al., 2014). Using the adjacency spectral embedding, we have that $\hat{P} = \hat{X}\hat{X}^T = \hat{U}\hat{S}\hat{U}^T$.

To compute $\hat{P}$, we need to specify what rank d to use and there are various ways of dealing with dimension selection. In this paper, we use Zhu and Ghodsi's elbow selection method Zhu and Ghodsi (2006) and the universal singular value thresholding (USVT) method Chatterjee et al. (2015). Details are discussed in Section 5.1.

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 $\hat{P}$ based only on $\bar{A}$. To compensate for this issue, we use an iterative method developed by Scheinerman and Tucker Scheinerman and Tucker (2010). Details are discussed in Section 5.2.

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

1: **Input:** $A^{(1)}, A^{(2)}, \cdots, A^{(M)}$, each $A^{(m)} \in \{0,1\}^{N \times N}$, $k_m ax$

2: Calculate the sample mean $\bar{A} = \frac{1}{M} \sum_{m=1}^{M} A^{(m)}$;

3: Calculate the scaled degree matrix $D^{(0)} = \text{diag}(\bar{A}\mathbf{1})/(n-1)$;

4: Select the dimension d using $\bar{A} + D^{(0)}$ (see Section 5.1);

5: Calculate the rank-d approximation $\hat{U}\hat{S}\hat{U}^T$ for $\bar{A} + D^{(0)}$;

6: Update the diagonal to $D^{(1)} \leftarrow \text{diag}(\hat{U}\hat{S}\hat{U}^T)$ (see Section 5.2);

7: Set $\tilde{P} \rightarrow \hat{U}\hat{S}\hat{U}^T$, the rank-$d$ approximation for $\bar{A} + D^{(1)}$;

8: **Output**: $\hat{P}$ equal to $\tilde{P}$ with values < 0 set to 0 and values > 1 set to 1.

Algorithm 1 gives the steps involved to compute the low-rank estimate $\hat{P}$. As we will see in the proceeding 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 next sections, we explore this estimator in the context of the stochastic blockmodel.

2.5   Stochastic Block Model as an 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} is independently sampled from the Bernoulli distribution with parameter B_{τ_i, τ_j}.

In order to analyze the estimator $\hat{P}$ motivated by RDPG, we move to another representation of SBM based on RDPG. Consider a positive semi-definite K-block SBM with a rank $d \leq K$ block probability matrix B, we can always decompose B into $\nu\nu^T$, where $\nu \in \mathbb{R}^{K \times d}$ and each row ν_k is the shared latent position for all vertices assigned to block k. For $X \in \mathbb{R}^{N \times d}$ with rows given by $X_i = \nu_{\tau_i}$, we have

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

In this way, the SBM can be seen as an RDPG where all vertices in the same block will have identical latent positions.

An example SBM is illustrated in Fig 1. We consider a 5-block SBM and plot the corresponding probability matrix and one adjacency matrix generated from it with 200 vertices. From 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. We also see that using the low-rank approximation $\hat{P}$ performs much better to estimate P compared to usingt he sample mean $\bar{A}$.
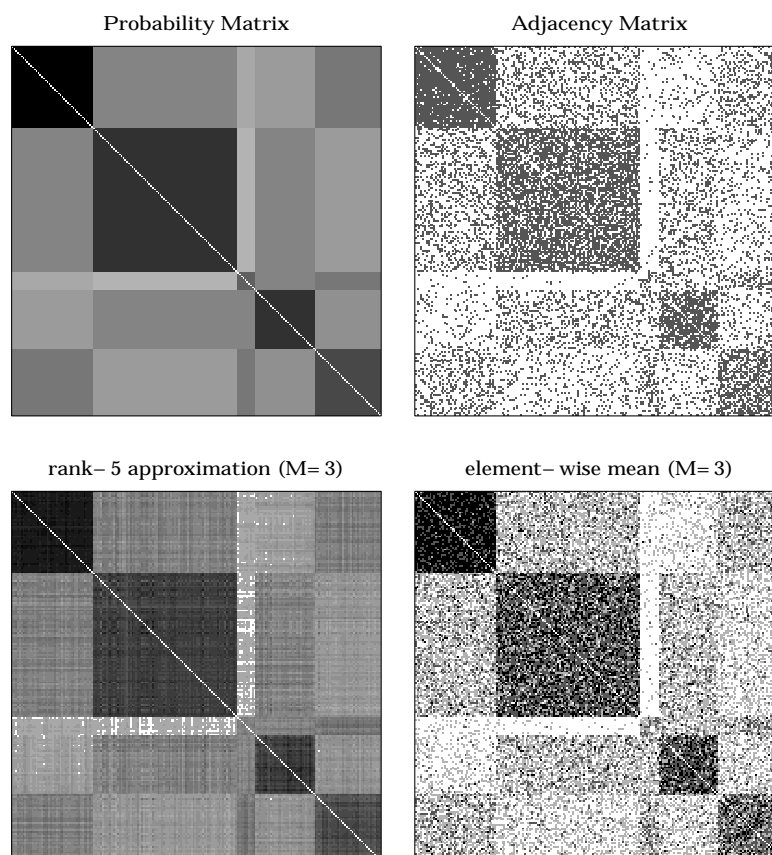
Fig 1. Example illustrating the stochastic blockmodel. The top left figure shows the mean graph P with $K = 5$ blocks and $N = 200$ vertices and the top right figure shows an adjacency matrix A sampled according to the probabilities from P. While A is a noisy version of P, much of the structure of P 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 2.2. Finally, by taking a rank-5 approximation of $\bar{A}$ and thresholding the values to be between O and 1, we construct our proposed estimate $\hat{P}$, shown in the lower left panel (see Section 2.4). By visual inspection, it is clear that the low-rank estimate $\hat{P}$ more closely approximates the probability matrix P as compared to $\bar{A}$.

