Robust LLG

June 1, 2017

Abstract

Estimation of the graph parameters based on a sample in a parametric setting is essential for a wide range of graph inferences. In particular, the graphs are generally observed with contaminations in practice. We consider an edge weight gross error model and propose an estimator based on the robust estimation followed by a low-rank decomposition. Under proper conditions, theoretical results show that our estimator not only inherits the robust property from robust estimators, but also wins the bias-variance tradeoff by exploiting the low-rank graph structure under proper conditions. We also demonstrate the improvement offered by our method via simulations and a human connectome data experiment.
Keywords: network, weighted, embedding, low-rank, robustness, estimation

1 Background and Overview

Network analysis is becoming more and more widely used recently. In a general parametric framework, $G \sim f \in \mathcal{F} = \{f_\theta : \theta \in \Theta\}$, selecting a reasonable estimator $\widehat{\theta}(G)$ for the unknown θ given a finite graph sample $\{G^{(1)}, \cdots, G^{(m)}\}$ is one of the most important tasks.

Consider the most basic setting, i.e. undirected and unweighted graphs with each edge independently distributed from a Bernoulli distribution, with the parameters to be the probabilities of the existence of edges between all pair of vertices. Then the maximum likelihood estimate, which happens to be the entry-wise sample mean in this situation, is the uniformly minimum-variance unbiased estimator under the independent edge graph model (IEM) [Bollobás et al., 2007] without taking any graph structure into account. However, in a high dimensional situation, unbiased estimators often leads to inaccurate estimates with very high variance when the sample size is small. And for this basic setting we considered, the graphs are high dimensional object with n^2 parameters since $\Theta = [0, 1]^{n \times n}$, where n denotes the number of vertices. Thus the MLE does not perform very well especially when there are only a few observations available, which is likely the case in real world.

Generally, by biasing towards low-rank structures carefully, the estimators will have a greatly reduced variances and win the bias-variance tradeoff [Trunk, 1979]. Such improvement is not only important for the estimation itself, but also can help with other statistical inference procedures. For example, Ginestet et al. [2014] proposed a method to test if there is a difference between the networks of

two groups of subjects. While hypothesis testing is the final goal, the estimation procedure is a key intermediate step and can be improved.

In order to improve the MLE considered above, the underlying graph structures are taken into account for exploiting the low-rank structure when constructing the estimator. One of the most important structures about the networks is the community structure in which vertices are clustered into different communities such that vertices within the same community behave similarly. The stochastic blockmodel (SBM) [Holland et al., 1983] captures such structural property strongly by assuming that vertices from the same block behave exactly the same.

As a generalization of the SBM, the latent position model (LPM) [Hoff et al., 2002] allows vertices to behave differently. Generally, the adjacencies between vertices depend on unobserved properties of the corresponding vertices. Thus under LPM, each vertex is associated with a latent position which influences the adjacencies for that vertex based on a link function. In this paper, we consider a special case of LPM, the random dot product graph (RDPG) [Nickel, 2007, Young and Scheinerman, 2007], which has the dot product as the link function.

Recently, Tang et al. [2016] considers an estimator based on a low-rank approximation of the sample mean graph motivated by the RDPG model and proves that in the basic Bernoulli setting under SBM, the new estimator outperforms the element-wise sample mean since it decreases the overall asymptotic variance dramatically by smoothing towards the low-rank structure.

While Tang et al. [2016] demonstrate the advantage of the low-rank estimator for the unweighted graphs with Bernoulli distribution, the results are based on the assumption that graphs are observed without contaminations. However in practice there will be noise in the observed graphs. In this situation, there is no guarantee that the performance of the low-rank estimator is still better.

In this work, we are going to improve the estimation procedure in a contaminated scenario. Before we introduce the contaminations, it is helpful to extend the unweighted graphs with Bernoulli distribution to weighted graphs with a general distribution f. All the models we mentioned above (IEM, RDPG, and SBM) could be extended naturally and are discussed in details in Section 3.1, Section 3.2, and Section 3.3 respectively.

One of the most popular contamination model is the gross error model [Bickel and Doksum, 2001, Mah and Tamhane, 1982]. In a gross error model, we observe good measurement $G^* \sim f_P \in \mathcal{F}$ most of the time, while there are a few wild values $G^{**} \sim h_C \in \mathcal{H}$ when the gross errors occur. As to the graphs, one way to generalize from the gross error model is to contaminate the entire graph with some small probability $\epsilon \in (0, 1)$, that is $G \sim (1-\epsilon)f_P + \epsilon h_C$. However, since all the models we consider are subsets of the IEM, it is more natural to consider the contaminations with respect to each edge, i.e. for $1 \leq i, j \leq n$, $G_{ij} \sim (1-\epsilon)f_{P_{ij}} + \epsilon h_{C_{ij}}$ with $f \in \mathcal{F}$ and $h \in \mathcal{H}$, where both $\mathcal{F}$ and $\mathcal{H}$ are one-parameter distribution families.

Under the contamination model, although we observe G instead of G^*, estimating the parameters P_{ij} $(1 \leq i, j \leq n)$ of $f_{P_{ij}}$ in $\mathcal{F}$ is still our goal. We first prove that the low-rank estimator ($\widetilde{P}^{(1)}$ in Figure 1) proposed in [Tang et al., 2016] is still much better than the entry-wise MLE ($\widehat{P}^{(1)}$ in Figure 1) in terms of mean squared error when the observations are contaminated under proper conditions.

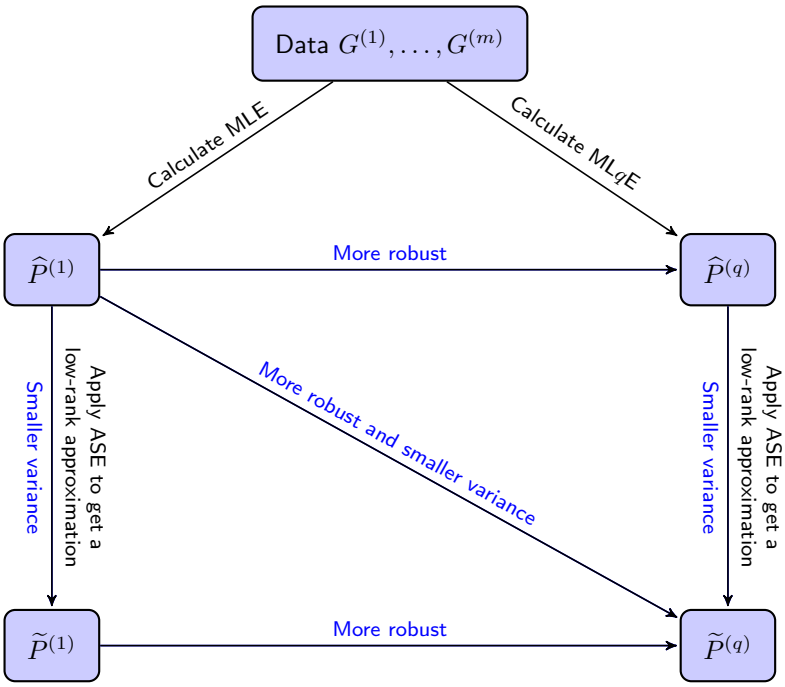

Figure 1: Roadmap among the data and four estimators.

Furthermore, with contaminations, it is more preferable to use robust methods, like MLqE [Ferrari and Yang, 2010, Qin and Priebe, 2013a]. We prove in this paper that entry-wise MLqE ($\widehat{P}^{(q)}$ in Figure 1) improves the performance compared to entry-wise MLE ($\widehat{P}^{(1)}$ in Figure 1) whenever contamination is relatively large.

Similarly, in order to take advantage of the low-rank structure, we enforce a low-rank approximation on the entry-wise MLqE. We prove that, under proper assumptions, the new estimator ($\widetilde{P}^{(q)}$ in Figure 1) not only inherits the robust property from MLqE ($\widehat{P}^{(q)}$ in Figure 1), but also wins the bias-variance tradeoff by taking advantage of the low-rank structure.

2 Introduction (Alternate)

Statistical analysis on networks is burgeoning in varies areas, for example neurosciences, social networks, politics, etc. It is vital to know the connections among the objects. As to each of the fields we mentioned, the connections are the white-matter tract among regions in a neural network, the friendship relations in a social network, and the interactions between members of a political party. Many models are proposed to capture such connections among objects within the random graphs. Goldenberg et al. [2010] gave a nice review of statistical models for networks. In particular, Bollobás et al. [2007] discussed inhomogeneous random graphs, which is more practical compared to the homogenous random graph models.

In a general parametric framework, $G \sim f \in \mathcal{F} = \{f_\theta : \theta \in \Theta\}$, selecting a good estimator $\widehat{\theta}(G)$ for the unknown θ given a finite graph sample $\{G^{(1)}, \cdots, G^{(m)}\}$ is the core to understand the specific connections. It is not only important for the estimation itself, but also can help with other statistical inference procedures. For example, Ginestet et al. [2014] proposed a method to test if there is a difference between the networks of two groups of subjects. While hypothesis testing is the final goal, the estimation procedure is a key intermediate step and can be improved.

Take the CoRR graphs experiment as an example. We have 114 brain scans. Each graph has 70 vertices representing different regions in the brain according to the Desikan atlas, while the weight of an edge between two vertices represents the number of white-matter tract connecting the corresponding two regions of the brain. Our goal in this situation is to estimate the average number of white-matter tract among different regions of the brain. A more accurate estimate can lead to a better understanding of the functionality of different parts of the brain. Also, it can potentially help with other tasks, such as diagnosis of brain disease. Details of this experiment are discussed in Section 7.2.

The maximum likelihood estimate, which happens to be the entry-wise sample mean in the example above, is a natural candidate of the estimation problem. However, it suffers from at least two major problems: high variance and non-robustness.

In a high dimensional situation, the maximum likelihood estimator often leads to inaccurate estimates with very high variance when the sample size is small. In a general setting, graphs are high dimensional object since n^2 parameters are needed to model n^2 possible edges, where n denotes the number of vertices. Thus the MLE does not perform very well especially when there are only a few observations available, which is likely the case in real world. However, if the graphs are low-rank, by biasing towards low-rank structures carefully, the estimators will have a greatly reduced variances and win the bias-variance tradeoff. In our CoRR graphs experiment, we see the quasi low-rank property by Figure 5. Recently, Tang et al. [2016] considers an estimator based on a low-rank approximation and proves that the new estimator outperforms the MLE since it decreases the overall asymptotic variance dramatically by smoothing towards the low-rank structure.

Another problem is that observations are usually contaminated in practice. So the weights of the edges are possibly observed with noise. The MLE is asymptotically efficient, i.e. when sample size is large enough, the MLE is at least as accurate as any other estimator. However, when the sample size is moderate, robust estimators always outperforms MLE in terms of mean squared error by winning the bias-variance tradeoff. Moreover, under contamination models, robust estimators can even beat MLE asymptotically since they are designed to be not unduly affected by the outliers. Thus, with contaminations, it is more preferable to use robust methods, like MLqE [Ferrari and Yang, 2010, Qin and Priebe, 2013a] considered in this paper.

To resolve these two problems simultaneously, we propose an estimator, which is a natural extension of [Tang et al., 2016]. It not only inherits the robust property from MLqE, but also wins the bias-variance tradeoff by taking advantage of the low-rank structure.

We organize the paper as follows. In Section 3, we extend the Independent

Edge Model, Random Dot Product Graph Model, and Stochastic Blockmodel to the weighted version naturally and define the contamination model we are going to consider. In Section 4, we propose two estimators other than the entry-wise MLE, in order to resolve the two issues we mentioned above respectively. And then we construct our final estimator by combining the two estimators, which improve in both aspects. In Section 5, we prove that our final estimator is the best among four estimators under some conditions, which is generalized in Section 6. In Section 7, we illustrate the improvement of our final estimator through experimental results on simulated and real data.

3 Models

For this work, we are in the scenario where m weighted graphs on n vertices are given in the adjacency matrices form $\{A^{(t)}\}(t = 1, \ldots, m)$. The graphs are undirected without self-loop, i.e. each $A^{(t)}$ is symmetric with zeros along the diagonal. Moreover, we assume the vertex correspondence is known across different graphs, so that vertex i of the t_1-th graph corresponds to vertex i of the t_2-th graph for any $i \in [n]$, $t_1, t_2 \in [m]$.

In this section, we present three nested models, the weighted independent edge model (WIEM) in Section 3.1, the weighted random dot product graph model (WRDPG) in Section 3.2, and the weighted stochastic blockmodel (WSBM) as a WRDPG in Section 3.3. Moreover, we introduce a contaminated model based on Section 3.3 in Section 3.4.

3.1 Weighted Independent Edge Model

As we know, in an independent edge model (IEM) [Bollobás et al., 2007] with probability matrix $P \in [0, 1]^{n\times n}$, every edge weight A_{ij} is distributed from a Bernoulli distribution with parameter P_{ij} independent of other edges. We first extend the definition of IEM to the weighted independent edge model (WIEM) with respect to a one-parameter family $\mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}\}$, e.g. f_θ could be the Poisson distribution with parameter θ. Denote the graph parameters as a matrix $P \in \Theta^{n\times n} \subset \mathbb{R}^{n\times n}$. Then under a WIEM, each edge between vertex i and vertex j ($i < j$ because of symmetry) has weight A_{ij} distributed from $f_{P_{ij}}$ independently.

To see that an IEM is a special case of WIEM, just let $\mathcal{F}$ be the collection of Bernoulli distributions and let the graph parameters be a symmetric and hollow matrix $P \in [0, 1]^{n\times n}$. Note that the graphs considered in this paper are undirected without self-loop, thus the parameter matrix P needs to be symmetric and hollow. However, for convenience, we still define the parameters to be an n-by-n matrix while only $\binom{n}{2}$ of them are effective.

3.2 Weighted Random Dot Product Graph

The connectivity between two vertices in a graph generally depends on some hidden properties of the corresponding vertices. The latent position model proposed by Hoff et al. [2002] captures such properties by assigning each vertex i with a corresponding latent vector $X_i \in \mathbb{R}^d$. Conditioned on the latent vectors

X_i and X_j, the edge weight between vertex i and vertex j is independent of all other edges and depends only on X_i and X_j through a link function.

A special case of the latent position model is the random dot product graph model (RDPG) in which the link function is the inner product [Nickel, 2007, Young and Scheinerman, 2007]. Now we give a definition of the weighted random dot product graph (WRDPG) as a special case of the weighted latent position model as following:

Definition 3.1 (Weighted Random Dot Product Graph Model) *Consider a collection of one-parameter distributions $\mathcal{F} = \{f_\theta, \theta \in \Theta \subset \mathbb{R}\}$. The weighted random dot product graph model (WRDPG) with respect to $\mathcal{F}$ is defined as following: Let $X \in \mathbb{R}^{n\times d}$ be such that $X = [X_1, X_2, \ldots, X_n]^\top$, where $X_i \in \mathbb{R}^d$ for all $i \in [n]$. The matrix X is random and satisfies $\mathbb{P}\left[X_i^\top X_j \in \Theta\right] = 1$ for all $i, j \in [n]$. Conditioned on X, the entries of the adjacency matrix A are independent and A_{ij} is a random variable following distribution $f_\theta \in \mathcal{F}$ with parameter $\theta = X_i^\top X_j$ for all $i \neq j \in [n]$.*

Under the WRDPG defined above, the parameter matrix $P = XX^T \in \mathbb{R}^{n\times n}$ is automatically symmetric because the link function is inner product. Moreover, to have symmetric graphs without self-loop, only A_{ij} $(i < j)$ are sampled while leaving the diagonals of A to be all zeros.

3.3 Weighted Stochastic Blockmodel as a Weighted Random Dot Product Graph

Community structure is an important property of graphs under which vertices are clustered into different communities such that vertices within the same community behave similarly. The stochastic blockmodel (SBM) proposed by Holland et al. [1983] captures such property, where each vertex is assigned to one block and the connectivity between two vertices depends only on their respective block memberships.

Formally, the SBM is determined by the number of blocks K (generally much smaller than the number of vertices n), block probability matrix $B \in [0,1]^{K\times K}$, and the block assignment vector $\tau \in [K]^n$, where $\tau_i = k$ represents that vertex i belongs to block k. Conditioned on the block membership τ, the connectivity between vertex i and vertex j follows a Bernoulli distribution with parameter B_{τ_i,τ_j}. This can be easily generalized to the weighted stochastic blockmodel (WSBM), with the Bernoulli distributions replaced by a general distribution family one-parameter distributions $\mathcal{F} = \{f_\theta, \theta \in \Theta \subset \mathbb{R}\}$ and the block probability matrix to be $B \in \Theta^{K\times K} \subset \mathbb{R}^{K\times K}$.

Since the RDPG/WRDPG setting motivates the low-rank estimator, all analysis in this work are based on such setting. In order to consider WSBM as a WRDPG, the block probability matrix B needs to be positive semi-definite by the structure of WRDPG. From now on, we will denote the sub-model of WSBM with positive semi-definite B as the WSBM.

Now consider the WSBM as a WRDPG with respect to $\mathcal{F} = \{f_\theta, \theta \in \Theta \subset \mathbb{R}\}$. Let $d = \text{rank}(B)$, then all vertices in block k have shared latent position $\nu_k \in \mathbb{R}^d$, where $B = \nu\nu^\top$ and $\nu = [\nu_1, \ldots, \nu_K]^\top \in \mathbb{R}^{K\times d}$. That is to say, $X_i = \nu_{\tau_i}$ and A_{ij} $(i < j)$ is distributed from f with parameter $B_{\tau_i,\tau_j} = \nu_{\tau_i}^\top \nu_{\tau_j}$.

Here the parameter matrix $P \in \mathbb{R}^{n\times n}$ is symmetric, hollow, and satisfies $P_{ij} = X_i^\top X_j = \nu_{\tau_i}^\top \nu_{\tau_j} = B_{\tau_i,\tau_j}$.

In order to generate m graphs under this model with known vertex correspondence, we first sample τ from the categorical distribution with parameter $\rho = [\rho_1, \cdots, \rho_K]^\top$ with $\rho_k \in (0,1)$ and $\sum_{k=1}^K \rho_k = 1$, and keep ρ fixed when sampling all m graphs. Then m symmetric and hollow graphs are sampled such that conditioning on τ, the adjacency matrices are distributed entry-wise independently as $A_{ij}^{(t)} \overset{ind}{\sim} f_{B_{\tau_i,\tau_j}} = f_{P_{ij}}$ for each $1 \le t \le m$, $1 \le i < j \le n$.

3.4 Weighted Stochastic Blockmodel as a Weighted Random Dot Product Graph with Contaminations

In practice, we can hardly get data accurately. So there will always be noise in the observations, which deviates from our general model assumptions. In order to incorporate this effect, a contamination model, the gross error model [Bickel and Doksum, 2001, Mah and Tamhane, 1982], is considered in this work.

Generally in a gross error model, we observe good measurement $G^* \sim f_P \in \mathcal{F}$ most of the time, while there are a few wild values $G^{**} \sim h_C \in \mathcal{H}$ when the gross errors occur. Here P and C represent the respective parameter matrices of the two distribution families. As to the graphs, one way to generalize from the gross error model is to contaminate the entire graph with some small probability $\epsilon \in (0,1)$, that is $G \sim (1-\epsilon)f_P + \epsilon h_C$. However, since all the models we consider are subsets of the WIEM, it is more natural to consider the contaminations with respect to each edge, i.e. for $1 \le i < j \le n$, $G_{ij} \sim (1-\epsilon)f_{P_{ij}} + \epsilon h_{C_{ij}}$ with $f \in \mathcal{F}$ and $h \in \mathcal{H}$, where both $\mathcal{F}$ and $\mathcal{H}$ are one-parameter distribution families.

