# Robust Estimation from Multiple Graphs under Gross Error Contamination

Runze Tang, Minh Tang, Joshua T. Vogelstein, Carey E. Priebe
Johns Hopkins University

July 1, 2017

**Abstract**

Estimation of graph parameters based on a collection of graphs is essential for a wide range of graph inference tasks. In practice, weighted graphs are generally observed with edge contamination. We consider a weighted latent position graph model contaminated via an edge weight gross error model and propose an estimation methodology based on robust L$q$ estimation followed by low-rank adjacency spectral decomposition. We demonstrate that, under appropriate conditions, our estimator both maintains L$q$ robustness and wins the bias-variance tradeoff by exploiting low-rank graph structure. We illustrate the improvement offered by our estimator via both simulations and a human connectome data experiment.

**Keywords**: weighted, network, low-rank, embedding

## 1 Background and Overview

Network analysis has emerged as an area of intense statistical theory and application activity. In the general parametric framework, $G \sim f \in \mathcal{F} = \{f_\theta : \theta \in \Theta\}$, and selecting a principled and productive estimator $\widehat{\theta}$ for the unknown graph parameter $\theta$ given a sample of graphs $\{G^{(1)}, \cdots, G^{(m)}\}$ is one of the most foundational and essential tasks, facilitating subsequent inference. For example, Ginestet et al. [2014] proposes a method to test for a difference between the networks of two groups of subjects in functional neuroimaging; while hypothesis testing is the ultimate goal, estimation is a key intermediate step. We propose a widely-applicable, robust, low-rank estimation procedure for a collection of weighted graphs.

Consider for illustration the connectome dataset made available through the Consortium for Reliability and Reproducibility[1] and investigated in Section 7.2 below. We have $m = 114$ brain graphs, each having $n = 70$ vertices representing different anatomical regions; the (errorfully observed) weight of an edge between two vertices represents the number of white-matter tracts connecting the corresponding two regions of the brain, as measured by diffusion tensor magnetic resonance imaging. Our goal in this situation is to estimate the average number of white-matter tracts between different regions of the brain. A more accurate

[1] `http://fcon_1000.projects.nitrc.org/indi/CoRR/html/`

estimate can lead to a better understanding of brain connectivity and hence functionality. Also, better estimates will improve performance on other tasks, such as diagnosis of brain disease.

The maximum likelihood estimate (MLE) – the edge-wise sample mean, without taking any graph structure into account, as in the (weighted extension of) the independent edge graph model (IEM) [Bollobás et al., 2007] (described in Section 3.1 below) – is a natural candidate for our estimation problem. However, the MLE suffers from at least two major deficiencies in our setting: high variance and non-robustness.

In our high dimensional setting (a large number of vertices, $n$), the edge-wise MLE leads to estimates with unacceptably high variance unless the sample size (the number of graphs, $m$) is exceedingly large. However, if the graphs can be assumed to be (approximately) low-rank, then by biasing towards low-rank structure, more elaborate estimators can have greatly reduced variance and win the bias-variance tradeoff. For our connectome data (Section 7.2 Figure 5) we observe this approximate low-rank property. Tang et al. [2016] develops an estimator based on a low-rank approximation and proves that this new estimator outperforms the edge-wise MLE, decreasing the overall asymptotic variance dramatically by smoothing towards the low-rank structure.

The second edge-wise MLE deficiency in our setting derives from the edge observations being subject to contamination. That is, the weights attributed to edges are possibly observed with noise. The sample mean is notoriously unrobust to outliers; thus, under the possibility of contamination, it is wise to use robust methods, such as the ML$q$E [Qin and Priebe, 2013a, **?**] considered in this paper.

To address these two deficiencies simultaneously, we propose an estimation methodology which is a natural extension of [Tang et al., 2016] to gross error contamination. Our proposed estimator both inherits ML$q$E robustness and wins the bias-variance tradeoff by taking advantage of 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 versions, and define the gross error contamination model we will consider. In Section 4, we present our estimation methodology in terms of two estimators designed to address the two edge-wise MLE deficiencies described above, and we construct our final estimator by combining the two estimators. In Section 5, we prove that our estimator is superior, under appropriate conditions, and this result is generalized in Section 6. In Section 7, we illustrate the performance of our estimator through experimental results on simulated and real data.

## 2 Models

For this work, we are in the scenario where $m$ weighted graphs on $n$ vertices are given as adjacency matrices $\{A^{(t)}\}(t = 1, \ldots, m)$. 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 $i \in [n]$ and $t_1, t_2 \in [m]$.

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.

## 2.1 Weighted Independent Edge Model

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 drawn from a Bernoulli distribution with parameter $P_{ij}$ independent of all other edges. We first extend the definition of IEM to the weighted independent edge model (WIEM) with respect to a one-parameter family $\mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}\}$; for example, $f_\theta$ may be the Poisson distribution with parameter $\theta$. Denote the graph parameters as a matrix $P \in \Theta^{n \times n} \subset \mathbb{R}^{n \times n}$. Then under a WIEM, the (weighted) edge between vertex $i$ and vertex $j$ ($i < j$ due to symmetry) has weight $A_{ij}$ drawn from $f_{P_{ij}}$ independent of all other edges. Thus IEM is a special case of WIEM, with $\mathcal{F}$ representing the collection of Bernoulli distributions and $\Theta = [0,1]$.

Note that the graphs considered in this paper are undirected without self-loops, and the parameter matrix $P$ can be considered to be symmetric and hollow. That is, for convenience, we still define the parameters to be an $n$-by-$n$ matrix while only $\binom{n}{2}$ of them are active.

## 2.2 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 to each vertex $i$ 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 [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 follows:

**Definition 2.1 (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 via consideration of latent position matrix $X \in \mathbb{R}^{n \times d}$ 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 $i < j \in [n]$.*

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

## 2.3 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 a 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 $\tau$, the connectivity between vertex $i$ and vertex $j$ follows a Bernoulli distribution with parameter $B_{\tau_i,\tau_j}$. This can be easily generalized to the weighted stochastic blockmodel (WSBM), with the Bernoulli distribution replaced by a one-parameter distribution family $\mathcal{F} = \{f_\theta, \theta \in \Theta \subset \mathbb{R}\}$ and the block probability matrix given by $B \in \Theta^{K \times K} \subset \mathbb{R}^{K \times K}$.