3 Results

3.1 Asymptotic Theory

To estimate the mean of a collection of graphs, we consider the two estimators from Section 2: the entry-wise sample mean $\bar{A}$ and the low-rank $\hat{P}$ motivated by the RDPG. In this section, we analyze the performance of these two estimators under the SBM by computing the entry-wise relative efficiency (RE), defined as $\text{RE}(\bar{A}_{ij}, \hat{P}_{ij}) = \frac{\text{MSE}(\hat{P}_{ij})}{\text{MSE}(\bar{A}_{ij})}$. Specifically, we consider the asymptotic relative efficiency as the number of vertices $N \to \infty$ but with the number of graphs M fixed.

For this asymptotic framework, we assume where the block memberships τ_i are drawn iid from a multinomial distribution with block membership probabilities given by $\rho \in [0, 1]^K$. We will also assume that for a given N, the block membership probability are fixed for all graphs. Denote block probability matrix $B = \nu\nu^T$. By definition, the mean of the collection of graphs generated from this SBM is P, 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 $\hat{P}$.

Lemma 3.1. *For the above setting, for any i, j, if $\text{rank}(B) = K = d$, we have*

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

and for large enough N, we have

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

The proof of this lemma is outlined in Section 5.5 and is based on results for the variance of the adjacency spectral embedding from Athreya et al. (2013). From the result, we can see that the MSE of $\hat{P}_{ij}$ is of order $O(M^{-1}N^{-1})$ approximately.

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

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

This yields the following result.

Theorem 3.2. *In the same setting as in Lemma A.3, for any i and j, if $\text{rank}(B) = K = d$, the asymptotic relative efficiency (ARE) is*

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

and for large enough N, we have

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

This theorem indicates that under the SBM, $\hat{P}$ is a much better estimate of the mean of the collection of graphs P than $\bar{A}$. From the result, we see that the relative efficiency is of order $O(N^{-1})$ and $N \cdot \text{RE}(\bar{A}_{ij}, \hat{P}_{ij})$ converges to $1/\rho_{\tau_i} + 1/\rho_{\tau_j}$ when N goes to infinity. If for example there is one block containing half the vertices, the relative efficiency for estimating the probabilities of edges between vertices in that block will be approximately $4/N$. For even large blocks the relative will decrease towards $1/N$ where as if a block tends to have very few vertices, so that the the asymptotic relative
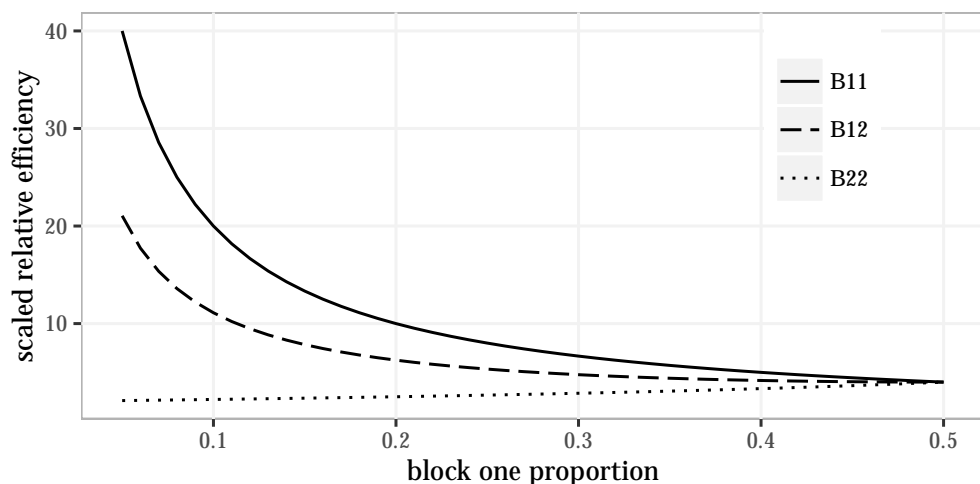
Fig 2. Theoretical scaled asymptotic relative efficiency $N \cdot RE(\bar{A}, \hat{P})$ based on Theorem 3.2 for the three distinct edge probabilities of a two-block SBM as a function of the proportion ρ_1 of vertices in the first block These lines all meet at a scaled relative efficiency of 4 when $\rho_1 = 1/2$. Improvements using low-rank methods are greater for larger blocks, such as for B_{22} when $\rho_2 = 1 - \rho_1$ is close to 1, while the improvements are smaller for blocks with relatively few vertex pairs such as B_{11} and B_{12} when ρ_1 is small.

efficiency will tends towards 1, but will always be less than one. Also, the ARE does not depend on the number of graphs M, so the larger the graphs are, the better $\hat{P}$ is relative to $\bar{A}$.