In this paper, we assume that when gross errors occur, the weights of the edges are also from the same one-parameter family $\mathcal{F}$. Moreover, we also assume that the connectivity follows the WSBM as a WRDPG. Thus, similar to the uncontaminated distribution $f_{P_{ij}}$ with $P_{ij} = B_{\tau_i,\tau_j}$ where B is the block probability matrix and τ is the block assignments, the contamination distribution $f_{C_{ij}}$ with $C_{ij} = B'_{\tau'_i,\tau'_j}$ also have the block structure, where B' is the block probability matrix and τ' is the block assignments. For clarity, we will introduce the sampling procedure when the contamination has the same block structure, i.e. $\tau = \tau'$. However, it is not required in our theories.

To generate m graphs under this contamination model with known vertex correspondence, we first sample τ from the categorical distribution with parameter ρ and keep it fixed for all m graphs as in Section 3.3. Then m symmetric and hollow graphs $G^{(1)}, \ldots, G^{(m)}$ are sampled such that conditioning on τ, the adjacency matrices are distributed entry-wise independently as $A_{ij}^{(t)} \overset{ind}{\sim} (1-\epsilon)f_{P_{ij}} + \epsilon f_{C_{ij}}$ for each $1 \le t \le m$, $1 \le i < j \le n$, where $P_{ij} = B_{\tau_i,\tau_j}$ and $C_{ij} = B'_{\tau_i,\tau_j}$. Here ϵ is the probability of an edge to be contaminated, P is the parameter matrix as in Section 3.3, and C is the parameter matrix for contaminations.

4 Estimators

Under any model introduced in Section 3, our goal is to estimate the parameter matrix P based on the m observations $A^{(1)}, \ldots, A^{(m)}$. Especially when under

the contamination model, although there are other parameters like ϵ and C, our goal is still to estimate the uncontaminated parameter matrix P. In this section, we present four estimators as in Figure 1, i.e. the standard entry-wise MLE $\widehat{P}^{(1)}$, the low-rank approximation of the entry-wise MLE $\widetilde{P}^{(1)}$, the entry-wise robust estimator MLqE $\widehat{P}^{(q)}$, and the low-rank approximation of the entry-wise MLqE $\widetilde{P}^{(q)}$. Since the observed graphs are symmetric and hollow with a symmetric parameter matrix of the model, we do not care about the estimate of the diagonal of P. However, the estimate itself should be at least symmetric.

4.1 Entry-wise Maximum Likelihood Estimator $\widehat{P}^{(1)}$

Under the WIEM, the most natural estimator is the MLE, which happens to be the element-wise MLE $\widehat{P}^{(1)}$ in this case. Moreover, when $\mathcal{F}$ is a one-parameter exponential family, for instance Bernoulli and Exponential, the entry-wise MLE $\widehat{P}^{(1)}$ is uniformly minimum-variance unbiased estimator, i.e. it has the smallest variance among all unbiased estimators. In addition, it satisfies many good asymptotic properties as the number of graphs m goes to infinity. However, in high dimensional situations like this, the entry-wise MLE often leads to inaccurate estimates with very high variance when the sample size m is small. Also, it does not exploit any graph structure. The performance will not get any better when the number of vertices in each graph n increases since it is an entry-wise estimator. Moreover, if the graphs are actually distributed under a WRDPG or a WSBM, then the entry-wise MLE is no longer the MLE any more and the performance can be very poor.

4.2 Estimator $\widetilde{P}^{(1)}$ Based on Adjacency Spectral Embedding of $\widehat{P}^{(1)}$

Motivated by the low-rank structure of the parameter matrix P in WRDPG, we consider the estimator $\widetilde{P}^{(1)}$ proposed by Tang et al. [2016] based on the spectral decomposition of $\widehat{P}^{(1)}$. We introduce the dimension selection technique in Section 4.2.2 and the diagonal augmentation procedure is discussed in Section 7.1.2. The construction procedure of $\widetilde{P}^{(1)}$ consists of several steps, which will be introduced respectively in the following subsections.

4.2.1 Rank-d Approximation

Given a dimension d, we consider $\widetilde{P}^{(1)} = \text{lowrank}_d(\widehat{P}^{(1)})$ as the best rank-d positive semi-definite approximation of $\widehat{P}^{(1)}$. To find such best approximation, first calculate the eigen-decomposition of the symmetric matrix $\widehat{P}^{(1)} = \widehat{U}\widehat{S}\widehat{U}^\top + \widetilde{U}\widetilde{S}\widetilde{U}^\top$, where $\widehat{S}$ is the diagonal matrix with the largest d eigenvalues of $\widehat{P}^{(1)}$, and $\widehat{U}$ has the corresponding eigenvectors as each column. Similarly, $\widetilde{S}$ is the diagonal matrix with non-increasing entries along the diagonal corresponding to the rest $n-d$ eigenvalues of $\widehat{P}^{(1)}$, and $\widetilde{U}$ has the columns given by the corresponding eigenvectors. The d-dimensional adjacency spectral embedding (ASE) of $\widehat{P}^{(1)}$ is given by $\widehat{X} = \widehat{U}\widehat{S}^{1/2} \in \mathbb{R}^{n\times d}$. Based on the ASE result, we have the best rank-d positive semi-definite approximation of $\widehat{P}^{(1)}$ to be $\widetilde{P}^{(1)} = \widehat{X}\widehat{X}^\top = \widehat{U}\widehat{S}\widehat{U}^\top$. In the RDPG setting, Sussman et al. [2014] proved that each row of $\widehat{X}$ can accurately estimate the the latent position for each vertex up to an

orthogonal transformation. We will analyze its performance under the WRDPG setting in Section 5.

Here, we restate the algorithm in [Tang et al., 2016] to give the detailed steps of computing this low-rank approximation of a general n-by-n symmetric matrix A in Algorithm 1.

Algorithm 1 Algorithm to compute the rank-d approximation of a matrix.

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

1: Compute the algebraically largest d eigenvalues of A, $s_1 \geq s_2 \geq \ldots \geq s_d$ and corresponding unit-norm eigenvectors $u_1, u_2, \ldots, u_d \in \mathbb{R}^n$;
2: Set $\widehat{S}$ to the $d \times d$ diagonal matrix $\text{diag}(s_1, \ldots, s_d)$;
3: Set $\widehat{U} = [u_1, \ldots, u_d] \in \mathbb{R}^{n\times d}$;
4: Set $\text{lowrank}_d(A)$ to $\widehat{U}\widehat{S}\widehat{U}^\top$;

4.2.2 Dimension Selection

Although Algorithm 1 gives us a way to calculate the best rank-d positive semi-definite approximation of a general symmetric matrix A, it does not tell us how to select a proper dimension d. If we choose a relatively small dimension d, the estimator based on approximation will fail to catch much important information. On the other hand, when d is too large, the approximation will contain too much noise and also lead to a bad estimate. So a carefully selected dimension d is a key part of a good estimation.

A general idea of selecting the dimension d is to analyze the ordered eigenvalues and looking for the "gap" or "elbow" in the scree-plot. In 2006, Zhu and Ghodsi [2006] proposed an automatic method for finding the gap in the scree-plot by only looking at the eigenvalues based on a Gaussian mixture model. This method provides multiple choices based on different elbow. In this paper, to avoid under-estimating the dimension, which is often much more harmful than over-estimating it, we always choose the 3rd elbow.

Although it is always challenge to select a proper dimension, based on the results of real data experiment in Section 7.2, a wide range of dimensions will lead to a fairly good results. Thus a proper dimension selection method can be applied directly without carefully tuning the parameter, which makes the estimator much more useful in practice.

Algorithm 2 Algorithm to compute $\widetilde{P}^{(1)}$

Input: Symmetric adjacency matrices $A^{(1)}, A^{(2)}, \ldots, A^{(m)}$, with each $A^{(t)} \in \mathbb{R}^{n\times n}$
Output: Estimate $\widetilde{P}^{(1)} \in \mathbb{R}^{n\times n}$

1: Calculate the entry-wise MLE $\widehat{P}^{(1)}$;
2: Select the dimension d based on the eigenvalues of $\widehat{P}^{(1)}$; (see Section 4.2.2)
3: Set Q to $\text{lowrank}_d(\widehat{P}^{(1)})$; (see Algorithm 1)
4: Set $\widetilde{P}^{(1)}$ with each entry $\widetilde{P}^{(1)}_{ij} = \max(Q_{ij}, 0)$.

By combining the key parts introduced above, we give the detailed descrip-

tion for calculating the estimator $\widetilde{P}^{(1)}$ with dimension selection method in Algorithm 2.

4.3 Entry-wise Maximum Lq-likelihood Estimator $\widehat{P}^{(q)}$

The MLE is asymptotically efficient, i.e. when sample size is large enough, the MLE is at least as accurate as any other estimator. However, when the sample size is moderate, robust estimators always outperforms MLE in terms of mean squared error by winning the bias-variance tradeoff. Moreover, under contamination models, robust estimators can even beat MLE asymptotically since they are designed to be not unduly affected by the outliers. And now we are going to consider one robust estimator, i.e. the maximum Lq-likelihood estimator (MLqE) proposed by Ferrari and Yang [2010].

Let $X_1, \ldots, X_m$ be sampled from $f_{\theta_0} \in \mathcal{F} = \{f_\theta, \theta \in \Theta\}$, $\theta_0 \in \Theta$. Then the maximum Lq-likelihood estimate ($q > 0$) of θ_0 based on the parametric model $\mathcal{F}$ is defined as

$$\widehat{\theta}_{\mathrm{ML}q\mathrm{E}} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{m} L_q[f_\theta(X_i)],$$

where $L_q(u) = (u^{1-q} - 1)/(1-q)$. Note that $L_q(u) \to \log(u)$ when $q \to 1$. Thus MLqE is a generalization of MLE. Moreover, define

$$U_\theta(x) = \nabla_\theta \log f_\theta(x)$$

and

$$U_\theta^\star(x; q) = U_\theta(x) f_\theta(x)^{1-q}.$$

Then the MLqE $\widehat{\theta}_{\mathrm{ML}q\mathrm{E}}$ can also be seen as a solution to the equation

$$\sum_{i=1}^{m} U_\theta^\star(X_i; q) = 0.$$

This form interprets $\widehat{\theta}_{\mathrm{ML}q\mathrm{E}}$ as a solution to the weighted likelihood equation. The weights $f_\theta(x)^{1-q}$ are proportional to the $(1-q)$th power of the corresponding probability. Specifically, when $0 < q < 1$, the MLqE puts less weight on the data points which do not fit the current distribution well. Equal weights happens when $q = 1$ and lead to the MLE.

Under the WIEM, we can calculate the robust entry-wise MLqE $\widehat{P}^{(q)}$ based on the adjacency matrices $A^{(1)}, \ldots, A^{(m)}$. Note that $\widehat{P}^{(1)}$, the entry-wise MLE, is a special case of entry-wise MLqE $\widehat{P}^{(q)}$ when $q = 1$. That is what the superscriptions q and 1 mean. There is also a bias-variance tradeoff in selecting the parameter q. Qin and Priebe [2013b] proposed a way to select q in general. In this work, we do not focus on how to select q.

4.4 Estimator $\widetilde{P}^{(q)}$ Based on Adjacency Spectral Embedding $\widehat{P}^{(q)}$

Intuitively, the low-rank structure of the parameter matrix P in WRDPG should be preserved more or less in the entry-wise MLqE $\widehat{P}^{(q)}$. Thus, in order to take advantage of such low-rank structure as well as the robustness, we apply the similar idea here as in building $\widetilde{P}^{(1)}$, i.e. enforce a low-rank approximation on

the entry-wise MLqE matrix $\widehat{P}^{(q)}$ to get $\widetilde{P}^{(q)}$. As in Algorithm 2, we apply the same dimension selection method and diagonal augmentation procedure. The only change is to substitute $\widehat{P}^{(1)}$ by $\widehat{P}^{(q)}$. The details of the algorithm is shown in Algorithm 3.

Algorithm 3 Algorithm to compute $\widetilde{P}^{(q)}$

Input: Symmetric adjacency matrices $A^{(1)}, A^{(2)}, \ldots, A^{(m)}$, with each $A^{(t)} \in \mathbb{R}^{n\times n}$

Output: Estimate $\widetilde{P}^{(q)} \in \mathbb{R}^{n\times n}$

1: Calculate the entry-wise MLqE $\widehat{P}^{(q)}$;
2: Select the dimension d based on the eigenvalues of $\widehat{P}^{(q)}$; (see Section 4.2.2)
3: Set Q to $\text{lowrank}_d(\widehat{P}^{(q)})$; (see Algorithm 1)
4: Set $\widetilde{P}^{(q)}$ with each entry $\widetilde{P}^{(q)}_{ij} = \max(Q_{ij}, 0)$.

5 Theoretical Results

In this section, for illustrative purpose, we are going to present theoretical results when the contamination model introduced in Section 3.4 is with respect to exponential distributions. That is $\mathcal{F} = \{f_\theta(x) = \frac{1}{\theta}e^{-x/\theta}, \theta \in [0, R] \subset \mathbb{R}\}$, where $R > 0$ is a constant. The results can be extended to a general situation with proper assumptions, which will be discussed in Section 6.

For clarity, we restate the model settings discussed in Section 3.4. Consider the SBM with parameter B and ρ. First sample the block membership τ from the categorical distribution with parameter ρ and keep it fixed for all m graphs. Conditioned on this τ we sampled, the probability matrix P then satisfies $P_{ij} = B_{\tau_i,\tau_j}$. In this section, we assume the contamination has the same block membership τ, thus the contamination matrix $C \in \mathbb{R}^{n\times n}$ has the same block structure as P. Note that this is not necessary for the result. Different block structure can lead to the same result since the rank is still finite. Denote ϵ as the probability of an edge to be contaminated. Then m symmetric graphs $G^{(1)}, \ldots, G^{(m)}$ are sampled such that conditioning on τ, the adjacency matrices are distributed entry-wise independently as $A^{(t)}_{ij} \overset{ind}{\sim} (1-\epsilon)f_{P_{ij}} + \epsilon f_{C_{ij}}$ for each $1 \le t \le m$, $1 \le i < j \le n$.

Under such setting, we now analyze the performance of all four estimators based on m adjacency matrices for estimating the probability matrix P in terms of the mean squared error. When comparing two estimators, we mainly focus on both asymptotic bias and asymptotic variance. Note that all the results in this section are entry-wise, which can easily lead to a result of the total MSE for the entire matrix.

We only present the main results in this section. The proofs are given in the Supplementary Materials.

5.1 $\widehat{P}^{(1)}$ vs. $\widehat{P}^{(q)}$

We first compare the performance between the entry-wise MLE $\widehat{P}^{(1)}$ and the entry-wise MLqE $\widehat{P}^{(q)}$. Without using the graphs structure, the asymptotic

results for these two estimators are in terms of the number of graphs m, not the number of vertices n within each graph.

Theorem 5.1 *For any $0 < q < 1$, there exists $C_0(P_{ij}, \epsilon, q) > 0$ such that under the contaminated model with $C > C_0(P_{ij}, \epsilon, q)$,*

$$\lim_{m\to\infty} \left| E[\widehat{P}_{ij}^{(q)}] - P_{ij} \right| < \lim_{m\to\infty} \left| E[\widehat{P}_{ij}^{(1)}] - P_{ij} \right|,$$

for $1 \le i, j \le n$ and $i \neq j$. Moreover, without any assumption on the contaminated model, for $1 \le i, j \le n$,

$$\mathrm{Var}(\widehat{P}_{ij}^{(1)}) = \mathrm{Var}(\widehat{P}_{ij}^{(q)}) = O(1/m).$$

And thus

$$\lim_{m\to\infty} \mathrm{Var}(\widehat{P}_{ij}^{(1)}) = \lim_{m\to\infty} \mathrm{Var}(\widehat{P}_{ij}^{(q)}) = 0.$$

Theorem 5.1 shows that the entry-wise MLqE $\widehat{P}^{(q)}$ has smaller bias for estimating P asymptotically compared to the entry-wise MLE $\widehat{P}^{(1)}$. Although we put restrictions on the parameter matrix C in the statement of the theorem, the result still holds provided that $\epsilon(C_{ij} - P_{ij}) > (1-q)P_{ij}$. This condition only requires the contamination of the model is large enough (either large contamination parameter matrix, or more likely to encounter an outlier). From a different perspective, it also requires $\widehat{P}^{(q)}$ to be robust enough with respect to the contamination. Thus besides the current condition for C, equivalently, we can also replace it by the assumption of a large enough ϵ or a small enough q.

Theorem 5.1 also indicates that both estimators have variances converge to zero as the number of graphs m goes to infinity, following the asymptotic properties of minimum contrast estimates. Thus the bias term will dominate in the comparison in terms of MSE.

As a result, $\widehat{P}^{(q)}$ reduces the bias while keeping variance the same asymptotically compared to $\widehat{P}^{(1)}$. Thus in terms of MSE, $\widehat{P}^{(q)}$ is a better estimator than $\widehat{P}^{(1)}$ when the number of graphs m is large with enough contamination.

5.2 $\widehat{P}^{(1)}$ vs. $\widetilde{P}^{(1)}$

We next analyze the effect of the ASE procedure applied to the entry-wise MLE $\widehat{P}^{(1)}$ under the contamination model, so that we can compare the performance between $\widehat{P}^{(1)}$ and $\widetilde{P}^{(1)}$.

Before proceeding to the comparison between the two estimators, we first recall the definition of the asymptotic relative efficiency (ARE) [Serfling, 2011], which is a very important and useful criterion to compare two estimators. Note that the original definition is for unbiased estimators. Here we adapt the definition to estimators with the same asymptotic bias.

Definition 5.2 *For any parameter θ of a distribution f, and for estimators $\widehat{\theta}^{(1)}$ and $\widehat{\theta}^{(2)}$ such that $E[\widehat{\theta}^{(1)}] = E[\widehat{\theta}^{(2)}] = \theta'$, $n \cdot \mathrm{Var}(\widehat{\theta}^{(1)}) \to V_1(f)$ and $n \cdot \mathrm{Var}(\widehat{\theta}^{(2)}) \to V_2(f)$, the ARE of $\widehat{\theta}^{(2)}$ to $\widehat{\theta}^{(1)}$ is given by*

$$\mathrm{ARE}(\widehat{\theta}^{(2)}, \widehat{\theta}^{(1)}) = \frac{V_1(f)}{V_2(f)}.$$

By the definition above, if $\mathrm{ARE}(\widehat{\theta}^{(2)}, \widehat{\theta}^{(1)}) < 1$, then $\widehat{\theta}^{(1)}$ has a smaller variance in its sampling distribution and thus is more efficient compared to $\widehat{\theta}^{(2)}$. Combine with the fact that both estimators have the same asymptotic bias, $\widehat{\theta}^{(1)}$ is a better estimate in this case.

To compare $\widehat{P}^{(1)}$ and $\widetilde{P}^{(1)}$, we will first show they have the same entry-wise asymptotic bias under proper conditions, and then use the ARE criterion to compare the performance in the following theorem.

