Law of Large Graphs

April 12, 2016

1 Introduction

Estimating the mean of a collection of graphs is becoming more and more important both in statistical inference and in various applications like connectomics, social networks, etc. Element-wise maximum likelihood estimate is a reasonable estimator if we only consider the independent graph model without taking any graph structure into account.

However, in a large graph, vertices are generally clustered into different communities such that vertices of the same community behave similarly. The stochastic blockmodel (SBM) introduced in Holland et al. (1983) [4] captures such structural property and is widely used in modeling networks. In this model, each of the N vertices is assigned to one of the K blocks. And the probability of an edge between two vertices only depends on their respective block memberships. For example, when modeling connectomics, vertices may represent neurons with edges indicating axon-synapse-dendrite connections, or vertices may represent brain regions with edges indicating connectivity between regions.

Also, latent positions graph model proposes a way to parameterize the graph structure by latent positions associated with each vertex. And random dot product graph, a special case of latent positions graph, is considered in this paper. In the RDPG, each vertex is associated with one latent vector. And the probability of an edge between two vertices only depends on the dot product of the two respective latent vectors. In particular, this paper considers SBM as a RDPG. So we will have exactly K different latent positions for N vertices.

Using the estimates of the latent positions in an RDPG based on a truncated eigen-decomposition of the adjacency matrix proposed by Sussman et al. (2012) [6], we invent a new estimator for the mean of the collection of graphs which captures the low-rank structure. Moreover, with the asymptotic result in Athreya et al. (2013) [1] which says the latent positions estimated using adjacency spectral graph embedding converge in distribution to a multivariate Gaussian mixture in the RDPG, we give a closed form representation for the asymptotic relative efficiency between our estimator and the element-wise MLE. Based on that, we theoretically prove that our estimator reduces the variance and is better than the element-wise MLE according to the relative efficiency when N is large enough.

2 Model

This section presents, with theory, a comparison of two estimators for the mean of a collection of graphs by observing the adjacency matrices. 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, i.e. each $A^{(m)}$ is a binary symmetric matrix with zeros along the diagonal. An example scenario of this arises in the field of connectomics, where

functional brain imaging data for each subject can be represented as a graph, with each vertex having a defined anatomical correspondence, and an edge between two regions is defined to exist if correlation in activity between the regions reaches a certain threshold. In this setting, we consider each random graph to be sampled from the independent edge model with parameter $P \in [0,1]^{N\times N}$, where each edge between vertex i and vertex j exists independently with probability P_{ij}. We aim to estimate the probability matrix P with our observations of the adjacency matrices $\{A^{(m)}\}$ of M graphs.

2.1 Entry-Wise Least Squares Estimate

The most intuitive approach in this scenario is the element-wise mean among the adjacency matrices:

$$\bar{A} = \frac{1}{M}\sum_{m=1}^{M} A^{(m)} \tag{1}$$

Since each element of the adjacency matrix A_{ij} is a sample from the Bernoulli distribution with probability P_{ij}, with each element examined in isolation, to estimate the mean graph P one would like to use the element-wise MLE, i.e. the element-wise mean, $\bar{A}$. Meanwhile, it is also the entry-wise least squares estimates.

2.2 Random Dot Product Graph

Hoff et. al. (2002) [3] proposed a model for random graphs called Latent Positions Graph Model. In this model, each vertex i has an associated latent vector $x_i \in \mathbb{R}^d$ (generally d is much smaller than the number of vertices N), and the probability of a edge being present between two vertices only depends on their latent vectors through a link function.

A specific instance of this model that we will examine is the random dot product graph model (RDPG) in which the link function is the dot product, i.e. the probability of an edge being present between two nodes is the dot product of their latent vectors. [5]. For example in the functional connectomics, components of the latent vectors may refer the relative importance of an anatomical region among a set of tasks. The magnitude then may refer to how active the region is generally. Therefore, active regions vital for a similar task are more likely to be functionally connected.

2.3 Stochastic Block Model as a Random Dot Product Graph

Generally, in a large graph, vertices are clustered into different communities such that vertices within the same community behave similarly. Such structural property is captured by the stochastic block model (SBM), 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. This imposes the idea of structural equivalence, where vertices are defined to be structurally equivalent if their connections to other nodes are similar. In the stochastic block model, groups of vertices, or blocks, are then structurally equivalent since the vertices contained have equal likelihood in their connections among the blocks. An example of block structure can be thought to exist in functional brain imaging, for instance the structures in the basal ganglia will likely behave similarly in their connections and may be considered a block.