Since the RDPG/WRDPG setting motivates low-rank estimation – $P$ is of rank less than or equal to $d$ – all analysis in this work is based on such a 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. Henceforth, 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 $\mathcal{F} = \{f_\theta, \theta \in \Theta \subset \mathbb{R}\}$. Letting $d = \mathrm{rank}(B)$, 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 as $f$ 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 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 $\tau$ 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 $\tau$ fixed when sampling all $m$ graphs. Then $m$ symmetric and hollow graphs are sampled such that conditioning on $\tau$, 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 $1 \le t \le m$, $1 \le i < j \le n$.

## 2.4 Weighted Stochastic Blockmodel as a Weighted Random Dot Product Graph with Contamination

In practice, completely accurate data is difficult to collect – there will almost 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 [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 contaminated values $G^{**} \sim h_C \in \mathcal{H}$ when gross errors occur. Here $P$ and $C$ represent the respective parameter matrices of the two distribution families. As for graphs, one way to generalize to 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 contamination 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.

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 $\tau$ is the block assignments, the contamination distribution $f_{C_{ij}}$ with $C_{ij} = B'_{\tau'_i,\tau'_j}$ also has the block structure, where $B'$ is the block probability matrix and $\tau'$ is the block assignment vector. For clarity, we will consider the sampling procedure when the contamination has the same block structure, i.e. $\tau = \tau'$. However, this simplification is not required in our theory.

To generate $m$ graphs under this contamination model with known vertex correspondence, we first sample $\tau$ from the categorical distribution with parameter $\rho$ and keep $\tau$ 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 $\tau$, 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 $1 \leq t \leq m$, $1 \leq i < j \leq n$, where $P_{ij} = B_{\tau_i,\tau_j}$ and $C_{ij} = B'_{\tau_i,\tau_j}$. Here $\epsilon$ 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 contamination.

# 3 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 such as $\epsilon$ and $C$, our goal is still to estimate the uncontaminated parameter matrix $P$. In this section, we present four estimators as depicted 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 ML$q$E $\widehat{P}^{(q)}$, and the low-rank approximation of the entry-wise ML$q$E $\widetilde{P}^{(q)}$. Since the observed graphs are symmetric and hollow with a symmetric parameter matrix of the model, we are not concerned with estimating the diagonal of $P$; however, the estimate itself should be at least symmetric.

## 3.1 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, such as Bernoulli, Poisson, or Exponential, the entry-wise MLE $\widehat{P}^{(1)}$ is the uniformly minimum-variance unbiased estimator, i.e. it has the smallest variance among all unbiased estimators. In addition, it has desirable asymptotic properties as the number of graphs $m$ goes to infinity. However, in high dimensional situations such as our graph setting, 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 improve as 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

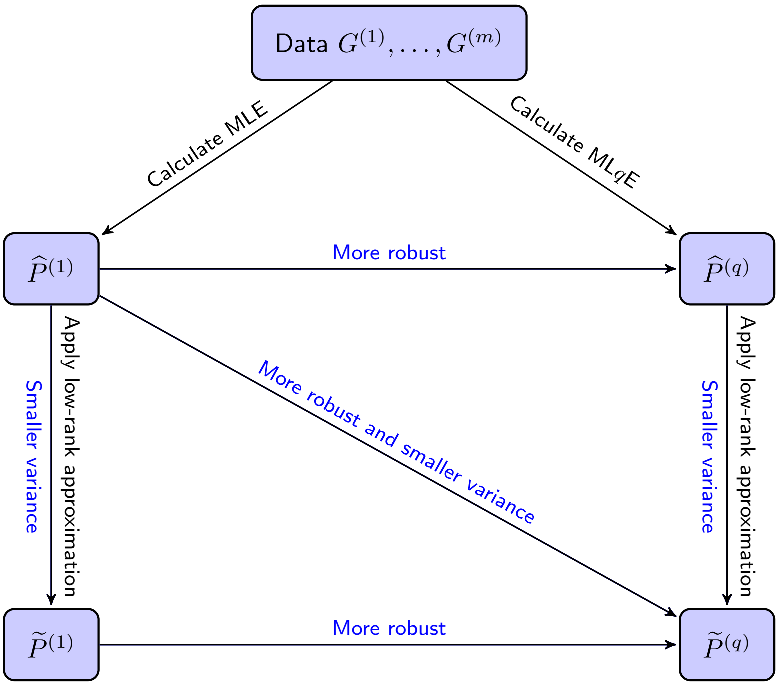

Figure 1: Roadmap among the data and four estimators.

WRDPG or a WSBM, then the entry-wise MLE is no longer the MLE and the performance can be improved by considering low-rank estimators.

## 3.2 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. The construction procedure of $\widetilde{P}^{(1)}$ consists of several steps, which will be introduced respectively in the following subsections.

### 3.2.1 Rank-$d$ Approximation

Given a dimension $d$, we consider $\widetilde{P}^{(1)} = \mathrm{lowrank}_d(\widehat{P}^{(1)})$ as the best rank-$d$ positive semi-definite approximation of $\widehat{P}^{(1)}$. To find this 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 remaining $n-d$ 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 given 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: 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 diag$(s_1, \ldots, s_d)$
3: Set $\widehat{U} = [u_1, \ldots, u_d] \in \mathbb{R}^{n \times d}$
4: Set lowrank$_d(A)$ to $\widehat{U}\widehat{S}\widehat{U}^\top$
```

### 3.2.2 Dimension Selection

Although Algorithm 1 provides 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 this approximation will fail to capture important information. On the other hand, when $d$ is too large, the approximation will be subject to substantial noise and also lead to a poor estimate. So a carefully selected dimension $d$ is an essential aspect of this approximation/estimation.

A general approach to selecting the dimension $d$ is to analyze the ordered eigenvalues and look for a "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 elbows. In this paper, to avoid under-estimating the dimension, which is often much more harmful than over-estimating it, we choose the 3rd Zhu and Ghodsi elbow.

Although it is always challenging to select a proper dimension, the results of our real data experiment in Section 7.2 demonstrate that a wide range of dimension choices will lead to a fairly good results. Thus a proper dimension selection method can be applied directly without excessively tuning parameters, 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 $\mathrm{lowrank}_d(\widehat{P}^{(1)})$; (see Algorithm 1)
4: Set $\widetilde{P}^{(1)}$ with each entry $\widetilde{P}^{(1)}_{ij} = \max(Q_{ij}, 0)$
```

By combining the key pieces introduced above, we give the detailed description for calculating the estimator $\widetilde{P}^{(1)}$ with dimension selection in Algorithm 2.