Theorem 5.3 *Assuming that $m = O(n^b)$ for any $b > 0$, then*

$$\lim_{n\to\infty} \mathrm{Bias}(\widetilde{P}_{ij}^{(1)}) = \lim_{n\to\infty} E[\widetilde{P}_{ij}^{(1)}] - P_{ij} = \lim_{n\to\infty} E[\widehat{P}_{ij}^{(1)}] - P_{ij} = \lim_{n\to\infty} \mathrm{Bias}(\widehat{P}_{ij}^{(1)}).$$

In addition, for $1 \le i, j \le n$ and $i \ne j$,

$$\mathrm{Var}(\widetilde{P}_{ij}^{(1)}) = O(m^{-1}n^{-1}(\log n)^3), \mathrm{Var}(\widehat{P}_{ij}^{(1)}) = O(m^{-1}).$$

And thus

$$\frac{\mathrm{Var}(\widetilde{P}_{ij}^{(1)})}{\mathrm{Var}(\widehat{P}_{ij}^{(1)})} = O(n^{-1}(\log n)^3),$$

$$\mathrm{ARE}(\widehat{P}_{ij}^{(1)}, \widetilde{P}_{ij}^{(1)}) = 0.$$

Theorem 5.3 says that when m is a constant, or m is going to infinity with order $m = O(n^b)$ for any $b > 0$, i.e. m is fixed or it grows not faster than any polynomial with respect to n, the ASE procedure applied to $\widehat{P}^{(1)}$ will not affect the asymptotic bias for estimating P. Combined with the fact that the ratio of the variances of two estimators is of order $O(n^{-1}(\log n)^3)$, we have that ARE goes to 0 when $n \to \infty$. Thus $\widetilde{P}_{ij}^{(1)}$ is much better than $\widehat{P}_{ij}^{(1)}$ for a large n. We emphasize that the order of the ratio of the variances does not depend on m.

As a result, the ASE procedure applied to the entry-wise MLE $\widehat{P}^{(1)}$ helps reduce the variance while keeping the bias unchanged asymptotically, leading to a better estimate $\widetilde{P}^{(1)}$ for P in terms of MSE.

5.3 $\widehat{P}^{(q)}$ vs. $\widetilde{P}^{(q)}$

We now proceed to analyze the effect of the ASE procedure applied to the entry-wise MLqE $\widehat{P}^{(q)}$ under the contamination model in order to compare the performance between $\widehat{P}^{(q)}$ and $\widetilde{P}^{(q)}$. Similarly, we first show that the two estimators have the same entry-wise asymptotic bias under proper conditions, and then use the ARE criterion to compare the performance in the following theorem.

Theorem 5.4 *Assuming that $m = O(n^b)$ for any $b > 0$, then the estimator based on ASE of MLqE has the same entry-wise asymptotic bias as MLqE, i.e.*

$$\lim_{n\to\infty} \mathrm{Bias}(\widetilde{P}_{ij}^{(q)}) = \lim_{n\to\infty} E[\widetilde{P}_{ij}^{(q)}] - P_{ij} = \lim_{n\to\infty} E[\widehat{P}_{ij}^{(q)}] - P_{ij} = \lim_{n\to\infty} \mathrm{Bias}(\widehat{P}_{ij}^{(q)}).$$

In addition, for $1 \le i, j \le n$ and $i \ne j$,

$$\mathrm{Var}(\widetilde{P}_{ij}^{(q)}) = O(n^{-1}(\log n)^3), \mathrm{Var}(\widehat{P}_{ij}^{(q)}) = O(m^{-1}).$$

And thus

$$\frac{\mathrm{Var}(\widetilde{P}_{ij}^{(q)})}{\mathrm{Var}(\widehat{P}_{ij}^{(q)})} = O(mn^{-1}(\log n)^3).$$

Moreover, if $m = o(n(\log n)^{-3})$*, then*

$$\mathrm{ARE}(\widehat{P}_{ij}^{(q)}, \widetilde{P}_{ij}^{(q)}) = 0.$$

The proof for Theorem 5.4 is almost the same as the proof for Theorem 5.3. But unlike the results for MLE, we are missing the term m^{-1} in the variance bound $\mathrm{Var}(\widetilde{P}^{(q)}) = O(n^{-1}(\log n)^3)$ due to the structure of maximum Lq likelihood equation. As a result, while the ASE procedure still does not affect the asymptotic bias, the ARE has an extra term m. This leads to a slight difference in the comparison. Specifically, when m is fixed, the order of the ARE is $O(n^{-1}(\log n)^3)$, which will goes to 0 as $n \to \infty$. Even if m also increases as n increases, as long as it grows in the order of $o(n(\log n)^{-3})$, the ARE still goes to 0.

Thus the ASE procedure applied to the entry-wise MLqE $\widehat{P}^{(q)}$ also helps reduce the variance while keeping the bias asymptotically, leading to a better estimate $\widetilde{P}^{(q)}$ for P in terms of MSE.

5.4 $\widetilde{P}^{(1)}$ vs. $\widetilde{P}^{(q)}$

To finish the last piece, we compare the performance between $\widetilde{P}^{(1)}$ and $\widetilde{P}^{(q)}$ by combining the previous results.

Theorem 5.5 *For sufficiently large* C *and any* $1 \le i, j \le n$*, if* $m = O(n^b)$ *for any* $b > 0$*, then*

$$\lim_{m,n\to\infty} \mathrm{Bias}(\widetilde{P}_{ij}^{(1)}) > \lim_{m,n\to\infty} \mathrm{Bias}(\widetilde{P}_{ij}^{(q)})$$

Moreover, if $m = O(n(\log n)^{-3})$*, then*

$$\lim_{m,n\to\infty} \mathrm{Var}(\widetilde{P}_{ij}^{(1)}) = \lim_{m,n\to\infty} \mathrm{Var}(\widetilde{P}_{ij}^{(q)}) = 0.$$

Theorem 5.5 is a direct result of Theorem 5.1, Theorem 5.3, and Theorem 5.4. It concludes that $\widetilde{P}^{(q)}$ inherits the robustness from the entry-wise MLqE $\widehat{P}^{(q)}$ and has a smaller asymptotic bias compared to $\widetilde{P}^{(1)}$ while both estimates have variance goes to 0 as $m \to \infty$. Thus in summary, $\widetilde{P}^{(q)}$ is the best among all four estimators.

5.5 Summary

We summarize all the four estimators and their relationship in Figure 2. From top to bottom of the figure, we apply ASE to construct low-rank approximations which preserve the asymptotic bias and reduce the asymptotic variance. From left to right, we underweight the outliers to construct robust estimators. So with enough contaminations, whenever the number of graphs m is large enough, the bias term which dominates the MSE will be improved.

In conclusion, when contamination is relatively large as well as m and n, $\widetilde{P}^{(q)}$ is the best among the four estimators.

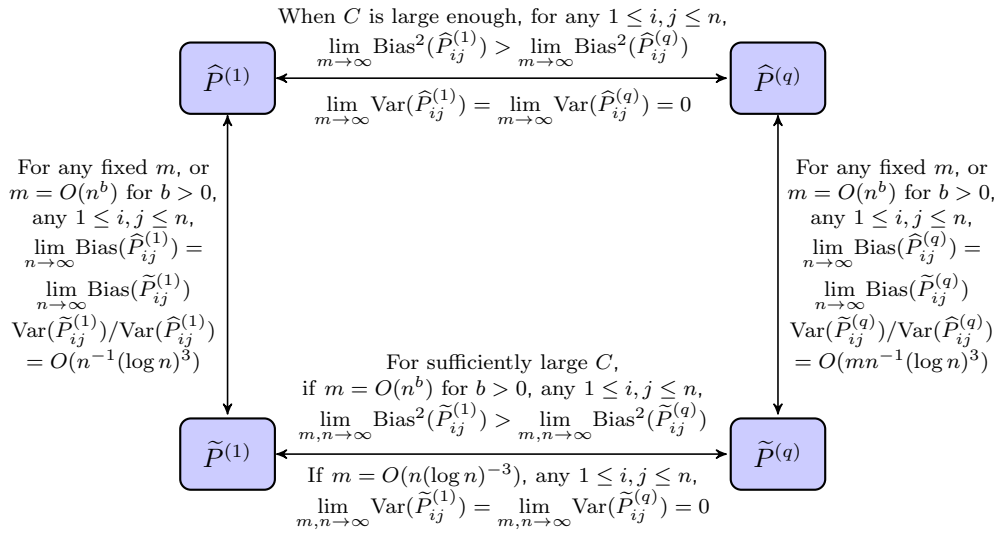

Figure 2: Relationship among four estimators.

6 Extensions

Results in Section 5 are presented in the setting of exponential distributions with MLqE estimator. However, the results can be generalized to a broader class of distribution families, and even a different entry-wise robust estimator (denoted as $\widehat{P}^{(R)}$) other than MLqE provided that the following conditions are satisfied:

1. Let $A_{ij} \overset{ind}{\sim} (1-\epsilon)f_{P_{ij}} + \epsilon f_{C_{ij}}$, then $E[(A_{ij} - E[\widehat{P}_{ij}^{(1)}])^k] \le \mathrm{const}^k \cdot k!$, where $\widehat{P}^{(1)}$ is the entry-wise MLE as defined before;

2. There exists $C_0(P_{ij}, \epsilon) > 0$ such that under the contaminated model with $C > C_0(P_{ij}, \epsilon)$,

$$\lim_{m\to\infty} \left| E[\widehat{P}_{ij}^{(R)}] - P_{ij} \right| < \lim_{m\to\infty} \left| E[\widehat{P}_{ij}^{(1)}] - P_{ij} \right|;$$

3. $\widehat{P}_{ij}^{(R)} \le \mathrm{const} \cdot \widehat{P}_{ij}^{(1)}$;

4. $\mathrm{Var}(\widehat{P}_{ij}^{(R)}) = O(m^{-1})$, where m is the number of observations.

Condition 1 is to ensure that observations will not deviate too far away from the expectation, so that the concentration inequalities hold; Condition 2 is discussed in Section 5.1. It requires the contamination of the model to be large enough (a restriction on the distribution) and $\widehat{P}$ to be sufficiently robust with respect to the contamination (a condition on the estimator); By taking advantage of Condition 1 which controls $\widehat{P}^{(1)}$, Condition 3 reuses Condition 1 to bound an arbitrary $\widehat{P}^{(R)}$; Condition 4 is to ensure that the variance of $\widehat{P}_{ij}^{(R)}$ is comparable to the variance of the entry-wise MLE $\widehat{P}_{ij}^{(1)}$, which is of

order $O(m^{-1})$. Nevertheless, should the variance be bigger, similar but weaker results can still be derived.

As an example of the distribution satisfying the above four conditions other than the exponential distribution mentioned in Section 5, we sketch the Poisson distribution as following. Poisson distribution is a commonly used distribution for nonnegative graphs with integer values. Lemma A.34 verifies Condition 1. Intuitively, since exponential distribution has a fatter tail compare to Poisson, we should have the bound for central moment of Poisson directly from the results for exponential distribution. Condition 2 is satisfied when it is the robust MLqE. For Condition 3, $\widehat{P}_{ij}^{(R)}/\widehat{P}_{ij}^{(1)}$ is maximized when there are m data $x_1, \cdots, x_m$ with $0 \leq x_1 = \cdots = x_k \leq \bar{x} \leq x_{k+1} = \cdots = x_m \leq m\bar{x}/(m-k)$. In order to have ML$q$E larger than MLE $\bar{x}$, we need the weights of the first m data to be smaller than the weights of the rest $m-k$ data. So $e^{-\bar{x}} < \bar{x}^{x_m} e^{-\bar{x}}/x_m!$. Then $x_m! < \bar{x}^{x_m}$. By the lower bound in Stirling's formula, we have $x_m < e\bar{x}$ when $x_m > 0$. Note that if $x_m = 0$ then MLE equals MLqE since all data equals zero. Thus MLqE is bounded by $e\bar{x}$. As a result, $\widehat{P}_{ij} \leq e\widehat{P}_{ij}^{(1)}$ and Condition 3 is satisfied. At last, Condition 4 follows directly from theory of minimum contrast estimators.

In summary, all theorems in Section 5 hold for the Poisson distribution. This section provides a general way to extend the theory to proper models and robust estimators.

7 Empirical Results

7.1 Simulation

In this section, we first illustrate the theoretical comparison among the four estimators discussed in Section 5 via various Monte Carlo simulation experiments in an idealized setting.

7.1.1 Simulation Setting

Here we consider the 2-block SBM with respect to the exponential distributions parameterized by

$$B = \begin{bmatrix} 4 & 2 \\ 2 & 7 \end{bmatrix}, \qquad \rho = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}.$$

Let the contamination also be a 2-block SBM with the same structure parameterized by

$$B' = \begin{bmatrix} 9 & 6 \\ 6 & 13 \end{bmatrix}, \qquad \rho = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}.$$

With these parameters specified, we sample graphs according to Section 3.4.

For the ease of presentation, in the simulation, we assume the true dimension $d = \text{rank}(B) = 2$ is known. So we ignore the dimension selection step in Algorithm 2 and Algorithm 3.

7.1.2 Diagonal Augmentation

Since the graphs considered in this paper have no self-loops, all the adjacency matrices $A^{(t)}$ $(1 \leq t \leq m)$ are hollow, i.e. all diagonal entries are zeros. Thus

the diagonal of the parameter matrix P does not matter since all off-diagonal entries are independent of them conditioned on the off-diagonal entries of P.

However, unlike the entry-wise estimators, e.g. $\widehat{P}^{(1)}$, the ones which take advantage of the graph structure need the information from the diagonals. As a result, the zero diagonals of the observed graphs will lead to unnecessary biases in those estimates.

To compensate for such unnecessary biases, Marchette et al. [2011] suggested to use the average of the non-diagonal entries of the corresponding row as the diagonal entry before embedding. Also, Scheinerman and Tucker [2010] proposed an iterative method, which gives a different approach to resolve such issue.

As suggested in [Tang et al., 2016], in this work we are going to combine both ideas by first using Marchette's row-averaging method and then another one-step Scheinerman's iterative method.

7.1.3 Simulation Results

In order to see how the performance of the four estimators varies with respect to the contaminations, we first run 1000 Monte Carlo replicates based on the contaminated SBM specified in Section 7.1.1 with a fixed number of vertices $n = 100$ and a fixed number of graphs $m = 20$ while varying the contamination probability ϵ from 0 to 0.4. Given each sample, four estimators can be calculated following Algorithm 2 and Algorithm 3. Since we are not focusing on how to select the parameter q in the MLqE estimator, we are going to use a fixed $q = 0.9$ throughout this paper. Then the MSE of each estimator can be estimated since the probability matrix P is known in this simulation.

The results are given in Figure 3. Different colors represent the simulated MSE associated with four different estimators. Firstly, we see MLE $\widehat{P}^{(1)}$ is the best estimator when there is little or no contamination (i.e. ϵ is small or $\epsilon = 0$); however it degrades dramatically as contamination increases. On the contrary, the MLqE $\widehat{P}^{(q)}$ is slightly less efficient than the MLE $\widehat{P}^{(1)}$ when the contamination is small, but is much more robust under a large contamination compared to the MLE. Next we see that even with a relative small number of vertices $n = 100$, the ASE procedure which takes advantage of the low rank structure already helps improve the performance of $\widehat{P}^{(1)}$ and let $\widetilde{P}^{(1)}$ win the bias-variance tradeoff. Since the MLqE $\widehat{P}^{(q)}$ preserves the low rank structure of the original graph more or less, the ASE procedure also helps and makes $\widetilde{P}^{(q)}$ a better estimate. Although both $\widetilde{P}^{(q)}$ and $\widetilde{P}^{(1)}$ take advantage of the low-rank structure and has reduced variances, $\widetilde{P}^{(q)}$ constructed based on MLqE inherits the robustness from MLqE in addition. So when the contamination is large enough, $\widetilde{P}^{(q)}$ outperforms $\widetilde{P}^{(1)}$ and degrades slower.

Figure 4 shows additional simulation results by varying the parameter q in MLqE with fixed $n = 100$, $m = 20$ and $\epsilon = 0.1$ based on 1000 Monte Carlo replicates. Different types of lines represent the simulated MSE associated with four different estimators. From the figure, we can see that the ASE procedure takes advantage of the graph structure and improves the performance of the corresponding estimators for a wide range of q. Moreover, within a large range of q, the MLqE wins the bias-variance tradeoff and shows the robustness property compare to the MLE. And as q goes to 1, MLqE goes to the MLE as expected.

By comparing the performance of the four estimators based on different setting, we demonstrate the theoretical results in Section 5.

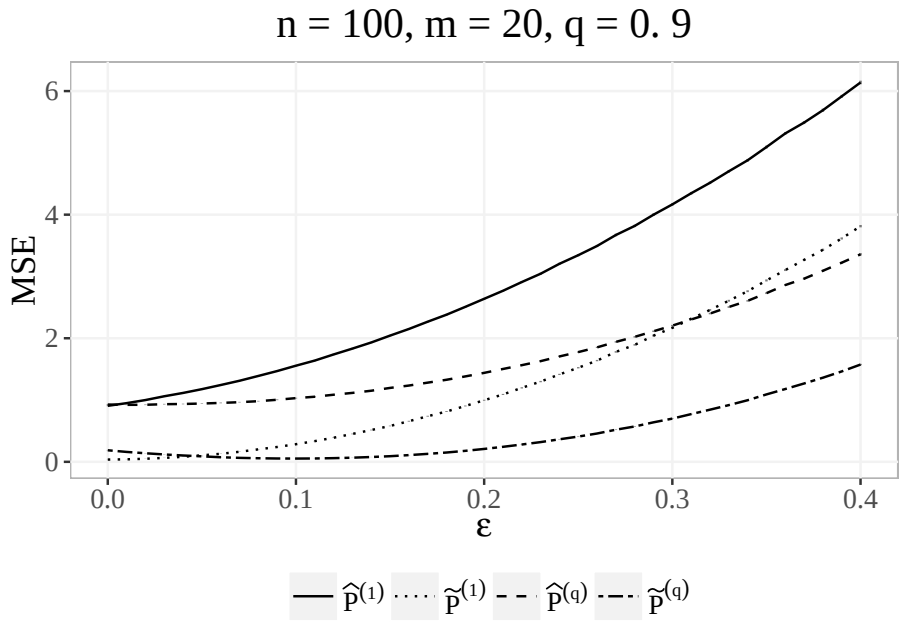

Figure 3: Mean squared error in average by varying contamination ratio ϵ with fixed $n = 100$ and $m = 20$ based on 1000 Monte Carlo replicates. And we use $q = 0.9$ when applying MLqE. Different colors represent the simulated MSE associated with four different estimators. **1. MLE $\widehat{P}^{(1)}$ vs MLqE $\widehat{P}^{(q)}$:** MLE outperforms a little bit when there is no contamination (i.e. $\epsilon = 0$), but it degrades dramatically when contamination increases; **2. MLE $\widehat{P}^{(1)}$ vs ASE $\circ$ MLE $\widetilde{P}^{(1)}$:** ASE procedure takes the low rank structure into account and $\widetilde{P}^{(1)}$ wins the bias-variance tradeoff; **3. MLqE $\widehat{P}^{(q)}$ vs ASE $\circ$ MLqE $\widetilde{P}^{(q)}$:** MLqE preserves the low rank structure of the original graph more or less, so ASE procedure still helps and $\widetilde{P}^{(q)}$ wins the bias-variance tradeoff; **4. ASE $\circ$ MLqE $\widetilde{P}^{(q)}$ vs ASE $\circ$ MLE $\widetilde{P}^{(1)}$:** When contamination is large enough, $\widetilde{P}^{(q)}$ based on MLqE is better, since it inherits the robustness from MLqE.

7.2 CoRR Graphs

We now compare the four estimators on a structural connectomic data. The graphs in this dataset are based on diffusion tensor MR images. There are 114 different brain scans, each of which was processed to yield an undirected, weighted graph with no self-loops, using the m2g pipeline described in [Kiar et al., 2016]. The vertices of the graphs represent different regions in the brain defined according to an atlas. We used the Desikan atlas with 70 vertices in this expereiment. The weight of an edge between two vertices represents the number of white-matter tract connecting the corresponding two regions of the brain.