Formally, the SBM is determined by the number of blocks K (generally way less than the number of vertices N), block proportion vector ρ, and block probability matrix B. In this model each vertex is assigned to one of K blocks and the fraction of vertices belonging

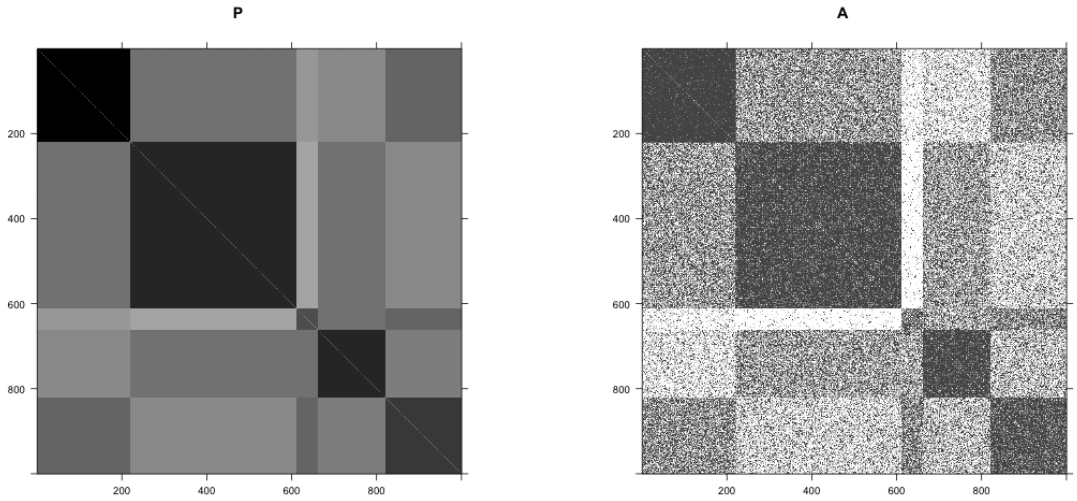

Figure 1: Example illustrating the SBM. (a) Probability matrix that follows a SBM with K = 7 blocks; (b) Adjacency matrix generated from the SBM with probability matrix in (a).

to the i th block is designated as ρ_i. The connection probabilities of this block structure are stored in the symmetric $K \times K$ block matrix B, where B_{ij} represents the probability of an edge existing between a vertex of block i and one of block j.

Now if we consider the SBM as a random dot product graph, all vertices in the same block would have identical latent positions.

2.4 Estimator $\hat{P}$ Based on Adjacency Spectral Embedding

In order to take advantage of the underlying low dimensions of the RDPG, we would like to use the adjacency spectral embedding (ASE) studied by Sussman et. al. to enforce a low rank approximation on the adjacency matrix A, which will decrease the variance if we embed it into the right dimension [6]. The adjacency spectral embedding creates an approximated RDPG representation of the adjacency matrix from its low rank eigen-decomposition. The latent vectors are stored as a $N \times d$ matrix X, where the columns are comprised of the eigenvectors associated with the d largest eigenvalues of the adjacency matrix. Then X_i, each row of X, is a latent vector for the corresponding vertex i.

In this work, rather than stopping at the element-wise MLE $\bar{A}$, we use ASE to embed the mean matrix $\bar{A}$ to X and then take $\hat{P} = XX^T$ as our estimate for P. Due to the underlying block-distibuted RDPG structure of graphs, enforcing this low rank approximation on $\bar{A}$ will provide a better estimate for the true mean matrix P. Details of this algorithm are presented in Section 5.1.

2.5 Performance Evaluation: Relative Efficiency

To compare the performance between $\hat{P}$ and $\bar{A}$, we examine the relative efficiency (RE), in mean squared error (MSE), among the two defined as:

$$RE_{ij} = \frac{MSE(\hat{P}_{ij})}{MSE(\bar{A}_{ij})} \tag{2}$$

3 Results

3.1 Theoretical Results

Theorem 3.1 *For any i and j, conditioning on $X_i = \nu_{\tau_i}$ and $X_j = \nu_{\tau_j}$, we have*

$$\text{ARE}(\bar{A}_{ij}, \hat{P}_{ij}) = 0.$$

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

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

This comes from a proof (outlined in section 5.6) for the variance of $\hat{P}_{ij}$ under the condition that N is large:

Lemma 3.2 *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

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

Further, knowing that $\bar{A}_{ij}$ is the average of M samples from the Bernoulli distribution with parameter P_{ij}, the variance of $\bar{A}_{ij}$ should be $P_{ij}(1 - P_{ij})/M$, which yields the above result.

