Law of Large Graphs

June 23, 2016

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.

However, 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 networks of two groups of subjects. While hypothesis testing is the goal of their work, estimation is a key stepping 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 in many situations, 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 is 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] 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) Young and Scheinerman [2007], Nickel [2007] which motivates our estimator. In this paper, we analyze our estimator in terms of RDPG specifically.

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. 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.

2 Models and Estimators

This work considers the scenario of having M graphs represented as adjacency matrices, $\{A^{(m)}\}$ $(m = 1, \cdots, M)$, each having N vertices with known correspondence. 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. An example application of this arises in the field of connectomics, where functional brain imaging data for each subject can be represented as a graph. Each vertex represent a well defined anatomical region, and an edge between two regions is defined to exist if correlation in activity between the regions surpasses a certain threshold.

We will also assume that the graphs are sampled independently and identically from some distribution. We consider three nested models for these distributions, the independent edge model, the random dot product model, and the stochastic blockmodel, and 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, the mean graph is the For this case, we aim to estimate the mean matrix $P = \mathbb{E}[A^{(m)}]$ base on the observed adjacency matrices $A^{(1)}, \ldots, A^{(M)}$.

2.2 Estimator $\bar{A}$

Under the IEM, the element-wise sample mean of the adjacency matrices $\bar{A} = \frac{1}{M}\sum_{m=1}^{M} A^{(m)}$ is the MLE for the mean graph P. It is an unbiased 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 $M \to \infty$.

However, $\bar{A}$ doesn't exploit any graph structure, and sometimes the performance is not very good especially when M is small. For example, when $M = 1$, $\bar{A}$ is exactly the binary graph we observe, which is an inaccurate estimate for an arbitrary P compared to estimates which exploit underlying structure.

2.3 Random Dot Product Graph

In graphs, the adjacencies between vertices always depend on the 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 graph model (LPG) proposed by Hoff et. al. (2002) Hoff et al. [2002] captures such structure, where each vertex is associated with a latent positions that influences the adjacencies for that vertex. In this model, each vertex i has an associated latent vector $X_i \in \mathbb{R}^d$. Based on those latent positions, the existence of edges are conditionally independent with probability that 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 then $\binom{N}{2}$.

A specific instance of a 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.

Formally, let $\mathcal{X} \subset \mathbb{R}^d$ be a set such that $x, y \in \mathcal{X}$ implies $\langle x, y \rangle \in [0, 1]$. 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 a latent position X_i. Conditioned on the latent positions X, the edges $A_{ij} \overset{iid}{\sim} \text{Bern}(\langle X_i, X_j \rangle)$. 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 $\text{rank}(P) = \text{rank}(X) \leq d$.

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 estiamtor 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_max
2: Calculate the sample mean adjacency $\bar{A} = \frac{1}{M} \sum_{m=1}^{M} A^{(m)}$;
3: Calculate the degree matrix $D^{(0)} = \text{diag}(\bar{A}\mathbf{1})$;
4: Select the dimension d using $\bar{A} + \bar{D}/(n-1)$ (see Section 5.1);
5: **for** $k = 0; k++; k \leq k_{max}$ **do**
6: Calculate the rank-d approximation $\hat{U}\hat{S}\hat{U}^T$ for $\bar{A} + D^{(k)}$;
7: $D^{(k+1)} \leftarrow \text{diag}(\hat{U}\hat{S}\hat{U}^T)$ (see Section 5.2);
8: **end for**
9: **Output:** Rank-d approximation $\hat{P} = \hat{U}\hat{S}\hat{U}^T$ of $\bar{A} + D^{(k_{max})}$.