Generally, we do not expect the graphs to perfectly follow an RDPG, or not even IEM. Before we calculate those estimators, we will perform some exploratory analysis to check whether the dataset could possibly have a low-rank

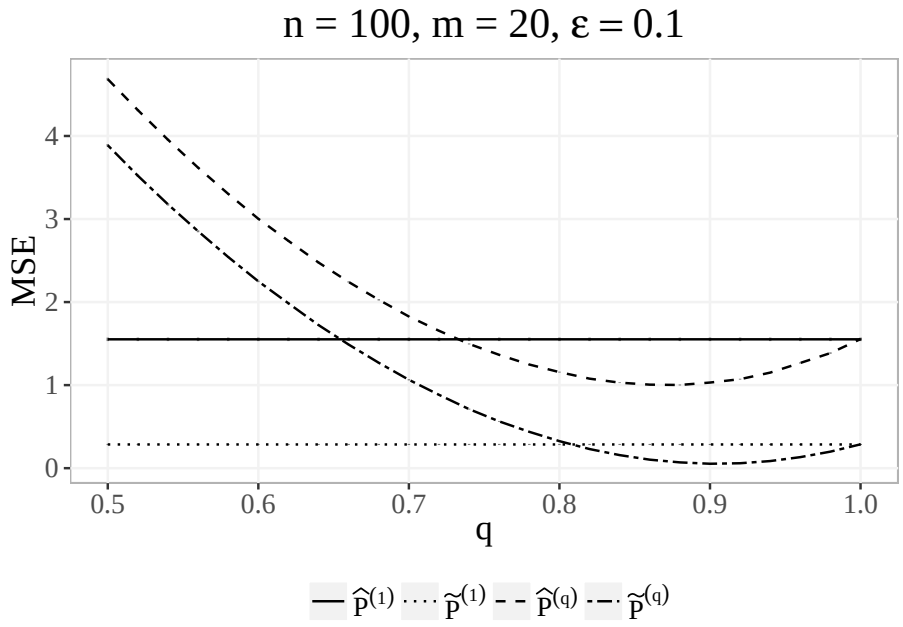

Figure 4: Mean squared error in average by varying the parameter q in MLqE with fixed $n = 100$, $m = 20$ and $\epsilon = 0.1$ based on 1000 Monte Carlo replicates. Different types of lines represent the simulated MSE associated with four different estimators. 1. ASE procedure takes advantage of the graph structure and improves the performance of the corresponding estimators independent of the selection of q; 2. Within a proper range of q, MLqE wins the bias-variance tradeoff and shows the robustness property compare to the MLE. Also as q goes to 1, MLqE goes to the MLE as expected.

structure. Indeed, without a low-rank structure, we will not expect the ASE procedure to improve the bias-variance tradeoff because of a potential high bias. In the left panel of Figure 5, we plot the eigenvalues of the mean graph of all 114 graphs (with diagonal augmentation) in decreasing algebraic order for the Desikan atlases based on the m2g pipeline. The eigenvalues first decrease dramatically and then stay around 0 for a large range of dimensions. In addition, we also plot the histograms in the right panel of Figure 5. From the figures we can see many eigenvalues are concentrated around zero. So the information is mostly contained in the first few dimensions. Such quasi low-rank property provides an opportunity to win the bias-variance tradeoff by applying ASE procedure.

We now discuss an important issue with respect to this current dataset. To compare the four estimators, we need a notion of the MSE, which requires the true parameter matrix P. However, unlike simulation experiment in Section 7.1, P is definitely not obtainable in practice since the 114 graphs themselves are also a sample from the population. We address this issue by finding a surrogate estimate for P and use it to calculate the MSE is a feasible way in this experiment. Recently, Kiar et al. [2016] proposed a better pipeline ndmg2 compared

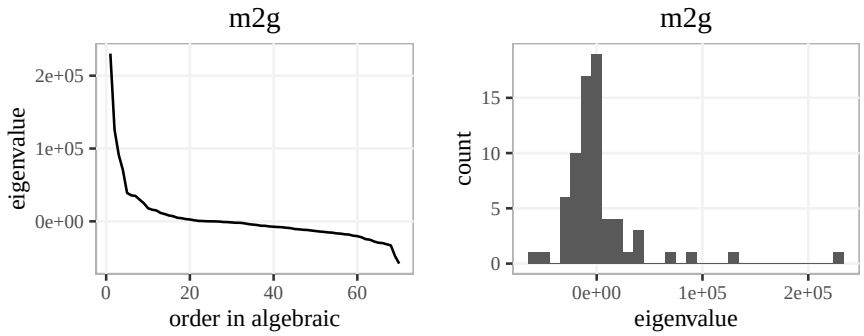

Figure 5: **Screeplot and the histogram of the eigenvalues of the mean of 114 graphs based on m2g pipeline.** The screeplot in the left panel shows the eigenvalues of the mean graph of all 114 graphs with diagonal augmentation in decreasing algebraic order for the Desikan atlas. The right panel shows the histogram of the eigenvalues of the mean graph of all 114 graphs with diagonal augmentation. Many eigenvalues are around zero, which lead to a quasi low-rank structure.

to m2g. Then the MLE derived from the 114 graphs in ndmg2 should be a relative more accurate estimate of the actual probability matrix P for the population. And we are going to use this as P when calculating the MSE. However, such P generally has full rank, which breaks the low-rank assumptions. So this setting makes it hard for $\widetilde{P}^{(1)}$ and $\widetilde{P}^{(q)}$ to improve and is favorable to the $\widehat{P}^{(1)}$ and $\widehat{P}^{(q)}$. Thus any improvement is conservative. Moreover, it is still possible that the 114 graphs from ndmg2 contain outliers. Thus by using the MLE as P, the performance of MLqE related estimators $\widehat{P}^{(q)}$ and $\widetilde{P}^{(q)}$ are underestimated.

In this experiment, we build the four estimates based on the samples with size m from the m2g dataset, while using the MLE of all 114 graphs from the ndmg2 dataset as the probability matrix P. Note that diagonal augmentation procedure introduced in Section 7.1.2 is also applied here to compensate for the unnecessary bias. We run 100 simulations on this dataset for different sample sizes $m = 2, 5, 10$. Specifically, in each Monte Carlo replicate, we sample m graphs out of the 114 from the m2g dataset and compute the four estimates based on the m sampled graphs. Once again for simplicity, we set q to be 0.9 without further exploiting. However, the results are consistent for many choices of q. We then compare these estimates to the MLE of all 114 graphs in the ndmg2 dataset. For those two low-rank estimators $\widetilde{P}^{(1)}$ and $\widetilde{P}^{(q)}$, we apply ASE for all possible dimensions, i.e. d ranges from 1 to n. The MSE results are shown in Figure 6.

When d is small, ASE procedure underestimates the dimension and fails to get important information, which leads to poor performance. In this work, we use Zhu and Ghodsi's method discussed in Section 4.2.2 to select the dimension d. We denote the selected dimensions by square and circle in the figure. We can see the algorithm does a pretty good job for selecting the dimension to embed. More importantly, there is a wide range of dimensions which could lead to a better performance when applying ASE. Although the P we are estimating is actually a high-rank matrix, ASE procedure still wins the bias-variance tradeoff

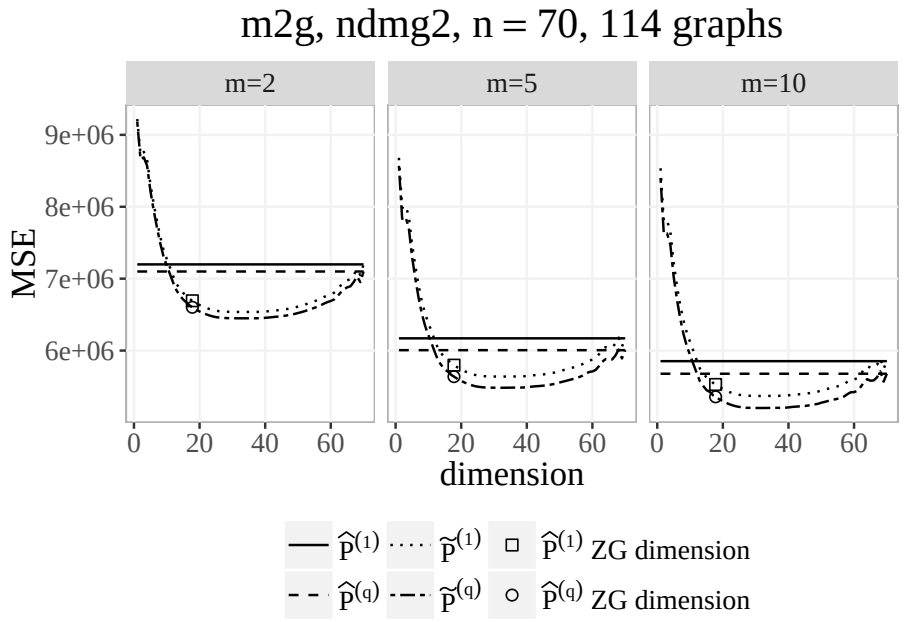

Figure 6: **Comparison of MSE of the four estimators for the Desikan atlases at three sample sizes.** The x-axis represents the dimensions to embed while y-axis is the MSE of each estimator. **1. MLE $\widehat{P}^{(1)}$ (horizontal solid line) vs MLqE $\widehat{P}^{(q)}$ (horizontal dotted line):** MLqE outperforms MLE since in practice observations are always contaminated and robust estimators are preferred; **2. MLE $\widehat{P}^{(1)}$ (horizontal solid line) vs ASE $\circ$ MLE $\widetilde{P}^{(1)}$ (dashed line):** $\widetilde{P}^{(1)}$ wins the bias-variance tradeoff when being embedded into a proper dimension; **3. MLqE $\widehat{P}^{(q)}$ (horizontal dotted line) vs ASE $\circ$ MLqE $\widetilde{P}^{(q)}$ (dashed dotted line):** $\widetilde{P}^{(q)}$ wins the bias-variance tradeoff when being embedded into a proper dimension; **4. ASE $\circ$ MLqE $\widetilde{P}^{(q)}$ (dashed dotted line) vs ASE $\circ$ MLE $\widetilde{P}^{(1)}$ (dashed line):** $\widetilde{P}^{(q)}$ is better, since it inherits the robustness from MLqE. The square and circle represent the dimensions selected by the Zhu and Ghodsi method. We can see it does a pretty good job. And more importantly, a wide range of dimensions could lead to an improvement.

and improves the performance while being suppressed in this setting.

Also, the robust estimator $\widehat{P}^{(q)}$ performs relatively better than $\widehat{P}^{(1)}$ in this experiment, even though P still contains outliers. This strongly indicates that there are many outliers in the original graphs from m2g pipeline. And $\widetilde{P}^{(q)}$ successfully inherits the robustness from MLqE and outperforms $\widetilde{P}^{(1)}$.

For all three sample sizes ($m = 2, 5, 10$), $\widetilde{P}^{(q)}$ estimates P most accurately while the target is preferable to the other three estimators more or less. So it should provide a even better estimate for the true but unknown P.

8 Discussion

In this work, analysis are mostly under the stochastic blockmodel setting. Noting that the results could be extended to the random dot product graph instead of SBM, i.e. our estimator does not necessarily have the block structure. The reason is that the SBM assumption is just to ensure rank($E[\widehat{P}^{(q)}]$) has an upper bound invariant of the number of vertices under the contamination model. With this assumption on the rank, all the theories still hold under the RDPG setting. In practice, graphs can hardly be exactly low rank. However, as shown in Figure 5 and Figure 6, our estimator still provides large improvement with a quasi low-rank structure. Thus our method can be applied to a much more general setting instead of being restricted to SBM.

In Section 5, we present theories based on the exponential distribution with MLqE for clarity. And Section 6 indicates that these results can be extended to other distributions and robust estimators. Note that the most important condition is Condition 1, which requires the MLE under the corresponding distribution is concentrated so that we can have all the matrix bounds we need. This generalization makes the theory more flexible and powerful.

Selecting a proper distortion parameter q in MLqE is complicated, and we use a fixed $q = 0.9$ throughout this work without presenting a formal way to do it. In order to have an improved performance, we might want to select a proper q based on a adaptive method proposed by Qin and Priebe [2013b].

In this work, we assume vertex correspondence is known across all graphs. However, in some applications, not all vertices are matched. In this case, our method can still apply after running the graph matching algorithms [Lyzinski et al., 2014, 2015, 2016].

The robust estimators $\widehat{P}^{(q)}$ and $\widetilde{P}^{(q)}$ outperform the non-robust ones $\widehat{P}^{(1)}$ and $\widetilde{P}^{(1)}$ mainly due to the reduced asymptotic bias in a contaminated model. However, robust estimators like MLqE should still provide improvement without contaminations based on finite samples. In this case, the embedding procedure will have a relatively larger impact, leading to a much better estimator $\widetilde{P}^{(q)}$ compared to $\widehat{P}^{(q)}$. More work is needed under the uncontaminated setting.

As we pointed out, improvement for estimation is not only important for the estimation itself, but also can help with other statistical inference procedures. Priebe et al. [2015] and Chen et al. [2016] both discussed vertex classification based on a single unweighted graph with contaminations. Moreover, to have a even more general setting, we might want to extend the current RDPG setting to latent positions graphs. Tang et al. [2013] showed the universally consistency of vertex classification method based on eigen-decomposition. More work is

needed for inference tasks other than estimation based on multiple weighted graphs.

Acknowledgments

This work is graciously supported by the XDATA program of the Defense Advanced Research Projects Agency (DARPA) administered through Air Force Research Laboratory contract FA8750-12-2-0303; the DARPA SIMPLEX program through SPAWAR contract N66001-15-C-4041; and DARPA GRAPHS contract N66001-14-1-4028.

References

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

P.J. Bickel and K.A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Number v. 1 in Holden-Day series in probability and statistics. Prentice Hall, 2001. ISBN 9780138503635. URL `https://books.google.co.uk/books?id=8poZAQAAIAAJ`.

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

Li Chen, Cencheng Shen, Joshua T Vogelstein, and Carey E Priebe. Robust vertex classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):578–590, 2016.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Davide Ferrari and Yuhong Yang. Maximum lq-likelihood estimation. *Ann. Statist.*, 38(2):753–783, 04 2010. doi: 10.1214/09-AOS687. URL `http://dx.doi.org/10.1214/09-AOS687`.

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.

Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, Edoardo M Airoldi, et al. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

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.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

G Kiar, W Gray Roncal, D Mhembere, E Bridgeford, R Burns, and J Vogelstein. ndmg: Neurodata's mri graphs pipeline, 2016.

Vince Lyzinski, Sancar Adali, Joshua T Vogelstein, Youngser Park, and Carey E Priebe. Seeded graph matching via joint optimization of fidelity and commensurability. *arXiv preprint arXiv:1401.3813*, 2014.

Vince Lyzinski, Daniel L Sussman, Donniell E Fishkind, Henry Pao, Li Chen, Joshua T Vogelstein, Youngser Park, and Carey E Priebe. Spectral clustering for divide-and-conquer graph matching. *Parallel Computing*, 47:70–87, 2015.

Vince Lyzinski, Donniell E Fishkind, Marcelo Fiori, Joshua T Vogelstein, Carey E Priebe, and Guillermo Sapiro. Graph matching: Relax at your own risk. *IEEE transactions on pattern analysis and machine intelligence*, 38(1): 60–73, 2016.

R. S. H. Mah and A. C. Tamhane. Detection of gross errors in process data. *AIChE Journal*, 28(5):828–830, 1982. ISSN 1547-5905. doi: 10.1002/aic.690280519. URL `http://dx.doi.org/10.1002/aic.690280519`.

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.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

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

Roberto Imbuzeiro Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600*, 2009.

Carey E Priebe, Daniel L Sussman, Minh Tang, and Joshua T Vogelstein. Statistical inference on errorfully observed graphs. *Journal of Computational and Graphical Statistics*, 24(4):930–953, 2015.

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

Yichen Qin and Carey E Priebe. Robust hypothesis testing via lq-likelihood. *arXiv preprint arXiv:1310.7278*, 2013b.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, pages 1878–1915, 2011.

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

Robert Serfling. Asymptotic relative efficiency in estimation. In *International encyclopedia of statistical science*, pages 68–72. Springer, 2011.

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.

Minh Tang, Daniel L Sussman, Carey E Priebe, et al. Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41 (3):1406–1430, 2013.

Runze Tang, Michael Ketcha, Joshua T Vogelstein, Carey E Priebe, and Daniel L Sussman. Law of large graphs. *arXiv preprint arXiv:1609.01672*, 2016.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

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

Stephen J Young and Edward R Scheinerman. Random dot product graph models for social networks. In *International Workshop on 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.

A Proofs for Theory Results

A.1 Outline of the Proofs

Firstly, in Section A.2, we prove in Lemma A.4 that when the contamination is large enough, the robust estimator $\widehat{P}^{(q)}$ has smaller asymptotic bias compared to $\widehat{P}^{(1)}$. By the results of minimum contrast estimator, we also show in Lemma A.8 that both estimators have variances go to zero as the number of graphs m goes to infinity.

In Section A.3, we mainly analyze the properties of the ASE procedure. We first prove Theorem A.9, which provides an upper bound for the 2-norm of the difference between the estimator $\widehat{P}^{(1)}$ and its expectation $H_{ij}^{(1)} = E[\widehat{P}_{ij}^{(1)}]$. Lemma A.11 shows that $U^\top \widehat{U}$ can be approximated by an orthogonal matrix $W^* = W_1 W_2^\top$, where U and $\widehat{U}$ are the eigenspaces with respect to the largest d eigenvalues of $H_{ij}^{(1)}$ and $\widehat{P}^{(1)}$ respectively. More conveniently, Lemma A.12 indicates that we can change the order of W^* in the matrix multiplications accordingly without affecting the result much. With these tool results, in Lemma A.13 we give an upper bound of $\|\widehat{Z} - ZW\|_F$, which controls the error of the $\widehat{Z}$ for estimating the true latent positions Z up to rotation. With the extent of Lemma A.13, we then give a bound of the $(2 \to \infty)$-norm of the $\widehat{Z} - ZW$, i.e. $\max_i \|\widehat{Z}_i - WZ_i\|_2$ in Theorem A.14.

In Section A.4, we give a bound of the estimation error $\left|\widehat{Z}_i^T\widehat{Z}_j - Z_i^T Z_j\right|$ in Lemma A.15 based on the results in Section A.3. In order to bound the variance of our estimator $\widetilde{P}^{(1)}$, all results in this section will be based on a truncated version of $\widetilde{P}^{(1)}$ defined in Definition A.16. This is just for technical reasons and will not affect the estimation procedure in practice, which is discussed in details in Remark A.17. Then we can bound the expectation (Lemma A.18) and variance (Theorem A.19) of the truncated $\widetilde{P}^{(1)}$ by carefully choosing a breakpoint a and analyzing separately. And as a direct result, we have the bound for the relative efficiency between $\widehat{P}_{ij}^{(1)}$ and $\widetilde{P}_{ij}^{(1)}$ in Theorem A.20.

In Section A.5, we compare the performance between $\widetilde{P}^{(q)}$ and $\widehat{P}^{(q)}$. The results in this section are proved in a similar way as in Section A.3 and Section A.4. However, since the MLqE in a mixture distribution model generally does not have a closed form, we explore the relationship between MLE and MLqE to give a relaxed bound which is used when MLqE is hard to analyze.

In Section A.6, we compare the performance between $\widetilde{P}^{(q)}$ and $\widetilde{P}^{(1)}$ by combining all the previous results.

In Section A.7, we provide proofs for all supplementary results mentioned in the manuscript.