This result implicates that for large random dot product graphs that follow a stochastic block model, a better estimate for the mean graph, under MSE, is the $\hat{P}$ estimate.

3.2 Validation with Simulations

We demonstrate the theoretical results in Section 3.1, the variance of $\hat{P}$ and the relative efficiency, via various Monte Carlo simulation experiments. Specifically, we consider the 2-block SBM parameterized by

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

The block proportion vector ρ shows that each vertex is uniformly assigned to either block. And probability matrix B indicates the probability of corresponding edges.

For each Monte Carlo replicate, we generate M random graphs with known vertex correspondence under the SBM described above. Specifically, we first draw the block assignment $\tau \in [K]^N$ for each vertex from a multinomial distribution with parameter ρ. Note that τ will be identical for all M graphs we are going to generate for vertex correspondence. Then we sample M conditionally independent graphs $A^{(1)}, \cdots, A^{(M)}$, which are binary, symmetric and hollow, such that $A^{(m)}_{ij} | \tau \overset{ind}{\sim} Bern(P_{ij})$, where $P_{ij} = B_{\tau_i, \tau_j}$, $1 \le m \le M$, $1 \le i < j \le N$.

Given M graphs, we can calculate $\bar{A}$ and $\hat{P}$ assuming that $d = \text{rank}(B) = 2$ is known. Also, $\text{MSE}(\hat{P}_{ij})$, $\text{MSE}(\bar{A}_{ij})$ and $\text{RE}(\bar{A}_{ij}, \hat{P}_{ij})$ with $1 \le i < j \le N$ for the data

can be derived as P is known in this simulation. Moreover, since vertices are equivalent in the same block under the SBM, we average over all the MSE and RE associated with edges corresponding to the same blocks as our estimates using the true labels τ. That is, $\mathrm{MSE}_{st}(\hat{P}) = (\sum_{\tau_i=s,\tau_j=t,i\neq j} \mathrm{MSE}(\hat{P}_{ij}))/(\sum_{\tau_i=s,\tau_j=t,i\neq j} 1)$, $\mathrm{MSE}_{st}(\bar{A}) = (\sum_{\tau_i=s,\tau_j=t,i\neq j} \mathrm{MSE}(\bar{A}_{ij}))/(\sum_{\tau_i=s,\tau_j=t,i\neq j} 1)$ and $\mathrm{RE}_{st}(\bar{A},\hat{P}) = \mathrm{MSE}_{st}(\hat{P})/\mathrm{MSE}_{st}(\bar{A})$ for $1 \leq s,t \leq K$.

By checking the averaging MSE and RE of the two estimates $\hat{P}$ and $\bar{A}$ over 1000 Monte Carlo replicates, we demonstrate that the theoretical results in Section 3.1.

Figure 2 plots the scaled average MSE with different N and M of 1000 Monte Carlo replicates. Colors denote the block membership associated with the edges we are averaging over. Lines in black represent the theoretical values. Figure 2(a) shows that with a fixed M, as N increases, $N \cdot \mathrm{MSE}_{st}(\hat{P})$ converges to $(1/\rho_s + 1/\rho_t)B_{st}(1-B_{st})/M$ represented as the black lines, as suggested in Lemma 3.2. Notice that this means $\mathrm{MSE}_{st}(\hat{P})$ is decreasing at rate $1/N$. Figure 2(b) illustrates that $M \cdot \mathrm{MSE}_{st}(\hat{P})$ holds to be $(1/\rho_s + 1/\rho_t)B_{st}(1-B_{st})/N$ approximately independent of the value of M while keeping N sufficiently large and fixed. Thus a sufficiently large M is not a necessary condition for Lemma 3.2 as expected.

