Outline for Robust LLG

December 4, 2016

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 is one of the most important tasks.

- In the most basic setting, i.e. undirected graphs with each edge distributed Bernoulli independently, the parameters are the probabilities of the existence of edges between all pair of vertices. Then the element-wise maximum likelihood estimate, which happens to be the sample mean in this situation, is the uniformly minimum-variance unbiased estimator if we only consider the independent edge graph model [3] without taking any graph structure into account. However, it does not perform very well especially when we only have a few observations, which is likely the case in real world.

- One of the most important structures about the networks is the community structure in which vertices are clustered into different communities such that vertices of the same community behave similarly. The stochastic blockmodel (SBM) [7] captures such structural property and is widely used in modeling networks. Meanwhile, the latent positions model (LPM) [6], a much more general model compared to SBM, proposes a way to parameterize the graph structure by latent positions associated with each vertex. However, the random dot product graph (RDPG) [16, 11] which is a special case of LPM stays in between and motivates our estimator.

- The Law of Large Graphs [15] considers a low-rank approximation of the sample mean graph as an estimator motivated by the RDPG model and proves that in the basic Bernoulli setting under SBM, the new estimator outperforms the element-wise MLE since it decreases the overall variance dramatically compare to the naive entry-wise sample mean by biasing towards the low-rank structure.

- While the Law of Large Graphs considers an estimator which captures the low-rank structure and demonstrates its improvement over the entry-wise MLE for the unweighted graphs with Bernoulli distribution, a more generalized setting is always preferred.

- Firstly, we extend the unweighted graphs with Bernoulli distribution to weighted graphs with a general distribution F. Then all the models we mentioned above could be modified naturally and are discussed in details in Section 2.1, 2.2 and 2.3.

- Also in the Law of Large Graphs paper, it is assumed that the adjacency matrix is observed without contamination, however in practice there will be noise in the observed graph. In this case, a contamination model, e.g. gross error model considered in this work [9, 2], is always preferred. 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 ϵ, that is $G \sim (1-\epsilon)F_P + \epsilon H_C$. However, since we are under the independent edge model, it is more natural to consider the contaminations on each edge, i.e. for $1 \le i < j \le n$, $G_{ij} \sim (1-\epsilon)F_{P_{ij}} + \epsilon H_{C_{ij}}$ with $F \in \mathcal{F}$ and $G \in \mathcal{G}$, where both $\mathcal{F}$ and $\mathcal{G}$ are one-parameter distribution families.

- Under the contamination model, although we observe G instead of G^*, estimating the parameters P_{ij} $(1 \le i < j \le n)$ of $F_{P_{ij}}$ in $\mathcal{F}$ is still our goal. We first proved that the estimator based on ASE ($\widetilde{P}^{(1)}$ in Figure 1) proposed in LLG is still better than entry-wise MLE ($\widehat{P}^{(1)}$ in Figure 1) in MSE when the observations are contaminated under proper conditions.

- Furthermore, with contaminations, it is more preferable to use robust methods, like MLqE [5, 12] considered in this paper. We proved 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 trade-off by taking advantage of the low-rank structure.

2 Models

- For this work, we are in the scenario where m weighted graphs each with n vertices are given in the adjacency matrices form $\{A^{(t)}\}(t = 1, \ldots, m)$. The graphs we consider in this work 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 throughout different graphs, so that vertex i of graph t_1 corresponds to vertex i of graph t_2 for any i, t_1, t_2.

- In this section, we present three nested models, the weighted independent edge model in Section 2.1, the weighted random dot product graph model 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.