Here we will first define the notation "with high probability", which is used through out the entire proofs.

Definition A.1 *We say a bound holds with high probability, if there exits a constant $n_0(c)$ such that if $n > n_0$, then for any η satisfying $n^{-c} < \eta < 1/2$, the bound holds with probability greater than $1 - \eta$.*

A.2 $\widehat{P}^{(q)}$ vs. $\widehat{P}^{(1)}$

Lemma A.2 *Consider the model $X_1, \cdots, X_m \overset{iid}{\sim} \text{Exp}(P)$ with $m \geq 2$ and $E[X_1] = P$. Given any data $x = (x_1, \cdots, x_m)$ such that $x_{(1)} > 0$ and not all x_i's are the same, then no matter how the data is sampled, we have*

- *There exists at least one solution to the MLq equation;*
- *All the solutions to the MLq equation are less than the MLE.*

Thus the MLqE $\widehat{P}^{(q)}$, the root closest to the MLE, is well defined.

Proof: The MLE is

$$\widehat{P}^{(1)}(x) = \bar{x}.$$

Consider the continuous function $g(\theta, x) = \sum_{i=1}^m e^{-\frac{(1-q)x_i}{\theta}}(x_i - \theta)$. Then the ML$q$ equation is $g(\theta, x) = 0$.

Let $x_{(1)} \leq \cdots \leq x_{(l)} \leq \bar{x} \leq x_{(l+1)} \leq \cdots \leq x_{(m)}$. Define $s_i = \bar{x} - x_{(i)}$ for $1 \leq i \leq l$, and $t_i = x_{(l+i)} - \bar{x}$ for $1 \leq i \leq m - l$. Note that $\sum_{i=1}^l s_i = \sum_{i=1}^{m-l} t_i$.

Then for any $\theta \geq \bar{x}$, we have

$$\begin{aligned}
g(\theta, x) &= \sum_{i=1}^{m} e^{-\frac{(1-q)x_{(i)}}{\theta}}(x_{(i)} - \theta) = \sum_{i=1}^{m} e^{-\frac{(1-q)x_{(i)}}{\theta}}(x_{(i)} - \bar{x} + \bar{x} - \theta) \\
&= -\sum_{i=1}^{l} e^{-\frac{(1-q)x_{(i)}}{\theta}} s_i + \sum_{i=1}^{m-l} e^{-\frac{(1-q)x_{(i+l)}}{\theta}} t_i + \sum_{i=1}^{m} e^{-\frac{(1-q)x_{(i)}}{\theta}}(\bar{x} - \theta) \\
&\leq -\sum_{i=1}^{l} e^{-\frac{(1-q)x_{(i)}}{\theta}} s_i + \sum_{i=1}^{m-l} e^{-\frac{(1-q)x_{(i+l)}}{\theta}} t_i \\
&\leq -e^{-\frac{(1-q)x_{(l+1)}}{\theta}} \sum_{i=1}^{l} s_i + \sum_{i=1}^{m-l} e^{-\frac{(1-q)x_{(i+l)}}{\theta}} t_i \\
&\leq -e^{-\frac{(1-q)x_{(l+1)}}{\theta}} \sum_{i=1}^{m-l} t_i + \sum_{i=1}^{m-l} e^{-\frac{(1-q)x_{(i+l)}}{\theta}} t_i \\
&\leq -\sum_{i=1}^{m-l} e^{-\frac{(1-q)x_{(i+l)}}{\theta}} t_i + \sum_{i=1}^{m-l} e^{-\frac{(1-q)x_{(i+l)}}{\theta}} t_i \\
&= 0,
\end{aligned}$$

and equality holds if and only if all x_i's are the same, which is excluded by the assumption. Thus $g(\theta, x) < 0$ for any $\theta \geq \bar{x}$.

Denote any solution to the MLq equation to be $\widehat{P}^{(q)}(x)$, then we also know:

- $g(\widehat{P}^{(q)}(x), x) = 0$;
- $\lim_{\theta \to 0^+} g(\theta, x) = 0$;
- $g(\theta, x) > 0$ when $\theta < x_{(1)}$;

Thus there exists at least one solution to the MLq equation. And all solutions to the MLq equation are between $x_{(1)}$ and $\bar{x}$, i.e. less than the MLE. ■

Lemma A.3 *Consider an exponential distribution model while the data is actually sampled from the contaminated model* $X, X_1, \cdots, X_m \overset{iid}{\sim} (1-\epsilon)\text{Exp}(P) + \epsilon\text{Exp}(C)$. *Denote such contaminated distribution as* F. *Then there exists exactly one real solution* $\theta(F)$ *of the population version of MLq equation, i.e.* $E_F[e^{-\frac{(1-q)X}{\theta(F)}}(X - \theta(F))] = 0$. *Moreover,* $\theta(F) < E_F[\bar{X}] = (1-\epsilon)P + \epsilon C$.

Proof: For the MLE, i.e. $\bar{X}$, we have $E[\bar{X}] = (1-\epsilon)P + \epsilon C$. According to Equation (3.2) in [Ferrari and Yang, 2010], $\theta(F)$ satisfies

$$\frac{\epsilon C}{(C(1-q)+\theta)^2} - \frac{\epsilon}{C(1-q)+\theta} + \frac{(1-\epsilon)P}{(P(1-q)+\theta)^2} - \frac{(1-\epsilon)}{P(1-q)+\theta} = 0,$$

i.e.

$$\frac{\epsilon(\theta - Cq)}{(C(1-q)+\theta)^2} = \frac{(1-\epsilon)(Pq-\theta)}{(P(1-q)+\theta)^2}.$$

Define $h(\theta) = (C(1-q)+\theta)^2(1-\epsilon)(Pq-\theta) - (P(1-q)+\theta)^2\epsilon(\theta - Cq)$. Then $\lim_{\theta\to\infty} h(\theta) = -\infty$, $h(0) > 0$, and $h(Cq) < 0$. Consider q as the variable and

solve the equation $h(E[\bar{X}]) = 0$, we have three roots and one of them is $q = 1$ obviously. The other two roots are

$$\frac{(P+C)\left((P-C)^2\epsilon(1-\epsilon)+2PC\right)}{2PC(P\epsilon+C(1-\epsilon))} \pm \sqrt{\frac{\epsilon(1-\epsilon)(C-P)^2\left(\epsilon(1-\epsilon)(C-P)^4-4P^2C^2\right)}{4P^2C^2(P\epsilon+C(1-\epsilon))^2}}.$$

To prove the roots are greater or equal to 1, we need to show

$$\frac{(P+C)\left((P-C)^2\epsilon(1-\epsilon)+2PC\right)}{2PC(P\epsilon+C(1-\epsilon))} - \sqrt{\frac{\epsilon(1-\epsilon)(C-P)^2\left(\epsilon(1-\epsilon)(C-P)^4-4P^2C^2\right)}{4P^2C^2(P\epsilon+C(1-\epsilon))^2}} > 1.$$

For the first part,

$$\frac{(P+C)\left((P-C)^2\epsilon(1-\epsilon)+2PC\right)}{2PC(P\epsilon+C(1-\epsilon))} > 1 + \frac{(P-C)^2\epsilon(1-\epsilon)(P+C)}{2PC(P\epsilon+C(1-\epsilon))}.$$

To prove the roots are greater or equal to 1, we just need to show

$$(P-C)^4\epsilon^2(1-\epsilon)^2(P+C)^2 \geq \epsilon^2(1-\epsilon)^2(C-P)^6.$$

Then it is sufficient to show that

$$(P+C)^2 \geq (C-P)^2,$$

which is true. Combined with the fact that when $q = 0$, $h(E[\bar{X}]) < 0$, we have for any $0 < q < 1$, $h(E[\bar{X}]) < 0$.

The equation $h(\theta) = 0$ is a cubic polynomial, so it has at most three real roots. In addition, by calculating we know there is only one real root, while the other two are complex roots. Combined with the fact that $h(Pq) > 0$, we have for any $0 < q < 1$, the only real root of the population version of MLq equation is less than $E[\bar{X}] = (1-\epsilon)P + \epsilon C$. ■

Lemma A.4 (Theorem 5.1) *For any $0 < q < 1$, there exists $C_0(P_{ij}, \epsilon, q) > 0$ such that under the contaminated model with $C > C_0(P_{ij}, \epsilon, q)$,*

$$\lim_{m\to\infty}\left|E[\widehat{P}_{ij}^{(q)}] - P_{ij}\right| < \lim_{m\to\infty}\left|E[\widehat{P}_{ij}^{(1)}] - P_{ij}\right|,$$

for $1 \leq i, j \leq n$ and $i \neq j$.

Proof: For the MLE $\widehat{P}_{ij}^{(1)} = \bar{A}_{ij}$,

$$E[\widehat{P}_{ij}^{(1)}] = E[\bar{A}_{ij}] = \frac{1}{m}\sum_{t=1}^{m} E[A_{ij}^{(t)}] = E[A_{ij}^{(1)}] = (1-\epsilon)P_{ij} + \epsilon C_{ij}.$$

As shown in Lemma A.3, $\theta(F)$ satisfies

$$\frac{\epsilon(\theta(F) - C_{ij}q)}{(C_{ij}(1-q)+\theta(F))^2} = \frac{(1-\epsilon)(P_{ij}q - \theta(F))}{(P_{ij}(1-q)+\theta(F))^2}.$$

Thus $\theta(F) - C_{ij}q$ and $\theta(F) - P_{ij}q$ should have different signs. Combined with $C_{ij} > P_{ij}$, we have

$$qP_{ij} < \theta(F).$$

To have a smaller asymptotic bias in absolute value, combined with Lemma A.7, we need

$$|\theta(F) - P_{ij}| < \epsilon(C_{ij} - P_{ij}).$$

Based on Lemma A.2, we need

$$qP_{ij} > P_{ij} - \epsilon(C_{ij} - P_{ij}),$$

i.e.

$$C_{ij} > P_{ij} + \frac{(1-q)P_{ij}}{\epsilon} = C_0(P_{ij}, \epsilon, q).$$

■

Lemma A.5 *The MLqE based on the model to be exponential distribution* $\text{Exp}(P)$ *while the data is actually sampled from the contaminated distribution* $(1-\epsilon)\text{Exp}(P) + \epsilon\text{Exp}(C)$ *is a minimum contrast estimator.*

Proof: Consider the contaminated distribution $F(x) = (1-\epsilon)f(x;P) + \epsilon f(x;C)$, where $f(x)$ represents the pdf of exponential distribution. By Lemma A.3, we know there is a one-to-one correspondence between the uncontaminated parameter P and the only real solution $\theta(F)$ of the population version of MLq equation, i.e. $E_F[e^{-\frac{(1-q)X}{\theta(F)}}(X - \theta(F))] = 0$. Let $r(\theta(F)) = P$. Then we can define $\rho(x;\theta) = \frac{f(x;r(\theta))^{1-q}}{1-q}$, where $q \in (0,1)$ is a constant. By reparameterizing $\rho(x;\theta)$ to $\widetilde{\rho}(x;r)$ such that $\widetilde{\rho}(x;r(\theta)) = \rho(x;\theta)$, we can use the proof of Lemma A.3 directly to prove that $D(\theta_0,\theta) = E_{\theta_0}[\rho(X,\theta)]$ is uniquely minimized at θ_0. Thus the MLqE is a minimum contrast estimator. ■

Lemma A.6 *Uniform convergence of the MLq equation, i.e.*

$$\sup_{\theta\in[0,R]}\left|\frac{1}{m}\sum_{i=1}^{m} e^{-\frac{(1-q)X_i}{\theta}}(X_i - \theta) - E_F[e^{-\frac{(1-q)X}{\theta}}(X-\theta)]\right| \overset{a.s.}{\to} 0.$$

Proof: Define $g(x,\theta) = e^{-\frac{(1-q)x}{\theta}}(x-\theta)$ and $d(x) = e^{-\frac{(1-q)x}{R}}(x+R)$. Then $E_F[d(X)] < \infty$ and $g(x,\theta) \le d(x)$ for all $\theta \in [0,R]$. Combined with the fact that $[0,R]$ is compact and the function $g(x,\theta)$ is continuous at each θ for all $x > 0$ and measurable function of x at each θ, we have the uniform convergence by Lemma 2.4 in [Newey and McFadden, 1994]. ■

Lemma A.7 $\widehat{P}_{ij}^{(q)} \overset{P}{\to} \theta(F_{ij})$ *as* $m \to \infty$*, where* F_{ij} *is the contaminated distribution* $(1-\epsilon)\text{Exp}(P_{ij}) + \epsilon\text{Exp}(C_{ij})$*, and* $\theta(F_{ij})$ *is defined in Lemma A.3.*

Proof: By the proof of Lemma A.3, we have

$$\inf\{D(\theta_0,\theta) : |\theta - \theta_0| \ge \epsilon\} > D(\theta_0,\theta_0)$$

for every $\epsilon > 0$. Combined with Lemma A.6, we know the MLq is consistent based on Theorem 5.2.3 in [Bickel and Doksum, 2001]. ■

Lemma A.8 (Theorem 5.1) *For* $1 \le i,j \le n$*,*

$$\text{Var}(\widehat{P}_{ij}^{(1)}) = \text{Var}(\widehat{P}_{ij}^{(q)}) = O(1/m).$$

And thus

$$\lim_{m\to\infty}\text{Var}(\widehat{P}_{ij}^{(1)}) = \lim_{m\to\infty}\text{Var}(\widehat{P}_{ij}^{(q)}) = 0.$$

Proof: Both MLE and MLqE are minimum constrast estimators. By consistency (shown in Lemma A.7) and other regularity conditions, we know the variances are both of order $1/m$ based on Theorem 5.4.2 in [Bickel and Doksum, 2001]. ■

A.3 ASE Procedure of $\widehat{P}^{(1)}$

Theorem A.9 *Let P and C be two n-by-n symmetric matrices satisfying element-wise conditions $0 < P_{ij} \le C_{ij} \le R$ for some constant $R > 0$. For $0 < \epsilon < 1$, we define m symmetric and hollow matrices as*

$$A^{(t)} \overset{iid}{\sim} (1-\epsilon)\mathrm{Exp}(P) + \epsilon \mathrm{Exp}(C),$$

for $1 \le t \le m$. Let $\widehat{P}^{(1)}$ be the element-wise MLE based on exponential distribution with m observations. Define $H^{(1)}_{ij} = E[\widehat{P}^{(1)}_{ij}] = (1-\epsilon)P_{ij} + \epsilon C_{ij}$, then for any constant $c > 0$, there exists another constant $n_0(c)$, independent of n, P, C and ϵ, such that if $n > n_0$, then for all η satisfying $n^{-c} \le \eta \le 1/2$,

$$P\left(\|\widehat{P}^{(1)} - H^{(1)}\|_2 \le 4R\sqrt{n\ln(n/\eta)/m}\right) \ge 1-\eta.$$

Remark: This is an extended version of Theorem 3.1 in [Oliveira, 2009].
Proof: Let $\{e_i\}_{i=1}^n$ be the canonical basis for $\mathbb{R}^n$. For each $1 \le i, j \le n$, define a corresponding matrix G_{ij}:

$$G_{ij} \equiv \begin{cases} e_i e_j^T + e_j e_i^T, & i \ne j; \\ e_i e_i^T, & i = j. \end{cases}$$

Thus

$$\widehat{P}^{(1)} = \sum_{1\le i<j\le n} \widehat{P}^{(1)}_{ij} G_{ij} = \frac{1}{m}\sum_{t=1}^{m} \sum_{1\le i<j\le n} A^{(t)}_{ij} G_{ij}$$

and

$$H^{(1)} = \sum_{1\le i<j\le n} H^{(1)}_{ij} G_{ij}.$$

Then we have $\widehat{P}^{(1)} - H^{(1)} = \frac{1}{m}\sum_{1\le t\le m, 1\le i<j\le n} X^{(t)}_{ij}$, where $X^{(t)}_{ij} = \left(A^{(t)}_{ij} - H^{(1)}_{ij}\right) G_{ij}$ for $1 \le t \le m$ and $1 \le i < j \le n$.

First bound the k-th moment of X_{ij} for $1 \le i < j \le n$ as following:

$$\begin{aligned} E[(A^{(t)}_{ij} - H^{(1)}_{ij})^k] \le & (1-\epsilon)\cdot \exp(-H_{ij}/P_{ij}) P^k_{ij} \Gamma(1+k, -H_{ij}/P_{ij}) \\ & + \epsilon \cdot \exp(-H_{ij}/C_{ij}) C^k_{ij} \Gamma(1+k, -H_{ij}/C_{ij}) \\ \le & \left((1-\epsilon)\cdot \exp(-H_{ij}/P_{ij}) P^k_{ij} + \epsilon \cdot \exp(-H_{ij}/C_{ij}) C^k_{ij}\right) k! \\ \le & \left((1-\epsilon)\cdot P^k_{ij} + \epsilon \cdot C^k_{ij}\right) k! \\ \le & R^k k!, \end{aligned} \tag{1}$$

Combined with

$$G^k_{ij} \equiv \begin{cases} e_i e_i^T + e_j e_j^T, & \text{k is even;} \\ e_i e_j^T + e_j e_i^T, & \text{k is odd,} \end{cases}$$

thus we have

1. When k is even,

$$E[(X_{ij}^{(t)})^k] = E[(A_{ij}^{(t)} - H_{ij}^{(1)})^k]G_{ij}^2 \preceq k!R^k G_{ij}^2;$$

2. When k is odd,

$$E[(X_{ij}^{(t)})^k] = E[(A_{ij}^{(t)} - H_{ij}^{(1)})^k]G_{ij} \preceq k!R^k G_{ij}^2.$$

So

$$E[(X_{ij}^{(t)})^k] \preceq k!R^k G_{ij}^2.$$

Let

$$\sigma^2 := \left\| \sum_{1\le t\le m, 1\le i<j\le n} (\sqrt{2}RG_{ij})^2 \right\|_2 = 2R^2 m\|(n-1)I\|_2 = 2R^2 m(n-1).$$

Notice that random matrices $X_{ij}^{(t)}$ are independent, self-adjoint and have mean zero, apply Theorem 6.2 in [Tropp, 2012] we have

$$\begin{aligned} P\left(\lambda_{\max}(\widehat{P}^{(1)} - H^{(1)}) \ge t\right) &= P\left(\lambda_{\max}\left(\frac{1}{m}\sum_{1\le t\le m, 1\le i<j\le n} X_{ij}^{(t)}\right) \ge t\right) \\ &= P\left(\lambda_{\max}\left(\sum_{1\le t\le m, 1\le i<j\le n} X_{ij}^{(t)}\right) \ge mt\right) \\ &\le n\exp\left(-\frac{(mt)^2/2}{\sigma^2 + Rmt}\right) \\ &\le n\exp\left(-\frac{mt^2/2}{2R^2 n + Rt}\right). \end{aligned}$$

Now consider $Y_{ij}^{(t)} \equiv \left(H_{ij}^{(1)} - A_{ij}^{(t)}\right)G_{ij}$, for $1 \le t \le m$ and $1 \le i < j \le n$. Then we have $H^{(1)} - \widehat{P}^{(1)} = \frac{1}{m}\sum_{1\le t\le m, 1\le i<j\le n} Y_{ij}^{(t)}$. Since

$$E[(H^{(1)} - \widehat{P}^{(1)})^k] = (-1)^k E[(\widehat{P}^{(1)} - H^{(1)})^k],$$

1. When k is even,

$$E[(Y_{ij}^{(t)})^k] = E[(\widehat{P}^{(1)} - H^{(1)})^k]G_{ij}^2 \preceq k!R^k G_{ij}^2;$$

2. When k is odd,

$$E[Y_{ij}^k] = -E[(\widehat{P}^{(1)} - H^{(1)})^k]G_{ij} \preceq k!R^k G_{ij}^2.$$

Thus by similar arguments,

$$\begin{aligned} P\left(\lambda_{\min}(\widehat{P}^{(1)} - H^{(1)}) \le -t\right) &= P\left(\lambda_{\max}(H^{(1)} - \widehat{P}^{(1)}) \ge t\right) \\ &\le n\exp\left(-\frac{mt^2/2}{2R^2 n + Rt}\right). \end{aligned}$$

Therefore we have

$$P\left(\|\widehat{P}^{(1)} - H^{(1)}\|_2 \geq t\right) \leq n \exp\left(-\frac{mt^2/2}{2R^2 n + Rt}\right).$$

Now let $c > 0$ be given and assume $n^{-c} \leq \eta \leq 1/2$. Then there exists a $n_0(c)$ independent of n, P, C and ϵ such that whenever $n > n_0(c)$,

$$t = 4R\sqrt{n \ln(n/\eta)/m} \leq 6Rn.$$

Plugging this t into the equation above, we get

$$P(\|\widehat{P}^{(1)} - H^{(1)}\|_2 \geq 4R\sqrt{n \ln(n/\eta)/m}) \leq n \exp\left(-\frac{t^2}{16R^2 n}\right) = \eta.$$

■

Define $H^{(1)} = E[\widehat{P}^{(1)}] = (1-\epsilon)P + \epsilon C$, where $P = XX^T$, $X \in \mathbb{R}^{n \times d}$, $C = YY^T$, $Y \in \mathbb{R}^{n \times d'}$. Let $d^{(1)} = \text{rank}(H^{(1)})$ be the dimension in which we are going to embed $\widehat{P}^{(1)}$. Then we can define $H^{(1)} = ZZ^T$ where $Z \in \mathbb{R}^{n \times d^{(1)}}$. Since $H^{(1)} = [\sqrt{1-\epsilon}X, \sqrt{\epsilon}Y][\sqrt{1-\epsilon}X, \sqrt{\epsilon}Y]^T$, we have $d^{(1)} \leq d + d'$.