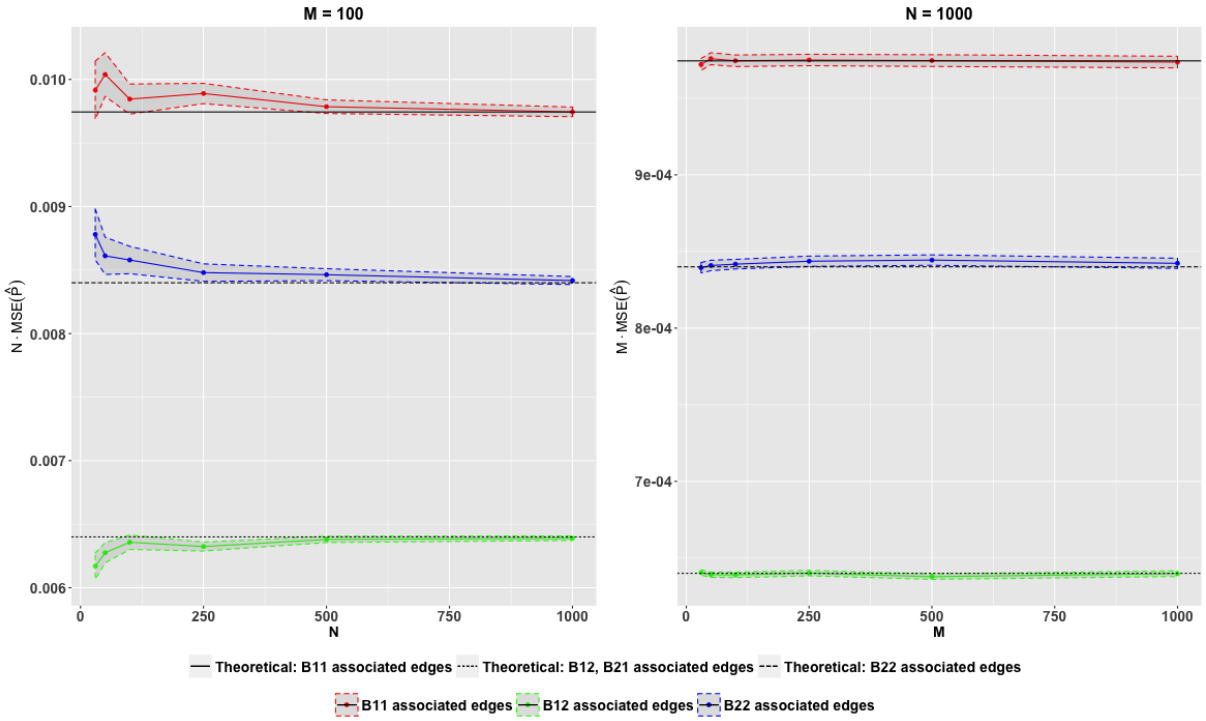

Figure 2: Simulation results for the scaled average MSE of $\hat{P}$ with different N and M of 1000 Monte Carlo replicates. Colors denote the block membership associated with the edges we are averaging over. Dashed lines represent the 95% confidence interval. Lines in black represent the theoretical values. (a) shows that $N \cdot \mathrm{MSE}_{st}(\hat{P})$ converges to $(1/\rho_s + 1/\rho_t)B_{st}(1-B_{st})/M$ represented as the dashed lines with a fixed M as N increases. (b) illustrates that $M \cdot \mathrm{MSE}_{st}(\hat{P})$ holds to be $(1/\rho_s + 1/\rho_t)B_{st}(1-B_{st})/N$ approximately independent of the value of M while keeping N sufficiently large and fixed.

Figure 3 plot the scaled average RE with different N and fixed M of 1000 Monte Carlo replicates. Colors denote the block membership associated with the edges we are averaging over. Solid line in black represents the theoretical value for scaled RE. From the figure, we see that $N \cdot \mathrm{RE}_{st}(\bar{A},\hat{P})$ converges to $1/\rho_s + 1/\rho_t$ represented as the black

solid line, as suggested in Lemma 3.1. Notice that this means $\mathrm{RE}_{st}(\bar{A}, \hat{P})$ is decreasing at rate $1/N$.