To illustrate Theorem 3.2, Fig 2 shows $1/\rho_s + 1/\rho_t$, the scaled asymptotic RE, for each $s, t \in \{1, 2\}$ as ρ_1 changes from 0.01 to 0.50 for a two block model. 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 3 for simulated data.

If instead of assuming that the graphs follow a SBM distribution we assume the graphs are distributed according to a RDPG distribution, similar gains in relative efficiency can be realized. While there is no compact analytical formula for the relative efficiency of $\hat{P}$ versus $\bar{A}$ in the general RDPG case, using the same ideas as in Theorem 3.2, we can show that $RE(\bar{A}_{ij}, \hat{P}_{ij}) = O(1/N)$.

Remark 3.3. *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 will increase, making $\bar{A}$ and and the low-rank approximation at the optimal dimension closer together. As a result, $RE(\bar{A}, \hat{P})$ will increase as M increases for full-rank models.*

3.2 Finite Sample Simulations

In this section, we will illustrate the theoretical results from Section 3.1 regarding the relative efficiency between $\bar{A}$ and $\hat{P}$ via Monte Carlo simulation experiments. These numerical simulations will also allow us to investigate the finite sample performance of the two estimators.
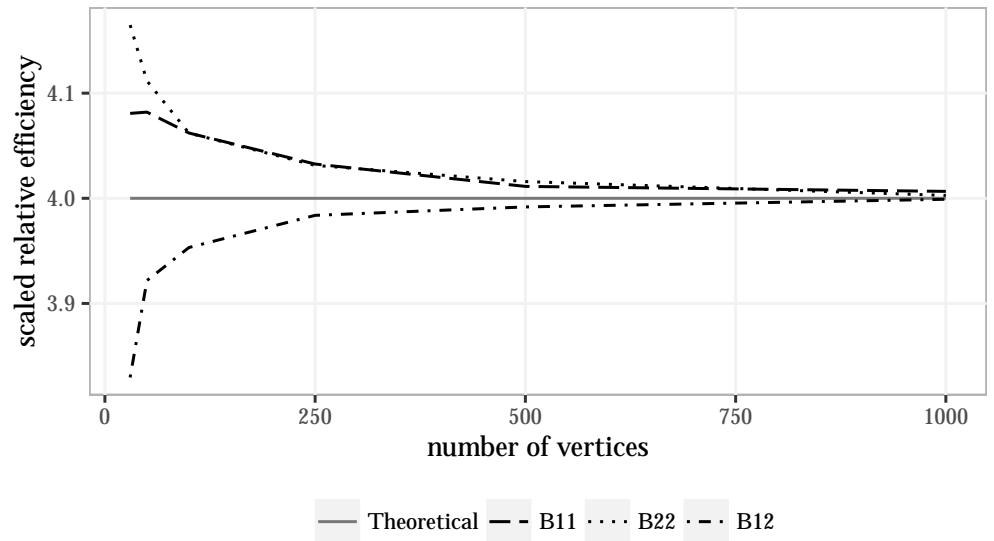
Fig 3. Scaled relative efficiency $N \cdot RE(\bar{A}, \hat{P})$ as a function N for fixed M based on 1000 Monte Carlo replicates of the SBM from Section 3.2. Each line represents the estimated scaled relative efficiency for one of the three distinct probabilities from the stochastic blockmodel. All the lines converge to four, the theoretical limit, as $N \to \infty$ since there are two blocks of equal size so $1/\rho_s + 1/\rho_t = 4$.

Here, we consider a 2-block SBM with parameters

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

When calculating $\hat{P}$, we omit the dimension selection step from Algorithm 1 and instead using the true dimension $d = \text{rank}(B) = 2$.

To investigate the finite sample relative efficiency, we first sample 1000 Monte Carlo replicates from the above SBM distribution with different number of vertices N and fixed number of graphs M. The scaled relative efficiency $N \cdot RE(\bar{A}_{ij}, \hat{P}_{ij})$ can be estimated since P is known for this simulation. Since the relative efficiency only depends on the blocks memberships of the pair i, j, we estimate the relative efficiency for each block pair using

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

for $s, t \in \{1, 2\}$, where $\hat{MSE}$ denotes the estimated mean square error based on the Monte Carlo replicates.

