DEPARTMENT OF STATISTICS
University of Wisconsin
1300 University Ave.
Madison, WI 53706

TECHNICAL REPORT NO. 1169
September 19, 2012

The Highest Dimensional Stochastic Blockmodel with a Regularized Estimator

Karl Rohe[1]
Department of Statistics
University of Wisconsin, Madison

Tai Qin[2]
Department of Statistics
University of Wisconsin, Madison

Haoyang Fan
Department of Statistics
University of Wisconsin, Madison

[1]Research of KR is supported by a grant from the University of Wisconsin.
[2]Research of TQ is supported by NSF Grant DMS-0906818 and NIH Grant EY09946.

The Highest Dimensional Stochastic Blockmodel with a Regularized Estimator

Karl Rohe
karlrohe@stat.wisc.edu
Department of Statistics
University of Wisconsin, Madison

Tai Qin
qin@stat.wisc.edu
Department of Statistics
University of Wisconsin, Madison

Haoyang Fan
haoyang@stat.wisc.edu
Department of Statistics
University of Wisconsin, Madison

Abstract

This paper advances the high dimensional frontier for network clustering. In the high dimensional Stochastic Blockmodel for a random network, the number of clusters (or blocks) K grows with the number of nodes N. Previous authors have studied the statistical estimation performance of spectral clustering and the maximum likelihood estimator under the high dimensional model. These authors do not allow K to grow faster than $N^{1/2}$. We study a model where, ignoring log terms, K can grow proportionally to N. Since the number of clusters must be smaller than the number of nodes, no reasonable model allows K to grow faster; thus, our asymptotic results are the "highest" dimensional. To push the asymptotic setting to this extreme, we make additional assumptions that are motivated from empirical observations in physical anthropology [1], and an in depth study of massive empirical networks[2]. Furthermore, we develop a regularized maximum likelihood estimator that performs well in the highest dimensional model. We prove that, under certain conditions, the proportion of nodes that the regularized estimator misclusters converges to zero. This is the first paper to explicitly introduce and demonstrate the advantages of statistical regularization in a parametric form for network analysis.

1 Introduction

Recent advances in information technology have produced a deluge of data on complex systems with myriad interacting elements. Depending on the area of interest, these interacting elements could be metabolites, people, or computers. Their interactions could be

represented in chemical reactions, friendship, or some type of communication. Networks (or graphs) appropriately describe these relationships. Therefore, the substantive questions in these various disciplines are, in essence, questions regarding the structure of a network. Communities or clusters of highly connected actors are an essential feature in a multitude of empirical networks, and identifying these clusters helps answer vital questions in various fields. A terrorist cell is a cluster in the communication network of terrorists; web pages that provide hyperlinks to each other form a community that might host discussions of a similar topic; a cluster in the network of biochemical reactions might contain metabolites with similar functions and activities.

Just as classical statisticians have studied when ordinary least-squares regression can estimate the "true regression model," it is timely and important to study the ability of clustering algorithms to estimate the "true clusters" in a network model. Understanding when and why a clustering algorithm correctly estimates the "true communities" would provide a rigorous understanding of the behavior of these algorithms and potentially lead to improved algorithms. The Stochastic Blockmodel is a model for a random network. The "blocks" in the model, correspond to the concept of "true communities" that we want to study. In the Stochastic Blockmodel, N actors (or nodes) each belong to one of K blocks and the probability of a connection between two nodes depends only on the memberships of the two nodes [3]. This paper aims to add to the rigorous understanding of the maximum likelihood estimator (MLE) under the Stochastic Blockmodel.

There has been significant interest in how the various clustering algorithms perform under the Stochastic Blockmodel [4, 5, 6, 7, 8, 9, 10]. Both [5] and [6] studied the high dimensional Stochastic Blockmodel, an asymptotic setting that allows the number of blocks K to grow with the number of nodes N. The impetus for this comes from several empirical observations. [2] studied a large corpus of empirical networks, of varying sizes and applications. Even though some of the networks had several million nodes, they found that in all the networks they analyzed, the tightest clusters[1] were no larger than 100 nodes. This corresponds to a finding in Physical Anthropology, which related the size of various primate's prefrontal cortex with the size of their natural communities [1]. Extrapolating this relationship to humans suggests that we do not have the social intellect to maintain stable communities larger than roughly 150 people (colloquially referred to as Dunbar's number). [2] found a similar pattern in several other networks that were not composed of humans. In the Stochastic Blockmodel, the population of the average block is N/K. The research of [2] and [1] suggests that this average block size should not grow. So, if N is growing, then K should also grow.

In a parallel line of research, several authors have studied clustering algorithms on the Planted Partition Model, a model nearly identical to the Stochastic Blockmodel. McSherry [11] provides a spectral algorithm to recover the planted partition and analyzes the estimation performance of this algorithm. [12] improved on this algorithm by introducing various forms of regularization. Several others [13] have proposed and studied other fast algorithms. [14, 15] have proposed other regularized algorithms. The work in this paper is differentiated from those in that we study a parametric method. This makes our regularization transparent

[1] as judged by several popular clustering criteria

because it appears in a parametric form.