### 3.3 Entry-wise Maximum L$q$-likelihood Estimator $\widehat{P}^{(q)}$

In the case of no contamination, 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 can outperform the MLE in terms of mean squared error by winning the bias-variance tradeoff. Moreover, under contamination models, robust estimators can even outperform the MLE asymptotically since they are designed to be not unduly affected by outliers. We consider one robust estimator, the maximum L$q$-likelihood estimator (ML$q$E) proposed by **?**.

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 L$q$-likelihood estimate ($q > 0$) of $\theta_0$ 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 ML$q$E is a generalization of MLE. Moreover, define

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

and

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

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

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

This form interprets $\widehat{\theta}_{\mathrm{ML}q\mathrm{E}}$ as a solution to a 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 ML$q$E puts less weight on the data points which do not fit the current distribution well. Equal weights are induced by $q = 1$ and lead to the standard MLE.

Under the WIEM, we can calculate the robust entry-wise ML$q$E $\widehat{P}^{(q)}$ 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 ML$q$E $\widehat{P}^{(q)}$ when $q = 1$. There is also a bias-variance tradeoff in selecting the parameter $q$. **?** proposed a way to select $q$ in general. In this work, we do not focus on automatic selection of $q$.

### 3.4 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 approximately in the entry-wise ML$q$E $\widehat{P}^{(q)}$. 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 ML$q$E matrix $\widehat{P}^{(q)}$ to get $\widetilde{P}^{(q)}$. As in Algorithm 2, we apply the same dimension selection method. The only change is to substitute $\widehat{P}^{(1)}$ by $\widehat{P}^{(q)}$. The details of the algorithm are 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 ML$q$E $\widehat{P}^{(q)}$
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 $\widetilde{P}^{(q)}_{ij} = \max(Q_{ij}, 0)$
```

# 4 Theoretical Results

In this section, for illustrative purposes, we present theoretical results for the case in which 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. These results can be extended beyond the exponential under appropriate conditions, which will be discussed in Section 6.

For clarity, we restate the model settings discussed in Section 3.4. Consider the WSBM with parameters $B$ and $\rho$. First we sample the block membership $\tau$ from the categorical distribution with parameter $\rho$ and keep it fixed for all $m$ graphs. Conditioned on this $\tau$, the uncontaminated probability matrix $P$ satisfies $P_{ij} = B_{\tau_i, \tau_j}$. In this section, we assume the contamination has the same block membership $\tau$, and so the contamination matrix $C \in \mathbb{R}^{n \times n}$ has the same block structure as $P$. Denote $\epsilon$ as the probability that an edge is contaminated. Then $m$ symmetric graphs $G^{(1)}, \ldots, G^{(m)}$ are sampled such that conditioning on $\tau$, 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 $1 \le t \le m$, $1 \le i < j \le n$. Note that our theoretical results do not require the contamination to have the same block structure and block membership $\tau$ as the uncontaminated probability matrix; different block structure will lead to similar results – but require higher embedding dimension – since the rank of $(1-\epsilon)P_{ij} + \epsilon C_{ij}$ is still finite.

In the setting outlined above, 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 easily leads to the result for the total MSE for the entire matrix.

We present the main results in this section. The proofs are given in the appendix.

## 4.1 $\widehat{P}^{(1)}$ vs. $\widehat{P}^{(q)}$

We first compare the performance between the entry-wise MLE $\widehat{P}^{(1)}$ and the entry-wise ML$q$E $\widehat{P}^{(q)}$. Without using the graph 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 4.1** *For any $0 < q < 1$, there exists $C_0(P_{ij}, \epsilon, q) > 0$ such that under the contaminated model with $C_{ij} > C_0(P_{ij}, \epsilon, q)$, MLqE has smaller entry-wise asymptotic bias compared to MLE, i.e.*

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

*for $1 \le i, j \le n$ and $i \neq j$. Moreover, without any assumptions 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 ML$q$E $\widehat{P}^{(q)}$ has smaller bias for estimating $P$ asymptotically compared to the entry-wise MLE $\widehat{P}^{(1)}$. Although we put restrictions on the contamination 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 requires only that the contamination of the model is large enough (either large contamination parameter matrix, or higher likelihood of encountering an outlier). From a different perspective, by putting a condition on $q$ with respect to the amount of contamination, 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 $\epsilon$ or a small enough $q$.

Theorem 5.1 also indicates that both estimators have variances converging 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, $\widehat{P}^{(q)}$ asymptotically reduces the bias while keeping the variance the same 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.

## 4.2 $\widehat{P}^{(1)}$ 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 an 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 4.2** *For any parameter $\theta$ 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 $\widehat{\theta}^{(2)}$. Combined with the fact that both estimators have the same asymptotic bias, we conclude that $\widehat{\theta}^{(1)}$ 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 appropriate conditions, and then use the ARE criterion to compare the performance in the following theorem.

**Theorem 4.3** *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.*

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