We plot the scaled relative efficiency $N \cdot \text{RE}_{st}(\bar{A}, \hat{P})$ in Fig 3. Based on the Theorem 3.2, we have that the scaled RE converges to $1/\rho_{\tau_i} + 1/\rho_{\tau_j} = 4$ as $N \to \infty$ for all i, j. This is plotted as a solid line. The different dashed lines denote the estimated scaled RE associated with different block pairs, either B_{11}, B_{12} or B_{22}. From the figure, we see that $N \cdot \text{RE}_{st}(\bar{A}, \hat{P})$ converges to scaled asymptotic RE quite rapidly. We omit error bars as the standard errors are very small for these estimates.

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 signicantly better than the estimates for edges within each block. The reason for this may be due the actual values of the true probability but it may also be do the fact there are approximately twice as many pairs of vertices in different blocks than there are in the same block. This could lead to an increase in effective sample size which may cause the larger differences displayed in the left parts of Fig. 3.

3.3 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 $\hat{P}$ is still useful in such cases, we test 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) (Gorgolewski et al., 2015; Zuo et al., 2014).

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 describe 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. 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 5.4.

In order to evaluate the performance of the two estimators, we use 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 $\hat{P}$ and the sample mean $\bar{A}$ using the M selected graphs. We then compare 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 3.4, it matches our definition as the proportion of times each edge appears in the population.

We run 1000 simulations on each of the three atlases for each sample size $M = 1, 5, 10$. We also considered all possible dimensions for $\hat{P}$ by ranging d from 1 to N in order to investigate the impact of the dimension selection procedures. We plot the MSE of $\bar{A}$ and $\hat{P}$ in Fig 4. The horizontal axis gives dimension d, which only impacs $\hat{P}$, which is why the MSE of $\bar{A}$ is shown as flat.

When d is small, $\hat{P}$ underestimate the dimension and throws away important information, which leads to relatively poor performance. When $d = N$, $\hat{P}$ is equal to $\bar{A}$, so that the curve of the MSE for $\hat{P}$ ends at the MSE for $\bar{A}$. In practice, we use algorithms like Zhu and Ghodsi's method or USVT to select the dimension d (see section 5.1). 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 square error.

When M is 1 or 5, $\bar{A}$ has large variance which leads to large MSE. Meanwhile, $\hat{P}$ reduces the variance by taking advantages of inherent low-rank structure of the mean graph. Additionally, we see that there is a large range of dimensions where the performance for $\hat{P}$ is superior to $\bar{A}$. With a larger M, the performance of $\bar{A}$ improves so that its performance is frequently superior to $\hat{P}$ but $\hat{P}$ still performs nearly as well.

For a more specific comparison of the performance of $\hat{P}$ and $\bar{A}$ in the most realistic situation where the rank d much be chosen based on data, Table 1 shows relative efficiencies of $\hat{P}$ versus $\bar{A}$. For each atlas and each sample size we compare the Zhu and Ghodsi (2006) with the USVT method and note that both perform reasonable well

relative to the full-dimensional $\bar{A}$. Again we can see that the largest improvements using $\hat{P}$ occur when M is small and N is lareg. On the other hand once $m = 10$ is , $\bar{A}$ tends to do nearly as well or better than the $\hat{P}$. Nonetheless, $\hat{P}$ offers certain advantages, especially since low-rank estimates can often be more easily interpretable by considering the latent position representation described in Section 2.5.

Table 1. Relative efficiencies for estimating the mean graph for each atlas for the CoRR data set. We compared different sample sizes m and different dimension selection procedures, ZG and USVT. Overall, the relative efficiencies are greater for smaller sample sizes m and larger number of vertices N.

JHU	$m = 1$	$m = 5$	$m = 10$
ZG	0.54	0.81	1.14
USVT	0.56	0.84	1.12
Desikan	$m = 1$	$m = 5$	$m = 10$
ZG	0.52	0.81	1.14
USVT	0.50	0.81	1.09
CPAC200	$m = 1$	$m = 5$	$m = 10$
ZG	0.51	0.68	0.89
USVT	0.52	0.79	0.99

To further illustrate the differences betweent the two estimators, we consider a single random sample of size $M = 5$ based on the Desikan atlas. We calculated $\bar{A}$ and $\hat{P}$, using Zhu and Ghodsi's 3rd elbow to select $d = 11$. In Fig 5, the estimates $\bar{A}$ and $\hat{P}$ as well as the sample mean of 454 graphs (as a close estimate of P) are plotted. 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, $\hat{P}$ has a finer gradient of values which in this case leads to a more accurate estimate.

Moreover, Fig 6 shows the values for the absolute estimation error $|\bar{A} - P|$ and $|\hat{P} - P|$. The lower triangular sections shows 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 $\hat{P}$ being highlighted in the figure, further indicating the superior performance of $\hat{P}$. Note that $\approx 13\%$ of the potential edges are present in all 454 graphs and hence $\bar{A}$ will always have zero error for those pairs of vertices. Nonetheless, $\hat{P}$ typically outperforms $\bar{A}$.