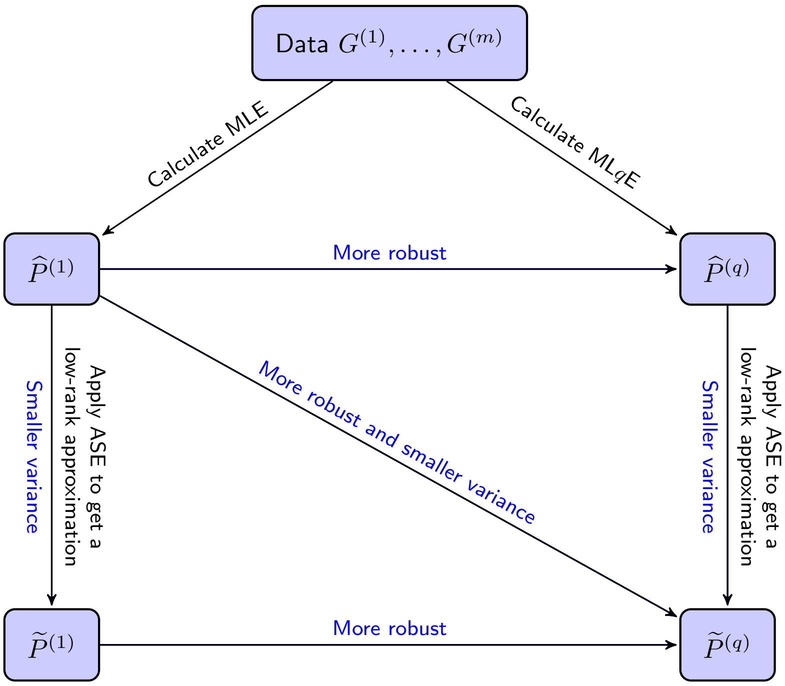


Figure 1: Roadmap among the data and four estimators.

2.1 Weighted Independent Edge Model

- We first extend the definition of independent edge model (IEM) [3] to the weighted independent edge model (WIEM) corresponding to a one-parameter family $\mathcal{F} = \{F_\theta : \theta \in \Theta \subset \mathbb{R}\}$. Denote the graph parameter 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.

- A simple example here is a WIEM for binary graphs with respect to $\mathcal{F}$ to be Bernoulli with a symmetric and hollow parameter 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 positions model proposed by Hoff et al. [6] 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 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. This can easily be generalized to the weighted version easily.

- 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 [16, 11]. In the following, we give a definition of the weighted random dot product graph (WRDPG) as a generalization of the weighted latent positions model.

  Definition 2.1 *Consider a collection of one-parameter distributions $\mathcal{F} = \{F_\theta, \theta \in \Theta \subset \mathbb{R}\}$. Then the weighted random dot product graph model (WRDPG) with respect to $\mathcal{F}$ is defined as following: Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ be such that $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T$, where $X_i \in \mathbb{R}^d$ for all $i \in [n]$. The matrix $\boldsymbol{X}$ is random and satisfies $\mathbb{P}[\langle X_i, X_j\rangle \in \Theta] = 1$ for all $i, j \in [n]$. Conditioned on $\boldsymbol{X}$, the entries of the adjacency matrix $\boldsymbol{A}$ are independent and $\boldsymbol{A}_{ij}$ is a random variable with distribution $F \in \mathcal{F}$ with parameter $\langle X_i, X_j\rangle$ 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) [7] 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 $B \in \Theta^{K\times K} \subset \mathbb{R}^{K\times K}$, and the block assignment vector $\tau \in [K]^n$, where $\tau_i = k$ represents 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 easily be generalized to the weighted stochastic blockmodel (WSBM), by substituting the Bernoulli distribution to a general distribution family $\mathcal{F}$.

- In order to consider WSBM as a WRDPG, the matrix B needs to be positive semi-definite. From now on, we will note the sub-model of WSBM with positive semi-definite B as the WSBM.

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

- In order to generate m graphs under this model with known vertex correspondence, we first sample τ from the categorical distribution with parameter ρ and keep it fixed for all m graphs. 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 $A_{ij}^{(t)} \overset{ind}{\sim} F_{B_{\tau_i, \tau_j}} = F_{P_{ij}}$ for each $1 \leq t \leq m$, $1 \leq i < j \leq n$. Here $P \in \mathbb{R}^{n \times n}$ is the actual parameter matrix for each pair of vertices with the same block structure as B.

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

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

- In practice, we always get noise in the observations, which deviates from our general model assumptions. In order to incorporate this effect, a contamination model, e.g. gross error model [9, 2] considered in this work, is always preferred.

