Robust LLG

March 12, 2017

1 Background and Overview

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

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

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

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

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

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

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

In this work, we are going to improve the estimation procedure in a contaminated scenario. Before we introduce the contaminations, it is helpful to extend the unweighted graphs with Bernoulli distribution to weighted graphs with a general distribution f. All the models we mentioned above (IEM, RDPG, and SBM) could be extended naturally and are discussed in details in Section 2.1, Section 2.2, and Section 2.3 respectively.

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

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

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

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

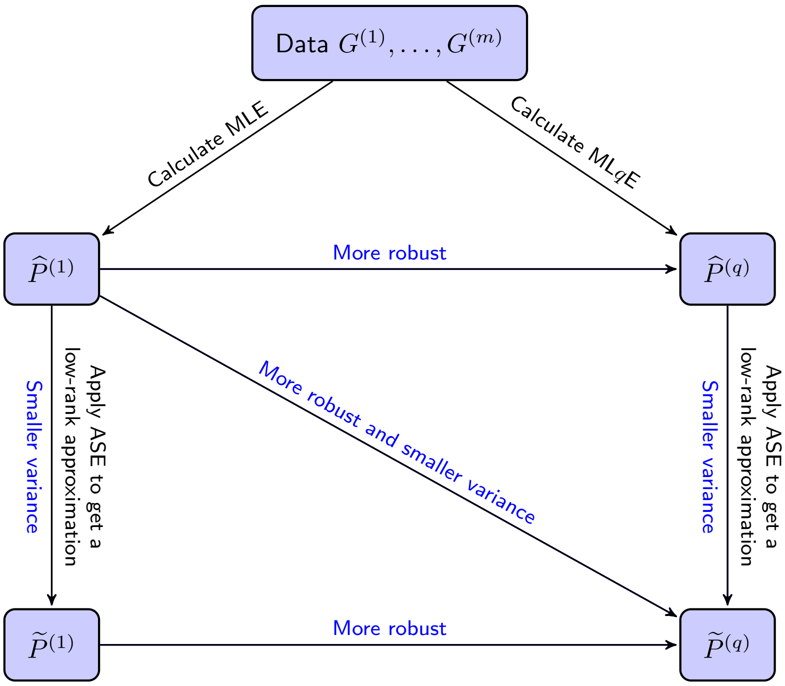

Figure 1: Roadmap among the data and four estimators.

2 Models

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

In this section, we present three nested models, the weighted independent edge model (WIEM) in Section 2.1, the weighted random dot product graph model (WRDPG) in Section 2.2, and the weighted stochastic blockmodel (WSBM) as a WRDPG in Section 2.3. Moreover, we introduce a contaminated model based on Section 2.3 in Section 2.4.

2.1 Weighted Independent Edge Model

We first extend the definition of independent edge model (IEM) [Bollobás et al., 2007] to the weighted independent edge model (WIEM) with respect to a one-parameter family $\mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}\}$. Denote the graph parameters as a matrix $P \in \Theta^{n \times n} \subset \mathbb{R}^{n \times n}$. Then under a WIEM, each edge between vertex i and vertex j ($i < j$ because of symmetry) is distributed from $f_{P_{ij}}$ independently.

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

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 each vertex i with a corresponding latent vector $X_i \in \mathbb{R}^d$. Conditioned on the latent vectors X_i and X_j, the edge weight between vertex i and vertex j is independent of all other edges and depends only on X_i and X_j through a link function.

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

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 as following: Let $X \in \mathbb{R}^{n \times d}$ be such that $X = [X_1, X_2, \dots, X_n]^\top$, 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 $f \in \mathcal{F}$ with parameter $X_i^\top X_j$ for all $i \neq j \in [n]$.*

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

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 property, where each vertex is assigned to one block and the connectivity between two vertices depends only on their respective block memberships.

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

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

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

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

- Here is an example of the WSBM. (Figure for simulation)

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

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

Generally in a gross error model, we observe good measurement $G^* \sim f_P \in \mathcal{F}$ most of the time, while there are a few wild values $G^{**} \sim h_C \in \mathcal{H}$ when the gross errors occur. Here P and C represent the respective parameter matrices of the two distribution families. As to the graphs, one way to generalize from the gross error model is to contaminate the entire graph with some small probability $\epsilon \in (0,1)$, that is $G \sim (1-\epsilon)f_P + \epsilon h_C$. However, since all the models we consider are subsets of the WIEM, it is more natural to consider the contaminations with respect to each edge, i.e. for $1 \leq i < j \leq n$, $G_{ij} \sim (1-\epsilon)f_{P_{ij}} + \epsilon h_{C_{ij}}$ 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 connectivity also follows the WSBM as a WRDPG. That is, the contamination distributions $h_{C_{ij}}$ are also from the same one-parameter family $\mathcal{F}$ as well as $f_{P_{ij}}$. For clarity, we will introduce the sampling procedure when the contamination has the same block structure. However, the parameter matrix C for the contamination does not need to have the same block structure as P in general.