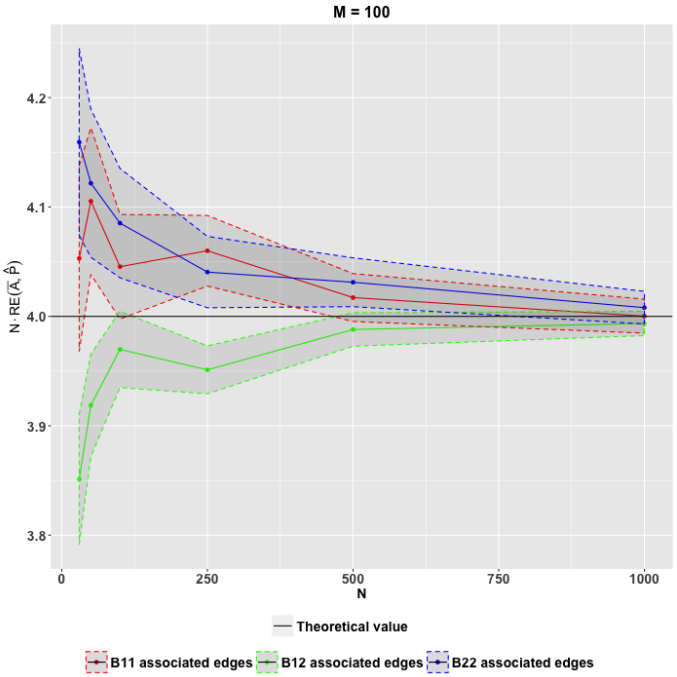

Figure 3: Scaled average RE with different N and fixed M of 1000 Monte Carlo replicates. Colors denote the block membership associated with the edges we are averaging over. Dashed lines represent the 95% confidence interval. Solid line in black represents the theoretical value for scaled RE. Observe that $N \cdot \mathrm{RE}_{st}(\bar{A}, \hat{P})$ converges to $1/\rho_s + 1/\rho_t$ as expected.

To verify Theorem 3.1 and Lemma 3.2 holds with different ρ, Figure 4 shows the average MSE and average RE with $N = 500$ and $M = 100$ while changing ρ_1 from 0.1 to 0.9. These simulated results again match well for the predictions from Theorem 3.1 and Lemma 3.2.

3.3 CoRR Brain Graphs

To demonstrate that the $\hat{P}$ estimate is valid under data that does not perfectly follow a SBM, we examine three datasets, JHU, desikan and CPAC200, which are sets of 454 brain connectomes with different number of nodes generated from fMRI scans available at the Consortium for Reliability and Reproducibility (CoRR). Details on these datasets and connectome generation can be seen in Section 5.4. The connectomes generated have 48(JHU), 70(desikan) and 200(CPAC200) vertices respectively, with anatomical correspondence. To compare $\bar{A}$ and $\hat{P}$ we perform a cross-validation study to examine the impact of the number of available graphs M. For each sample size M, we randomly sample M graphs from the 454 graphs in the CoRR dataset and estimate the mean with both $\bar{A}$ and $\hat{P}$ with some proper embedding dimension. Also, we assure M to be relatively small such that the mean of the $(454 - M)$ remaining graphs is a valid approximation to the true probability matrix P we are estimating. Then we can calculate the MSE of the two estimators based on the estimated probability matrix similarly as long as we know which dimension we should embed the graphs into.

Figure 5, Figure 6 and Figure 7 demonstrate that our algorithm gives a better estimate $\hat{P}$ according to all three datasets. When M is small, $\bar{A}$ has large variance which leads to