To investigate the difference in performance with respect to the geometry of brain, in Fig 7, we plot the 50 edges with the largest difference $|\bar{A} - P| - |\hat{P} - P|$ according to the location of the corresponding regions in the brain. Red edges indicate that $\hat{P}$ overestimate P while blue means that $\hat{P}$ underestimate P. The edge width is determined by the estimation error for $\hat{P}$, where pairs with larger estimation error are represented by thicker lines. We also highlight the regions corresponding to vertices that contribute most to the difference, meaning the vertices i with the largest value of $\sum_j (|\bar{A} - P| - |\hat{P} - P|)_{ij}$. Notably, these regions of improvement appear to be concentrated in three distinct regions of the brain. 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.

These results demonstrate that overall $\hat{P}$ gives a better estimate than $\bar{A}$ for the CoRR dataset with all three atlases. Importantly, this improvement is robust to the embedding dimension as long as provided the dimension is not underestimated. We

JHU, N=48

Desikan, N=70

CPAC200, N=200

—— Abar   - - Phat   ■ USVT c=0.7   ▲ ZG 3rd
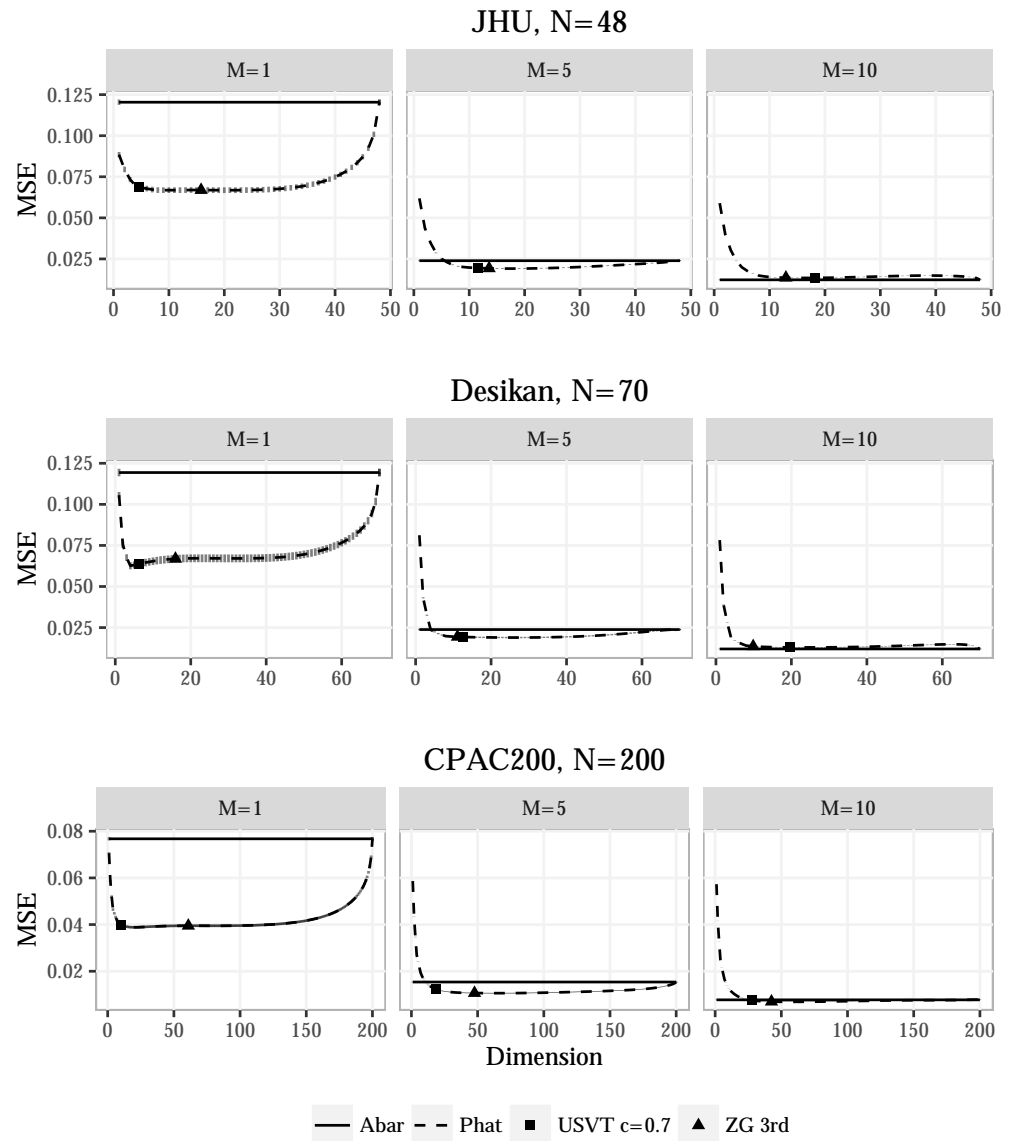
Fig 4. A comparison of the mean square error for $\bar{A}$ (solid line) and $\hat{P}$ (dashed line) for three dataset (JHU, Desikan and CPAC200) while embedding the graphs into different dimensions and with different sample sizes M. The average dimensions chosen by the 3rd elbow of Zhu and Ghodsi is denoted by a triangle and chosen by USVT with threshold equals 0.7 is denoted in square. Vertical intervals, visible mainly in the $N = 48, 70$ and $M = 1$ plots, represent the 95% confidence interval for the mean square errors. When M is small, $\hat{P}$ outperforms $\bar{A}$ with a flexible range of the embedding dimension including the average of the dimensions selected by Zhu and Ghodsi and USVT.

should note that though the total error for $\hat{P}$ is smaller in this case, for certain entries $\bar{A}$ does perform better. For example, there are some pairs of vertices that are adjacent for all samples in the population and for these pairs $\bar{A}$ will always perform better than $\hat{P}$.