For simplicity, from now on, we will use $\widehat{P}$ to represent $\widehat{P}^{(1)}$, use H to represent $H^{(1)}$ and use k to represent the dimension $d^{(1)}$ we are going to embed. Assume $H = USU^T = ZZ^T$, where $Z = [Z_1, \cdots, Z_n]^T$ is a n-by-k matrix. Then our estimate for Z up to rotation is $\widehat{Z} = \widehat{U}\widehat{S}^{1/2}$, where $\widehat{U}\widehat{S}\widehat{U}^T$ is the rank-k spectral decomposition of $|\widehat{P}| = (\widehat{P}^T\widehat{P})^{1/2}$.

Furthermore, we assume that the second moment matrix $E[Z_1 Z_1^T]$ is rank k and has distinct eigenvalues $\lambda_i(E[Z_1 Z_1^T])$. In particular, we assume that there exists $\delta > 0$ such that

$$\delta < \lambda_k(E[Z_1 Z_1^T])$$

Lemma A.10 *Under the above assumptions, $\lambda_i(H) = \Theta(n)$ with high probability when $i \leq k$, i.e. the largest k eigenvalues of H is of order n. Moreover, we have $\|S\|_2 = \Theta(n)$ and $\|\widehat{S}\|_2 = \Theta(n)$ with high probability.*

Remark: This is an extended version of Proposition 4.3 in [Sussman et al., 2014].

Proof: Note that $\lambda_i(H) = \lambda_i(ZZ^T) = \lambda_i(Z^T Z)$ when $i \leq k$. Since each entry of $Z^T Z$ is a sum of n independent random variables each in $[0, R]$, i.e. $(Z^T Z)_{ij} = \sum_{l=1}^n Z_{li} Z_{lj}$. By Hoeffding's inequality,

$$P(|(Z^T Z - nE[Z_1 Z_1^T])_{ij}| \geq t) \leq 2\exp(-\frac{2t^2}{nR^2}).$$

Now let $c > 0$ and assume $n^{-c} \leq \eta \leq 1/2$. Let

$$t = R\sqrt{n \ln(\sqrt{2/\eta})},$$

we have

$$P\left(|(Z^T Z - nE[Z_1 Z_1^T])_{ij}| \geq R\sqrt{n \ln(\sqrt{2/\eta})}\right) \leq \eta.$$

By the union bound, we have

$$P\left(\|Z^T Z - nE[Z_1 Z_1^T]\|_F \geq kR\sqrt{n \ln(\sqrt{2/\eta})}\right) \leq k^2 \eta.$$

Then by Weyl's Theorem [Horn and Johnson, 2012], we have

$$|\lambda_i(H) - n\lambda_i(E[Z_1 Z_1^T])| \leq \|Z^T Z - nE[Z_1 Z_1^T]\|_2 = O(\sqrt{n \log(1/\eta)})$$

with probability at least $1 - k^2\eta$. Thus $\lambda_i(H) = S_{ii} = \Theta(n)$ with probability at least $1 - \frac{2k^2}{n^2}$ when $i \leq k$. Moreover,

$$\|H\|_2 - \|H - \widehat{P}\|_2 \leq \|\widehat{S}\|_2 \leq \|\widehat{P} - H\|_2 + \|H\|_2.$$

Combined with Theorem A.9, with high probability we have $\|\widehat{S}\|_2 = \Theta(n)$. ■

Lemma A.11 *Let $W_1 \Sigma W_2^T$ be the singular value decomposition of $U^T \widehat{U}$. Then for sufficiently large n,*

$$\|U^T \widehat{U} - W_1 W_2^T\|_F = O(m^{-1} n^{-1} \log n)$$

with high probability.

Proof: Let $\sigma_1, \cdots, \sigma_k$ denote the singular values of $U^T \widehat{U}$. Then $\sigma_i = \cos(\theta_i)$ where the θ_i are the principal angles between the subspaces spanned by $\widehat{U}$ and U. Furthermore, by the Davis-Kahan $\sin(\Theta)$ theorem [Davis and Kahan, 1970], combined with Theorem A.9 and Lemma A.10,

$$\begin{aligned} &\|\widehat{U}\widehat{U}^T - UU^T\|_2 = \max_i |\sin(\theta_i)| \\ \leq &\frac{\|\widehat{P} - H\|_2}{\lambda_k(H)} \leq \frac{C\sqrt{n \log n/m}}{n} \\ = &O(m^{-1/2} n^{-1/2} \sqrt{\log n}) \end{aligned} \tag{2}$$

for sufficiently large n with high probability. Here $\lambda_k(H)$ denotes the k-th largest eigenvalue of H. Thus with high probability,

$$\begin{aligned} \|U^T \widehat{U} - W_1 W_2^T\|_F = \|\Sigma - I\|_F &= \sqrt{\sum_{i=1}^k (1 - \sigma_i)^2} \\ &\leq \sum_{i=1}^k (1 - \sigma_i) \leq \sum_{i=1}^k (1 - \sigma_i^2) \\ &= \sum_{i=1}^k \sin^2(\theta_i) \leq k \|\widehat{U}\widehat{U}^T - UU^T\|_2^2 \\ &= O(m^{-1} n^{-1} \log n). \end{aligned}$$

■

We will denote the orthogonal matrix $W_1 W_2^T$ by W^*.

Lemma A.12 *For sufficiently large n,*

$$\|W^* \widehat{S} - S W^*\|_F = O(m^{-1/2} \log n),$$

$$\|W^* \widehat{S}^{1/2} - S^{1/2} W^*\|_F = O(m^{-1/2} n^{-1/2} \log n)$$

and

$$\|W^* \widehat{S}^{-1/2} - S^{-1/2} W^*\|_F = O(m^{-1/2} n^{-3/2} \log n)$$

with high probability.

Proof: By Proposition 2.1 in [Rohe et al., 2011] and Equation (2), we have for some orthogonal matrix W,

$$\begin{aligned}&\|\widehat{U} - UW\|_F^2 \leq \frac{2\|\widehat{U}\widehat{U}^T - UU^T\|_F^2}{\delta^2} \leq \frac{8k^2\|\widehat{U}\widehat{U}^T - UU^T\|_2^2}{\delta^2}\\ =&O(m^{-1}n^{-1}\log n),\end{aligned}$$

with high probability. Let $Q = \widehat{U} - UU^T\widehat{U}$. And Q is the residual after projecting $\widehat{U}$ orthogonally onto the column space of U, we have

$$\|Q\|_F = \|\widehat{U} - UU^T\widehat{U}\|_F \leq \|\widehat{U} - UT\|_F = O(m^{-1/2}n^{-1/2}\sqrt{\log n}). \quad (3)$$

for all $k \times k$ matrices T with high probability. Then

$$\begin{aligned}W^*\widehat{S} =&(W^* - U^T\widehat{U})\widehat{S} + U^T\widehat{U}\widehat{S} = (W^* - U^T\widehat{U})\widehat{S} + U^T\widehat{P}\widehat{U}\\ =&(W^* - U^T\widehat{U})\widehat{S} + U^T(\widehat{P} - H)\widehat{U} + U^TH\widehat{U}\\ =&(W^* - U^T\widehat{U})\widehat{S} + U^T(\widehat{P} - H)Q + U^T(\widehat{P} - H)UU^T\widehat{U} + U^TH\widehat{U}\\ =&(W^* - U^T\widehat{U})\widehat{S} + U^T(\widehat{P} - H)Q + U^T(\widehat{P} - H)UU^T\widehat{U} + SU^T\widehat{U}.\end{aligned}$$

Combined with Theorem A.9, Lemma A.10, Lemma A.11, we have

$$\begin{aligned}&\|W^*\widehat{S} - SW^*\|_F\\ =&\|(W^* - U^T\widehat{U})\widehat{S} + U^T(\widehat{P} - H)Q + U^T(\widehat{P} - H)UU^T\widehat{U} + S(U^T\widehat{U} - W^*)\|_F\\ \leq&\|W^* - U^T\widehat{U}\|_F(\|\widehat{S}\|_2 + \|S\|_2) + \|U^T\|_F\|\widehat{P} - H\|_2\|Q\|_F + \|U^T(\widehat{P} - H)U\|_F\\ \leq&O(m^{-1}\log n) + O(m^{-1/2}\log n) + \|U^T(\widehat{P} - H)U\|_F\end{aligned}$$

with high probability. And we know $U^T(\widehat{P} - H)U$ is a $k \times k$ matrix with ij-th entry to be

$$u_i^T(\widehat{P} - H)u_j = \sum_{s=1}^{n}\sum_{t=1}^{n}(\widehat{P}_{st} - H_{st})u_{is}u_{jt} = 2\sum_{s<t}(\widehat{P}_{st} - H_{st})u_{is}u_{jt}$$

where u_i and u_j are the i-th and j-th columns of U. Thus, conditioned on H, U is fixed and $u_i^T(\widehat{P} - H)u_j$ is a sum of independent mean 0 random variables.

By Equation (1), we have

$$\begin{aligned}&E\left[\left((A_{st}^{(t')} - H_{st})u_{is}u_{jt}\right)^k\right]\\ \leq&k!R^ku_{is}^ku_{jt}^k\\ \leq&\frac{k!}{2}R^{k-2}(\sqrt{2}u_{is}u_{jt}R)^2.\end{aligned}$$

Also we have

$$\sigma^2 := \left|\sum_{t',s<t} 2R^2u_{is}^2u_{jt}^2\right| \leq mR^2,$$

then by Theorem 6.2 in [Tropp, 2012], we have

$$P\left(\left|2\sum_{s<t}(\widehat{P}_{st} - H_{st})u_{is}u_{jt}\right| \geq t\right) \leq \exp\left(\frac{-mt^2/8}{R^2 + Rt/2}\right).$$

Let $t = cRm^{-1/2}\log n$ for any $c > 0$, we have

$$P\left(\left|2\sum_{s<t}(\widehat{P}_{st} - H_{st})u_{is}u_{jt}\right| \geq Cm^{-1/2}\log n\right) \leq n^{-c}.$$

Thus each entry of $U^T(\widehat{P}-H)U$ is of order $O(m^{-1/2}\log n)$ with high probability and

$$\|U^T(\widehat{P} - H)U\|_F = O(m^{-1/2}\log n) \tag{4}$$

with high probability. Hence

$$\|W^*\widehat{S} - SW^*\|_F = O(m^{-1/2}\log n)$$

with high probability. Also, since

$$W^*_{ij}(\lambda_j^{1/2}(\widehat{P}) - \lambda_i^{1/2}(H)) = W^*_{ij}\frac{\lambda_j(\widehat{P}) - \lambda_i(H)}{\lambda_j^{1/2}(\widehat{P}) + \lambda_i^{1/2}(H)}$$

and the eigenvalues $\lambda_j^{1/2}(\widehat{P})$ and $\lambda_i^{1/2}(H)$ are both of order $\Theta(\sqrt{n})$, we have

$$\|W^*\widehat{S}^{1/2} - S^{1/2}W^*\|_F = O(m^{-1/2}n^{-1/2}\log n)$$

with high probability. Similarly, since

$$W^*_{ij}(\lambda_j^{-1/2}(\widehat{P}) - \lambda_i^{-1/2}(H)) = W^*_{ij}\frac{\lambda_i(H) - \lambda_j(\widehat{P})}{(\lambda_j^{-1/2}(\widehat{P}) + \lambda_i^{-1/2}(H))\lambda_j(\widehat{P})\lambda_i(H)}$$

and the eigenvalues $\lambda_j(\widehat{P})$ and $\lambda_i(H)$ are both of order $\Theta(n)$, with high probability we have

$$\|W^*\widehat{S}^{-1/2} - S^{-1/2}W^*\|_F = O(m^{-1/2}n^{-3/2}\log n).$$

■

Lemma A.13 *There exists a rotation matrix W such that for sufficiently large n,*

$$\|\widehat{Z} - ZW\|_F = \|(\widehat{P} - H)US^{-1/2}\|_F + O(m^{-1/2}n^{-1/2}(\log n)^{3/2})$$

with high probability.

Proof: Let $Q_1 = UU^T\widehat{U} - UW^*$, $Q_2 = W^*\widehat{S}^{1/2} - S^{1/2}W^*$ and $Q_3 = \widehat{U} - UW^* = \widehat{U} - UU^T\widehat{U} + Q_1 = Q + Q_1$. Then since $UU^TH = H$ and $\widehat{U}\widehat{S}^{1/2} = \widehat{P}\widehat{U}\widehat{S}^{-1/2}$,

$$\begin{aligned}
\widehat{Z} - US^{1/2}W^* =& \widehat{U}\widehat{S}^{1/2} - UW^*\widehat{S}^{1/2} + U(W^*\widehat{S}^{1/2} - S^{1/2}W^*) \\
=& (\widehat{U} - UU^T\widehat{U})\widehat{S}^{1/2} + Q_1\widehat{S}^{1/2} + UQ_2 \\
=& (\widehat{P} - H)\widehat{U}\widehat{S}^{-1/2} - UU^T(\widehat{P} - H)\widehat{U}\widehat{S}^{-1/2} + Q_1\widehat{S}^{1/2} + UQ_2 \\
=& (\widehat{P} - H)UW^*\widehat{S}^{-1/2} - UU^T(\widehat{P} - H)UW^*\widehat{S}^{-1/2} \\
&+ (I - UU^T)(\widehat{P} - H)Q_3\widehat{S}^{-1/2} + Q_1\widehat{S}^{1/2} + UQ_2.
\end{aligned}$$

By Lemma A.11, with high probability,

$$\|Q_1\|_F \le \|U\|_F \|U^T\widehat{U} - W^*\|_F = O(m^{-1}n^{-1}\log n).$$

By Lemma A.12, with high probability,

$$\|Q_2\|_F = O(m^{-1/2}n^{-1/2}\log n).$$

By Equation (3), with high probability,

$$\|Q_3\|_F \le \|Q\|_F + \|Q_1\|_F = O(m^{-1/2}n^{-1/2}(\log n)^{1/2}).$$

By Equation (4), with high probability,

$$\|UU^T(\widehat{P}-H)UW^*\widehat{S}^{-1/2}\|_F \le \|U^T(\widehat{P}-H)U\|_F \|\widehat{S}^{-1/2}\|_2 = O(m^{-1}n^{-1/2}\log n).$$

By Lemma A.12, with high probability,

$$\|W^*\widehat{S}^{-1/2} - S^{-1/2}W^*\|_F = O(m^{-1/2}n^{-3/2}\log n).$$

Therefore, with high probability,

$$\begin{aligned}
&\|\widehat{Z} - US^{1/2}W^*\|_F \\
=&\|(\widehat{P}-H)UW^*\widehat{S}^{-1/2}\|_F + O(m^{-1}n^{-1/2}\log n) + \|I - UU^T\|_2\|\widehat{P}-H\|_2 O(m^{-1/2}n^{-1}(\log n)^{1/2}) \\
&+ O(m^{-1}n^{-1/2}\log n) + O(m^{-1/2}n^{-1/2}\log n) \\
=&\|(\widehat{P}-H)UW^*\widehat{S}^{-1/2}\|_F + O(m^{-1/2}n^{-1/2}\log n) \\
\le&\|(\widehat{P}-H)US^{-1/2}W^*\|_F + \|(\widehat{P}-H)U(W^*\widehat{S}^{-1/2} - S^{-1/2}W^*)\|_F + O(m^{-1/2}n^{-1/2}\log n) \\
=&\|(\widehat{P}-H)US^{-1/2}\|_F + O(m^{-1}n^{-1}(\log n)^{3/2}) + O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) \\
=&\|(\widehat{P}-H)US^{-1/2}\|_F + O(m^{-1/2}n^{-1/2}(\log n)^{3/2}).
\end{aligned}$$

Note that $Z = US^{1/2}W$ for some orthogonal matrix W. As W^* is also orthogonal, therefore $Z\widetilde{W} = US^{1/2}W^*$ for some orthogonal $\widetilde{W}$, which completes the proof. ■

Theorem A.14 *There exists a rotation matrix W such that for sufficiently large n,*

$$\max_i \|\widehat{Z}_i - WZ_i\|_2 = O(m^{-1/2}n^{-1/2}(\log n)^{3/2})$$

with high probability.