*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 fixed or grows no 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 the two estimators is of order $O(n^{-1}(\log n)^3)$, we have that ARE is 0. Thus $\widetilde{P}_{ij}^{(1)}$ is much better than $\widehat{P}_{ij}^{(1)}$ for 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.

## 4.3 $\widehat{P}^{(q)}$ vs. $\widetilde{P}^{(q)}$

We now proceed to analyze the effect of the ASE procedure applied to the entry-wise ML$q$E $\widehat{P}^{(q)}$ 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 appropriate conditions, and then use the ARE criterion to compare the performance in the following theorem.

**Theorem 4.4** *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 $1 \leq i, j \leq n$ and $i \neq j$,*

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

*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 L$q$ likelihood equation. As a result, while the ASE procedure still does not affect the asymptotic bias, the ratio of variances has an extra term $m$. This leads to a slight difference in the comparison. Specifically, when $m$ is fixed, the order of the ratio of variances is $O(n^{-1}(\log n)^3)$, which goes to 0 as $n \to \infty$. Even if $m$ also increases as $n$ increases, as long as it grows on the order of $o(n(\log n)^{-3})$, the ARE is still 0.

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

## 4.4 $\widetilde{P}^{(1)}$ 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 4.5** *For sufficiently large $C$ and any $1 \leq i, j \leq n$, if $m = O(n^b)$ for any $b > 0$, then the estimator based on ASE of MLqE has smaller entry-wise asymptotic bias compared to the estimator based on ASE of MLE, i.e.*

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

*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 ML$q$E $\widehat{P}^{(q)}$ and has a smaller asymptotic bias compared to $\widetilde{P}^{(1)}$ while both estimates have variance going to 0 as $m \to \infty$. Thus in summary, $\widetilde{P}^{(q)}$ is the best among all four estimators.

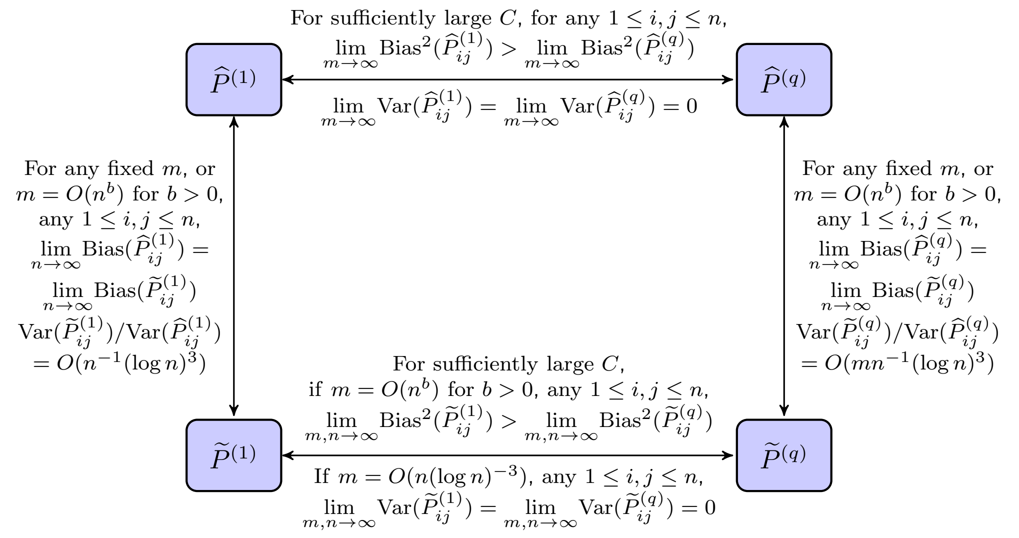
For sufficiently large $C$, for any $1 \le i, j \le n$,
$\lim_{m\to\infty} \mathrm{Bias}^2(\widehat{P}_{ij}^{(1)}) > \lim_{m\to\infty} \mathrm{Bias}^2(\widehat{P}_{ij}^{(q)})$
$\widehat{P}^{(1)}$
$\widehat{P}^{(q)}$
$\lim_{m\to\infty} \mathrm{Var}(\widehat{P}_{ij}^{(1)}) = \lim_{m\to\infty} \mathrm{Var}(\widehat{P}_{ij}^{(q)}) = 0$
For any fixed $m$, or $m = O(n^b)$ for $b > 0$, any $1 \le i, j \le n$, $\lim_{n\to\infty} \mathrm{Bias}(\widehat{P}_{ij}^{(1)}) = \lim_{n\to\infty} \mathrm{Bias}(\widetilde{P}_{ij}^{(1)})$ $\mathrm{Var}(\widetilde{P}_{ij}^{(1)})/\mathrm{Var}(\widehat{P}_{ij}^{(1)}) = O(n^{-1}(\log n)^3)$
For any fixed $m$, or $m = O(n^b)$ for $b > 0$, any $1 \le i, j \le n$, $\lim_{n\to\infty} \mathrm{Bias}(\widehat{P}_{ij}^{(q)}) = \lim_{n\to\infty} \mathrm{Bias}(\widetilde{P}_{ij}^{(q)})$ $\mathrm{Var}(\widetilde{P}_{ij}^{(q)})/\mathrm{Var}(\widehat{P}_{ij}^{(q)}) = O(mn^{-1}(\log n)^3)$
For sufficiently large $C$, if $m = O(n^b)$ for $b > 0$, any $1 \le i, j \le n$,
$\lim_{m,n\to\infty} \mathrm{Bias}^2(\widetilde{P}_{ij}^{(1)}) > \lim_{m,n\to\infty} \mathrm{Bias}^2(\widetilde{P}_{ij}^{(q)})$
$\widetilde{P}^{(1)}$
$\widetilde{P}^{(q)}$
If $m = O(n(\log n)^{-3})$, any $1 \le i, j \le n$,
$\lim_{m,n\to\infty} \mathrm{Var}(\widetilde{P}_{ij}^{(1)}) = \lim_{m,n\to\infty} \mathrm{Var}(\widetilde{P}_{ij}^{(q)}) = 0$

Figure 2: Relationship among four estimators.

### 4.5 Summary

We summarize all four estimators and their relationships in Figure 2. From top to bottom in 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 contamination, whenever the number of graphs $m$ is large enough, the bias term which dominates the MSE will be improved. Thus in Figure 2 we have quantified the qualitative roadmap introduced in Figure 1.

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

## 5 Extensions

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

1. Letting $A_{ij} \overset{ind}{\sim} (1-\epsilon) f_{P_{ij}} + \epsilon f_{C_{ij}}$, then we require $f_\theta$ to satisfy $E[(A_{ij} - E[\widehat{P}_{ij}^{(1)}])^k] \le \mathrm{const}^k \cdot k!$, 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. $\widehat{P}_{ij}^{(R)} \leq \text{const} \cdot \widehat{P}_{ij}^{(1)}$;

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

Condition 1 is to ensure that observations will not deviate too far from the expectation so that the concentration inequalities hold. Condition 2 is discussed in Section 5.1 and requires the contamination of the model to be large enough (a restriction on the distribution) and $\widehat{P}^{(R)}$ 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 $O(m^{-1})$. (Note that in the absence of Condition 4, similar but weaker results can still be derived.)

As an example of a (distribution, robust estimator)-pair satisfying the above four conditions, other than (exponential distribution, ML$q$E) as presented above in Section 5, we sketch the proof that the (Poisson distribution, ML$q$E) satisfies. The Poisson distribution is a commonly used distribution for nonnegative graphs with integer weights. Lemma A.34 verifies Condition 1; intuitively, since the exponential distribution has a fatter tail compared to the Poisson, we should have the bound for the central moment of the Poisson directly from the results for the exponential distribution. Condition 2 is satisfied when we use the ML$q$E with the Poisson. For Condition 3, $\widehat{P}_{ij}^{(R)}/\widehat{P}_{ij}^{(1)}$ is maximized when there are $m$ data points $x_1, \cdots, x_m$ 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 points to be smaller than the weights of the remaining $m-k$ points. Thus $e^{-\bar{x}} < \bar{x}^{x_m} e^{-\bar{x}}/x_m!$. But 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 ML$q$E since all data points equal zero. Thus ML$q$E is bounded by $e\bar{x}$. As a result, $\widehat{P}_{ij} \leq e\widehat{P}_{ij}^{(1)}$ and Condition 3 is satisfied. Finally, Condition 4 follows directly from the theory of minimum contrast estimators.