To generate m graphs under this contamination model with known vertex correspondence, we first sample τ from the categorical distribution with parameter ρ and keep it fixed for all m graphs as in Section 2.3. Then m symmetric and hollow graphs $G^{(1)}, \ldots, G^{(m)}$ are sampled such that conditioning on τ, the adjacency matrices are distributed entry-wise independently as $A_{ij}^{(t)} \overset{ind}{\sim} (1-\epsilon)f_{P_{ij}} + \epsilon f_{C_{ij}}$ for each $1 \leq t \leq m$, $1 \leq i < j \leq n$. Here ϵ is the probability of an edge to be contaminated, P is the parameter matrix as in Section 2.3, and C is the parameter matrix for contaminations.

3 Estimators

Under any model introduced in Section 2, our goal is to estimate the parameter matrix P based on the m observations $A^{(1)}, \dots, A^{(m)}$. Especially when under the contamination model, although there are other parameters like ϵ and C, our goal is still to estimate the uncontaminated parameter matrix P. In this section, we present four estimators as in Figure 1, i.e. the standard entry-wise MLE $\widehat{P}^{(1)}$, the low-rank approximation of the entry-wise MLE $\widetilde{P}^{(1)}$, the entry-wise robust estimator MLqE $\widehat{P}^{(q)}$, and the low-rank approximation of the entry-wise MLqE $\widetilde{P}^{(q)}$. Since the observed graphs are symmetric and hollow with a symmetric parameter matrix of the model, we do not care about the estimate of the diagonal of P. However, the estimate itself should be at least symmetric.

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

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 $\widehat{P}^{(1)}$. The estimator $\widetilde{P}^{(1)}$ is similar to the estimator proposed by Chatterjee et al. [2015] with adjustment for the specific estimation task, including a different dimension selection technique discussed in Section 3.2.2 and a diagonal augmentation procedure discussed in Section 3.2.3. 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 such best approximation, first calculate the eigen-decomposition of the symmetric matrix $\widehat{P}^{(1)} = \widehat{U}\widehat{S}\widehat{U}^\top + \widetilde{U}\widetilde{S}\widetilde{U}^\top$, where $\widehat{S}$ is the diagonal matrix with the largest d eigenvalues of $\widehat{P}^{(1)}$, and $\widehat{U}$ has the corresponding eigenvectors as each column. Similarly, $\widetilde{S}$ is the diagonal matrix with non-increasing entries along the diagonal corresponding to the rest $n-d$ eigenvalues of $\widehat{P}^{(1)}$, and $\widetilde{U}$ has the columns given by the corresponding eigenvectors. The d-dimensional adjacency spectral embedding (ASE) of $\widehat{P}^{(1)}$ is given by $\widehat{X} = \widehat{U}\widehat{S}^{1/2} \in \mathbb{R}^{n \times d}$. Based on the ASE result, we have the best rank-d positive semi-definite approximation of $\widehat{P}^{(1)}$ to be $\widetilde{P}^{(1)} = \widehat{X}\widehat{X}^\top = \widehat{U}\widehat{S}\widehat{U}^\top$. In the RDPG setting, Sussman et al. [2014] proved that each row of $\widehat{X}$ can accurately estimate the the latent position for each vertex up to an orthogonal transformation. We will analyze its performance under the WRDPG setting in Section 4.

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

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

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

3.2.2 Dimension Selection

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

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

Although it is always challenge to select a proper dimension, based on the results of real data experiment in Section 6.2, a wide range of dimensions will lead to a fairly good results. Thus a proper dimension selection method can be applied directly without carefully tuning the parameter, which makes the estimator much more useful in practice.

3.2.3 Diagonal Augmentation

Since the graphs considered in this paper have no self-loops, all the adjacency matrices $A^{(t)}$ $(1 \leq t \leq m)$ are hollow, i.e. all diagonal entries are zeros. Thus the diagonal of the parameter matrix P does not matter since all off-diagonal entries are independent of them conditioned on the off-diagonal entries of P.

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

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

As suggested in [Tang et al., 2016], in this work we are going to combine both ideas by first using Marchette's row-averaging method (see Step 3 of Algorithm 2) and then another one-step Scheinerman's iterative method (see Step 6 of Algorithm 2).

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

```
Input: Symmetric adjacency matrices $A^{(1)}, A^{(2)}, \dots, A^{(m)}$, 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: Calculate the scaled degree matrix $D = \mathrm{diag}(\widehat{P}^{(1)}\mathbf{1})/(n-1)$;
3: Select the dimension $d$ based on the eigenvalues of $\widehat{P}^{(1)} + D$; (see Section 3.2.2)
4: Set $Q$ to $\mathrm{lowrank}_d(\widehat{P}^{(1)} + D)$; (see Algorithm 1)
5: Set $D'$ to $\mathrm{diag}(Q)$, the diagonal matrix with diagonal matching $Q$;
6: Set $Q'$ to $\mathrm{lowrank}_d(Q + D')$; (see Algorithm 1)
7: Set $\widetilde{P}^{(1)}$ with each entry $\widetilde{P}^{(1)}_{ij} = \max(Q'_{ij}, 0)$.
```

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

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

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

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

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

where $L_q(u) = (u^{1-q} - 1)/(1-q)$. Note that $L_q(u) \to \log(u)$ when $q \to 1$. Thus

MLqE is a generalization of MLE. Moreover, define

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

and

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

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

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

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

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

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 more or less in the entry-wise MLqE $\widehat{P}^{(q)}$. Thus, in order to take advantage of such low-rank structure as well as the robustness, we apply the similar idea here as in building $\widetilde{P}^{(1)}$, i.e. enforce a low-rank approximation on the entry-wise MLqE matrix $\widehat{P}^{(q)}$ to get $\widetilde{P}^{(q)}$. As in Algorithm 2, we apply the same dimension selection method and diagonal augmentation procedure. The only change is to substitute $\widehat{P}^{(1)}$ by $\widehat{P}^{(q)}$. The details of the algorithm is shown in Algorithm 3.