Proof: By Lemma A.13, we have

$$\|\widehat{Z} - ZW\|_F = \|(\widehat{P}-H)US^{-1/2}\|_F + O(m^{-1/2}n^{-1/2}(\log n)^{3/2})$$

with high probability and similarly we could have the bound for each column vector with high probability that

$$\begin{aligned}
\max_i \|\widehat{Z}_i - WZ_i\|_2 \le& \frac{1}{\lambda_k^{1/2}(H)} \max_i \|((\widehat{P}-H)U)_i\|_2 + O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) \\
\le& \frac{k^{1/2}}{\lambda_k^{1/2}(H)} \max_j \|(\widehat{P}-H)u_j\|_\infty + O(m^{-1/2}n^{-1/2}(\log n)^{3/2})
\end{aligned}$$

where $((\widehat{P}-H)U)_i$ represents the i-th row of $(\widehat{P}-H)U$ and u_j denotes the j-th column of U. Now given i and j, the i-th element of the vector $(\widehat{P}-H)u_j$ is of the form

$$\sum_{s=1}^{n}(\widehat{P}_{is}-H_{is})u_{js}=\sum_{s\neq i}(\widehat{P}_{is}-H_{is})u_{js}.$$

Thus, conditioned on H, the i-th element of the vector $(\widehat{P}-H)u_j$ is a sum of independent mean 0 random variables. By Equation (1), we have

$$\begin{aligned}&E\left[\left((A_{is}^{(t)}-H_{is})u_{js}\right)^k\right]\\ \leq& k!R^k u_{js}^k\\ \leq& \frac{k!}{2}R^{k-2}(\sqrt{2}Ru_{js})^2.\end{aligned}$$

Also we have

$$\sigma^2:=|\sum_{t,s\neq i}2R^2u_{js}^2|\leq 2R^2m,$$

then by Theorem 6.2 in [Tropp, 2012], we have

$$P\left(\left|\sum_{s\neq i}(\widehat{P}_{is}-H_{is})u_{js}\right|\geq t\right)\leq\exp\left(\frac{-mt^2/2}{2R^2+Rt}\right).$$

Let $t=3cRm^{-1/2}\log n$, we have

$$P\left(\left|\sum_{s\neq i}(\widehat{P}_{is}-H_{is})u_{js}\right|\geq 3cRm^{-1/2}\log n\right)\leq n^{-c},$$

i.e. it is of order $O(m^{-1/2}\log n)$ with high probability. Taking the union bound over all i and j, with high probability we have,

$$\begin{aligned}\max_i\|\widehat{Z}_i-WZ_i\|_2&\leq\frac{Ck^{1/2}}{\lambda_k^{1/2}(H)}m^{-1/2}(\log n)^{3/2}+O(m^{-1/2}n^{-1/2}(\log n)^{3/2})\\ &=O(m^{-1/2}n^{-1/2}(\log n)^{3/2}).\end{aligned}$$

■

A.4 $\widetilde{P}^{(1)}$ vs. $\widehat{P}^{(1)}$

Lemma A.15 $\left|\widehat{Z}_i^T\widehat{Z}_j-Z_i^TZ_j\right|=O(m^{-1/2}n^{-1/2}(\log n)^{3/2})$ *with high probability.*

Proof: Let W be the rotation matrix in Theorem A.14, then

$$\begin{aligned}\left|\widehat{Z}_i^T\widehat{Z}_j-Z_i^TZ_j\right|&=\left|\widehat{Z}_i^T\widehat{Z}_j-\widehat{Z}_i^TWZ_j+\widehat{Z}_i^TWZ_j-(WZ_i)^TWZ_j\right|\\ &\leq\left|\widehat{Z}_i^T(\widehat{Z}_j-WZ_j)+(\widehat{Z}_i^T-(WZ_i)^T)WZ_j\right|\\ &\leq\|\widehat{Z}_i\|_2\|\widehat{Z}_j-WZ_j\|_2+\|Z_j\|_2\|\widehat{Z}_i^T-(WZ_i)^T\|_2.\end{aligned}$$

Since $\|Z_i\|_2^2 = Z_i^T Z_i = H_{ii}^{(1)} = E[\widehat{P}_{ii}^{(1)}] = (1-\epsilon)P_{ij} + \epsilon C_{ij} \leq R$, we have $\|Z_i\|_2 = O(1)$. Combined with Theorem A.14,

$$\begin{aligned}\left|\widehat{Z}_i^T\widehat{Z}_j - Z_i^T Z_j\right| =& (\|\widehat{Z}_i\|_2 + \|Z_j\|_2)O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) \\ \leq& (\|\widehat{Z}_i - WZ_i\|_2 + \|WZ_i\|_2 + \|Z_j\|_2)O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) \\ =& O(m^{-1/2}n^{-1/2}(\log n)^{3/2})\end{aligned}$$

with high probability. ■

Definition A.16 *Define $\widetilde{P}_{ij}^{(1)} = (\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}}$, our estimator for P_{ij}, to be a projection of $\widehat{Z}_i^T\widehat{Z}_j$ onto $[0, \min(\widehat{P}_{ij}^{(1)}, R)]$.*

Remark A.17 *The truncation step above to construct estimator is only for technical reasons. Since the constant R could be arbitrarily large, we do not need this truncation step in practice. Note that Theorem 5.3 still holds with this modified estimator. And all our simulation and real data experiment do not contain this truncation procedure.*

Lemma A.18 (Theorem 5.3 Part 1) *Assuming that $m = O(n^b)$ for any $b > 0$, then the estimator based on ASE of MLE has the same entry-wise asymptotic bias as MLE, i.e.*

$$\lim_{n\to\infty}\text{Bias}(\widetilde{P}_{ij}^{(1)}) = \lim_{n\to\infty}E[\widetilde{P}_{ij}^{(1)}] - P_{ij} = \lim_{n\to\infty}E[\widehat{P}_{ij}^{(1)}] - P_{ij} = \lim_{n\to\infty}\text{Bias}(\widehat{P}_{ij}^{(1)}).$$

Proof: Fix some $a > 0$, we have

$$\begin{aligned}&E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j|] \\ =& E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij} \leq a\}] + E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij} > a\}]\end{aligned}$$

For the first term, we have

$$\begin{aligned}&E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij} \leq a\}] \\ \leq& E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij} \leq a\}\mathbb{I}\{\text{Lemma } A.15 \text{ holds}\}] \\ &+ E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij} \leq a\}\mathbb{I}\{\text{Lemma } A.15 \text{ does not hold}\}] \\ \leq& E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij} \leq a\}|\text{Lemma } A.15 \text{ holds}] \\ &+ n^{-c}E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij} \leq a\}|\text{Lemma } A.15 \text{ does not hold}] \\ \leq& O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) \\ &+ n^{-c}E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - \widehat{P}_{ij}|\mathbb{I}\{\widehat{P}_{ij} \leq a\}|\text{Lemma } A.15 \text{ does not hold}] \\ &+ n^{-c}E[|\widehat{P}_{ij} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij} \leq a\}|\text{Lemma } A.15 \text{ does not hold}] \\ \leq& O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) + n^{-c}E[\widehat{P}_{ij}\mathbb{I}\{\widehat{P}_{ij} \leq a\}|\text{Lemma } A.15 \text{ does not hold}] \\ &+ n^{-c}E[(\widehat{P}_{ij} + R)\mathbb{I}\{\widehat{P}_{ij} \leq a\}|\text{Lemma } A.15 \text{ does not hold}] \\ \leq& O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) + an^{-c} + (a+R)n^{-c} \\ \leq& O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) + 2n^{-c}(a+R).\end{aligned}$$

Notice that

$$
\begin{aligned}
&E[\widehat{P}_{ij}\mathbb{I}\{\widehat{P}_{ij}>a\}] = E[\left(\frac{1}{m}\sum_{1\le t\le m} A_{ij}^{(t)}\right)\mathbb{I}\{\widehat{P}_{ij}>a\}] \\
=&\frac{1}{m}E[\sum_{1\le t\le m} A_{ij}^{(t)}\mathbb{I}\{\widehat{P}_{ij}>a\}] \le \frac{1}{m}E[\sum_{1\le t\le m} A_{ij}^{(t)}\mathbb{I}\{\max_{1\le s\le m} A_{ij}^{(s)}>a\}] \\
\le&\frac{1}{m}E[\sum_{1\le t\le m} A_{ij}^{(t)}\left(\sum_{1\le s\le m}\mathbb{I}\{A_{ij}^{(s)}>a\}\right)] = E[A_{ij}^{(1)}\left(\sum_{1\le s\le m}\mathbb{I}\{A_{ij}^{(s)}>a\}\right)] \\
=&E[A_{ij}^{(1)}\mathbb{I}\{A_{ij}^{(1)}>a\})] + (m-1)E[A_{ij}^{(1)}\mathbb{I}\{A_{ij}^{(2)}>a\})] \\
=&E[A_{ij}^{(1)}\mathbb{I}\{A_{ij}^{(1)}>a\})] + (m-1)E[A_{ij}^{(1)}]P(A_{ij}^{(1)}>a),
\end{aligned}
$$

and similarly

$$
\begin{aligned}
&E[(\widehat{P}_{ij}+R)\mathbb{I}\{\widehat{P}_{ij}>a\}] \\
=&E[\widehat{P}_{ij}\mathbb{I}\{\widehat{P}_{ij}>a\}] + R\cdot P(\widehat{P}_{ij}>a) \\
\le&E[A_{ij}^{(1)}\mathbb{I}\{A_{ij}^{(1)}>a\})] + (m-1)E[A_{ij}^{(1)}]P(A_{ij}^{(1)}>a) + R\cdot m\cdot P(A_{ij}^{(1)}>a).
\end{aligned}
$$

Thus for the second term,

$$
\begin{aligned}
&E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij}>a\}] \\
\le&E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - \widehat{P}_{ij}|\mathbb{I}\{\widehat{P}_{ij}>a\}] + E[|\widehat{P}_{ij} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij}>a\}] \\
\le&E[\widehat{P}_{ij}\mathbb{I}\{\widehat{P}_{ij}>a\}] + E[(\widehat{P}_{ij}+R)\mathbb{I}\{\widehat{P}_{ij}>a\}] \\
\le&2E[A_{ij}^{(1)}\mathbb{I}\{A_{ij}^{(1)}>a\})] + 2(m-1)E[A_{ij}^{(1)}]P(A_{ij}^{(1)}>a) \\
&+ R\cdot m\cdot P(A_{ij}^{(1)}>a) \\
\le&2e^{-a/R}(a+2mR).
\end{aligned}
$$

Thus

$$
\begin{aligned}
&E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j|] \\
\le&O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) + 2n^{-c}(a+R) + 2e^{-a/R}(a+2mR).
\end{aligned}
$$

Let $a = m^{-1}n^{2b}$ for any $b>0$, and $c = 2b+3$, combined with the assumption $m = O(n^b)$, we have

$$
\begin{aligned}
&E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j|] \\
=&O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) + O(m^{-1}n^{-3}) + O(m^{-1}n^{2b})\cdot O(e^{-m^{-1}n^{2b}}) \\
=&O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) + O(m^{-1}n^{-3}) + O(m^{-1}n^{2b})\cdot O(e^{-n^b}) \\
=&O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) + O(m^{-1}n^{-3}) + O(m^{-1}n^{2b})\cdot O(n^{-2b-3}) \\
=&O(m^{-1/2}n^{-1/2}(\log n)^{3/2}) + O(m^{-1}n^{-3}) \\
=&O(m^{-1/2}n^{-1/2}(\log n)^{3/2}).
\end{aligned}
$$

■

Theorem A.19 *Assuming that $m = O(n^b)$ for any $b > 0$, then* $\mathrm{Var}((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}}) = O(m^{-1}n^{-1}(\log n)^3)$.

Proof: By Lemma A.15,

$$\begin{aligned}
\mathrm{Var}((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}}) =& E[((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - E[(\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}}])^2] \\
=& E[((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j + Z_i^T Z_j - E[(\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}}])^2] \\
=& E[((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2] + E[(Z_i^T Z_j - E[(\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}}])^2] \\
&+ 2E[((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)(Z_i^T Z_j - E[(\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}}])] \\
\leq& E[((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2] + E[(Z_i^T Z_j - E[(\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}}])^2] \\
&+ 2\sqrt{E[((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2] E[(Z_i^T Z_j - E[(\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}}])^2]} \\
\leq& 4E[((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2].
\end{aligned}$$

Fix some $a > 0$, we have

$$\begin{aligned}
& E[((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2] \\
=& E[((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij} \leq a\}] + E[((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij} > a\}].
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
& E[((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij} \leq a\}] \\
\leq& E[((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij} \leq a\} \mathbb{I}\{\text{Lemma } A.15 \text{ holds}\}] \\
&+ E[((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij} \leq a\} \mathbb{I}\{\text{Lemma } A.15 \text{ does not hold}\}] \\
\leq& E[((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij} \leq a\} | \text{Lemma } A.15 \text{ holds}] \\
&+ n^{-c} E[((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij} \leq a\} | \text{Lemma } A.15 \text{ does not hold}] \\
\leq& O(m^{-1}n^{-1}(\log n)^3) \\
&+ 2n^{-c} E[((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}} - \widehat{P}_{ij})^2 \mathbb{I}\{\widehat{P}_{ij} \leq a\} | \text{Lemma } A.15 \text{ does not hold}] \\
&+ 2n^{-c} E[(\widehat{P}_{ij} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij} \leq a\} | \text{Lemma } A.15 \text{ does not hold}] \\
\leq& O(m^{-1}n^{-1}(\log n)^3) + 2n^{-c} E[\widehat{P}_{ij}^2 \mathbb{I}\{\widehat{P}_{ij} \leq a\} | \text{Lemma } A.15 \text{ does not hold}] \\
&+ 2n^{-c} E[(\widehat{P}_{ij} + R)^2 \mathbb{I}\{\widehat{P}_{ij} \leq a\} | \text{Lemma } A.15 \text{ does not hold}] \\
\leq& O(m^{-1}n^{-1}(\log n)^3) + 2a^2 n^{-c} + 2(a+R)^2 n^{-c} \\
\leq& O(m^{-1}n^{-1}(\log n)^3) + 4n^{-c}(a+R)^2.
\end{aligned}$$

Notice that

$$\begin{aligned}
& E[\widehat{P}_{ij}^2 \mathbb{I}\{\widehat{P}_{ij} > a\}] = E[(\frac{1}{m} \sum_{1 \leq t \leq m} A_{ij}^{(t)})^2 \mathbb{I}\{\widehat{P}_{ij} > a\}] \\
\leq& \frac{1}{m} E[\sum_{1 \leq t \leq m} A_{ij}^{(t)2} \mathbb{I}\{\widehat{P}_{ij} > a\}] \leq \frac{1}{m} E[\sum_{1 \leq t \leq m} A_{ij}^{(t)2} \mathbb{I}\{\max_{1 \leq s \leq m} A_{ij}^{(s)} > a\}] \\
\leq& \frac{1}{m} E[\sum_{1 \leq t \leq m} A_{ij}^{(t)2} (\sum_{1 \leq s \leq m} \mathbb{I}\{A_{ij}^{(s)} > a\})] = E[A_{ij}^{(1)2} (\sum_{1 \leq s \leq m} \mathbb{I}\{A_{ij}^{(s)} > a\})] \\
=& E[A_{ij}^{(1)2} \mathbb{I}\{A_{ij}^{(1)} > a\})] + (m-1) E[A_{ij}^{(1)2} \mathbb{I}\{A_{ij}^{(2)} > a\})] \\
=& E[A_{ij}^{(1)2} \mathbb{I}\{A_{ij}^{(1)} > a\})] + (m-1) E[A_{ij}^{(1)2}] P(A_{ij}^{(1)} > a),
\end{aligned}$$

and similarly

$$\begin{aligned}
&E[(\widehat{P}_{ij}+R)^2\mathbb{I}\{\widehat{P}_{ij}>a\}]\\
=&E[\widehat{P}_{ij}^2\mathbb{I}\{\widehat{P}_{ij}>a\}]+2R\cdot E[\widehat{P}_{ij}\mathbb{I}\{\widehat{P}_{ij}>a\}]+R^2P(\widehat{P}_{ij}>a)\\
\leq&E[A_{ij}^{(1)2}\mathbb{I}\{A_{ij}^{(1)}>a\})]+(m-1)E[A_{ij}^{(1)2}]P(A_{ij}^{(1)}>a)\\
&+2R\left(E[A_{ij}^{(1)}\mathbb{I}\{A_{ij}^{(1)}>a\})]+(m-1)E[A_{ij}^{(1)}]P(A_{ij}^{(1)}>a)\right)\\
&+R^2\cdot m\cdot P(A_{ij}^{(1)}>a).
\end{aligned}$$

Thus for the second term,

$$\begin{aligned}
&E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}-Z_i^TZ_j)^2\mathbb{I}\{\widehat{P}_{ij}>a\}]\\
\leq&2E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}-\widehat{P}_{ij})^2\mathbb{I}\{\widehat{P}_{ij}>a\}]+2E[(\widehat{P}_{ij}-Z_i^TZ_j)^2\mathbb{I}\{\widehat{P}_{ij}>a\}]\\
\leq&2E[\widehat{P}_{ij}^2\mathbb{I}\{\widehat{P}_{ij}>a\}]+2E[(\widehat{P}_{ij}+R)^2\mathbb{I}\{\widehat{P}_{ij}>a\}]\\
\leq&4E[A_{ij}^{(1)2}\mathbb{I}\{A_{ij}^{(1)}>a\})]+4(m-1)E[A_{ij}^{(1)2}]P(A_{ij}^{(1)}>a)\\
&+4R\cdot E[A_{ij}^{(1)}\mathbb{I}\{A_{ij}^{(1)}>a\})]+2R(m-1)E[A_{ij}^{(1)}]P(A_{ij}^{(1)}>a)\\
&+2R^2\cdot m\cdot P(A_{ij}^{(1)}>a)\\
\leq&4e^{-a/R}\left(a^2+3Ra+3(m+1)R^2\right)\\
\leq&4e^{-a/R}(a+2m^{1/2}R)^2.
\end{aligned}$$

Thus,

$$\mathrm{Var}((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}})\leq O(m^{-1}n^{-1}(\log n)^3)+16(a+R)^2n^{-c}+16(a+2m^{1/2}R)^2e^{-a/R}.$$

Let $a=m^{-1/2}n^b$ for any $b>0$, and $c=2b+3$, combined with the assumption $m=O(n^b)$, we have

$$\begin{aligned}
\mathrm{Var}((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}})=&O(m^{-1}n^{-1}(\log n)^3)+O(m^{-1}n^{-3})+O(m^{-1}n^{2b})\cdot O(e^{-m^{-1/2}n^b})\\
=&O(m^{-1}n^{-1}(\log n)^3)+O(m^{-1}n^{-3})+O(m^{-1}n^{2b})\cdot O(e^{-n^{b/2}})\\
=&O(m^{-1}n^{-1}(\log n)^3)+O(m^{-1}n^{-3})+O(m^{-1}n^{2b})\cdot O(n^{-2b-3})\\
=&O(m^{-1}n^{-1}(\log n)^3)+O(m^{-1}n^{-3})\\
=&O(m^{-1}n^{-1}(\log n)^3).
\end{aligned}$$

■

Theorem A.20 (Theorem 5.3 Part 2) *Assuming that $m=O(n^b)$ for any $b>0$, then for $1\leq i,j\leq n$ and $i\neq j$,*

$$\frac{\mathrm{Var}(\widetilde{P}_{ij}^{(1)})}{\mathrm{Var}(\widehat{P}_{ij}^{(1)})}=O(n^{-1}(\log n)^3).$$

And thus

$$\mathrm{ARE}(\widehat{P}_{ij}^{(1)},\widetilde{P}_{ij}^{(1)})=0.$$

Proof: The results are direct from Theorem A.19 and Theorem 5.1. ■

A.5 $\widetilde{P}^{(q)}$ vs. $\widehat{P}^{(q)}$

Theorem A.21 *Let P and C be two n-by-n symmetric and hollow matrices satisfying element-wise conditions $0 < P_{ij} \le C_{ij} \le R$ for some constant $R > 0$. For $0 < \epsilon < 1$, we define m symmetric and hollow matrices as*

$$A^{(t)} \overset{iid}{\sim} (1-\epsilon)\mathrm{Exp}(P) + \epsilon \mathrm{Exp}(C)$$

for $1 \le t \le m$. Let $\widehat{P}^{(q)}$ be the entry-wise MLqE based on exponential distribution with m observations. Define $H^{(q)} = E[\widehat{P}^{(q)}]$, then for any constant $c > 0$ there exists another constant $n_0(c)$, independent of n, P, C and ϵ, such that if $n > n_0$, then for all η satisfying $n^{-c} \le \eta \le 1/2$,

$$P\left(\|\widehat{P}^{(q)} - H^{(q)}\|_2 \le 8R\sqrt{2n\ln(n/\eta)})\right) \ge 1 - \eta.$$

Proof: Similar to the proof of Theorem A.9.