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 $i \in [n]$, $\tau_i = k$

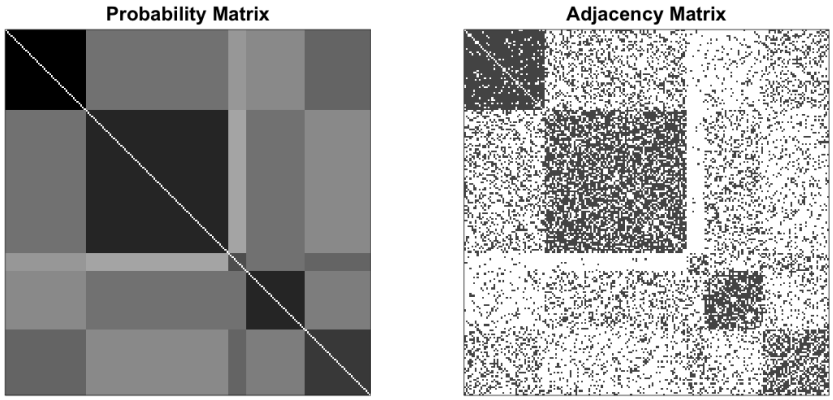

Figure 1: Example illustrating the stochastic blockmodel. The left figure shows the mean graph P with $K = 5$ blocks and $N = 200$ vertices and the right figure shows an adjacency matrix A sampled according to the probabilities from P.

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 Figure 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.

3 Results

3.1 Theoretical Results

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 $\mathrm{RE}(\bar{A}_{ij}, \hat{P}_{ij}) = \frac{\mathrm{MSE}(\hat{P}_{ij})}{\mathrm{MSE}(\bar{A}_{ij})}$. 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 5.4 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 3.1, 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. 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 $\bar{A}$, regardless of M.

RESULT FOR RDPG

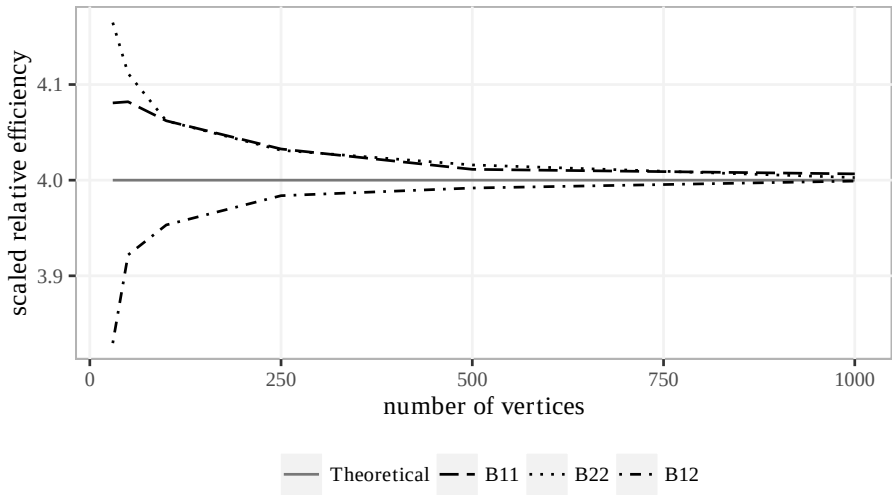

Figure 2: Scaled relative efficiency $N \cdot RE(\bar{A}, \hat{P})$ for each distinct edge probability as a function N for fixed M based on 1000 Monte Carlo replicates of the SBM from Section 3.2.

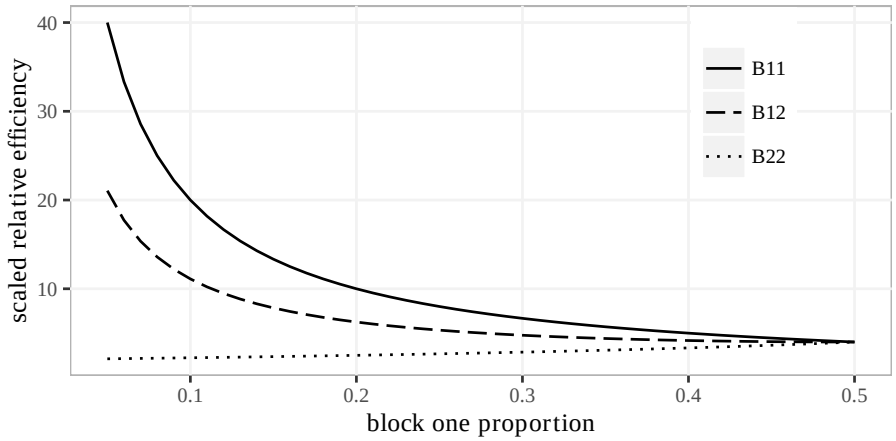

Figure 3: Theoretical scaled relative efficiency $N \cdot RE(\bar{A}, \hat{P})$ for the three distinct edge probabilities as a function the size of the first block based on the SBM described in Section 3.2.