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

```
Input: Symmetric adjacency matrices $A^{(1)}, A^{(2)}, \ldots, A^{(m)}$, with each $A^{(t)} \in \mathbb{R}^{n \times n}$
Output: Estimate $\widetilde{P}^{(q)} \in \mathbb{R}^{n \times n}$
Calculate the entry-wise MLqE $\widehat{P}^{(q)}$;
Calculate the scaled degree matrix $D = \mathrm{diag}(\widehat{P}^{(q)}\mathbf{1})/(n-1)$;
Select the dimension $d$ based on the eigenvalues of $\widehat{P}^{(q)} + D$; (see Section 3.2.2)
Set $Q$ to $\mathrm{lowrank}_d(\widehat{P}^{(q)} + D)$; (see Algorithm 1)
Set $D'$ to $\mathrm{diag}(Q)$, the diagonal matrix with diagonal matching $Q$;
Set $Q'$ to $\mathrm{lowrank}_d(Q + D')$; (see Algorithm 1)
Set $\widetilde{P}^{(q)}$ with each entry $\widetilde{P}^{(q)}_{ij} = \max(Q'_{ij}, 0)$.
```

4 Theoretical Results

In this section, for illustrative purpose, we are going to present theoretical results when the contamination model introduced in Section 2.4 is with respect to exponential distributions. That is $\mathcal{F} = \{f_\theta(x) = \frac{1}{\theta} e^{-x/\theta}, \theta \in [0, R] \subset \mathbb{R}\}$, where $R > 0$ is a constant. The results can be extended to a general situation with proper assumptions, which will be discussed in Section 5.

For clarity, we restate the model settings discussed in Section 2.4. Consider the SBM with parameter B and ρ. First sample the block membership τ from the categorical distribution with parameter ρ and keep it fixed for all m graphs. Conditioned on this τ we sampled, the probability matrix P then satisfies $P_{ij} = B_{\tau_i, \tau_j}$. In this section, we assume the contamination has the same block membership τ, thus the contamination matrix $C \in \mathbb{R}^{n \times n}$ has the same block structure as P. Note that this is not necessary for the result. Different block structure can lead to the same result since the rank is still finite. Denote ϵ as the probability of an edge to be contaminated. Then m symmetric graphs $G^{(1)}, \ldots, G^{(m)}$ are sampled such that conditioning on τ, the adjacency matrices are distributed entry-wise independently as $A_{ij}^{(t)} \overset{ind}{\sim} (1-\epsilon) f_{P_{ij}} + \epsilon f_{C_{ij}}$ for each $1 \le t \le m$, $1 \le i < j \le n$.

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

We only present the most important results in this section. The proofs with more results are 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 MLqE $\widehat{P}^{(q)}$. Without using the graphs structure, the asymptotic results for these two estimators are in terms of the number of graphs m, not the number of vertices n within each graph.

Lemma 4.1 *For any $0 < q < 1$, there exists $C_0(P_{ij}, \epsilon, q) > 0$ such that under the contaminated model with $C > C_0(P_{ij}, \epsilon, q)$,*

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

for $1 \le i, j \le n$ and $i \ne j$.

Lemma 4.1 shows that the entry-wise MLqE $\widehat{P}^{(q)}$ has smaller bias for estimating P asymptotically compared to the entry-wise MLE $\widehat{P}^{(1)}$ under proper conditions. Although we put restrictions on the parameter matrix C in the assumption, to have the current result about asymptotic bias, we only need to satisfy the inequality $\epsilon(C_{ij} - P_{ij}) > (1-q)P_{ij}$. This condition actually requires the contamination of the model is large enough (either large contamination parameter matrix, or more likely to encounter an outlier). From a different perspective, it also requires $\widehat{P}^{(q)}$ to be robust enough with respect to the contamination. Thus besides the current condition for C, equivalently, we can also replace it by the assumption of a large enough ϵ or a small enough q.

Lemma 4.2
$$\lim_{m\to\infty} \mathrm{Var}(\widehat{P}_{ij}^{(1)}) = \lim_{m\to\infty} \mathrm{Var}(\widehat{P}_{ij}^{(q)}) = 0,$$
for $1 \leq i, j \leq n$.

By this lemma, both estimators have asymptotic variances equal zero as the number of graphs m goes to infinity. Under the mixture model, the result for MLE follows the central limit theorem immediately, while the MLqE result is not trivial. Our proof is based on the minimum contrast estimates (more details later).

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

4.2 $\widehat{P}^{(1)}$ vs. $\widetilde{P}^{(1)}$

Then we 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 $\widetilde{P}^{(1)}$. Firstly, we show that the two estimators have the same entry-wise asymptotic bias under proper conditions in the following lemma.