335
336
337

P for Desikan $\bar{A}$ for Desikan with M=5 $\hat{P}$ for Desikan with M=5
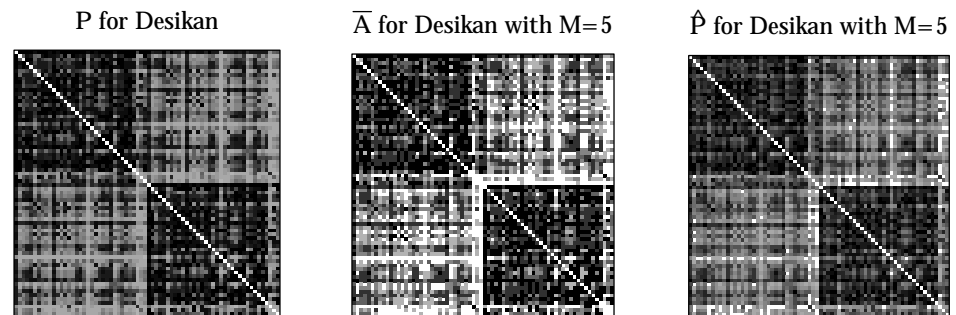
Fig 5. Heat maps indicatign the sample mean for the remaining $454 - 5$ graphs (left), sample mean for the 5 sampled graphs (center), and $\hat{P}$ for the 5 sampled graphs with dimension $d = 11$ selected using the Zhu and Ghodsi method (right). Darker pixels indicated a higher probability of an edge between the given vertices. Note that $\hat{P}$ appears to be better estimate the true probabilities P, especially for edges between the two hemispheres, in the upper right (and lower left) block.

## 3.4 Simulation under the Full Rank Independent Edge Model

While the theory we have developed is based on the assumption that the mean graph is low rank, as we have seen in Section 3.3, $\hat{P}$ often performs well even when this assumption is false. To further illuminate this point, we perform a simulation under a full-rank independent edge model where we use the sample mean of the 454 graphs in the Desikan dataset as the probability matrix P. As in the previous section, we simulated data sets of size $M = 1, 5$, and 10 and used $\bar{A}$ and $\hat{P}$, where for $\hat{P}$ we varied the rank from 1 to 70.

Fig 8 shows resulting estimated MSE for $\bar{A}$ (solid line) and $\hat{P}$ (dashed line) for simulated data based on the full rank probability matrix P shown in the left panel of Fig 5. We see that the results are very similar to those presented in section 3.3, though overall $\hat{P}$ performs even better than in the real data experiments. When M is small, $\hat{P}$ outperforms $\bar{A}$ with a flexible range of the embedding dimension including those selected by the Zhu and Ghodsi method while 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 $\hat{P}$ to deviations from the RDPG model, specifically if the probability matrix is full-rank. We also see that the degradation in the performance of $\hat{P}$ in real data can at least partially be due to the fact that the independent edge assumption does not hold for real data.

4 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 $\hat{P}$ in the case of a stochastic blockmodel, demonstrating that when N is sufficiently low-rank methods provide a substantial improvement. 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 $\hat{P}$ performs well even when the low-rank assumption is violated and that $\hat{P}$ is robust and can be applied in practice.