3.2 Validation with Simulations

To illustrate Theorem 3.2 as ρ changes, Figure 3 shows $1/\rho_s + 1/\rho_t$, the scaled asympotic RE, as ρ changes from 0.01 to 0.50, for fixed $N = 500$ and $M = 100$. For $N = 500$ and $M = 100$, estimates of the scaled RE based on simulations agree very closely with their corresponding theoretical values. Notice that when $\rho_1 = 0.5$, the scaled RE has value 4.0, which agrees with the result in Figure 2.

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.

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 $N \cdot \text{RE}(\bar{A}_{ij}, \hat{P}_{ij})$ can be estimated since P is known for this simulation. Since the relative efficiency only depends on the blocks memberships of the pair i, j, we combine the relative estimates and compute

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

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 Figure 2. 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 the estimates. Say something about why $B_{11}, B_{22} > 4$ and $B_{12} < 4$

3.3 CoRR Brain Graphs: Cross-Validation

In practice, graphs do not follow the independent edge model, let alone an RDPG or SBM, but we are still interested in the mean of a collection of graphs. To demonstrate that the estimator $\hat{P}$ is still useful in such cases, we test its performance on strcutural 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].

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 Gray et al. [2012] (I think?). 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 regions whenever there is at least one white-matter tract connecting the corresponding two parts of the brain. Further details of the dataset are provided in Section 5.3.

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 different 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 Figure 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 a relative 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 (with largest 95% confidence interval length to be 3.5), and denote the dimension selected by USVT with threshold 0.7 by a square (with largest 95% confidence interval length to be 0.7). Both algorithms dimension selection algorithms select dimensions which nearly minimize minimize.

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-like structure of the mean graph. Additionally, we see that there is a large range of dimensions where perfomance 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 illustration, we consider a 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 Figure 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, Figure 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 $\bar{A}$.

We also investigate the difference in performance with respect to the geometry of brain. In Figure 7, we plot the physical locations for the 50 edges between regions with the 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. We also highlight the regions corresponding to vertices that contribute most to the difference, meaning the vertices i with the largest value of $\sum_j (|\bar{A} - P| - |\hat{P} - P|)_{ij}$.