Lemma 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.*

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

To prove the the lemma, we first give a bound on $\|\widehat{P}^{(1)} - E[\widehat{P}^{(1)}]\|_2$ by the matrix Bernstein inequality [Tropp, 2012]. Since we are under the WRDPG setting, $E[\widehat{P}^{(1)}]$ still has finite rank, which is the target embedding dimension. Then we approximate $U^T\widehat{U}$ by an orthogonal matrix $W^\star$ based on Davis-Kahan theorem [Davis and Kahan, 1970], where U and $\widehat{U}$ are the eigen-spaces for $E[\widehat{P}^{(1)}]$ and $\widehat{P}^{(1)}$ respectively. Then we can analyze the bounds much more easily by interchanging $U^T\widehat{U}$ in the matrix multiplications. As a result, we can get the bound between the estimates of the latent positions and the true latent positions up to an orthogonal transformation in terms of $2 \to \infty$ norm. Lemma 4.3 then follows by selecting the parameter carefully in the analysis.

Lemma 4.3 says that when m is a constant, or m is going to infinity with order $m = O(n^b)$ for any $b > 0$, the ASE procedure applied to $\widehat{P}^{(1)}$ will not affect the asymptotic bias for estimating P. In this case, the asymptotic relative efficiency (ARE) [Serfling, 2011] provides a good way to compare these two estimators. The original definition is based on unbiased estimators. Here we extend it a little bit such that it can measure the performance between two estimators with the same asymptotic bias.

Definition 4.4 *For any parameter θ of a distribution f, and for estimators $\widehat{\theta}^{(1)}$ and $\widehat{\theta}^{(2)}$ approximately $N(\theta', V_1(f)/n)$ and $N(\theta', V_2(f)/n)$ respectively, the ARE of $\widehat{\theta}^{(2)}$ to $\widehat{\theta}^{(1)}$ is given by*

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

In our situation, we can compare the performance between $\widetilde{P}^{(1)}$ and $\widehat{P}^{(1)}$ based on the entry-wise ARE, which can be written as $\mathrm{ARE}(\widetilde{P}_{ij}^{(1)}, \widehat{P}_{ij}^{(1)}) = \lim_{n\to\infty} \mathrm{Var}(\widetilde{P}_{ij}^{(1)})/\mathrm{Var}(\widehat{P}_{ij}^{(1)})$. The order of $\mathrm{Var}(\widetilde{P}_{ij}^{(1)})$ is analyzed in the following theorem.

Theorem 4.5 *Assuming that $m = O(n^b)$ for any $b > 0$, then* $\mathrm{Var}(\widetilde{P}_{ij}^{(1)}) = O(m^{-1}n^{-1}(\log n)^3)$.

The tools we use to prove this theorem is the same as for Lemma 4.3. Combined with the fact that $\mathrm{Var}(\widehat{P}_{ij}^{(1)}) = O(m^{-1})$, we have the ARE results directly as following:

Theorem 4.6 *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

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

Theorem 4.6 tells us that whenever $m = O(n^b)$ for any $b > 0$, i.e. m is fixed or it grows not faster than any polynomial with respect to n, the order of the ARE is $O(n^{-1}(\log n)^3)$, which goes to 0 when $n \to \infty$. An interesting fact here is that this bound of the ARE 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.

Remark 4.7 *To prove the upper bound of the variance as in Theorem 4.5, we slightly modify our estimator $\widetilde{P}^{(1)}$ to be $\min(\widetilde{P}^{(1)}, \alpha\widehat{P}^{(1)})$ entry-wise for any constant $\alpha > 0$, i.e. truncate our estimator $\widetilde{P}_{ij}^{(1)}$ whenever it is larger than $\alpha\widehat{P}^{(1)}$. Note that Theorem 4.3 still holds with this modified estimator. Since the constant α can be arbitrarily large, this modification is just for technical reason in the proof. Importantly, we do not have this truncation step in the algorithms, or any simulation/real data results in Section 6.*

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

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

Lemma 4.8 *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.*

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

The proof for Lemma 4.8 is almost the same as the proof for Lemma 4.3. But unlike the results for MLE, we cannot get the term $m^{-1/2}$ in the results due to the structure of maximum Lq likelihood equation. Although it does not affect the result for asymptotic bias as in Lemma 4.8, the order of the variance of MLqE is missing m^{-1} compared to MLE as in the following theorem. Moreover, later we will see that this will cause a slight difference in the comparison.

Theorem 4.9 *Assuming that $m = O(n^b)$ for any $b > 0$, then $\mathrm{Var}(\widetilde{P}^{(q)}) = O(n^{-1}(\log n)^3)$.*

The proofs are similar to Theorem 4.5. Combined with the fact that $\mathrm{Var}(\widehat{P}_{ij}^{(q)}) = O(m^{-1})$, we have the ARE results directly as following:

Theorem 4.10 *Assuming that $m = O(n^b)$ for any $b > 0$, then for $1 \le i, j \le n$ and $i \neq j$,*

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

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

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

Theorem 4.10 tells us that whenever $m = O(n^b)$ for any $b > 0$, i.e. m is fixed or it grows not faster than polynomial with respect to n, the order of the ARE is $O(mn^{-1}(\log n)^3)$. Specifically, when m is fixed, the order of the ARE is $O(n^{-1}(\log n)^3)$, which will goes to 0 as $n \to \infty$. Even if m also increases as n increases, as long as it grows in the order of $o(n(\log n)^{-3})$, the ARE still goes to 0.