- 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 for the two distribution families. As to the graphs, one way is to contaminate the entire graph with some small probability ϵ, that is $G \sim (1-\epsilon)F_P + \epsilon H_C$. However, since we are under the independent edge model, it is more natural to consider the contaminations on 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 $G \in \mathcal{G}$, where both $\mathcal{F}$ and $\mathcal{G}$ 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 $F_{P_{ij}}$ do. Moreover, the parameter matrix C for the contamination has the same block structure. (Note: not necessarily the same block structure)

- 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 each of the models mentioned in Section 2, our goal is to estimate the parameter matrix P based on the m observations $A^{(1)}, \ldots, A^{(m)}$. In this section, we present four estimators as in Figure 1, 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 don't really care about the estimate of the diagonal of P, but 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 element-wise MLE $\widehat{P}^{(1)}$ based on the adjacency matrices $A^{(1)}, \ldots, A^{(m)}$.

- In some cases, for instance Bernoulli, Exponential, and Poisson, the entry-wise MLE happens to be the mean graph $\bar{A}$, which is the UMVUE under WIEG with no other constraints. So it has the smallest variance among all unbiased estimators. In addition, as being an MLE, it satisfies many good asymptotic properties as the number of graphs $m \to \infty$. But the performance will not get any better when the number of vertices in each graph n increases. In addition, it doesn't exploit any graph structure. If the graphs are actually distributed under a WRDPG or a WSBM, then the entry-wise MLE is no longer the MLE any more. As a result, the performance can be very poor, especially when m is small.

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. [15] based on the spectral decomposition of $\widehat{P}^{(1)}$. The estimator $\widetilde{P}^{(1)}$ is similar to the estimator proposed by Chaterjee [4] with adjustment for the specific estimation task, including a different dimension selection technique and a diagonal augmentation procedure.

3.2.1 Rank-d Approximation

- Given the dimension d, we consider the best rank-d positive semi-definite approximation of $\widehat{P}^{(1)}$, denoted as $\widetilde{P}^{(1)}$. 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. Then the best rank-d positive semi-definite approximation is $\widetilde{P}^{(1)} = \widehat{U}\widehat{S}\widehat{U}^{\top}$. Note that the d-dimensional adjacency spectral embedding (ASE) here is defined as $\widehat{X} = \widehat{U}\widehat{S}^{1/2} \in \mathbb{R}^{n \times d}$. Thus the low rank approximation of $\widehat{P}^{(1)}$ can also be represented as $\widehat{X}\widehat{X}^{\top}$. In the RDPG setting, Sussman et al. [14] 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.

- We restate the algorithm in [15] to give the steps of computing this low-rank approximation of a general n-by-n symmetric matrix A in Algorithm 1.

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

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

3.2.2 Dimension Selection

- A general way to choose the dimension d in a dimension reduction setting is based on analyzing the ordered eigenvalues and looking for the "gap" or "elbow" in the scree-plot.

- In 2006, Zhu and Ghodsi proposed an automatic method for finding the gap in the scree-plot by only looking at the eigenvalues based on a Gaussian mixture model [17]. Generally, the method provides multiple results 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 choose a proper dimension, based on the (try simulation?) and real data experiment, a wide range of dimensions will lead to results nearly optimal. Thus a proper dimension selection method can be applied directly without carefully tuning the parameter, which is much more practical.

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. And thus the diagonal of the parameter matrix P doesn't 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 biased estimates.

- To compensate for such inaccurate estimate, Marchette et al. [10] suggested to use the average of the non-diagonal entries of the corresponding row as the diagonal entry before embedding. Also, Scheinerman and Tucker [13] proposed an iterative method, which gives a different approach to resolve such issue.