In summary, all theorems in Section 5 hold for the Poisson distribution together with the ML$q$E. The four conditions presented in this section provide a general framework for extending the theory to more general models and robust estimators.

# 6 Empirical Results

## 6.1 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.

### 6.1.1 Simulation Setting

Here we consider a 2-block WSBM 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 WSBM 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 ease of presentation, in the simulation we assume the true dimension $d = \text{rank}(B) = 2$ is known, and thus we eliminate the dimension selection step in Algorithm 2 and Algorithm 3.

### 6.1.2 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 the diagonal conditioned on the off-diagonal entries of $P$.

However, unlike the entry-wise estimators, e.g. $\widehat{P}^{(1)}$, the estimators which take advantage of the graph structure can benefit from 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 using 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 resolving this issue.

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

### 6.1.3 Simulation Results

In order to see how the performance of the four estimators vary with respect to contamination, we first run 1000 Monte Carlo replicates based on the contaminated WSBM 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 $\epsilon$ 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 ML$q$E estimator, we 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 presented in Figure 3. Different curves represent the simulated MSE associated with the four different estimators. Firstly, we see MLE $\widehat{P}^{(1)}$ is the best estimator when there is little or no contamination (i.e. $\epsilon$ is small or $\epsilon = 0$); however this estimator degrades dramatically as the contamination probability increases. On the other hand, the ML$q$E $\widehat{P}^{(q)}$ is slightly less efficient than the MLE $\widehat{P}^{(1)}$ when the contamination probability is small, but is much more robust under a large contamination probability compared to the MLE. Next, we see that even with a relatively 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 lets $\widetilde{P}^{(1)}$ win the bias-variance tradeoff. Since the ML$q$E $\widehat{P}^{(q)}$ approximately preserves the low-rank structure of the original graph, 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 have reduced variances, $\widetilde{P}^{(q)}$ constructed based on ML$q$E inherits the robustness from ML$q$E, so when the contamination probability is large enough, $\widetilde{P}^{(q)}$ outperforms $\widetilde{P}^{(1)}$ and degrades more slowly.

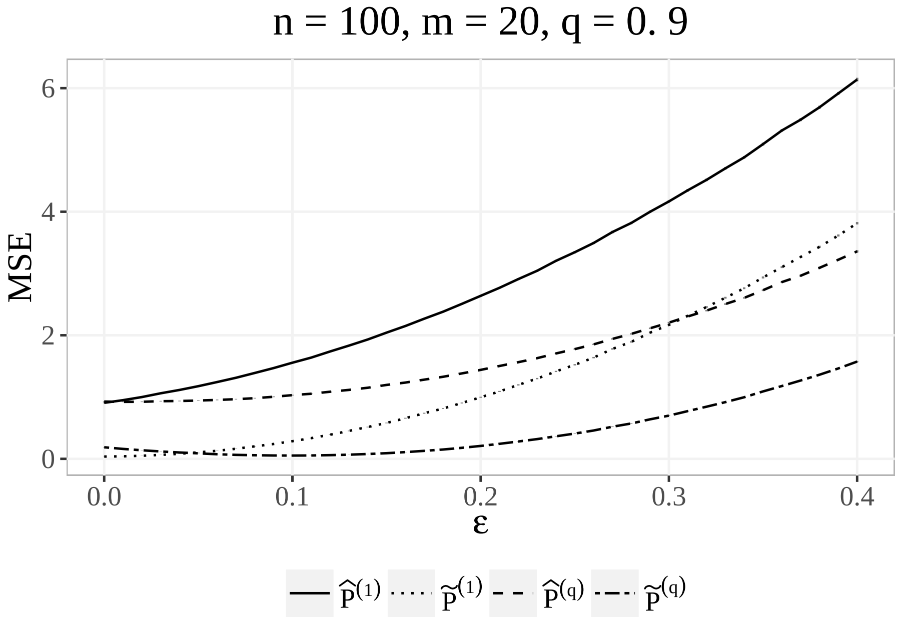

Figure 3: Mean squared error in average by varying contamination ratio $\epsilon$ with fixed $n = 100$ and $m = 20$ based on 1000 Monte Carlo replicates, using $q = 0.9$ when applying ML$q$E. Different curves represent the simulated MSE associated with four different estimators. 1. MLE $\widehat{P}^{(1)}$ vs ML$q$E $\widehat{P}^{(q)}$: MLE outperforms by a small amount when there is no contamination (i.e. $\epsilon = 0$), but it degrades dramatically when contamination probability 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. ML$q$E $\widehat{P}^{(q)}$ vs ASE $\circ$ ML$q$E $\widetilde{P}^{(q)}$: ML$q$E approximately preserves the low rank structure of the original graph, so ASE procedure still helps and $\widetilde{P}^{(q)}$ wins the bias-variance tradeoff; 4. ASE $\circ$ ML$q$E $\widetilde{P}^{(q)}$ vs ASE $\circ$ MLE $\widetilde{P}^{(1)}$: When contamination probability is large enough, $\widetilde{P}^{(q)}$ based on ML$q$E is better, since it inherits the robustness from ML$q$E.

Figure 4 shows additional simulation results by varying the parameter $q$ in ML$q$E with fixed $n = 100$, $m = 20$ and $\epsilon = 0.1$ based on 1000 Monte Carlo replicates. 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, for a wide range of $q$, the ML$q$E wins the bias-variance tradeoff and exhibits the robustness property compared to the MLE. And as $q$ goes to 1, ML$q$E goes to the MLE as expected.

Figures 3 and 4 provide a tangible demonstration of the theoretical results presented in Section 5.

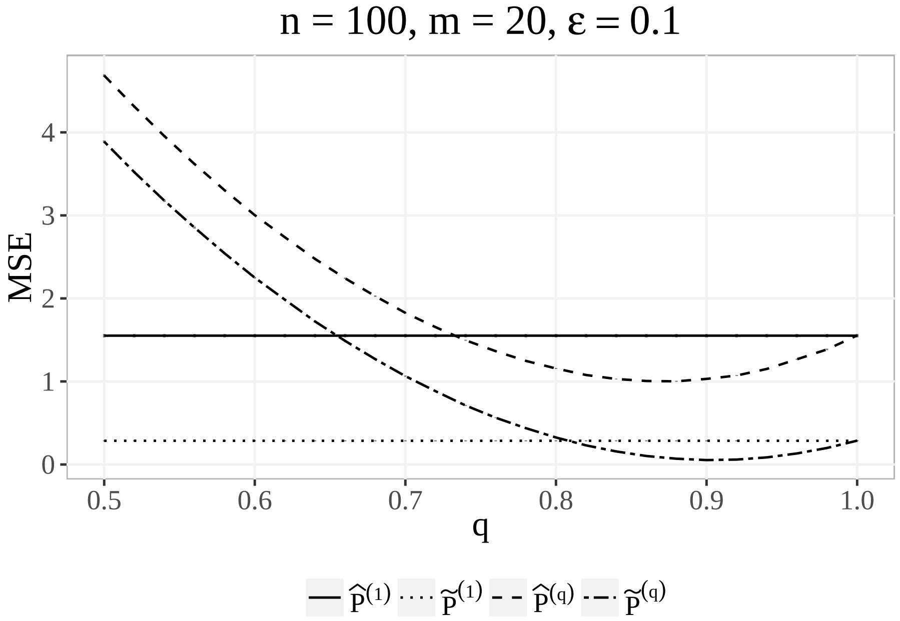

Figure 4: Mean squared error in average by varying the parameter $q$ in ML$q$E with fixed $n = 100$, $m = 20$ and $\epsilon = 0.1$ based on 1000 Monte Carlo replicates. Different curves represent the simulated MSE associated with the four different estimators. 1. ASE procedure takes advantage of the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

## 6.2 CoRR Graphs

We now compare the four estimators on a structural connectomic dataset. 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 experiment. The weight of an edge between two vertices represents the number of white-matter tracts connecting the corresponding two regions of the brain.