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

Remark 4.11 *Similar to Remark 4.7, we prove the theorems above based on the modified estimator $\min(\widetilde{P}^{(q)}, \alpha\widehat{P}^{(q)})$ entry-wise for any constant $\alpha > 0$. This modification is simply for the proof and is not included in algorithms or empirical results.*

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

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

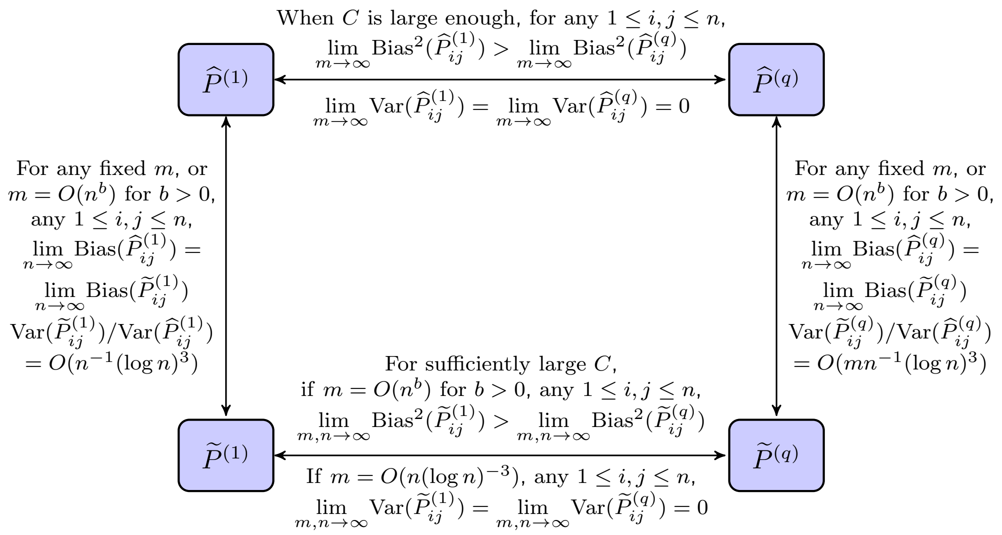
$\widehat{P}^{(1)}$
$\widehat{P}^{(q)}$
$\widetilde{P}^{(1)}$
$\widetilde{P}^{(q)}$
When C is large enough, for any $1 \le i, j \le n$,
$\lim_{m\to\infty} \mathrm{Bias}^2(\widehat{P}^{(1)}_{ij}) > \lim_{m\to\infty} \mathrm{Bias}^2(\widehat{P}^{(q)}_{ij})$
$\lim_{m\to\infty} \mathrm{Var}(\widehat{P}^{(1)}_{ij})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

Figure 2: Relationship among four estimators.

Theorem 4.13 *For sufficiently large C and any $1 \le i, j \le n$, if $m = O(n(\log n)^{-3})$, then*

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

Theorem 4.12 is a direct result of Lemma 4.1, Lemma 4.3, and Lemma 4.8, while Theorem 4.13 is based on Lemma 4.2, Theorem 4.6, and Theorem 4.10.

So $\widetilde{P}^{(q)}$ inherits the robustness from the entry-wise MLqE $\widehat{P}^{(q)}$ and has a smaller asymptotic bias compared to $\widetilde{P}^{(1)}$ while both estimates have variance goes to 0 as $m \to \infty$.

4.5 Summary

We summarize all the comparison in this section and plot the relationship among four estimators in Figure 2. From the upper layer to the low layer, we apply ASE to construct low-rank approximations which keep the asymptotic bias and reduce the asymptotic variance. So as long as the number of vertices n is large enough, ASE will improve the performance with a proper number of graphs m. From the left part to the right part, we underweight the outliers to construct robust estimators. So with enough contaminations, whenever the number of graphs m is large enough, the bias term which dominates the MSE will be improved.

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

5 Extensions

Although in Section 4, we only present the results under exponential distributions, the results can be generalized to a broader class of distribution families, and even a different entry-wise robust estimator other than MLqE with the following conditions:

1. Let $A_{ij} \overset{ind}{\sim} (1-\epsilon) f_{P_{ij}} + \epsilon f_{C_{ij}}$, then $E[(A_{ij} - E[\widehat{P}^{(1)}_{ij}])^k] \leq \text{const} \cdot k!$, where $\widehat{P}^{(1)}$ is the entry-wise MLE as defined before;
   This is to ensure the that observations will not deviate from the expectation too far away, such that the concentration inequality can apply.

2. There exists $C_0(P_{ij}, \epsilon) > 0$ such that under the contaminated model with $C > C_0(P_{ij}, \epsilon)$,
   $$\lim_{m \to \infty} \left| E[\widehat{P}_{ij}] - P_{ij} \right| < \lim_{m \to \infty} \left| E[\widehat{P}^{(1)}_{ij}] - P_{ij} \right| ;$$
   This condition is discussed in Section 4.1. It requires the contamination of the model to be large enough (a restriction on the distribution) and $\widehat{P}$ to be robust enough with respect to the contamination (a condition on the estimator).

3. $\widehat{P}_{ij} \leq \text{const} \cdot \widehat{P}^{(1)}_{ij}$; (This might be generalized to with high probability later)
   Since we use the results of $\widehat{P}^{(1)}$ to bound $\widehat{P}^{(q)}$, the proof can apply directly with this condition for an arbitrary $\widehat{P}$.