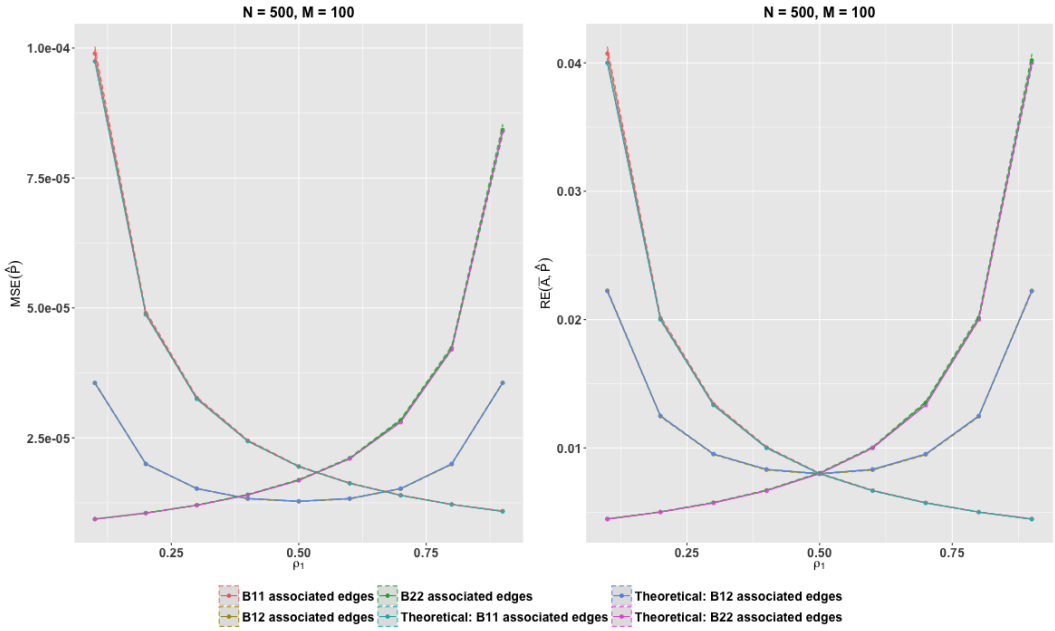

Figure 4: Simulated results for (a) $\text{MSE}_{st}(\hat{P})$ and (b) $\text{RE}_{st}(\bar{A}, \hat{P})$ with $N = 500$ and $M = 100$ of 1000 Monte Carlo replicates while changing ρ_1 from 0.1 to 0.9. Dashed lines represent the 95% confidence interval. The simulated values for the average MSE and average RE measurements match perfectly with the theoretical values.

large MSE. Meanwhile, $\hat{P}$ reduces the variance by taking advantages of the graph structure and outperforms $\bar{A}$ dramatically. With a relatively large M, $\bar{A}$ has enough information to performs well, leaving less space for improvement. And we can see both estimators perform almost perfect when M is large. Moreover, Zhu and Ghodsi's algorithm does a good job for selecting the dimension to embed. As we can see, the result is insensitive to the embedding dimension we choose, which makes our estimator more robust and useful in analyzing real data. The results justify that $\hat{P}$ is a valid and likely more accurate estimate of P even when the data does not perfectly follow an SBM.

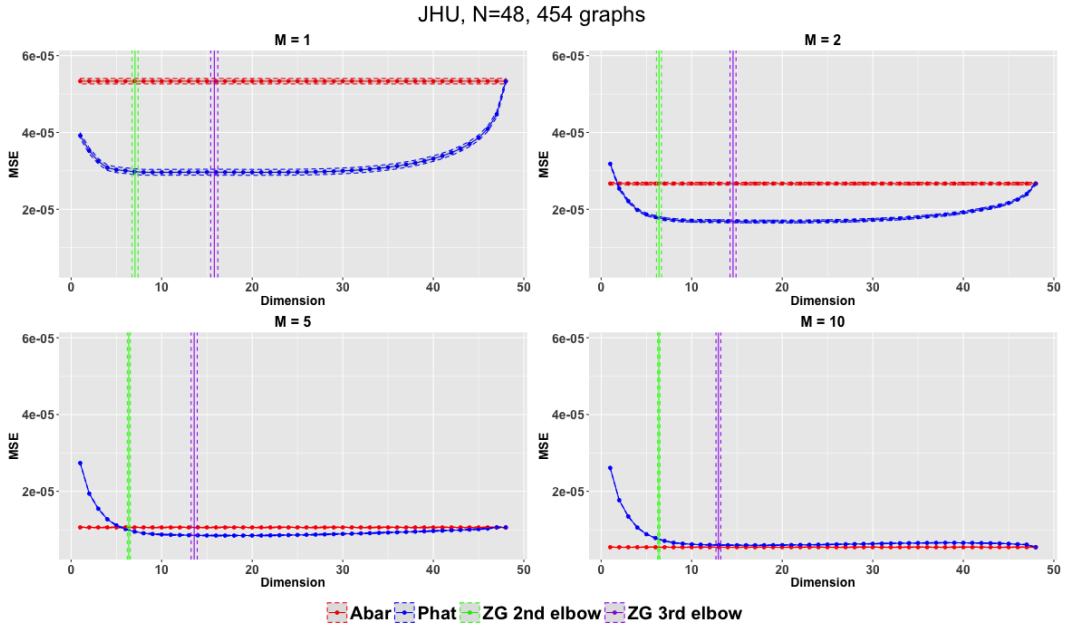

Figure 5: Comparison of MSE between $\bar{A}$ (red) and $\hat{P}$ (blue) for JHU dataset while embedding the graphs into different dimensions with different size M of the subsamples. The dimension chosen by Zhu and Ghodsi is denoted in green (2nd elbow) and purple (3rd elbow). Dashed lines represent 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.