Generally, we do not expect the graphs to perfectly follow an RDPG, or even IEM. Before we perform our estimation, we will perform some exploratory analysis to check whether the data can reasonably be assumed to have approximate low-rank structure. Indeed, without at least approximately 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 histogram in the right panel of Figure 5. From the figure, we see that many eigenvalues are concentrated around zero. This exploration suggests that the information is mostly contained in the first few dimensions. Such approximate low-rank property provides an opportunity to win the bias-variance tradeoff by applying the ASE procedure.

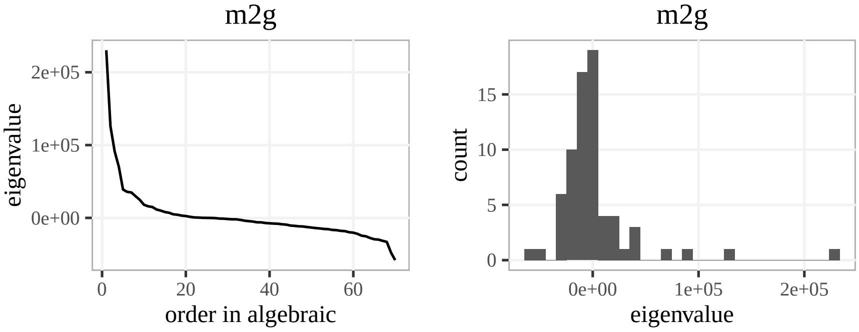

Figure 5: Screeplot and the histogram of the eigenvalues of the mean of 114 graphs based on m2g pipeline. The screeplot in the left panel shows the eigenvalues of the mean graph of all 114 graphs with diagonal augmentation in decreasing algebraic order for the Desikan atlas. The right panel shows the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

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 available 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 using it to calculate the MSE. Recently, Kiar et al. [2016] proposed a better pipeline ndmg2 compared to m2g. So the MLE derived from the 114 graphs in ndmg2 should be a relatively more accurate estimate of the actual probability matrix $P$ for the population. We use this as our surrogate for $P$ when calculating the MSE. However, such a $P$ 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 $\widehat{P}^{(1)}$ 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 of the ndmg2 data as $P$, the performance of ML$q$E-related estimators $\widehat{P}^{(q)}$ and $\widetilde{P}^{(q)}$ are underestimated. In summary, our approach to constructing a workable surrogate for $P$ relies on the availability of a better pipeline ndmg2, but is biased against both ASE-based and ML$q$E-based estimators; still, as we will see, ASE $\circ$ ML$q$E is our winner.

In this experiment, we build the four estimates based on the sample of size $m$ from the m2g dataset, while using the MLE of all 114 graphs from the ndmg2 dataset as the surrogate 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 $m$

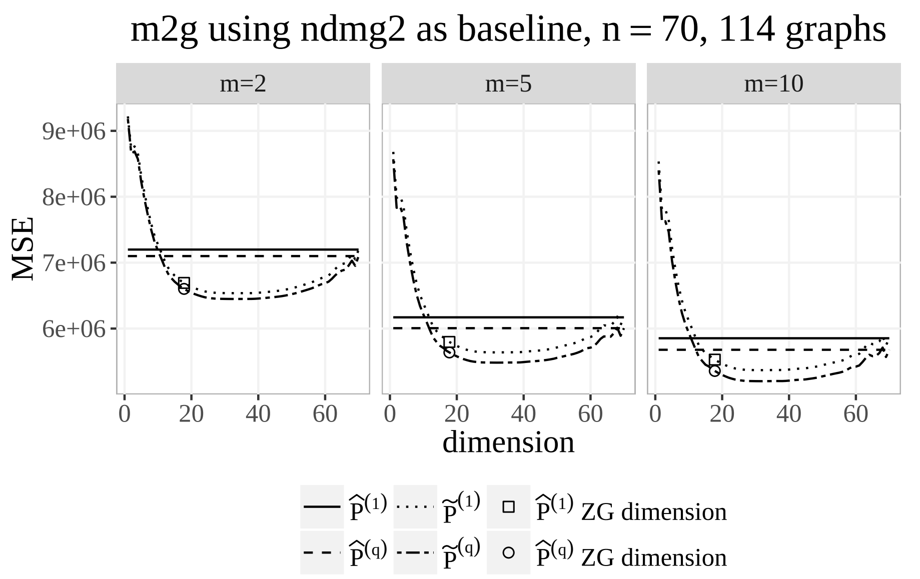

Figure 6: Comparison of MSE of the four estimators for the Desikan atlases at three sample sizes. The x-axis represents the dimensions to embed while y-axis is the MSE of each estimator. 1. MLE $\widehat{P}^{(1)}$ (horizontal solid line) vs MLqE $\widehat{P}^{(q)}$ (horizontal dotted line): ML$q$E outperforms 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 $\circ$ ML$q$E $\widetilde{P}^{(q)}$ (dashed dotted line): $\widetilde{P}^{(q)}$ wins the bias-variance tradeoff when being embedded into a proper dimension; 4. ASE $\circ$ ML$q$E $\widetilde{P}^{(q)}$ (dashed dotted line) vs ASE $\circ$ MLE $\widetilde{P}^{(1)}$ (dashed line): $\widetilde{P}^{(q)}$ is better, since it inherits the robustness from ML$q$E. The squares and circles represent the dimensions selected by the Zhu and Ghodsi method, which we see are reasonable choices. And more importantly, a wide range of dimensions lead to an improvement.

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 exploration. 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 the two low-rank estimators $\widetilde{P}^{(1)}$ and $\widetilde{P}^{(q)}$, we apply ASE into all possible dimensions, i.e. $d$ ranges from 1 to $n$. The MSE results are shown in Figure 6.

When $d$ is small, the ASE procedures underestimate the dimension and fail to capture 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 performs adequately for selecting a dimension in which to embed. More importantly, there is a wide range of dimensions which lead to a better performance when applying ASE. Although the $P$ we are estimating is actually a high-rank matrix, ASE procedures still win the bias-variance tradeoff and improve performance even in this unfavorable setting.

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

For all three sample sizes ($m = 2, 5, 10$), $\widetilde{P}^{(q)}$ estimates $P$ most accurately while the target – the surrogate $P$ – is biased in favor of the other three estimators. As such, we expect $\widetilde{P}^{(q)}$ to provide an even better estimate for the true but unknown $P$.

# 7 Discussion

In this work, our theoretical analysis is performed mostly in the weighted stochastic blockmodel setting. Note that the results can be extended to the random dot product graph instead of SBM, i.e. our estimator does not require the block structure. The reason is that the SBM assumption is just to ensure $\mathrm{rank}(E[\widehat{P}^{(q)}])$ has an upper bound invariant of the number of vertices under the contamination model. With this assumption on the rank, all the theory still holds in the RDPG setting. In practice, graphs are not exactly low rank. However, as shown in Figure 5 and Figure 6, our estimator still provides large improvement with approximate 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 theory based on the exponential distribution with ML$q$E for clarity. 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 that the MLE under the corresponding distribution is concentrated so that we obtain the matrix bounds we need. This generalization makes the theory more flexible and powerful.

Selecting a proper distortion parameter $q$ in ML$q$E is complicated, and we use a fixed $q = 0.9$ throughout this work without presenting a formal automatic selection methodology. In order to have improved performance, we might want to select a proper $q$ based on an adaptive method such as that proposed by **?**.