4. $\text{Var}(\widehat{P}_{ij}) = O(m^{-1})$, where m is the number of observations.
   We will get exactly the same results as in Section 4. However, even if the variance of the new estimator is not of order $O(m^{-1})$, we will get similar results with a different term related to m.

6 Empirical Results

6.1 Simulation

In this section, we will illustrate the theoretical results of four estimators discussed in Section 4 via various Monte Carlo simulation experiments in an idealized setting.

6.1.1 Simulation Setting

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

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

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

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

The contamination probability $\epsilon = 0.1$ if not specified. However, we will vary ϵ in the following simulation. With these parameters specified, we can sample graphs according to Section 2.4.

When calculating the two low-rank estimators $\widetilde{P}^{(1)}$ and $\widetilde{P}^{(q)}$, we skip the dimension selection step in Algorithm 2 and Algorithm 3. Instead embed the graphs into the true dimension $d = \text{rank}(B) = 2$. Note that the simplified setting in this simulation is primarily for interpretability.

6.1.2 Simulation Results

In order to see how the performance of the four estimators varies with respect to the contaminations, we first run 1000 Monte Carlo replicates based on the contaminated SBM specified in Section 6.1.1 with a fixed number of vertices $n = 100$ and a fixed number of graphs $m = 20$ while varying the contamination probability ϵ from 0 to 0.4. Given each sample, four estimators can be calculated following Algorithm 2 and Algorithm 3. Since we are not focusing on how to select the parameter q in the MLqE estimator, we are going to use a fixed $q = 0.9$ throughout this paper. Then the MSE of each estimator can be estimated since the probability matrix P is known in this simulation.

We plot the MSE in average in Figure 3. Different colors represent the simulated MSE associated with four different estimators. Firstly, MLE $\widehat{P}^{(1)}$ outperforms a little bit when there is no contamination (i.e. $\epsilon = 0$), but it degrades dramatically when contamination increases. On the contrary, the MLqE $\widehat{P}^{(q)}$ lose a little efficiency when the contamination is small, but shows its robustness under a large contamination. Even with a relative small number of vertices $n = 100$, the ASE procedure which takes advantage of the low rank structure already helps improve the performance of $\widehat{P}^{(1)}$ and let $\widetilde{P}^{(1)}$ win the bias-variance tradeoff. Since the MLqE $\widehat{P}^{(q)}$ preserves the low rank structure of the original graph more or less, the ASE procedure also helps and makes $\widetilde{P}^{(q)}$ a better estimate. Although both $\widetilde{P}^{(q)}$ and $\widetilde{P}^{(1)}$ take advantage of the low-rank structure and has reduced variances, $\widetilde{P}^{(q)}$ constructed based on MLqE inherits the robustness from MLqE in addition. So when the contamination is large enough, $\widetilde{P}^{(q)}$ outperforms $\widetilde{P}^{(1)}$ and degrades slower.

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

By comparing the performance of the four estimators based on different setting, we demonstrate the theoretical results in Section 4.

6.2 CoRR Graphs

Generally, the graphs in real life may not perfectly follow an RDPG, or not even follow IEM. But we are still interested in how the four estimators perform under those situations. In this work, we examine the structural connectomic data and compare the four estimators. 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

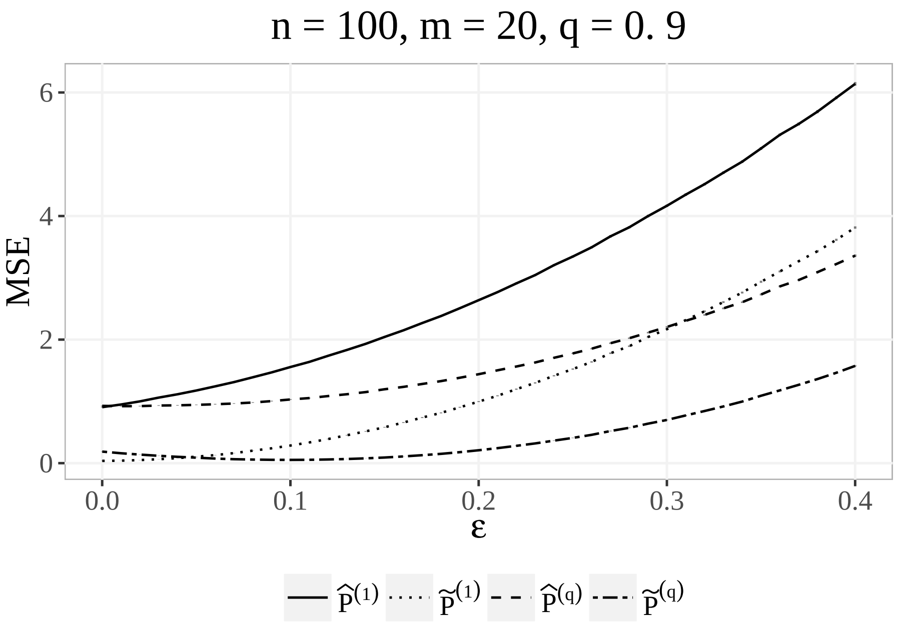

Figure 3: Mean squared error in average by varying contamination ratio ϵ with fixed $n = 100$ and $m = 20$ based on 1000 Monte Carlo replicates. And we use $q = 0.9$ when applying MLqE. Different colors represent the simulated MSE associated with four different estimators. **1. MLE $\widehat{P}^{(1)}$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

the m2g pipeline described in [Kiar et al., 2016]. The vertices of the graphs represent different regions in the brain defined according to an atlas. We used the Desikan atlas with 70 vertices in this expereiment. The weight of an edge between two vertices represents the number of white-matter tract connecting the corresponding two regions of the brain.