In the previous research of [5] and [6], the average block size grows at least as fast as $N^{3/4}$ and $N^{1/2}$ respectively. Even though these asymptotic results allow for K to grow with N, K does not grow fast enough. The average block size quickly surpass Dunbar's number. In this paper, we introduce the Highest Dimensional Stochastic Blockmodel (HSBM), where $K = N \log^{-4} N$ and $N/K = \log^4 N$. Thus, under the HSBM, the size of the clusters grows much more slowly. We call it the "highest" dimensional because, ignoring the log term, K cannot grow any faster. If it did, then eventually $K > N$ and there would necessarily be blocks containing zero nodes. To create a sparse graph, the out-of-block probabilities decay roughly like polylog(N)/N in the HSBM. To ensure that the subnetwork formed by each of the blocks is connected, the in-block probabilities do not decay. We show that under this asymptotic setting, a regularized maximum likelihood estimator (RMLE) can estimate the block partition for most nodes in the HSBM.

High dimensional learning–in regression, covariance estimation, matrix completion, and elsewhere–requires some type of low dimensional structure. This paper breaks from the previous high dimensional clustering results of [5] and [6] by restricting the parameter space of the Stochastic Blockmodel. In several high dimensional settings, regularization restricts the full parameter space providing a path to consistent estimators [16]. If the true parameter setting is close to the restricted parameter space, then regularization trades a small amount of bias for a potentially large reduction in variance. For example, in the high dimensional regression literature, sparse regression techniques such as the LASSO restrict the parameter space to produce sparse regression estimators [17]. Several authors have also suggested parameter space restrictions for high dimensional covariance estimation, e.g. [18, 19, 20]. When applying Linear Discriminant Analysis to high dimensional data, it is impossible to estimate the whole covariance matrix Σ; the Nearest Shrunken Centroids approach restricts the parameter space by setting every off diagonal elements of $\hat{\Sigma}$ equal to zero and shrinks the classwise mean toward the overall mean [21]. In this paper, we similarly propose a restricted parameter space to fit the Stochastic Blockmodel. These restrictions are supported by empirical observations [1, 2], and they require a statistically regularized estimator. We will show that our RMLE is suitable in the HSBM setting.

2 Preliminaries

2.1 Highest Dimensional Stochastic Blockmodel (HSBM)

In the Stochastic Blockmodel (SBM), each node belongs to one of K blocks. Each edge corresponds to an independent Bernoulli random variable where the probability of an edge between any two nodes depends only on the two nodes' block memberships [3]. The formal definition is as follows.

Definition 2.1. *For a node set $\{1, 2, ..., N\}$, let P_{ij} denote the probability of including an edge linking node i and j. Let $\tilde{z} : \{1, 2, ..., N\} \to \{1, 2, ..., K\}$ partition the N nodes into K blocks. So, $\tilde{z}_i$ equals the block membership for node i. Let $\boldsymbol{\theta}$ be a $K \times K$ matrix where*

$\theta_{ab} \in [0,1]$ for all a, b. Then $P_{ij} = \theta_{z_i z_j}$ for any $i, j = 1, 2, ..., n$. So under the SBM, the the probability of observing adjacency matrix A is

$$P(A) = \prod_{i<j} \theta_{\tilde{z}_i \tilde{z}_j}{}^{A_{ij}} (1 - \theta_{\tilde{z}_i \tilde{z}_j})^{(1-A_{ij})}.$$

The distribution factors over $i < j$ because we only consider undirected graphs without self-loops.

The Highest Dimensional Stochastic Blockmodel (HSBM) is not a single model. Rather, it defines an asymptotic setting for the SBM. We will discuss the HSBM as an actual model. However, it is important to remember that it is actually an asymptotic setting. The HSBM defined in Definition 2.2 restricts the parameters of the SBM in two ways. First, because empirical evidence suggests that community sizes do not grow with the size of the network, the HSBM allows s, defined to be the population of the smallest block, to grow very slowly. The second restriction ensures that a network sampled from the HSBM will contain communities. At a high level, there are two types of edges, "in-block edges" that connect nodes in the same block and "out-of-block edges" that connect nodes in different blocks. In the high dimensional setting, the number of possible out-of-block edges far exceeds the number of possible in-block edges. If the blocks all have the same population s, and s is roughly constant, then the former grows like n^2 and the latter grows like n. For the sampled network to contain communities, the out-of-block edges should not dramatically out number the in-block edges. For this reason, the HSBM restricts the off diagonal elements of $\boldsymbol{\theta}$; they must shrink as the network grows. The set Q prevents this restriction from becoming too stringent; it allows some pairs of blocks to have a tighter connection. If $(a, b) \in Q$, then θ_{ab} is not required to shrink as the network grows. As a result, blocks a and b will share more edges.