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 may still be applicable after running the graph matching algorithms of [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 such as ML$q$E may still provide improvement without contamination 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 investigation 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 contamination. Moreover, to have an 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 a 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 D3M program of the Defense Advanced Research Projects Agency (DARPA) administered through AFRL contract FA8750-17-2-0112; DARPA XDATA contract FA8750-12-2-0303; DARPA SIMPLEX 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 Appendix: 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 going to zero as the number of graphs $m$ goes to infinity.

In Section A.3, we analyze the properties of the ASE procedure. We first prove Theorem A.9, which provides an upper bound for the spectral 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 orthogonal transformation. With the extent of Lemma A.13, we then give a bound for the $2 \to \infty$-norm of $\widehat{Z} - ZW$, i.e. we bound $\max_i \|\widehat{Z}_i - WZ_i\|_2$ in Theorem A.14.

In Section A.4, we give a bound of the estimation error $\left|\widehat{Z}_i^\top \widehat{Z}_j - Z_i^\top Z_j\right|$ 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 purely for technical reasons and will not affect the estimation procedure in practice, which is discussed in details in Remark A.17. We then bound the expectation (Lemma A.18) and variance (Theorem A.19) of $\widetilde{P}^{(1)}$ by carefully choosing a truncation point $a$ and applying the above truncation argument. As a direct result, we obtain 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 manner to those in Section A.3 and Section A.4. However, since the ML$q$E estimator for a mixture distribution model does not have a closed form expression, we explore a relationship between MLE and ML$q$E to bound $\widetilde{P}^{(q)}$ and $\widehat{P}^{(q)}$; this technique could be of independent interest. Finally, in Section A.6, we compare the performance between $\widetilde{P}^{(q)}$ and $\widetilde{P}^{(1)}$.

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

Before presenting the proofs, we first define the following notion of "with high probability" that is used throughout this appendix.

**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 $\eta$ 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** *Let $X_1, \cdots, X_m \overset{iid}{\sim} \text{Exp}(P)$ with $m \geq 2$ and $E[X_1] = P$. Then with probability 1,*

- *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:** Let $x_1, \cdots, x_m$ be the observed values of $X_1, X_2, \ldots, X_m$. Then with probability 1, the $x_i$ are unique and $x_{(1)} = \min_i x_i > 0$. The MLE is

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

Let $g(\theta, x) = \sum_{i=1}^{m} e^{-\frac{(1-q)x_i}{\theta}}(x_i - \theta)$. Then the ML$q$ equation is $g(\theta, x) = 0$. Now let $l$ be the smallest index such that $x_{(1)} \leq \cdots \leq x_{(l)} \leq \bar{x} \leq x_{(l+1)} \leq \cdots$. 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 occurs with probability 0. Thus with probablity 1, $g(\theta, x) < 0$ for all $\theta \geq \bar{x}$.

Denote any solution to the ML$q$E equation as $\widehat{P}^{(q)}(x)$; we then have that

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

Thus there exists at least one solution to the ML$q$E equation. And since all solutions to the ML$q$E equation are in the interval $(x_{(1)}, \bar{x})$, we have t$\widehat{P}^{(q)}(x) \leq \widehat{P}^{(1)}(x)$. ■

**Lemma A.3** *Consider an exponential distribution model while the data is actually sampled from the contaminated model* $X, X_1, \cdots, X_m \overset{iid}{\sim} (1-\epsilon)\mathrm{Exp}(P) + \epsilon\mathrm{Exp}(C)$. *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 [?], $\theta(F)$ 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 ML$q$ 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 ML$q$ 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 $\theta_0$. Thus the ML$q$E 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 $\theta$ for all $x > 0$ and measurable function of $x$ at each $\theta$, we have the uniform convergence by Lemma 2.4 in [Newey and McFadden, 1994]. ■

**Lemma A.7** *$\widehat{P}_{ij}^{(q)} \xrightarrow{P} \theta(F_{ij})$ 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

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

for every $\epsilon > 0$. Combined with Lemma A.6, we know the ML$q$ is consistent based on Theorem 5.2.3 in [**?**]. ■

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

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

*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 ML$q$E 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 [**?**]. ■

## 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 $0 < P_{ij} \leq C_{ij} \leq R$ for some constant $R > 0$. For $0 < \epsilon < 1$, we define $m$ symmetric and hollow matrices as*

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

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

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

**Remark:** This is an extended version of Theorem 3.1 in [Oliveira, 2009].

Define $H^{(1)} = E[\widehat{P}^{(1)}] = (1-\epsilon)P + \epsilon C$, where $P = XX^\top$, $X \in \mathbb{R}^{n\times d}$, $C = YY^\top$, $Y \in \mathbb{R}^{n\times d'}$. Let $d^{(1)} = \text{rank}(H^{(1)})$ be the dimension in which we are going to embed $\widehat{P}^{(1)}$. Then we can define $H^{(1)} = ZZ^\top$ where $Z \in \mathbb{R}^{n\times d^{(1)}}$. Since $H^{(1)} = [\sqrt{1-\epsilon}X, \sqrt{\epsilon}Y][\sqrt{1-\epsilon}X, \sqrt{\epsilon}Y]^\top$, 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 $H = USU^\top = ZZ^\top$, where $Z = [Z_1, \cdots, Z_n]^\top$ is a $n$-by-$k$ matrix. Then our estimate for $Z$ up to rotation is $\widehat{Z} = \widehat{U}\widehat{S}^{1/2}$, where $\widehat{U}\widehat{S}\widehat{U}^\top$ is the rank-$k$ spectral decomposition of $|\widehat{P}| = (\widehat{P}^\top\widehat{P})^{1/2}$.