Before we calculate those estimators, we want to check whether the dataset has a low-rank property. Without that, the ASE procedure cannot win the bias-variance tradeoff because of a high bias. In the left panel of Figure 5, we plot the eigenvalues of the mean graph of all 114 graphs (with diagonal augmentation) in decreasing algebraic order for the Desikan atlases based on the m2g pipeline. The eigenvalues first decrease dramatically and then stay around 0 for a large range of dimensions. In addition, we also plot the histograms in the right panel of Figure 5. From the figures we can see many eigenvalues are around zero. So the information is mostly contained in the first few dimensions. Such quasi

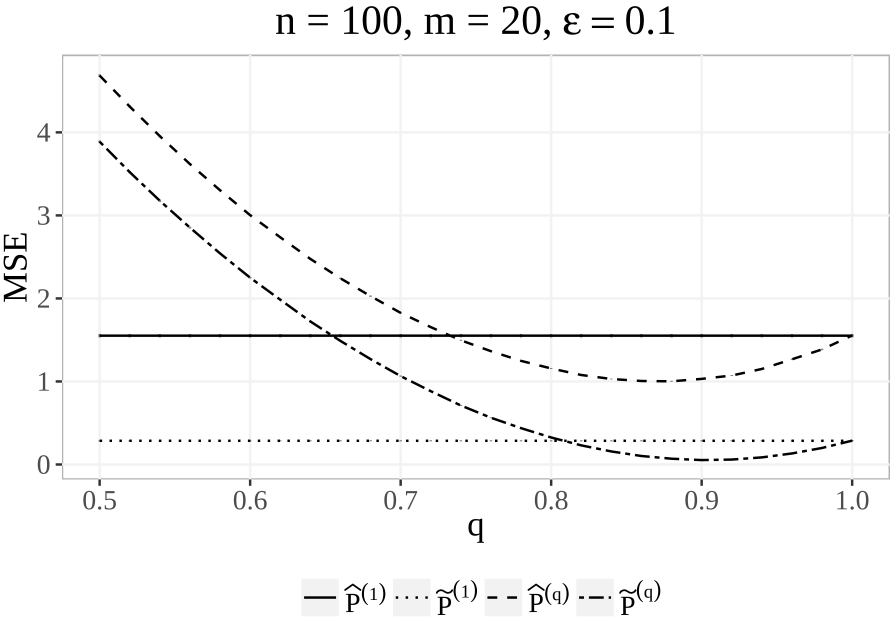

Figure 4: Mean squared error in average by varying the parameter q in MLqE with fixed $n = 100$, $m = 20$ and $\epsilon = 0.1$ based on 1000 Monte Carlo replicates. Different types of lines represent the simulated MSE associated with four different estimators. 1. ASE procedure takes advantage of the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

low-rank property provides an opportunity to win the bias-variance tradeoff by applying ASE procedure.

To compare the four estimators, we need to calculate the MSE, which requires the true parameter matrix P. However, unlike simulation experiment in Section 6.1, P is definitely not obtainable in practice since the 114 graphs are also a sample from the population. So finding a good estimate for P and use it to calculate the MSE is a feasible way in this experiment. Recently, Kiar et al. [2016] proposed a better pipeline ndmg2 compared to m2g. Then the MLE derived from the 114 graphs in ndmg2 should be a relative more accurate estimate of the actual probability matrix P for the population. And we are going to use this as P when calculating the MSE. Such P generally has full rank, which breaks the low-rank assumptions. So this setting makes it hard for $\widetilde{P}^{(1)}$ and $\widetilde{P}^{(q)}$ to improve and is favorable to the $\widehat{P}^{(1)}$ and $\widehat{P}^{(q)}$. Moreover, it is still possible that the 114 graphs from ndmg2 contain outliers. Thus by using the MLE as P, the performance of MLqE related estimators $\widehat{P}^{(q)}$ and $\widetilde{P}^{(q)}$ are underestimated.

In this experiment, we build the four estimates based on the samples with size m from the m2g dataset, while using the MLE of all 114 graphs from the

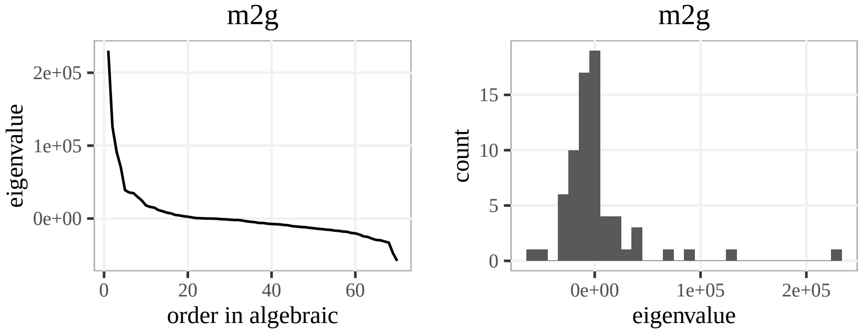

Figure 5: **Screeplot and the histogram of the eigenvalues of the mean of 114 graphs based on m2g pipeline.** The screeplot in the left panel shows the eigenvalues of the mean graph of all 114 graphs with diagonal augmentation in decreasing algebraic order for the Desikan atlas. The right panel shows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