- As suggested in [15], 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: Adjacency matrices $A^{(1)}, A^{(2)}, \dots, A^{(m)}$, with each $A^{(t)} \in \{0,1\}^{n \times n}$
Output: Estimate $\widetilde{P}^{(1)} \in \mathbb{R}^{n \times n}$
Calculate the entry-wise MLE $\widehat{P}^{(1)}$;
Calculate the scaled degree matrix $D = \mathrm{diag}(\widehat{P}^{(1)}\mathbf{1})/(n-1)$;
Select the dimension $d$ based on the eigenvalues of $\widehat{P}^{(1)} + D$; (see Section 3.2.2)
Set $Q$ to $\mathrm{lowrank}_d(\widehat{P}^{(1)} + D)$; (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 $\widetilde{P}^{(1)}$ to $Q'$ with values $< 0$ set to 0.
```

- Detailed description of the estimator $\widetilde{P}^{(1)}$ are given in Algorithm 2.

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

- Motivated by the 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 always outperforms MLE in terms of mean squared error by winning the bias-variance tradeoff.
- Moreover, in practice, the observations are generally contaminated. In this case, robust estimators can even beat MLE asymptotically.
- The class of parametric estimators based on the q-entropy function, the MLqE, is considered in this paper.

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

- Intuitively, the low rank structure of the WRDPG should be preserved more or less in the entry-wise MLqE.
- Similarly, in order to take advantage of such low rank structure, we apply the same idea for building $\widetilde{P}^{(1)}$, i.e. enforce a low rank approximation on the entry-wise MLqE matrix $\widehat{P}^{(q)}$ to get $\widetilde{P}^{(q)}$.
- Detailed description of the algorithm for our estimator $\widetilde{P}^{(q)}$.

4 Theoretical Results

In this section, for illustrative purpose, we are going to present theoretical results when the contaminated model is based on exponential distributions, i.e. $\mathcal{F} = \{f_\theta(x) = \frac{1}{\theta}e^{-x/\theta}, \theta \in [0,R] \subset \mathbb{R}\}$, 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.

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

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

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

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

Lemma 4.2

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

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

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

Corollary 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.*