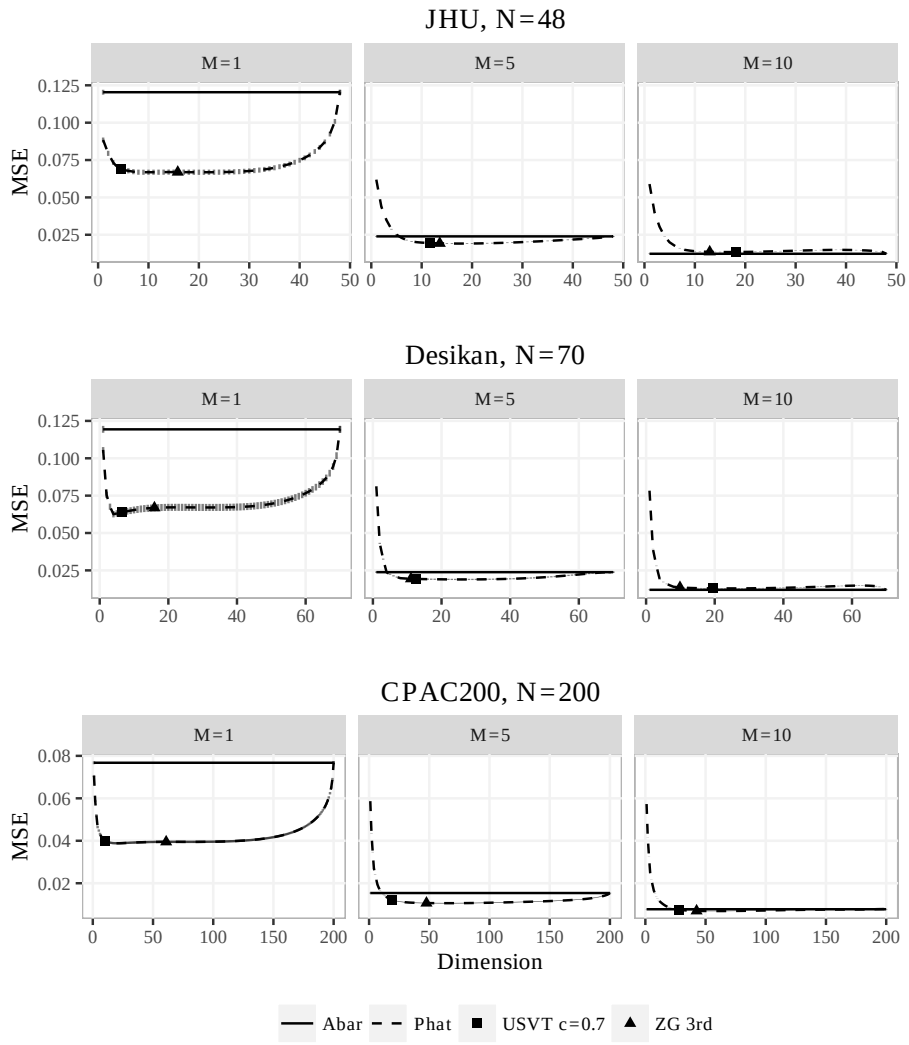

Figure 4: A comparison of the mean square error for $\bar{A}$ (solid line) and $\hat{P}$ (dashed line) for three dataset (JHU, Desikan and CPAC200) while embedding the graphs into different dimensions and with different sample sizes M. The dimension chosen by the 3rd elbow of Zhu and Ghodsi is denoted in triangle (with largest 95% confidence interval length to be 3.5), and chosen by USVT with threshold equals 0.7 is denoted in square (with largest 95% confidence interval length to be 0.7). Vertical intervals represent the 95% confidence interval. When M is small, $\hat{P}$ outperforms $\bar{A}$ with a flexible range of the embedding dimension including what Zhu and Ghodsi selects.

The result demonstrates that $\hat{P}$ 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.

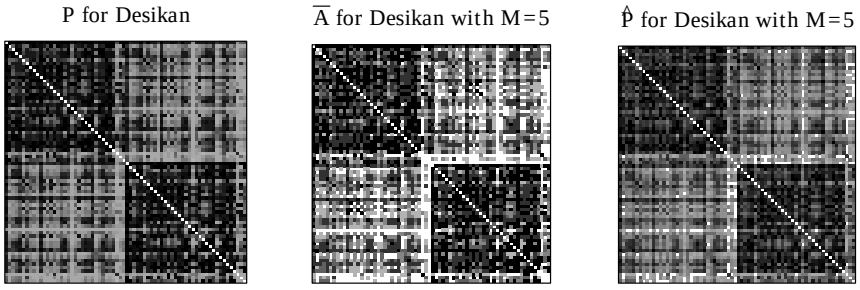

Figure 5: Left: Sample mean for the remaining $454 - 5$ graphs. Center: Sample mean for the 5 sampled graphs. Right: $\hat{P}$ for the 5 sampled graphs with dimension $d = 11$ selected using the Zhu and Ghodsi method.

3.4 Simulation under the Full Rank Independent Edge Model

While the theory we have devloped 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.

Figure 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 Figure 5. We see that the results are very similar to those presented in section 3.3. 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 very well. So $\hat{P}$ does a good job even when the low rank assumption of the model is violated. This simulation shows again the robustness of $\hat{P}$.