Definition 2.2. *An HSBM is an SBM with the following asymptotic restrictions.*

(R1) For s equal to the population of the smallest block and $x_n = \omega(y_n) \Leftrightarrow y_n/x_n = o(1)$,

$$s = \omega(\log^\beta N), \ \beta > 3.$$

(R2) Let Q contain a subset of the indices for $\boldsymbol{\theta}$. For constants C and $\Delta < 0.5$ and $f(N) = o(s/\log N)$,

$$\theta_{ab} = \theta_{ba} \in \begin{cases} (\Delta, 1-\Delta) & a = b \\ (1/N^2, Cf(N)/N) & a < b, \{a, b\} \notin Q \\ (\Delta, 1-\Delta) & a < b, \{a, b\} \in Q. \end{cases}$$

In the next sections we will introduce the RMLE and then show that it can identify the blocks under the HSBM's asymptotic settings.

2.2 Regularized Maximum Likelihood Estimator

Under the HSBM, the number of parameters in $\boldsymbol{\theta}$ is quadratic in K and the sample size available for estimating each parameter in $\boldsymbol{\theta}$ is as small as s^2. For tractable estimation in the "large K small s" setting, we propose an RMLE.

Recall that $\tilde{z}$ denotes the true partition. Let z denote any arbitrary partition. The log-likelihood for an observed adjacency matrix A under the SBM w.r.t node partition z is

$$L(A; z, \boldsymbol{\theta}) = \log P(A; z, \boldsymbol{\theta}) = \sum_{i<j} \{A_{ij} \log \theta_{z_i z_j} + (1 - A_{ij}) \log(1 - \theta_{z_i z_j})\}.$$

For fixed class assignment z, let N_a denote the number of nodes assigned to class a, and let n_{ab} denote the maximum number of possible edges between class a and b; i.e., $n_{ab} = N_a N_b$ if $a \neq b$ and $n_{aa} = \binom{N_a}{2}$. For an arbitrary partition z, the MLE of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}^{(\boldsymbol{z})} = \arg \max_{\boldsymbol{\theta} \in [0,1]^{K \times K}} L(A; z, \boldsymbol{\theta}).$$

This is a symmetric matrix in the parameter space $\Theta = [0, 1]^{K \times K}$. It is straightforward to show

$$\hat{\theta}^{(z)}_{ab} = \frac{1}{n_{ab}} \sum_{i<j} A_{ij} 1\{z_i = a, z_j = b\}, \quad \forall a, b = 1, 2, ..., K$$

By substituting $\hat{\boldsymbol{\theta}}^{(\boldsymbol{z})}$ into $L(A; z, \boldsymbol{\theta})$, we can get the profiled log-likelihood. Define

$$L(A; z) = L(A; z, \hat{\boldsymbol{\theta}}^{(\boldsymbol{z})}).$$

Define $\hat{z} = \arg\max_z L(A; z)$ as the MLE of $\tilde{z}$. To define the RMLE, define the restricted parameter space, $\Theta^R \subset \Theta$, by the following regularization:

$$\Theta^R = \left\{ \boldsymbol{\theta} \in [0, 1]^{K \times K} : \theta_{ab} = c, \ \forall \ a \neq b \ \text{ and for } c \in [0, 1] \right\}.$$

If $\theta \in \Theta^R$, then all the off-diagonal elements of θ are equal. We call the new estimator "regularized" because, where Θ has $K(K + 1)/2$ free parameters, Θ^R has only $K + 1$ free parameters.

Given class assignment z, The RMLE $\boldsymbol{\theta}^{R,(z)}$ is the maximizer of $L(A; z, \boldsymbol{\theta})$ within Θ^R.

$$\boldsymbol{\theta}^{R,(z)} = \arg \max_{\boldsymbol{\theta} \in \Theta^R} L(A; z, \boldsymbol{\theta}).$$

The optimization problem within Θ^R can be treated as an unconstrained optimization problem within $[0, 1]^{K+1}$ since we force the off-diagonal elements of $\boldsymbol{\theta}$ to be equal to some number r. It has a closed form solution:

$$\hat{\theta}^{R,(z)}_{ab} = \begin{cases} \hat{\theta}^{(z)}_{aa} = \frac{1}{n_{aa}} \sum_{i<j} A_{ij} 1\{z_i = a, z_j = b\} & a = b, \\ \hat{r}^{(z)} = \frac{1}{n_{out}} \sum_{i<j} A_{ij} 1\{z_i \neq z_j\} & a \neq b. \end{cases}$$