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

Theorem 4.4 *Assuming that $m = O(n^b)$ for any $b > 0$, then $\mathrm{Var}((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}}) = O(m^{-1}n^{-2}(\log n)^6)$.*

Theorem 4.5 *For fixed m, $1 \leq i, j \leq n$ and $i \neq j$,*

$$\frac{\mathrm{Var}(\widetilde{P}^{(1)}_{ij})}{\mathrm{Var}(\widehat{P}^{(1)}_{ij})} = O(n^{-2}(\log n)^6).$$

Thus

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

Furthermore, as long as m goes to infinity of order $O(n^b)$ for any $b > 0$,

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

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

Corollary 4.6 *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.*

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

Theorem 4.7 *Assuming that $m = O(n^b)$ for any $b > 0$, then $\mathrm{Var}((\widehat{Z}_i^T \widehat{Z}_j)_{\mathrm{tr}}) = O(n^{-2}(\log n)^6)$.*

Theorem 4.8 *For fixed m, $1 \leq i, j \leq n$,*

$$\frac{\mathrm{Var}(\widetilde{P}^{(q)}_{ij})}{\mathrm{Var}(\widehat{P}^{(q)}_{ij})} = O(mn^{-2}(\log n)^6).$$

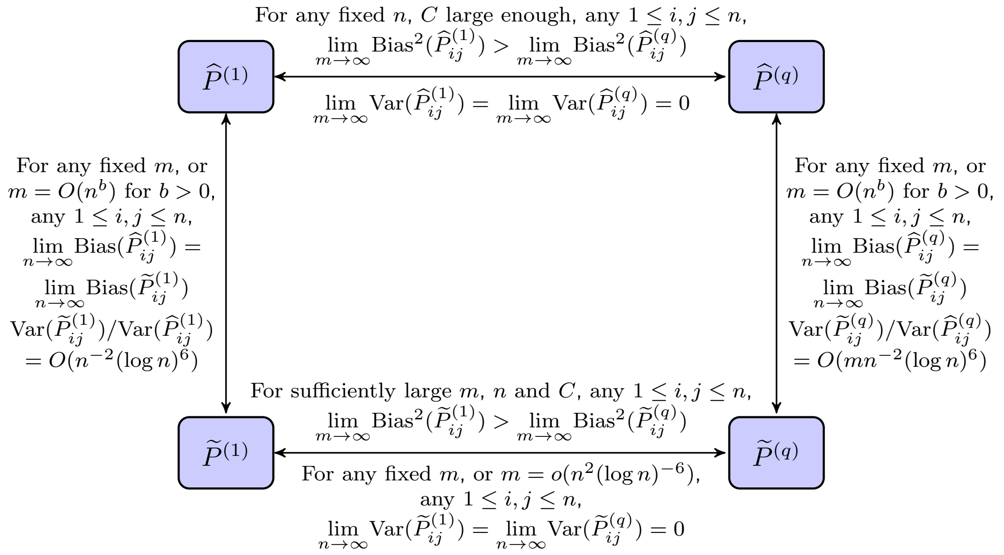

For any fixed n, C large enough, any $1 \le i,j \le n$,
$\lim_{m\to\infty} \mathrm{Bias}^2(\widehat{P}_{ij}^{(1)}) > \lim_{m\to\infty} \mathrm{Bias}^2(\widehat{P}_{ij}^{(q)})$
$\widehat{P}^{(1)}$
$\widehat{P}^{(q)}$
$\lim_{m\to\infty} \mathrm{Var}(\widehat{P}_{ij}^{(1)}) = \lim_{m\to\infty} \mathrm{Var}(\widehat{P}_{ij}^{(q)}) = 0$
For any fixed m, or $m = O(n^b)$ for $b > 0$, any $1 \le i,j \le n$, $\lim_{n\to\infty} \mathrm{Bias}(\widehat{P}_{ij}^{(1)}) = \lim_{n\to\infty} \mathrm{Bias}(\widetilde{P}_{ij}^{(1)})$ $\mathrm{Var}(\widetilde{P}_{ij}^{(1)})/\mathrm{Var}(\widehat{P}_{ij}^{(1)}) = O(n^{-2}(\log n)^6)$
For any fixed m, or $m = O(n^b)$ for $b > 0$, any $1 \le i,j \le n$, $\lim_{n\to\infty} \mathrm{Bias}(\widehat{P}_{ij}^{(q)}) = \lim_{n\to\infty} \mathrm{Bias}(\widetilde{P}_{ij}^{(q)})$ $\mathrm{Var}(\widetilde{P}_{ij}^{(q)})/\mathrm{Var}(\widehat{P}_{ij}^{(q)}) = O(mn^{-2}(\log n)^6)$
For sufficiently large m, n and C, any $1 \le i,j \le n$,
$\lim_{m\to\infty} \mathrm{Bias}^2(\widetilde{P}_{ij}^{(1)}) > \lim_{m\to\infty} \mathrm{Bias}^2(\widetilde{P}_{ij}^{(q)})$
$\widetilde{P}^{(1)}$
$\widetilde{P}^{(q)}$
For any fixed m, or $m = o(n^2(\log n)^{-6})$, any $1 \le i,j \le n$,
$\lim_{n\to\infty} \mathrm{Var}(\widetilde{P}_{ij}^{(1)}) = \lim_{n\to\infty} \mathrm{Var}(\widetilde{P}_{ij}^{(q)}) = 0$

Figure 2: Relationship among four estimators.

Thus

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