By Lemma A.2 we have

$$\begin{aligned}
\left|\widehat{P}_{ij}^{(q)} - H_{ij}^{(q)}\right| =& \left|\widehat{P}_{ij}^{(q)} - \widehat{P}_{ij}^{(1)} + \widehat{P}_{ij}^{(1)} - H_{ij}^{(1)} + H_{ij}^{(1)} - H_{ij}^{(q)}\right| \\
\le& \left|\widehat{P}_{ij}^{(q)} - \widehat{P}_{ij}^{(1)}\right| + \left|\widehat{P}_{ij}^{(1)} - H_{ij}^{(1)}\right| + \left|H_{ij}^{(1)} - H_{ij}^{(q)}\right| \\
\le& \widehat{P}_{ij}^{(1)} + \left|\widehat{P}_{ij}^{(1)} - H_{ij}^{(1)}\right| + H_{ij}^{(1)} \\
\le& 2\left(\left|\widehat{P}_{ij}^{(1)} - H_{ij}^{(1)}\right| + H_{ij}^{(1)}\right).
\end{aligned}$$

Also,

$$\begin{aligned}
E[(\widehat{P}_{ij}^{(q)} - H_{ij}^{(q)})^k] \le& E\left[\left|\widehat{P}_{ij}^{(q)} - H_{ij}^{(q)}\right|^k\right] \\
\le& 2^k E\left[\left(\left|\widehat{P}_{ij}^{(1)} - H_{ij}^{(1)}\right| + H_{ij}^{(1)}\right)^k\right] \\
\le& 2^k \sum_{s=0}^{k} \binom{k}{s} E\left[\left|\widehat{P}_{ij}^{(1)} - H_{ij}^{(1)}\right|^s\right] \left(H_{ij}^{(1)}\right)^{k-s} \\
\le& 2^k \sum_{s=0}^{k} \binom{k}{s} R^s s! \left(H_{ij}^{(1)}\right)^{k-s} \\
\le& 2^k k! \sum_{s=0}^{k} \binom{k}{s} R^s \left(H_{ij}^{(1)}\right)^{k-s} \\
=& 2^k k! \left(R + H_{ij}^{(1)}\right)^k \\
\le& 2^{2k} k! R^k. \qquad (5)
\end{aligned}$$

Therefore we have

$$P\left(\|\widehat{P}^{(q)} - H^{(q)}\| \ge t\right) \le n \exp\left(-\frac{t^2/2}{32R^2 n + Rt}\right).$$

Now let $c > 0$ be given and assume $n^{-c} \le \eta \le 1/2$. Then there exists a $n_0(c)$ independent of n, P, C and ϵ such that whenever $n > n_0(c)$,

$$t = 8R\sqrt{2n\ln(n/\eta)} \le 32Rn.$$

Plugging this t into the equation above, we get

$$P(\|\widehat{P}^{(q)} - H^{(q)}\| \geq 8R\sqrt{2n\ln(n/\eta)}) \leq n\exp\left(-\frac{t^2}{64R^2n}\right) = \eta.$$

■

As we define $H^{(q)} = E[\widehat{P}^{(q)}]$, let $d^{(q)} = \text{rank}(H^{(q)})$ be the dimension in which we are going to embed $\widehat{P}^{(q)}$. Notice that it is less than or equal to $K \times K'$ based on the SBM assumption. Then we can define $H^{(q)} = ZZ^T$ where $Z \in \mathbb{R}^{n\times d^{(q)}}$.

For simplicity, from now on, we will use $\widehat{P}$ to represent $\widehat{P}^{(q)}$, use H to represent $H^{(q)}$ and use k to represent the dimension $d^{(q)}$ we are going to embed. Assume $H = USU^T = ZZ^T$, where $Z = [Z_1, \cdots, Z_n]^T$ is a n-by-k matrix. Then our estimate for Z up to rotation is $\widehat{Z} = \widehat{U}\widehat{S}^{1/2}$, where $\widehat{U}\widehat{S}\widehat{U}^T$ is the rank-d spectral decomposition of $|\widehat{P}| = (\widehat{P}^T\widehat{P})^{1/2}$.

Furthermore, we assume that the second moment matrix $E[Z_1Z_1^T]$ is rank k and has distinct eigenvalues $\lambda_i(E[Z_1Z_1^T])$. In particular, we assume that there exists $\delta > 0$ such that

$$\delta < \lambda_k(E[Z_1Z_1^T])$$

Lemma A.22 *Under the above assumptions, $\lambda_i(H) = \Theta(n)$ with high probability when $i \leq k$, i.e. the largest k eigenvalues of H is of order n. Moreover, we have $\|S\|_2 = \Theta(n)$ and $\|\widehat{S}\|_2 = \Theta(n)$ with high probability.*

Proof: Exactly the same as proof for Lemma A.10. ■

Lemma A.23 *Let $W_1\Sigma W_2^T$ be the singular value decomposition of $U^T\widehat{U}$. Then for sufficiently large n,*

$$\|U^T\widehat{U} - W_1W_2^T\|_F = O(n^{-1}\log n)$$

with high probability.

Proof: Exactly the same as proof for Lemma A.11. ■

We will denote the orthogonal matrix $W_1W_2^T$ by W^*.

Lemma A.24 *For sufficiently large n,*

$$\|W^*\widehat{S} - SW^*\|_F = O(\log n),$$

$$\|W^*\widehat{S}^{1/2} - S^{1/2}W^*\|_F = O(n^{-1/2}\log n)$$

and

$$\|W^*\widehat{S}^{-1/2} - S^{-1/2}W^*\|_F = O(n^{-3/2}\log n)$$

with high probability.

Proof: Similar to the proof of Lemma A.12. ■

Lemma A.25 *There exists a rotation matrix W such that for sufficiently large n,*

$$\|\widehat{Z} - ZW\|_F = \|(\widehat{P} - H)US^{-1/2}\|_F + O(n^{-1/2}(\log n)^{3/2})$$

with high probability.

Proof: Exactly the same as proof for Lemma A.13. ■

Theorem A.26 *There exists a rotation matrix W such that for sufficiently large n,*

$$\max_i \|\widehat{Z}_i - W Z_i\|_2 = O(n^{-1/2}(\log n)^{3/2})$$

with high probability.

Proof: Similar to the proof of Theorem A.14. ■

Lemma A.27 $\left|\widehat{Z}_i^T \widehat{Z}_j - Z_i^T Z_j\right| = O(n^{-1/2}(\log n)^{3/2})$ *with high probability.*

Proof: Similar to the proof of Lemma A.15. ■

Definition A.28 *Define $\widetilde{P}_{ij}^{(q)} = (\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}}$, our estimator for P_{ij}, to be a projection of $\widehat{Z}_i^T \widehat{Z}_j$ onto $[0, \min(\widehat{P}_{ij}^{(q)}, R)]$.*

Lemma A.29 (Theorem 5.4 Part 1) *Assuming that $m = O(n^b)$ for any $b > 0$, then the estimator based on ASE of MLqE has the same entry-wise asymptotic bias as MLqE, i.e.*

$$\lim_{n\to\infty} \text{Bias}(\widetilde{P}_{ij}^{(q)}) = \lim_{n\to\infty} E[\widetilde{P}_{ij}^{(q)}] - P_{ij} = \lim_{n\to\infty} E[\widehat{P}_{ij}^{(q)}] - P_{ij} = \lim_{n\to\infty} \text{Bias}(\widehat{P}_{ij}^{(q)}).$$

Proof: Fix some $a > 0$, we have

$$\begin{aligned}
&E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j|] \\
=&E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j| \mathbb{I}\{\widehat{P}_{ij}^{(1)} \le a\}] + E[|(\widehat{Z}_i^T \widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j| \mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}].
\end{aligned}$$

Note that we are thresholding according to $\widehat{P}^{(1)}$ instead of $\widehat{P}^{(q)}$. By Lemma

A.2, we know $\widehat{P}^{(q)} < \widehat{P}^{(1)}$ given any data. For the first term, we have

$$\begin{aligned}
&E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij}^{(1)} \le a\}] \\
\le& E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij}^{(1)} \le a\}\mathbb{I}\{\text{Lemma } A.27 \text{ holds}\}] \\
&+ E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij}^{(1)} \le a\}\mathbb{I}\{\text{Lemma } A.27 \text{ does not hold}\}] \\
\le& E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij}^{(1)} \le a\}|\{\text{Lemma } A.27 \text{ holds}] \\
&+ n^{-c}E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij}^{(1)} \le a\}|\text{Lemma } A.27 \text{ does not hold}] \\
\le& O(n^{-1/2}(\log n)^{3/2}) \\
&+ n^{-c}E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - \widehat{P}_{ij}^{(q)}|\mathbb{I}\{\widehat{P}_{ij}^{(1)} \le a\}|\text{Lemma } A.27 \text{ does not hold}] \\
&+ n^{-c}E[|\widehat{P}_{ij}^{(q)} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij}^{(1)} \le a\}|\text{Lemma } A.27 \text{ does not hold}] \\
\le& O(n^{-1/2}(\log n)^{3/2}) \\
&+ n^{-c}E[\widehat{P}_{ij}^{(q)}\mathbb{I}\{\widehat{P}_{ij}^{(1)} \le a\}|\text{Lemma } A.27 \text{ does not hold}] \\
&+ n^{-c}E[(\widehat{P}_{ij}^{(q)} + R)\mathbb{I}\{\widehat{P}_{ij}^{(1)} \le a\}|\text{Lemma } A.27 \text{ does not hold}] \\
\le& O(n^{-1/2}(\log n)^{3/2}) \\
&+ n^{-c}E[\widehat{P}_{ij}^{(1)}\mathbb{I}\{\widehat{P}_{ij}^{(1)} \le a\}|\text{Lemma } A.27 \text{ does not hold}] \\
&+ n^{-c}E[(\widehat{P}_{ij}^{(1)} + R)\mathbb{I}\{\widehat{P}_{ij}^{(1)} \le a\}|\text{Lemma } A.27 \text{ does not hold}] \\
\le& O(n^{-1/2}(\log n)^{3/2}) + an^{-c} + (a+R)n^{-c} \\
\le& O(n^{-1/2}(\log n)^{3/2}) + 2n^{-c}(a+R).
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
&E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] \\
\le& E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - \widehat{P}_{ij}^{(q)}|\mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] + E[|\widehat{P}_{ij}^{(q)} - Z_i^T Z_j|\mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] \\
\le& E[\widehat{P}_{ij}^{(q)}\mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] + E[(\widehat{P}_{ij}^{(q)} + R)\mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] \\
\le& E[\widehat{P}_{ij}^{(1)}\mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] + E[(\widehat{P}_{ij}^{(1)} + R)\mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] \\
\le& 2e^{-a/R}(a + 2mR).
\end{aligned}$$

Similarly, assuming $m = O(n^b)$ for any $b > 0$, we have

$$E[|(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^T Z_j|] = O(n^{-1/2}(\log n)^{3/2}).$$

■

Theorem A.30 *Assuming that $m = O(n^b)$ for any $b > 0$, then* $\mathrm{Var}((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}) = O(n^{-1}(\log n)^3)$.

Proof: By Lemma A.27,

$$\begin{aligned}
\mathrm{Var}((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}) =& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])^2] \\
=& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j + Z_i^TZ_j - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])^2] \\
=& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2] + E[(Z_i^TZ_j - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])^2] \\
&+ 2E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)(Z_i^TZ_j - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])] \\
\leq& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2] + E[(Z_i^TZ_j - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])^2] \\
&+ 2\sqrt{E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2]E[(Z_i^TZ_j - E[(\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}}])^2]} \\
\leq& 4E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2].
\end{aligned}$$

Fix some $a > 0$, we have

$$\begin{aligned}
&E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2] \\
=& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2\mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\}] + E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2\mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}].
\end{aligned}$$

Note that we are thresholding according to $\widehat{P}^{(1)}$ instead of $\widehat{P}^{(q)}$. By Lemma A.2, we know $\widehat{P}^{(q)} < \widehat{P}^{(1)}$ given any data. For the first term, we have

$$\begin{aligned}
&E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2\mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\}] \\
\leq& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2\mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\}\mathbb{I}\{\text{Lemma } A.27 \text{ holds}\}] \\
&+ E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2\mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\}\mathbb{I}\{\text{Lemma } A.27 \text{ does not hold}\}] \\
\leq& E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2\mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\}|\{\text{Lemma } A.27 \text{ holds}] \\
&+ n^{-c}E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - Z_i^TZ_j)^2\mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\}|\text{Lemma } A.27 \text{ does not hold}] \\
\leq& O(n^{-1}(\log n)^3) \\
&+ 2n^{-c}E[((\widehat{Z}_i^T\widehat{Z}_j)_{\mathrm{tr}} - \widehat{P}_{ij}^{(q)})^2\mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\}|\text{Lemma } A.27 \text{ does not hold}] \\
&+ 2n^{-c}E[(\widehat{P}_{ij}^{(q)} - Z_i^TZ_j)^2\mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\}|\text{Lemma } A.27 \text{ does not hold}] \\
\leq& O(n^{-1}(\log n)^3) \\
&+ 2n^{-c}E[\widehat{P}_{ij}^{(q)2}\mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\}|\text{Lemma } A.27 \text{ does not hold}] \\
&+ 2n^{-c}E[(\widehat{P}_{ij}^{(q)} + R)^2\mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\}|\text{Lemma } A.27 \text{ does not hold}] \\
\leq& O(n^{-1}(\log n)^3) + 2n^{-c}E[\widehat{P}_{ij}^{(1)2}\mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\}|\text{Lemma } A.27 \text{ does not hold}] \\
&+ 2n^{-c}E[(\widehat{P}_{ij}^{(1)} + R)^2\mathbb{I}\{\widehat{P}_{ij}^{(1)} \leq a\}|\text{Lemma } A.27 \text{ does not hold}] \\
\leq& O(n^{-1}(\log n)^3) + 2a^2n^{-c} + 2(a+R)^2n^{-c} \\
\leq& O(n^{-1}(\log n)^3) + 4n^{-c}(a+R)^2.
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
&E[((\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] \\
\leq& 2E[((\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}} - \widehat{P}_{ij}^{(q)})^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] + 2E[(\widehat{P}_{ij}^{(q)} - Z_i^T Z_j)^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] \\
\leq& 2E[\widehat{P}_{ij}^{(q)2} \mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] + 2E[(\widehat{P}_{ij}^{(q)} + R)^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] \\
\leq& 2E[\widehat{P}_{ij}^{(1)2} \mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] + 2E[(\widehat{P}_{ij}^{(1)} + R)^2 \mathbb{I}\{\widehat{P}_{ij}^{(1)} > a\}] \\
\leq& 4e^{-a/R}(a + 2m^{1/2}R)^2.
\end{aligned}$$

Similarly, assuming $m = O(n^b)$ for any $b > 0$, we have

$$\text{Var}((\widehat{Z}_i^T\widehat{Z}_j)_{\text{tr}}) = O(n^{-1}(\log n)^3).$$

■

Theorem A.31 (Theorem 5.4 Part 2) *Assuming that $m = O(n^b)$ for any $b > 0$, then for $1 \leq i, j \leq n$ and $i \neq j$,*

$$\frac{\text{Var}(\widetilde{P}_{ij}^{(q)})}{\text{Var}(\widehat{P}_{ij}^{(q)})} = O(mn^{-1}(\log n)^3).$$

Moreover, if $m = o(n(\log n)^{-3})$, then

$$\text{ARE}(\widehat{P}_{ij}^{(q)}, \widetilde{P}_{ij}^{(q)}) = 0.$$

Proof: The results are direct from Theorem A.30 and Theorem 5.1. ■

A.6 $\widetilde{P}^{(q)}$ vs. $\widetilde{P}^{(1)}$

Theorem A.32 *For sufficiently large C and any $1 \leq i, j \leq n$, if $m = O(n^b)$ for any $b > 0$, then*

$$\lim_{m,n\to\infty} \text{Bias}(\widetilde{P}_{ij}^{(1)}) > \lim_{m,n\to\infty} \text{Bias}(\widetilde{P}_{ij}^{(q)})$$

Proof: Direct result from Theorem 5.1, Theorem 5.3 and Theorem 5.4. ■

Theorem A.33 *For sufficiently large C and any $1 \leq i, j \leq n$, if $m = O(n(\log n)^{-3})$, then*

$$\lim_{m,n\to\infty} \text{Var}(\widetilde{P}_{ij}^{(1)}) = \lim_{m,n\to\infty} \text{Var}(\widetilde{P}_{ij}^{(q)}) = 0.$$

Proof: Direct result from Theorem 5.3 and Theorem 5.4. ■

A.7 Other Proofs

Lemma A.34 *Let $A_{ij} \overset{ind}{\sim} (1-\epsilon)f_{P_{ij}} + \epsilon f_{C_{ij}}$ with f to be Poisson, then $E[(A_{ij} - E[\widehat{P}_{ij}^{(1)}])^k] \leq \text{const}^k \cdot k!$, where $\widehat{P}^{(1)}$ is the entry-wise MLE as defined before.*

Proof: First we prove $(x-\theta)^k \leq k!(e^{x-\theta} + e^{\theta-x})$.

1. k is even. Then by Taylor expansion, $e^{x-\theta} + e^{\theta-x} \geq \frac{(x-\theta)^k}{k!}$

2. k is odd. When $x \geq \theta$, still by Taylor expansion, $(x-\theta)^k \leq k! e^{x-\theta}$. When $x < \theta$, $(x-\theta)^k < 0 \leq k! e^{x-\theta}$.

Thus $(x-\theta)^k \leq k!(e^{x-\theta} + e^{\theta-x})$. So the k-th central moment of Poisson distribution with parameter θ is bounded by

$$\begin{aligned} E[(X-\theta)^k] &\leq k! \left(E[e^{X-\theta}] + E[e^{\theta-X}]\right) \\ &= k! \left(e^{-\theta} E[e^X] + e^{\theta} E[e^{-X}]\right) \\ &= k! \left(e^{\theta(e-2)} + e^{\theta e^{-1}}\right). \end{aligned}$$

Let $X_1 \sim \text{Poisson}(P_{ij})$ and $X_2 \sim \text{Poisson}(C_{ij})$. Then if A_{ij} is distributed from a mixture model as in the statement, we have

$$\begin{aligned} & E[(A_{ij} - E[\widehat{P}_{ij}^{(1)}])^k] \\ =& (1-\epsilon) E[(X_1 - P_{ij} + P_{ij} - E[\widehat{P}_{ij}^{(1)}])] + \epsilon E[(X_2 - C_{ij} + C_{ij} - E[\widehat{P}_{ij}^{(1)}])] \\ =& (1-\epsilon) \sum_{j=0}^{k} \binom{k}{j} (P_{ij} - E[\widehat{P}_{ij}^{(1)}])^{k-j} E[(X_1 - P_{ij})^j] \\ & + \epsilon \sum_{j=0}^{k} \binom{k}{j} (C_{ij} - E[\widehat{P}_{ij}^{(1)}])^{k-j} E[(X_2 - C_{ij})^j] \\ \leq& (1-\epsilon) \sum_{j=0}^{k} \binom{k}{j} (P_{ij} - E[\widehat{P}_{ij}^{(1)}])^{k-j} \cdot j! \cdot \text{const} \\ & + \epsilon \sum_{j=0}^{k} \binom{k}{j} (C_{ij} - E[\widehat{P}_{ij}^{(1)}])^{k-j} \cdot j! \cdot \text{const} \\ \leq& (1-\epsilon) k! \cdot \text{const}^k + \epsilon k! \cdot \text{const}^k \\ \leq& \text{const}^k \cdot k!. \end{aligned}$$

■