Here $n_{out} = \sum_{a<b} n_{ab}$ is the maximum number of possible edges between all different blocks. The Regularized MLE for θ_{aa} is exactly the same as ordinary MLE, while the Regularized MLE for $\theta_{ab}, a \neq b$ is set to be equal to the total off-diagonal average. Finally, by substituting $\hat{\boldsymbol{\theta}}^{R,(z)}$ into $L(A; z, \boldsymbol{\theta})$, define the regularized profile log-likelihood to be

$$L^R(A; z) = L(A; z, \hat{\boldsymbol{\theta}}^{R,(z)}) = \sup_{\boldsymbol{\theta} \in \Theta^R} L(A; z, \boldsymbol{\theta}),$$

and denote the RMLE of the true partition $\tilde{z}$ to be

$$\hat{z}^R = \arg\max_z L^R(A; z). \tag{1}$$

3 The asymptotic performance of the RMLE on the HSBM

Our main result shows that most nodes are correctly clustered by the RMLE under the HSBM. This result requires the definition of correctly clustered which comes from [6].

Definition 3.1. *For any estimated class assignment z, define $N_e(z)$ as the number of incorrect class assignments under z, counted for every node whose true class under $\tilde{z}$ is not in the majority within its estimated class under z.*

The main result, Theorem 3.2, uses the KL divergence between two Bernoulli distributions. This is defined as,

$$D(p\|q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

Recall that under the HSBM, Q denotes the off diagonal indices of $\boldsymbol{\theta}$ that do not asymptotically decay. Additionally, n_{ab} denotes the total number of possible edges between nodes in block a and nodes in block b. Define $|Q|$ as the number of possible tight edges across different blocks,

$$|Q| = \sum_{\{a,b\} \in Q} n_{ab}. \tag{2}$$

The following theorem is our main result. It shows that under the HSBM, the proportion of nodes that the RMLE misclusters converges to zero.

Theorem 3.2. *Under the HSBM in Definition 2.2, N is the total number of nodes, and s is the population of the smallest block. Assume that the set of friendly block pairs Q (defined in R2 of Definition 2.2) is small enough that $|Q| = o(Ns)$, where $|Q|$ is defined in Equation 2. Further, for the matrix of probabilities $\boldsymbol{\theta}$, assume that for any distinct class pairs (a, b), there exists a class c such that the following condition holds:*

$$D\left(\theta_{ac} \| \frac{\theta_{ac} + \theta_{bc}}{2}\right) + D\left(\theta_{bc} \| \frac{\theta_{ac} + \theta_{bc}}{2}\right) \geq C \frac{MK}{N^2} \tag{3}$$

Under these assumptions, RMLE $\hat{z}^R$ defined in Equation 1 satisfies

$$\frac{N_e(\hat{z}^R)}{N} = o_p(1),$$

where $N_e(z)$ is the number of misclustered nodes defined in Definition 3.1.

This theorem requires two main assumptions. The first main assumption is $Q^* = o(Ns)$. Define the number of expected edges $M = \sum_{i<j} EA_{ij}$. Under the HSBM, this first assumption implies that M grows slowly, $M = \omega(N(\log N)^{3+\delta})$. The second main assumption says that every distinct class pair (a, b) has at least one class c that satisfies Equation 3. This assumption ensures the identifiability of $\tilde{z}$ under the HSBM. For example, if $(a, b) \notin Q$, then choosing $c = a$ satisfies the assumption in Equation 3, because θ_{aa} is large and θ_{ba} is small. However, if $(a, b) \in Q$, then there should exist at least one class c to make θ_{ac},θ_{bc} identifiable. Otherwise, blocks a and b should be merged into the same block.

4 The proof of the main result

The proof requires some additional definitions. After giving these definitions, we will outline the proof.

Define the expectation of $\hat{\boldsymbol{\theta}}^{(z)}$ and $\hat{\boldsymbol{\theta}}^{R,(z)}$ to be $\bar{\boldsymbol{\theta}}^{(z)}$ and $\bar{\boldsymbol{\theta}}^{R,(z)}$. Define the expectation of $L(A; z, \boldsymbol{\theta})$ to be

$$\bar{L}_P(z, \boldsymbol{\theta}) = E[L(A; z, \boldsymbol{\theta})] = \sum_{i<j} \{P_{ij} \log \theta_{z_i z_j} + (1 - P_{ij}) \log(1 - \theta_{z_i z_j})\}.$$

Let $\bar{L}_P(z)$ to be the maximizer of $\bar{L}_P(z, \boldsymbol{\theta})$ over Θ, and let $\bar{L}_P^R(z)$ to be the maximizer of $\bar{L}_P(z, \boldsymbol{\theta})$ over Θ^R. That is,

$$\bar{L}_P(z) = \bar{L}_P(z, \bar{\boldsymbol{\theta}}^{(z)}) = \sup_{\boldsymbol{\theta} \in \Theta} \bar{L}_P(z, \boldsymbol{\theta}), \tag{4}$$

$$\bar{L}_P^R(z) = \bar{L}_P(z, \bar{\boldsymbol{\theta}}^{R,(z)}) = \sup_{\boldsymbol{\theta} \in \Theta^R} \bar{L}_P(z, \boldsymbol{\theta}). \tag{5}$$