Furthermore, as long as m goes to infinity of order $o(n^2(\log n)^{-6})$,

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

4.4 $\widetilde{P}^{(q)}$ vs $\widetilde{P}^{(1)}$

Theorem 4.9 *For sufficiently large n and C, any* $1 \le i, j \le n$,

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

Theorem 4.10 *For any fixed m, any* $1 \le i, j \le n$,

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

Furthermore, as long as m goes to infinity of order $o(n^2(\log n)^{-6})$, *any* $1 \le i, j \le n$,

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

4.5 Summary

In summary, we plot the relationship among four estimators in Figure 2.

5 Extensions

We can generalize both the exponential distribution and the entry-wise MLqE to an arbitrary F and any entry-wise estimator $\widehat{P}$ with the following assumptions:

- Let $A \overset{iid}{\sim} (1-\epsilon)F(P) + \epsilon F(C)$ and $H_{ij}^{(1)} = E[\widehat{P}_{ij}^{(1)}] = (1-\epsilon)E_F(P_{ij}) + \epsilon E_F(C_{ij})$, then $E[(A_{ij} - H_{ij}^{(1)})^k] \leq \text{const} \cdot k!$.
- There exists $C_0(P_{ij}, \epsilon) > 0$ such that under the contaminated model with $C > C_0(P_{ij}, \epsilon)$,
$$\lim_{m \to \infty} \left| E[\widehat{P}_{ij}] - P_{ij} \right| < \lim_{m \to \infty} \left| E[\widehat{P}_{ij}^{(1)}] - P_{ij} \right|,$$
for $1 \leq i, j \leq n$.
- $\widehat{P}_{ij} \leq \text{const} \cdot \widehat{P}_{ij}^{(1)}$. This might be generalized to with high probability later.
- $\text{Var}(\widehat{P}_{ij}) = O(m^{-1})$, where m is the number of observations.

6 Empirical Results

6.1 Simulation Results

- We demonstrate the theoretical results in Section 3.1, the relative efficiency of $\widehat{P}$, via various Monte Carlo simulation experiments.

6.1.1 Simulation Setting

- Here we consider the 2-block SBM parameterized by
$$B = \begin{bmatrix} 4.2 & 2 \\ 2 & 7 \end{bmatrix}, \qquad \rho = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}.$$
- The contamination is also a 2-block SBM parameterized by
$$B = \begin{bmatrix} 20 & 18 \\ 18 & 25 \end{bmatrix}, \qquad \rho = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}.$$
- And we embed the graphs into the dimension $d = \text{rank}(B) = 2$.

6.1.2 Simulation Results