One of the key observations from our real data analysis was that the largest improvements using the low-rank method occurred when the number of graphs M was
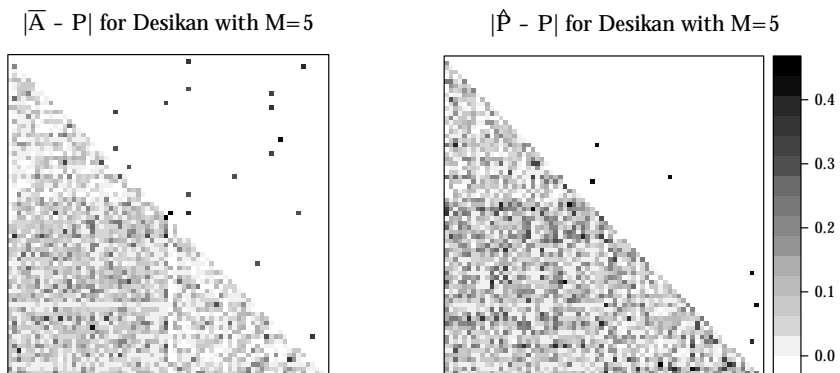
Fig 6. Heat plot of the absolute estimation error $|\bar{A} - P|$ and $|\hat{P} - P|$ for a sample of size $M = 5$ from Desikan dataset while embedding the graphs into dimension $d = 11$ selected by the 3rd elbow of ZG method. The lower triangular matrix shows the actual absolute difference, while the 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 while only six edges from $\hat{P}$ are highlighted indicates that $\hat{P}$ suffers from fewer large outliers.
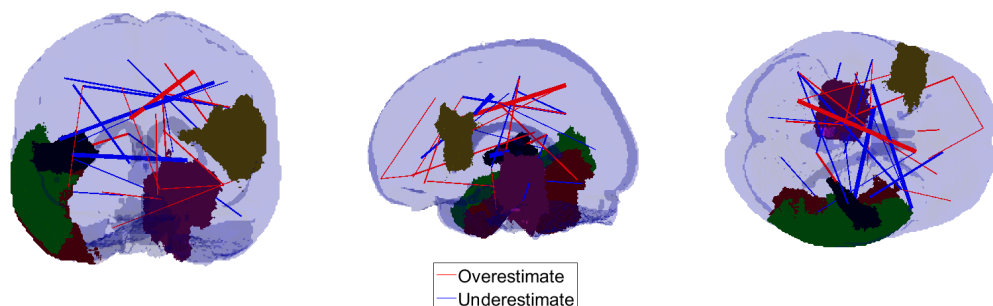
Fig 7. Top 5 regions of the brain (vertices in graphs) and top 50 connections between regions (edges in graphs) with largest difference $|\bar{A} - P| - |\hat{P} - P|$. Red edges indicate that $\hat{P}$ overestimate P while blue means that $\hat{P}$ underestimate P. The edge width is determined by the estimation error. Connections with larger estimation error are represented by thicker lines. This figure shows the regions and connections of the brain where $\hat{P}$ outperforms $\bar{A}$ the most for estimating P.

small and that it provided only minor improvements or degrade performance slightly when M is large. However, even in large scale studies the low-rank methods will be useful for estimating graph means for subpopulations, for example 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. In similar direction, 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 these ideas.

We also note that low-rank methods can often be more easily interpreted. By representing a low-rank matrix in terms of the latent position, where each vertex is represented as a vectors in $\Re^d$ and the entries of the matrix are given by inner products of these vectors (see Section 2.5, 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
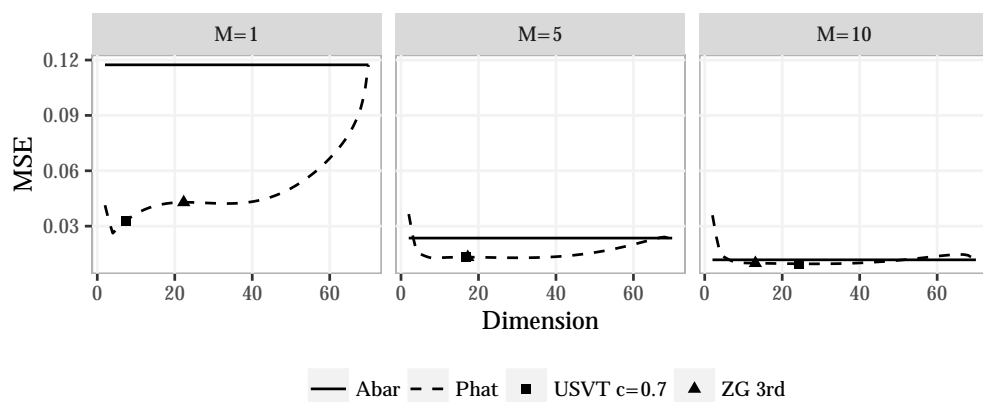
Fig 8. Comparison of MSE between $\bar{A}$ (solid line) and $\hat{P}$ (dashed line) for simulated data with different sample sizes M based on the sample mean for the Desikan dataset. Again, the average of dimensions selected by the USVT method (square) and the ZG method (triangle) tend to nearly approximate the optimal dimension. Overall, we 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.

graph. Additionally, such a representation allows the use of techniques from multivariate analysis to further study the estimated population mean.

While the low-rank methods considered in this paper will often offer substantial improvements, further refinements of these methods are still necessary. For example, we assume that the adjacency matrix is observed without contamination, however, in practice there will be noise in the observed graph and accounting for this noise with other more robust methods is necessary. 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.

Another issue that arose in our analysis of the connectome data set 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 data set, 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 $\hat{P}$ to the mean for the entire population including the subsample. Both of these analyses lead to very similar results in the cases presented above but for various reasons one may prefer one analysis over another. The cross validation method is most reasonable from a prediction perspective where prediction about new samples is of interest. If instead, one is interested in learning directly about the mean of a population, especially a finite population, the sub-sampling approach may be the most logical choice.

5 Methods

5.1 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 use the 3rd elbow in the experiments performed in this work.

Universal Singular Value Thresholding (USVT) is a simple estimation procedure proposed by Chatterjee et al. (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 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 square errors and so efforts to ensure the selected dimension are in the appropriate range are more important than finding the singular best dimension.