The proof of the main theorem is divided into five lemmas. The first step is to bound the difference between $\bar{L}_P(\tilde{z})$ and $\bar{L}_P^R(\hat{z}^R)$ (Lemma 4.3). Lemma 4.1 and Lemma 4.2 are two building blocks of Lemma 4.3 . Lemma 4.1 establishes a union bound of $|L^R(A; z) - \bar{L}_P^R(z)|$ for any partition z. Lemma 2 shows that under the true partition $\tilde{z}$, the expectation of regularized likelihood is close to the expectation of the ordinary likelihood. Lemma 4.3 divides $\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R)$ into three parts and controls them respectively. We can see this as a bias-variance tradeoff; we sacrifice some bias $\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\tilde{z})$ to decrease the variance $\max_z |L^R(A; z) - L_P^R(z)|$. After Lemma 4.3, it is necessary to develop the concept of regularized refinement, an extension of the refinement idea proposed in [6]. Using the concept of regularized refinement, we can bound the error rate $N_e(\hat{z}^R)/N$ with a function of

$\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R)$. Lemma 4.5 and Lemma 4.6 use a new concept of regularized refinement to connect the bounds on the log-likelihood with the error rate $N_e(\hat{z}^R)/N$. From here on, we write $\hat{\theta}$ and $\bar{\theta}$ instead of $\hat{\theta}^{(z)}$ and $\bar{\theta}^{(z)}$ when the choice of z is understood.

Lemma 4.1. *Let M to be the total expected degree of A. That is, $M = \sum_{i<j} EA_{ij}$.*

$$\max_z |L^R(A; z) - \bar{L}_P^R(z)| = o_p(M). \tag{6}$$

This proof follows a similar argument made in [6].

Proof. Let $H(p) = -p \log p - (1-p)\log(1-p)$, which is the entropy of a Bernoulli random variable with parameter p. Define $X = \sum_{i<j} A_{ij} \log\{\bar{\theta}_{z_i z_j}/(1-\bar{\theta}_{z_i z_j})\}$. Let n_{ab} denote the maximum number of possible edges between all different blocks.

$$\begin{aligned} L^R(A; z) - L_P^R(z) &= -\sum_{a=1}^{K} n_{aa}(H(\hat{\theta}_{aa}) - H(\bar{\theta}_{aa})) - n_{out}(H(\hat{r}) - H(\bar{r})) \\ &= \sum_{a=1}^{K} n_{aa} D(\hat{\theta}_{aa} \| \bar{\theta}_{aa}) + n_{out} D(\hat{r} \| \bar{r}) + X - E(X). \end{aligned}$$

For the first part $\sum_{a=1}^{K} n_{aa} D(\hat{\theta}_{aa} \| \bar{\theta}_{aa}) + n_{out} D(\hat{r} \| \bar{r})$, by the same argument as in [6], we have that for every regularized estimator $\hat{\theta}^R$:

$$pr(\hat{\theta}^R) \le \exp\left\{ -\sum_{a=1}^{K} n_{aa} D(\hat{\theta}_{aa} \| \bar{\theta}_{aa}) + n_{out} D(\hat{r} \| \bar{r}) \right\}.$$

Let $\hat{\Theta}$ denote the range of $\hat{\theta}^R$ for fixed z. Then the total number of sets of values $\hat{\theta}^R$ can take is $|\hat{\Theta}| = (n_{out}+1) \cdot \Pi_{a=1}^{K}(n_{aa}+1)$. Notice that $\sum_{a=1}(n_{aa}+1) + (n_{out}+1) = \frac{N(N-1)}{2} + K + 1$, we have $|\hat{\Theta}| \le (\frac{N(N-1)}{2(K-1)} + 1)^{K+1} \le (\frac{N^2}{2K})^{(K+1)}$. Then $\forall \epsilon > 0$,

$$\begin{aligned} pr\left\{ \sum_{a=1}^{K} n_{aa} D(\hat{\theta}_{aa} \| \bar{\theta}_{aa}) + n_{out} D(\hat{r} \| \bar{r}) > \epsilon \right\} &\le |\hat{\Theta}| e^{-\epsilon} \le (\frac{N^2}{2K})^{(K+1)} e^{-\epsilon} \\ &\le \exp\left\{ 2(K+1)\log N - (K+1)\log(2K) - \epsilon \right\}. \end{aligned}$$

For the second part $X - E(X)$, each $X_{ij} = A_{ij} \log\{\bar{\theta}_{z_i z_j}/(1-\bar{\theta}_{z_i z_j})\}$ is bounded in magnitude by $C = 2\log N$. By the following concentration inequality:

$$pr\{|X - E(X)| \ge \epsilon\} \le 2\exp\left\{ -\frac{\epsilon^2}{2\sum_{i<j} E(X_{ij}^2) + (2/3)C\epsilon} \right\}.$$