- Figure 3 plots the mean squared error in average by varying contamination ratio ϵ with fixed $n = 100$ and $m = 10$ based on 1000 Monte Carlo replicates. And we use $q = 0.8$ 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 o 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 o 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 o MLqE $\widetilde{P}^{(q)}$ vs ASE o MLE $\widetilde{P}^{(1)}$:** When contamination is large enough, $\widetilde{P}^{(q)}$ based on MLqE is better, since it inherits the robustness from MLqE.

- Figure 4 show the mean squared error in average by varying the parameter q in MLqE with fixed $n = 100$, $m = 10$ and $\epsilon = 0.2$ based on 1000 Monte Carlo replicates. Different colors 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; 2. MLqE 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

- In practice, the graphs may not perfectly follow an RDPG, or even not IEM. But we are still interested in the discussed approach. To demonstrate that the estimates are still valid in such cases, we examine the datasets, CPAC200, which is a set of 454 brain connectomes with different number of nodes generated from fMRI scans available at the Consortium for Reliability and Reproducibility (CoRR).

- The dataset has 454 different brain scans in the form of weighted, undirected graph with no self loop, based on the pipeline described in [8] and [1].

- To compare the four estimators, we perform a cross-validation study on 454 graphs.

- We run 1000 simulations on the dataset for each sample size $m = 1$, $m = 2$, $m = 5$. And we apply ASE for all possible dimensions, i.e. d ranges from 1 to n. The result is shown in Figure 5.

- Since it is real data, MLqE outperforms MLE because of the robustness property. Moreover, as suggested in the previous theorems, such property is kept after the ASE procedure.

- When d is small, ASE procedure underestimates the dimension and fail to get important information, which leads to poor performance. In practice, we use algorithms like Zhu and Ghodsi's method to select the dimension d. We can see Zhu and Ghodsi's algorithm does a pretty good job for selecting the dimension to embed.

- When m is small, MLE and MLqE have large variances which lead to large MSE. Meanwhile, the ASE procedure reduces the variance by taking advantages of the graph structure.

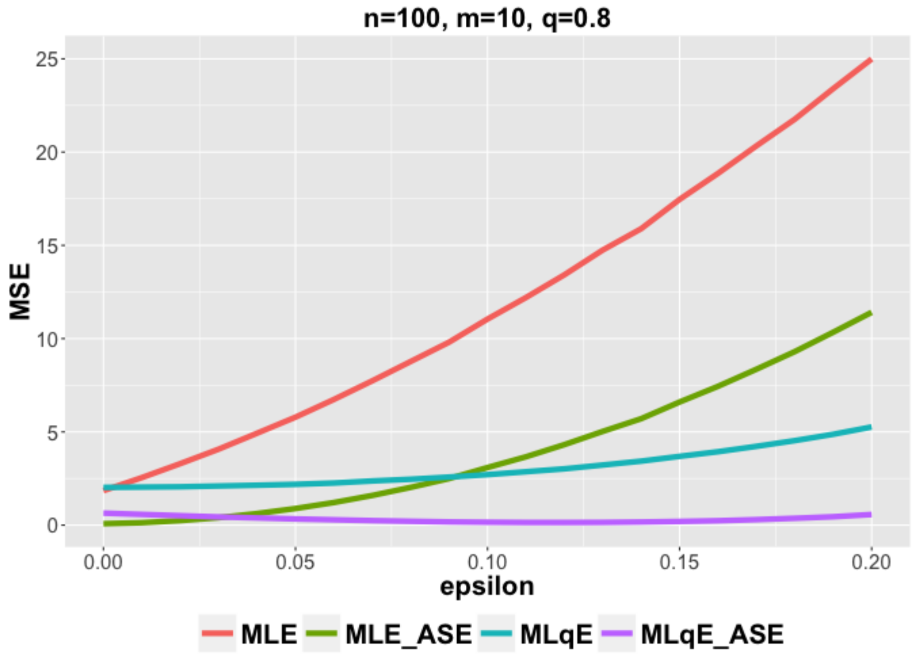


Figure 3: Mean squared error in average by varying contamination ratio ϵ with fixed $n = 100$ and $m = 10$ based on 1000 Monte Carlo replicates. And we use $q = 0.8$ when applying MLqE. Different colors represent the simulated MSE associated with four different estimators. **1. MLE $\widehat{P}^{(1)}$ 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 o 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 o 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 o MLqE $\widetilde{P}^{(q)}$ vs ASE o MLE $\widetilde{P}^{(1)}$:** When contamination is large enough, $\widetilde{P}^{(q)}$ based on MLqE is better, since it inherits the robustness from MLqE.