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

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

**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].

We ignore the proofs of the following results since they are similar to the proofs in [**?**].

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

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

*with high probability.*

We will denote the orthogonal matrix $W_1 W_2^\top$ 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.*

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

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

*with high probability.*

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

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

*with high probability.*

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

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

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

$$\begin{aligned}
\left|\widehat{Z}_i^\top \widehat{Z}_j - Z_i^\top Z_j\right| &= \left|\widehat{Z}_i^\top \widehat{Z}_j - \widehat{Z}_i^\top W Z_j + \widehat{Z}_i^\top W Z_j - (W Z_i)^\top W Z_j\right| \\
&\leq \left|\widehat{Z}_i^\top (\widehat{Z}_j - W Z_j) + (\widehat{Z}_i^\top - (W Z_i)^\top) W Z_j\right| \\
&\leq \|\widehat{Z}_i\|_2 \|\widehat{Z}_j - W Z_j\|_2 + \|Z_j\|_2 \|\widehat{Z}_i^\top - (W Z_i)^\top\|_2.
\end{aligned}$$

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

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

with high probability. ■

**Definition A.16** *Define $\widetilde{P}_{ij}^{(1)} = (\widehat{Z}_i^\top \widehat{Z}_j)_{\mathrm{tr}}$, our estimator for $P_{ij}$, to be a projection of $\widehat{Z}_i^\top \widehat{Z}_j$ 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.*

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

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

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

For the first term, we have

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

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,

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

Thus

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

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

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

■

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

**Proof:** By Lemma A.15,

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

Fix some $a > 0$, we have

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

For the first term, we have

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

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,

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

Thus,

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

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

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

■

**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 $\epsilon$, such that if $n > n_0$, then for all $\eta$ 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,

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

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 $\epsilon$ 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 $H^{(q)} = ZZ^\top$ 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 $H = USU^\top = ZZ^\top$, where $Z = [Z_1, \cdots, Z_n]^\top$ is a $n$-by-$k$ matrix. Then our estimate for $Z$ up to rotation is $\widehat{Z} = \widehat{U}\widehat{S}^{1/2}$, where $\widehat{U}\widehat{S}\widehat{U}^\top$ is the rank-$d$ spectral decomposition of $|\widehat{P}| = (\widehat{P}^\top\widehat{P})^{1/2}$.

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

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

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

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

*with high probability.*

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

We will denote the orthogonal matrix $W_1W_2^\top$ 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 $n$,*

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

*with high probability.*

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

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

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

**Definition A.28** *Define $\widetilde{P}_{ij}^{(q)} = (\widehat{Z}_i^\top \widehat{Z}_j)_{\mathrm{tr}}$, our estimator for $P_{ij}$, to be a projection of $\widehat{Z}_i^\top \widehat{Z}_j$ 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.*

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

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

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

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

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

For the second term, we have

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

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

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

■

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

**Proof:** By Lemma A.27,

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

Fix some $a > 0$, we have

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

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

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

For the second term, we have

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

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

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

■

**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 $\theta$ 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

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

■