Here $E(X_{ij}^2) \leq 4M \log^2 N$. Finally by a union bound inequality over all partition z, we have:

$$pr\{\max_z |L^R(A;z) - L_P^R(z)| \geq 2\epsilon M\} \leq \exp\{N \log K + 2(K+1)\log N - (K+1)\log(2K) - M\epsilon\} + 2\exp\left\{N \log K - \frac{\epsilon^2 M}{8\log^2 N + (4/3)\epsilon \log N}\right\}.$$

Notice that in this asymptotic setting , the total expected degree $M = \omega(N(\log N)^{3+\delta})$. Then, $\max_z |L^R(A;z) - L_P^R(z)| = o_p(M)$. □

Lemma 4.2. *Under the true partition $\tilde{z}$, $\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\tilde{z}) = o(M)$.*

Proof. When N is sufficiently large,

$$\begin{aligned}
\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\tilde{z}) &= \sum_{a<b} D(\theta_{ab}\|\bar{r}) = \sum_{a<b,\{a,b\}\in Q} n_{ab} D(\theta_{ab}\|\bar{r}) + \sum_{a<b,\{a,b\}\notin Q} n_{ab} D(\theta_{ab}\|\bar{r}) \\
&\leq |Q|C(\Delta) + (N(N-1)/2 - \sum_{a=1}^{K} n_{aa} - |Q|)\frac{Cf(N)}{N}(\log(C * Nf(N))) \\
&\leq |Q|C(\Delta) + N^2 \frac{Cf(N)}{N}(\log N + \log Cf(N)) = o(M).
\end{aligned}$$

Here $C(\Delta)$ is some constant related only to Δ. □

Lemma 4.3. *Under the true partition $\tilde{z}$ and the RMLE $\hat{z}^R$, $\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R) = o_p(M)$.*

Proof. First notice that the left hand side is a nonnegative value since $\tilde{z}$ maximizes $\bar{L}_P(\cdot)$ and $\bar{L}_P(\hat{z}^R) \geq \bar{L}_P^R(\hat{z}^R)$.

By adding another positive term, and using Lemma 4.1 and Lemma 4.2:

$$\begin{aligned}
\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R) &\leq \bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R) + L^R(A;\hat{z}^R) - L^R(A,\tilde{z}) \\
&\leq |\bar{L}_P(\tilde{z}) - L^R(A,\tilde{z})| + |\bar{L}_P^R(\hat{z}^R) - L^R(A;\hat{z}^R)| \\
&\leq |\bar{L}_P(\tilde{z}) - \bar{L}_p^R(\tilde{z})| + |\bar{L}_P^R(\tilde{z}) - L^R(A,\tilde{z})| + |\bar{L}_P^R(\hat{z}^R) - L^R(A;\hat{z}^R)| \\
&= o_p(M).
\end{aligned}$$

□

To make $N_e(z)$ mathematically tractable, [6] introduced the concept of block refinements. The next paragraphs reintroduce the definition from [6]. We then extend this definition to the regularized block refinement.

4.1 Partitions and refinements

The refinement is the key concept to connect $\max_z |L^R(A;z) - L_P^R(z)|$ with the error rate $N_e(\hat{z}^R)/N$. For this subsection, we first review the concept of partition and refinement from [6]. Then, we give its regularized version. Second, we state the fact that a refinement's log-likelihood is no less than the original partition's. Then, the distance between log-likehood of its (regularized) refinement and log-likelihood of true $\tilde{z}$ can be bounded by the distance between (regularized) log-likelihood of arbitrary z partitions and log-likelihood true $\tilde{z}$. At last, the connection between (regularized) refinement log-likelihood and the error rate is established (Lemma 4.6.)

For positive integer N, define $[N]$ as the set $\{1, \ldots, N\}$. The partition log-likelihood $\bar{L}_P^*$ is defined for any partition Π of the indices of a lower triangular matrix,

$$\Pi : \{(i,j)\}_{i\in[N],j\in[N],i<j} \to (1, \ldots, L).$$

Define

$$S_\ell = \{(i,j) : \Pi(i,j) = \ell \text{ and } i < j\} \quad \text{and} \quad \bar{\theta_\ell} = |S_\ell|^{-1} \sum_{i<j:\Pi(i,j)=\ell} P_{ij}.$$

The partition log-likelihood is defined as

$$\bar{L}_P^*(\Pi) = \sum_{i<j} \{P_{ij} \log \bar{\theta}_{\Pi(i,j)} + (1 - P_{ij}) \log(1 - \bar{\theta}_{\Pi(i,j)})\}.$$

Notice that any class assignment z induces a corresponding partition Π^z,

$$\Pi^z(i,j) = \ell, \quad \text{where } \ell = z_i + (z_j - 1) \cdot K.$$

It is straightforward to show that $\bar{L}_P^*(\Pi^z) = \bar{L}_P(z)$.