7 Discussion

8 Appendix

All proofs here.

References

[1] Neurodata's mri to graphs pipeline. `http://m2g.io`. Accessed: 2016-05-23.

[2] 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.

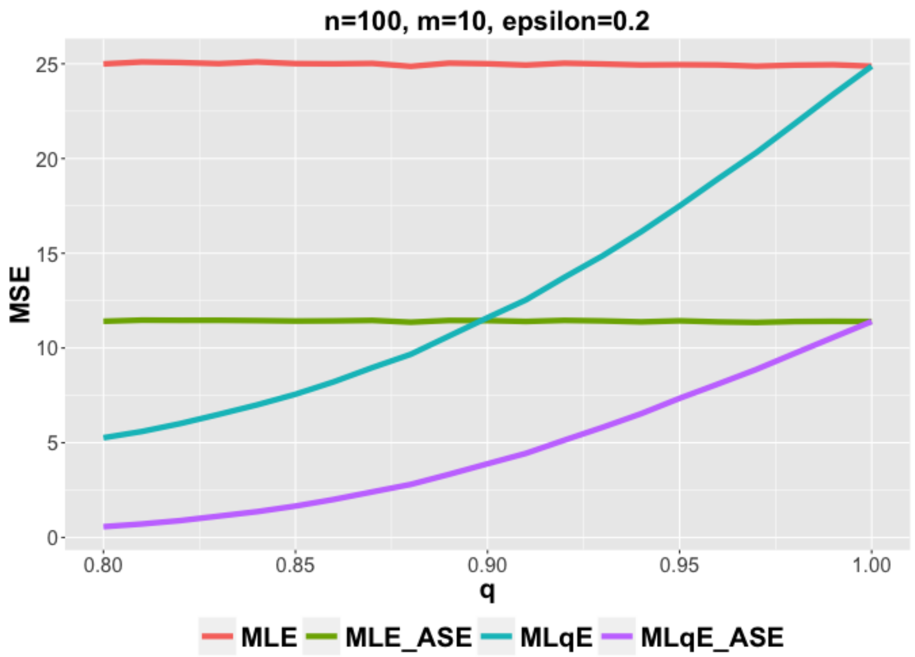


Figure 4: Mean squared error in average by varying the parameter q in MLqE with fixed $n = 100$, $m = 10$ and $\epsilon = 0.2$ based on 1000 Monte Carlo replicates. Different colors represent the simulated MSE associated with four different estimators. 1. ASE procedure takes advantage of the graph structure and improves the performance of the corresponding estimators; 2. MLqE shows the robustness property compare to the MLE. And as q goes to 1, MLqE goes to the MLE as expected.

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

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

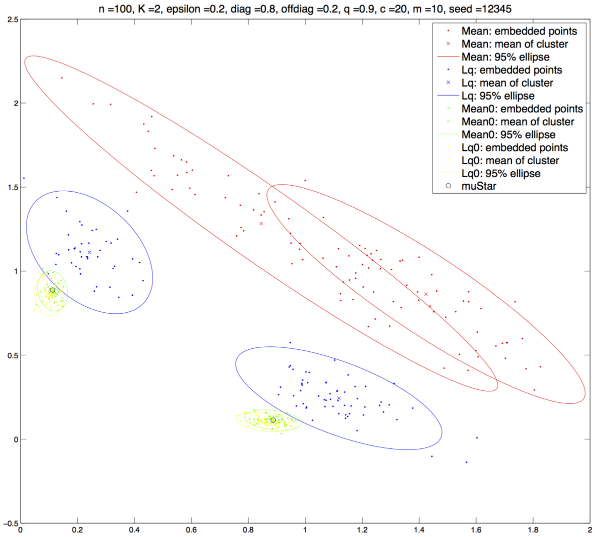
[5] Davide Ferrari and Yuhong Yang. Maximum lq-likelihood estimation. *Ann. Statist.*, 38(2):753–783, 04 2010.

[6] 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.

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

[8] Gregory Kiar. Gremlin: Graph estimation from mr images leading to inference in neuroscience. 2016.

[9] R. S. H. Mah and A. C. Tamhane. Detection of gross errors in process data. *AIChE Journal*, 28(5):828–830, 1982.



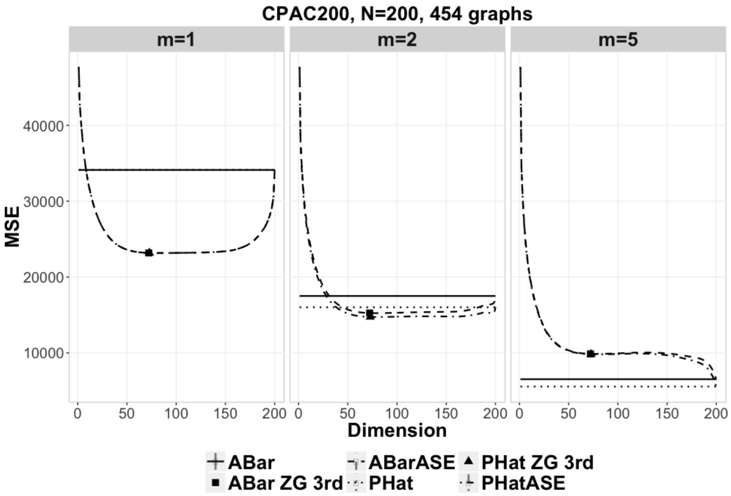


Figure 5: Comparison of mean squared error in average among four estimators while embedding the graphs into different dimensions with different size m of the subsamples. The dimensions chosen by the 3d elbow of Zhu and Ghodsi are denoted in triangle and square. When m is small, both robust estimation and ASE procedure help improving the performance, making $\widetilde{P}^{(q)}$ the best among four estimators.

[10] 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.

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

[12] 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, 2013.

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

[14] 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.

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

[16] 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.

[17] 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.