5.2 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 exasperated since the variance of the others entries is lower, so the relative impact of the bias in the diagonal entries is higher.

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.

5.3 Mean Square Error and Relative Efficiency

We evaluate our estimators in terms of mean square error, either $\mathrm{MSE}(\hat{P}_{ij}) = \mathbb{E}[\hat{P}_{ij} - P]^2$ or $\mathbb{E}[\bar{A}_{ij} - P]^2$. While we can directly compare the difference in mean square errors between the two estimators, it is frequently useful to consider the relative efficiency between two estimators. In our case this is $\mathrm{RE}(\bar{A}_{ij}, \hat{P}_{ij}) = \frac{\mathrm{MSE}(\hat{P}_{ij})}{\mathrm{MSE}(\bar{A}_{ij})}$, with values above 1 indicating $\bar{A}$ should be preferred while values below 1 indicate $\hat{P}$ should be preferred. We also use asymptotic relative efficiency, which is the limit of the relative efficiency as the number of vertices increase to infinity, and the scaled relative efficiency, $N \cdot \mathrm{RE}(\bar{A}_{ij}, \hat{P}_{ij})$ which in our case normalizes the relative efficiency so that the asymptotic scaled relative efficiency is non-zero and finite. 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.

5.4 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) Gorgolewski et al. (2015); Zuo et al. (2014). 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 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.

5.5 Outline for the Proof of the Theorems

Here we provide an outline of the proof Lemma A.3 which provides the approximate MSE of $\hat{P}$ in the stochastic blockmodel case. The result depends on using the asymptotic results 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 $\text{Var}(A_{ij}) = \frac{1}{M} P_{ij}(1 - P_{ij})$. The results of Athreya et al. (2013) provide a central limit theorem for estimates of the latent position in a RDPG model for a single graph. Since the variance of each entry is scaled by $1/M$ in $\bar{A}$, the analogous result for $\bar{A}$ is that the estimated latent positions will follow an approximately normal distribution with variance scaled by $1/M$ compared to the variance for a single graph.

Since $\hat{P}_{ij} = \hat{X}_i^T \hat{X}_j$ is a noisy version of the dot product of $\nu_s^T \nu_t$ from section 2.5 and each $\hat{X}_i$ are approximately independent and normal, we can use common results for the variance of the inner product of two independent multivariate normals Brown and Rutemiller (1977). After simplifications that occur in the stochastic blockmodel setting we can derive that the variance of $\hat{P}_{ij}$ converges to $\left(1/\rho_{\tau_i} + 1/\rho_{\tau_j}\right) P_{ij}(1 - P_{ij})/(N \cdot M)$ as $N \to \infty$. Since the variance of $\bar{A}_{ij}$ is $P_{ij}(1 - P_{ij})/M$, the relative efficiency between $\hat{P}_{ij}$ and $\bar{A}_{ij}$ is approximately $(\rho_{\tau_i}^{-1} + \rho_{\tau_j}^{-1})/N$ when N is sufficiently large.

The full proof is provided at https://www.overleaf.com/2776898cydwhv.

Acknowledgments

References

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

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 et al. 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.

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.

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.

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. doi: 10.1109/GlobalSIP.2013.6736878.

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.

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 from Section 3.1. 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 multinomial distribution with block membership probabilities given by $\rho \in [0,1]^K$ where $\sum_i \rho_= 1$. 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 $B = \nu\nu^T \in [0,1]^{K \times K}$ which we assume has rank k and is positive definite. By definition, the mean of the collection of graphs generated from this SBM is P, where $P_{ij} = B_{\tau_i, \tau_j}$.

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

The proofs presented here will rely on a central limit theorem developed in Athreya et al. (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 $X = [X_1, \ldots, X_N]^T \in \Re^{K \times d}$ have row i equal to $X_i = \nu_{\tau_i}$ where recall that τ_i are*

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

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

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 stochatic blockmodel.

Lemma A.2. *In the same setting as Theorem 3.2, for any $1 \le s, t \le K$, we have*

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

Proof. Under the stochastic block model with parameters (B, ρ), we have $X_i \overset{iid}{\sim} \sum_{k=1}^K \rho_k \delta_{\nu_k}$, where $\nu = [\nu_1, \cdots, \nu_K]^T$ satisfies $B = \nu\nu^T$. Without loss of generality, we could assume that $\nu = US$ where $U = [u_1, \cdots, u_K]^T$ is orthonormal in columns and S is a diagonal matrix. Here we can conclude that $\nu_s^T = u_s^T S$. Also define $R = \text{diag}(\rho_1, \cdots, \rho_K)$, then we have

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

Thus

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

□

Lemma A.3. *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*

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

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

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

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

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

and

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

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

Since $\hat{P}_{ij} = \hat{X}_i^T \hat{X}_j$ is a noisy version of the dot product of $\nu_s^T \nu_t$, 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

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

and

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

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

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

$\square$

The proof for Theorem 3.2 is now a simple application of the above Lemmas to the ratio of the mean square errors for $\bar{A}$ and $\hat{P}$.