A refinement Π' of partition Π further divides the partitions in Π into subgroups. Formally,

Definition 4.4. *A **refinement** Π' of partition Π satisfies the following condition.*

$$\Pi'(i_1,j_1) = \Pi'(i_2,j_2) \quad \Longrightarrow \quad \Pi(i_1,j_1) = \Pi(i_2,j_2), \quad \textit{for any } i_1 < j_1 \textit{ and } i_2 < j_2.$$

From Lemma A2 in [6],

$$\bar{L}_P^*(\Pi) \leq \bar{L}_P^*(\Pi') \tag{7}$$

This will be essential for for Lemma 4.6.

To define Π^*, a specific refinement partition Π^z, we first need to define a set of triples T. The following construction comes directly from [6]. "For a given membership class under z, partition the corresponding set of nodes into subclasses according to the true class assignment $\tilde{z}$ of each node. Then remove one node from each of the two largest subclasses so obtained, and group them together as a pair; continue this pairing process until no more than one nonempty subclass remains. Then, terminate. If pair (i,j) is chosen from the above procedure, then $z_i = z_j$ and $\tilde{z}_i \neq \tilde{z}_j$." Define C_1 as the number of (i,j) pairs

selected by the above routine. Notice that at least one of i or j is misclustered. In fact, $N_e(z)/2 \leq C_1 \leq N_e(z)$. This will be important for Lemma 4.5 which connects the error rate $N_e(z)/N$ with the refinement.

Define the set T to contain the triple (i, j, k) if the pair (i, j) was tallied in C_1, and $k \in [N]$ satisfies

$$D\left(P_{ik} \| \frac{P_{ik} + P_{jk}}{2}\right) + D\left(P_{jk} \| \frac{P_{ik} + P_{jk}}{2}\right) \geq C\frac{MK}{N^2}.$$

From assuming Equation 3 , if (i, j) is tallied in C_1, then there exists at least one such k. Further, if $z_k = z_\ell$, then (i, j, ℓ) is also in T. The set T is essential to defining the refinement partition Π^* and later the refined regularized partition Π^{*R}.

For each $(i, j, k) \in T$, remove (i, k) and (j, k) from their previous subset under Π^z, and place them into their own, distinct two-element set. Define the resulting partition as Π^*. Notice that it is a refinement of Π^z.

4.2 Regularized partition and regularized refinement

To extend the analysis to the RMLE, we will define the regularized partition Π^{zR} and the associated refinement partition Π^{*R}. Π^{zR} partitions the nodes into $K + 1$ groups; if $z_i = z_j$, then $\Pi^{zR}(i, j) = z_i$ and if $z_i \neq z_j$, then $\Pi^{zR}(i, j) = K + 1$. It follows from the definition of $\bar{L}_p^*$ that $\bar{L}_p^R(z) = \bar{L}_p^*(\Pi^{zR})$.

Construct Π^{*R} in the following way. For each $(i, j, k) \in T$, remove (i, k) and (j, k) from their previous subset under Π^{zR}, and place them into their own, distinct two-element set. Define the resulting partition as Π^{*R}. Notice that Π^{*R} is constructed from Π^{zR} in the same way that Π^* is constructed from Π^z. Define R as the set of elements in the off-diagonal block partition that where not removed by the set T,

$$R = \big\{(q, k) \in [N] \times [N] : z_q \neq z_k,\ (q, x, k) \notin T,\ (x, q, k) \notin T,\ \text{ for any } x \in [N]\big\}.$$

Notice that R is one group in Π^{*R}. Make a refinement Π' by subdividing R into $\binom{K}{2}$ new groups:

$$\text{For } u < v, u \in [K], v \in [K],\ \text{define } G_{uv} = \{(i, j) \in R : z_i = u, z_j = v \ \text{ or } z_i = v, z_j = u\}.$$

It follows that $\Pi' = \Pi^*$. So, Π^* is a refinement of Π^{*R} and Π^{*R} is a refinement for Π^{zR}.

Lemma 4.5. *(Theorem 3 in [6]) For any partition z and Π^* being its refinement, if the size of the smallest block $s = \Omega(\frac{MK}{N^2})$, and for any distinct class pairs (a, b), there exists a class c such that Equation 3 holds, then*

$$\bar{L}_P(\tilde{z}) - \bar{L}_P^*(\Pi^*) = \frac{N_e(z)}{N}\Omega(M). \tag{8}$$

Proof.

$$\bar{L}_p(\tilde{z}) - \bar{L}_p^*(\Pi^*) = \sum_{i<j} D(P_{ij}||\bar{\theta}_{\Pi(i,j)}) = C_1\Omega\left(s\ \frac{MK}{N^2}\right) = \frac{N_e(z)}{N}\Omega(M)$$