ndmg2 dataset as the probability matrix P. We run 100 simulations on this dataset for different sample sizes $m = 2, 5, 10$. Specifically, in each Monte Carlo replicate, we sample m graphs out of the 114 from the m2g dataset and compute the four estimates based on the m sampled graphs. As mentioned before, q is set to be 0.9 without further exploiting. We then compare these estimates to the MLE of all 114 graphs in the ndmg2 dataset. For those two low-rank estimators $\widetilde{P}^{(1)}$ and $\widetilde{P}^{(q)}$, we apply ASE for all possible dimensions, i.e. d ranges from 1 to n. The MSE results are shown in Figure 6.

When d is small, ASE procedure underestimates the dimension and fails to get important information, which leads to poor performance. In this work, we use Zhu and Ghodsi's method discussed in Section 3.2.2 to select the dimension d. We denote the selected dimensions by square and circle in the figure. We can see the algorithm does a pretty good job for selecting the dimension to embed. More importantly, there is a wide range of dimensions which could lead to a better performance when applying ASE. Although the P we are estimating is actually a high-rank matrix, ASE procedure still wins the bias-variance tradeoff and improves the performance while being suppressed in this setting.

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

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

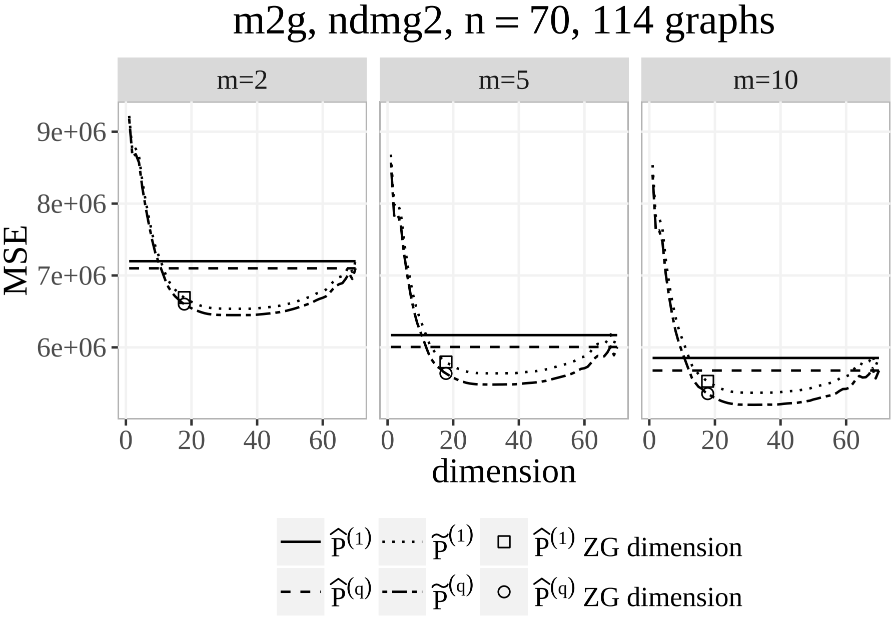

Figure 6: **Comparison of MSE of the four estimators for the Desikan atlases at three sample sizes.** The x-axis represents the dimensions to embed while y-axis is the MSE of each estimator. **1. MLE $\widehat{P}^{(1)}$ (horizontal solid line) vs MLqE $\widehat{P}^{(q)}$ (horizontal dotted line):** MLqE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

7 Discussion

Acknowledgments

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

References

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.

Sourav Chatterjee et al. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

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.

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.

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

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

Yichen Qin and Carey E Priebe. Maximum l q-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.

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.

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

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

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

Stephen J Young and Edward R Scheinerman. Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 138–149. Springer, 2007.

Mu Zhu and Ali Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930, 2006.

A Proofs for Theory Results

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

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

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

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

Proof: The MLE is

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

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

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

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

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

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

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

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

Lemma A.2 *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 at least one solution $\theta(F)$ of the population version of MLq equation, i.e. $E_F[e^{-\frac{(1-q)X}{\theta(F)}}(X - \theta(F))] = 0$, such that $\theta(F) < E_F[\bar{X}] = (1-\epsilon)P + \epsilon C$. So we can define $\theta(F_{ij})$ to be the largest root which is less than $E_F[\bar{X}]$.*

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

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

i.e.

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

Define $h(\theta) = (C(1-q)+\theta)^2(1-\epsilon)(Pq-\theta) - (P(1-q)+\theta)^2\epsilon(\theta - Cq)$. Then $\lim_{\theta\to\infty} h(\theta) = -\infty$, $h(0) > 0$, and $h(Cq) < 0$. Consider q as the variable and solve the equation $h(E[\bar{X}]) = 0$, we have three roots and one of them is $q = 1$ obviously. The other two roots are

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

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

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

For the first part,

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

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

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

Then it is sufficient to show that

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

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

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

Lemma A.3 (Lemma 4.1) *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.2, $\theta(F)$ 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 **??**, we need

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

Based on Lemma A.1, we need

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

i.e.

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

■