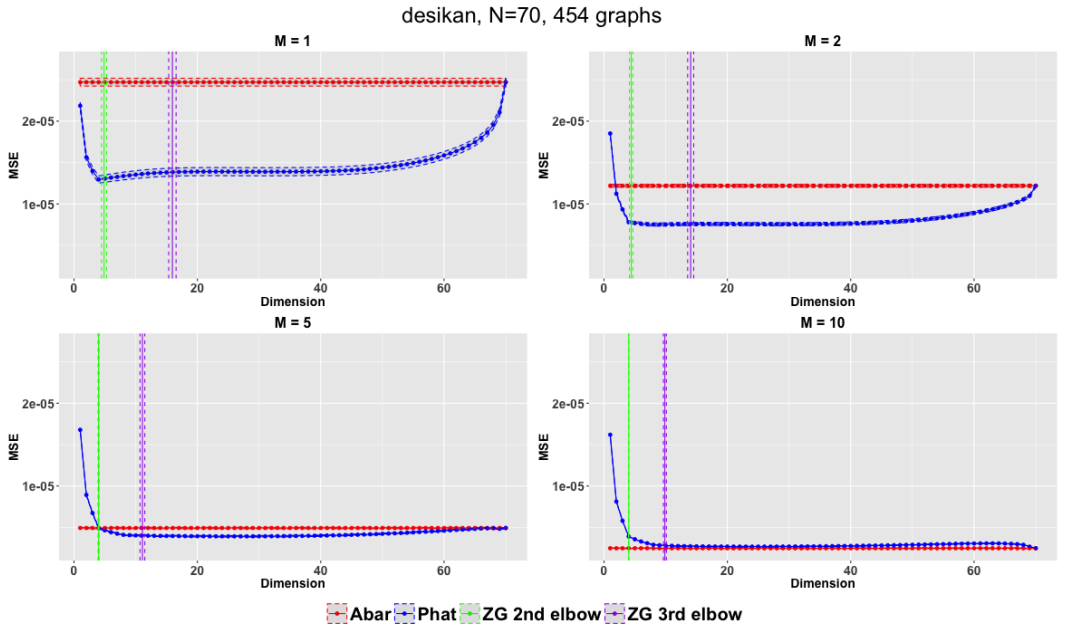

Figure 6: Comparison of MSE between $\bar{A}$ (red) and $\hat{P}$ (blue) for desikan dataset while embedding the graphs into different dimensions with different size M of the subsamples. The dimension chosen by Zhu and Ghodsi is denoted in green (2nd elbow) and purple (3rd elbow). Dashed lines represent 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.