□

Lemma 4.6. *Let $\Pi^{\hat{z}^R}$ be the partition corresponding to $\hat{z}^R$ (the regularized block estimator). Let Π' be the refinement of $\Pi^{\hat{z}^R}$, and let $\tilde{\Pi}'^R$ be the regularized refinement of $\Pi^{\hat{z}^R}$.*

$$\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R) \geq \bar{L}_P(\tilde{z}) - \bar{L}_P^*(\Pi'^R) \geq \bar{L}_P(\tilde{z}) - \bar{L}_P^*(\Pi'). \tag{9}$$

Proof. Recall that taking a refinement increases the partition log-likelihood (see the inequality in Equation 7 or Lemma A2 in [6]).The first inequality is due to the fact that Π'^R is a refinement of the partition $\Pi^{\hat{z}^R}$. The second inequality follows from the fact that Π' is a refinement of Π'^R. □

Proof of main theorem: The conditions in Lemma 4.5 are satisfied by the HSBM assumption. By Lemma 4.3, 4.5, 4.6, we have:

$$o_p(M) = \bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R) \geq \bar{L}_P(\tilde{z}) - \bar{L}_P^*(\tilde{\Pi}) = \frac{N_e(\hat{z}^R)}{N}\Omega(M). \text{ Hence } \frac{N_e(\hat{z}^R)}{N} = o_p(1).$$

5 Discussion

The focus of this paper is on the theoretical properties of the regularized maximum likelihood estimator (RMLE) under the Highest Dimensional Stochastic Blockmodel (HSBM). Ideally, the insights gained from our analysis could be extended to computationally tractable estimators (e.g. spectral clustering) because the MLE and the RMLE are computationally intractable. In this paper we do not present any simulations because our initial attempts with simulated annealing were either highly sensitive to the initialization or extremely slow to converge to an acceptable local maximum. The development of acceptable algorithmic approximations for the MLE and the RMLE is an area for future research.

This paper proposes a new asymptotic framework (the HSBM) that aligns with several empirical observations. Most importantly, the size of the communities in the HSBM grow at a poly-logarithmic rate, not at a polynomial rate. When the community sizes grow this slowly, the number of possible out-of-block edges grows nearly quadratically with N, while the number of in-block edges grows linearly with N. If the the probability of the out-of-block edges does not decay with the size of the network, then a network sampled from the model will have drastically more out-of-block edges than in-block edges. Not only will estimation be extremely difficult (if not impossible), the sampled networks will not display the type of communities that we would find informative. Stated another way, if the nodes in a cluster are more tightly connected to the rest of the graph than with each other, is it really a cluster at all? In our highest dimensional setting, the number of free parameters

in the unrestricted parameter space grows quadratically in the number of blocks. To make estimation tractable, we maximize the likelihood over a restricted parameter space that corresponds to our assumption that out-of-block edge probabilities decay. The parameter in the restricted parameter space that maximizes the likelihood is the RMLE. Theorem 3.2 shows that under the HSBM and certain identifiability conditions, the RMLE can estimate the correct block for most nodes. Overall, this paper represents the first step in applying statistically regularized estimators towards high dimensional network analysis.

Acknowledgments

Thanks to Sara Taylor for helpful comments. Research of KR is supported by a grant from the University of Wisconsin. Research of TQ is supported by NSF Grant DMS-0906818 and NIH Grant EY09946.

References

[1] R.I.M. Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493, 1992.

[2] J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceeding of the 17th international conference on World Wide Web*, pages 695–704. ACM, 2008.

[3] P.W. Holland and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[4] P.J. Bickel and A. Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[5] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

[6] D.S. Choi, P.J. Wolfe, and E.M. Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284, 2012.

[7] P.J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):38–59, 2011.

[8] Y. Zhao, E. Levina, and J. Zhu. On consistency of community detection in networks. *Arxiv preprint arXiv:1110.3854*, 2011.

[9] C.J. Flynn and P.O. Perry. Consistent biclustering. *Arxiv preprint arXiv:1206.6927*, 2012.

[10] P. Bickel, D. Choi, X. Chang, and H. Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Arxiv preprint arXiv:1207.0865*, 2012.

[11] F. McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

[12] K. Chaudhuri, F. Chung, and A. Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research*, 2012:1–23.

[13] N.H. Bshouty and P.M. Long. Finding planted partitions in nearly linear time using arrested spectral clustering.

[14] M.W. Mahoney. Approximate computation and implicit regularization for very large-scale data analysis. *Arxiv preprint arXiv:1203.0786*, 2012.

[15] A. Chen, A.A. Amini, P.J. Bickel, and E. Levina. Fitting community models to large sparse networks. *Arxiv preprint arXiv:1207.2340*, 2012.

[16] S. Negahban, P. Ravikumar, M.J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Arxiv preprint arXiv:1010.2731*, 2010.

[17] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[18] J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

[19] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[20] P. Ravikumar, M.J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1 penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[21] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567, 2002.