4 Discussion

4.1 Summary

In this paper, we propose a better way to estimate the mean of a collection of graphs.Motivated by RDPG, our methodology take advantage of the low-rank structure of the graphs by applying ASE to the entry-wise MLE. We then give a closed form for the asymptotical relative efficiency between the entry-wise MLE $\bar{A}$ and our estimator $\hat{P}$, which theoretically proves that $\hat{P}$ has smaller MSE when N is sufficiently large. These results are demonstrate by various simulations. Moreover, our estimator also outperforms the entry-wise MLE in the CoRR brain graphs and in the full-rank simulation, which shows that $\hat{P}$ still performs well even when the low-rank assumption is violated. It demonstrate that $\hat{P}$ is robust and can be applied in practice.

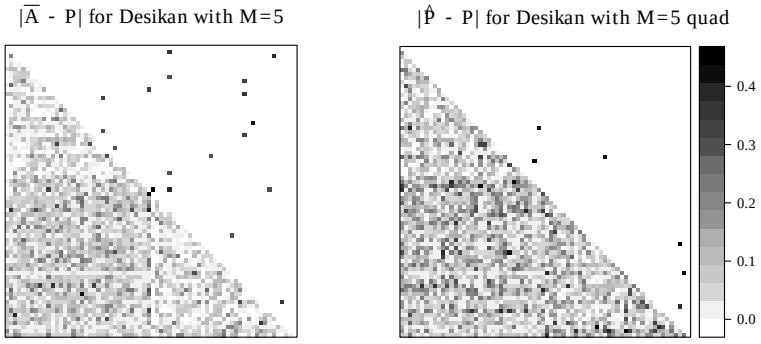

Figure 6: Heat plot of the absolute estimation error $|\bar{A} - P|$ and $|\hat{P} - P|$ for a sample of size $M = 5$ from Desikan dataset while embedding the graphs into dimension $d = 11$ selected by the 3rd elbow of ZG method. The lower triangular matrix shows the actual absolute difference, while the upper triangular matrix only highlights the edges with absolute differences larger than 0.4. The fact that 18 edges from $\bar{A}$ and 6 edges from $\hat{P}$ being highlighted shows the better performance of $\hat{P}$.

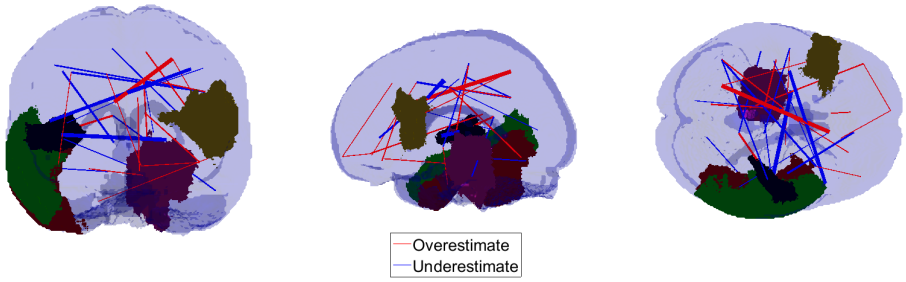

Figure 7: Top 5 regions of the brain (vertices in graphs) and top 50 connections between regions (edges in graphs) with largest difference $|\bar{A} - P| - |\hat{P} - P|$. Red edges indicate that $\hat{P}$ overestimate P while blue means that $\hat{P}$ underestimate P. The edge width is determined by 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.

4.2 Future Work

In this paper, we assume that the adjacency matrix is observed without contamination. However, generally there will be noise in practice. With contaminations, robust estimators like MLqE is preferred. If an estimator can not only inherit robustness from the robust estimators but also has small variance by taking advantage of the low rank structure of the graphs, it will be very useful.