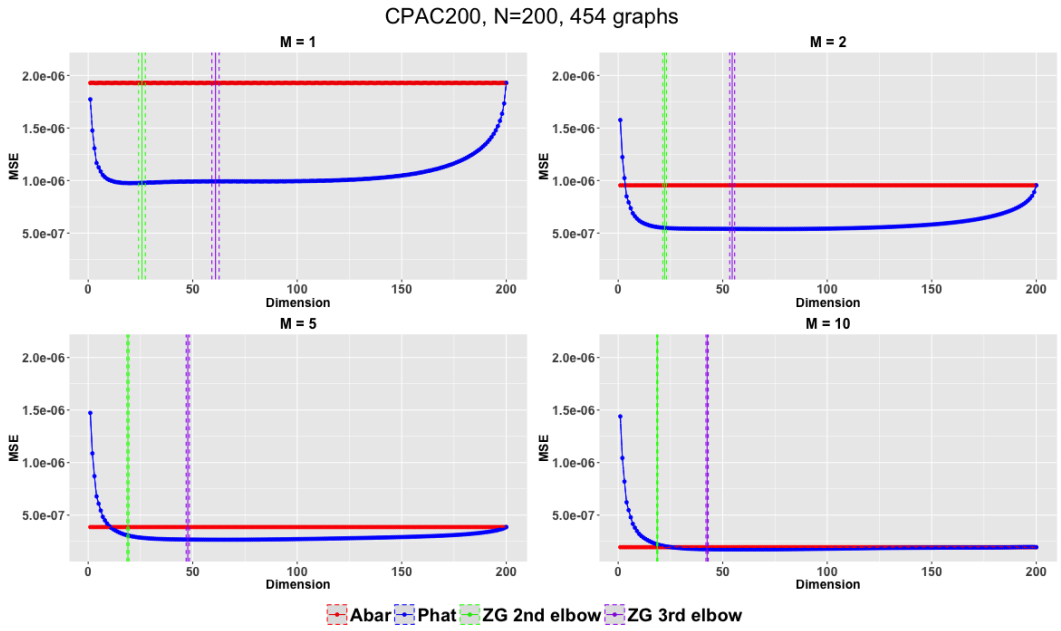

Figure 7: Comparison of MSE between $\bar{A}$ (red) and $\hat{P}$ (blue) for CPAC200 dataset while embedding the graphs into different dimensions with different size M of the subsamples. The dimension chosen by Zhu and Ghodsi is denoted in green (2nd elbow) and purple (3rd elbow). Dashed lines represent 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.

4 Discussion

In this paper we have proposed a better way to estimate the mean of a collection of graphs sampling from the SBM. Our methodology is motivated by the asymptotical distribution of the adjacency spectral embedding of RDPG graphs. To take advantage of the low-rank structure of the graphs, adjacency spectral embedding, a rank-reduction procedure, is applied to the element-wise MLE. We then give a closed form for asymptotical relative efficiency between our estimator and the element-wise MLE, which theoretically proves that our estimator has smaller variance with sufficiently large N while keeping to be asymptotically unbiased. These results are demonstrate by various simulations. Moreover, our estimator also outperforms element-wise MLE for the CoRR brain graphs, which shows our estimator is valid even when the data does not perfectly follow a SBM.

5 Methods

5.1 Algorithm: $\hat{P}$

Algorithm 1

1: **Input:** $A^{(1)}, A^{(2)}, \cdots, A^{(M)}$, with each $A^{(m)} \in \{0,1\}^{N \times N}$ sampling from SBM with vertex correspondence;
2: Calculate $\bar{A} = \frac{1}{M} \sum_{m=1}^{M} A^{(m)}$;
3: Estimate the dimension in the SBM d (see Section 5.2);
4: Obtain estimated latent positions $\hat{X} \in \mathbb{R}^{N \times d}$ by applying adjacency spectral embedding to $\bar{A}$ with diagonal augmentation. The columns of $\hat{X}$ consist of the eigenvectors corresponding to the d largest eigenvalues of diagonal augmented $\bar{A}$ (see Section 5.3);
5: $\hat{P} = \hat{X}\hat{X}^T$ is our estimator.

5.2 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 [7] 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 under-estimating d, which is much more harmful than over-estimating, we initialize $d_0 = 0$ and iterate over the Zhu and Ghodsi algorithm by removing the first d_{i-1} eigenvalues from calculation at the ith iteration to determine the location of the "next elbow". For the experiments performed in this work, we choose d to be the 2nd and 3rd elbow under this approach.

5.3 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. We minimize this bias by using an iterative method developed by Scheinerman and Tucker [5]. In this method, Steps 4 and 5 of Algorithm 1 are repeated, each time replacing the diagonal component of $\bar{A}$ with the diagonal of $\hat{P}$, until $\hat{P}$ converges.

5.4 Dataset Description RT: Needs update

The connectomes analyzed were created from resting state functional MRI (fMRI) and Diffusion Tensor Imaging (DTI) scans from the Consortium for Reliability and Reproducibility (CoRR) and are available via the International Neuroimaging Data-sharing Initiative (INDI). The SWU 4 - Southwest University image collection was used to generate 454 connectomes with 788 anatomically corresponding vertices. (Need to describe how graphs were made with reference, etc.)

5.5 Source code and data

5.6 Outline for Proof of Relative Efficiency

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 Corollary 4.11 in Athreya et al. (2013) [1], 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) [2], combined with asymptotic independence between $\hat{X}_i$ and $\hat{X}_j$, and the covariance matrices given by extended version of Corollary 4.11 in Athreya et al. (2013) [1], 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 in the appendix.

6 Appendix

Proof is provided here: https://www.overleaf.com/2776898cydwhv. Feel free to edit it.

References

[1] 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.

[2] 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.

[3] 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.

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

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

[6] Daniel L Sussman, Minh Tang, Donniell E Fishkind, and Carey E Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

[7] 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.