Meanwhile, estimating the rank of the graph structure accurately will certainly help improve the performance of the estimator $\hat{P}$. Now we are using Zhu and Ghodsi's method and USVT, but there is still a lot of space for improve-

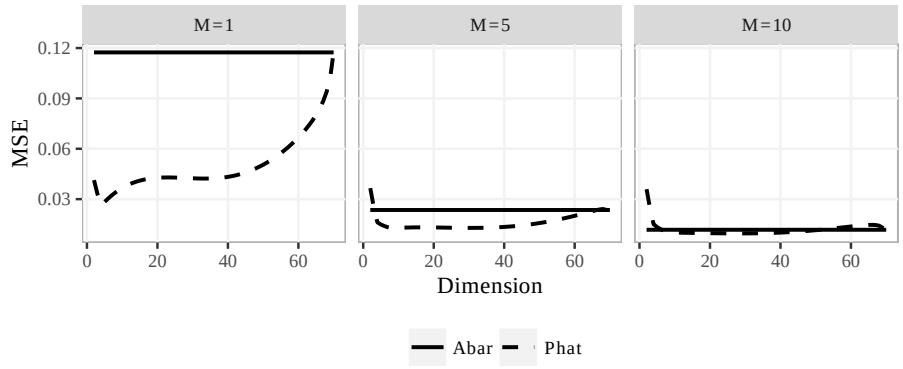

Figure 8: Comparison of MSE between $\bar{A}$ (solid line) and $\hat{P}$ (dashed line) for simulated data with different sample sizes M based on the sample mean for the Desikan dataset.

ment, especially in this particular case.

5 Methods

5.1 Choosing Dimension

Often in dimensionality reduction techniques, the choice for dimension d, relies on visually analyzing a plot of the ordered eigenvalues, looking for a "gap" or "elbow" in the scree-plot. Zhu and Ghodsi 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. In order to prevent underestimating d, which is much more harmful than overestimating, we use the 3rd elbow in the experiments performed in this work.

Universal Singular Value Thresholding (USVT) Chatterjee et al. [2015] is a simple estimation procedure proposed by Chatterjee that works for any matrix that has "a little bit of structure". Basically in our setting, it selects the dimension d as the number singular values that are greater than a constant c times $\sqrt{N/M}$. In the simulation, we set $c = 0.7$.

5.2 Graph Diagonal Augmentation

The graphs examined in this work are hollow, in that there are no self-loops and thus the diagonal entries of the adjacency matrix are 0. This leads to a bias in the calculation of the eigenvectors. To compensate such bias, we use an iterative method developed by Scheinerman and Tucker Scheinerman and Tucker [2010] to augment the diagonal before ASE. In the experiments, we are using 1 iteration of Scheinerman's method to do the diagonal augmentation.

5.3 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 the incomplete data, the true number of scans is 454.

When deriving MR connectomes, the NeuroData team parcellate the brain into groups of nodes as defined by anatomical atlases neu, Kiar [2016]. The atlases are defined either physiologically or structurally by neuroanatomists (Desikan and JHU), or are generated using a segmentation algorithm looking for certain features or groupings (CPAC200).

The graphs we are using are processed by NeuroData team from DTI data of the original dataset generated with different atlases (Desikan, JHU and CPAC200), each containing different region/node definitions. The graphs are undirected, unweighted and with no self-loops. An edge exists between two regions when there is at least one white-matter tract connecting the corresponding two parts of the brain.

5.4 Outline for the Proof of the Theorems

Here we provide an outline of the proof for the MSE($\hat{P}$) result presented in Section 3.1.

When comparing two estimators, the first thing we need to consider is consistency. It is easy to see that $\bar{A}$ is unbiased as an estimate of P. Moreover, since two latent positions are conditionally asymptotically independent by extended version of Theorem 1 in Athreya et al. (2013) Athreya et al. [2013], we know $\hat{P}$ is consistent, as well as $\bar{A}$.

Thus the relative efficiency between $\hat{P}$ and $\bar{A}$, which is equivalent to the ratio of mean square errors in this case, is a good indicate in comparison.

Since $\hat{P}_{ij} = \hat{X}_i^T \hat{X}_j$ is a noisy version of the dot product of $\nu_s^T \nu_t$, by Equation 5 in Brown and Rutemiller (1977) Brown and Rutemiller [1977], combined with asymptotic independence between $\hat{X}_i$ and $\hat{X}_j$, and the covariance matrices given by extended version of Theorem 1 in Athreya et al. (2013) Athreya et al. [2013], we have 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 (relative) full proof is provided here:
https://www.overleaf.com/2776898cydwhv. Feel free to edit it.

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.

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

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

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

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

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.

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

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.