Statistica Sinica **24** (2014), 1771-1786
doi:http://dx.doi.org/10.5705/ss.2013.066

THE HIGHEST DIMENSIONAL STOCHASTIC BLOCKMODEL WITH A REGULARIZED ESTIMATOR

Karl Rohe, Tai Qin and Haoyang Fan

University of Wisconsin, Madison

Abstract: In the high-dimensional Stochastic Blockmodel for a random network, the number of clusters (or blocks) K grows with the number of nodes N. Two previous studies have examined the statistical estimation performance of spectral clustering and the maximum likelihood estimator under the high-dimensional model; neither of these results 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 by empirical observations in physical anthropology (Dunbar (1992)), and in an in-depth study of massive empirical networks (Leskovec et al. (2008)). We develop a regularized maximum likelihood estimator that leverages these insights and prove that, under certain conditions, the proportion of nodes that the regularized estimator misclusters converges to zero. We thus introduce and demonstrate the advantages of statistical regularization in a parametric form for network analysis.

Key words and phrases: Consistency, high dimensional, stochastic block model, regularization, clustering.

1. Introduction

Recent advances in information technology have produced a deluge of data on complex systems with myriad interacting elements that can be represented by networks. Communities or clusters of highly connected actors are an essential feature in a multitude of empirical networks, and identifying them helps answer vital questions in various fields. Interacting elements could be metabolites, people, or computers, their interactions represented in chemical reactions, friendships, or some type of communication. For example, a terrorist cell is a cluster in the communication network of terrorists; web pages that provide hyperlinks to each other form a community that may host discussions of a similar topic; a cluster in the network of biochemical reactions might contain metabolites with similar functions and activities. Networks (or graphs) appropriately describe these relationships. Therefore, substantive questions in these various disciplines regard structure of networks. To make statistical inference from an observed network, it

is essential to evaluate the ability of clustering algorithms to estimate the "true clusters" in a network model.

The Stochastic Blockmodel is a model for a random network, the "blocks" in the model correspond to the concept of "true communities". 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 (Holland and Leinhardt (1983)). We study the maximum likelihood estimator (MLE) under the Stochastic Blockmodel.

There has been significant interest in how various clustering algorithms perform under the Stochastic Blockmodel (for example, Bickel and Chen (2009); Rohe, Chatterjee, and Yu (2011); Choi, Wolfe, and Airoldi (2012); Bickel, Chen, and Levina (2011); Zhao, Levina, and Zhu (2011); Celisse, Daudin, and Pierre (2011); Channarond, Daudin, and Robin (2011); Flynn and Perry (2013); Bickel et al. (2012); Sussman et al. (2012)). In a parallel line of research, several authors have studied clustering algorithms on the Planted Partition Model, a nearly identical model. For example, McSherry (2001) studies a spectral algorithm to recover the planted partition and analyzes the estimation performance of this algorithm. Chaudhuri, Chung, and Tsiatas (2012) improve upon this algorithm by introducing a type of regularization and proving consistency results under the planted partition model.

Two papers have studied the high-dimensional Stochastic Blockmodel, where the number of blocks K grows with the number of nodes N (Rohe, Chatterjee, and Yu (2011); Choi, Wolfe, and Airoldi (2012)). Impetus for a high-dimensional model comes from two empirical observations. Leskovec et al. (2008) found that in a large corpus of empirical networks, the tightest clusters (as judged by several popular clustering criteria) were no larger than 100 nodes, even though some of the networks had several million nodes. This result echoes similar findings in Physical Anthropology. Dunbar (1992) took various measurements of brain size in 38 different primates and found that the size of the neocortex divided by the size of the rest of the brain had a log-linear relationship with the size of the primate's natural communities. In humans, the neocortex is roughly four times larger than the rest of the brain. Extrapolating the log-linear relationship estimated from the 38 other primates, Dunbar (1992) suggested that humans do not have the social intellect to maintain stable communities larger than roughly 150 people (colloquially referred to as Dunbar's number). Leskovec et al. (2008) found a similar result in several other networks that were not composed of humans. These researches suggest that the block sizes in the Stochastic Blockmodel should not grow asymptotically. Rather, block sizes should remain fixed (or grow very slowly).

In this paper, we introduce the highest dimensional asymptotic setting that allows $K = N \log^{-5} N$. We call it the "highest" dimensional because, ignoring

the log term, K cannot grow any faster. To create a sparse graph, the out-of-block probabilities decay roughly as $\log^{\gamma} N/N$ in the highest dimensional setting, where $\gamma > 0$ is some constant. To ensure that a block's induced subgraph remains connected, the in-block probabilities are only allowed to decay slowly, like $\log^{-1} N$. We show that under this asymptotic setting, a regularized maximum likelihood estimator (RMLE) can estimate the block partition for most nodes.

This paper departs from the previous high-dimensional estimators of Rohe, Chatterjee, and Yu (2011) and Choi, Wolfe, and Airoldi (2012) by introducing a restricted parameter space for the Stochastic Blockmodel. In several high-dimensional settings, regularization restricts the full parameter space providing a path to consistent estimators (Negahban et al. (2010)). 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. In the high-dimensional regression literature, sparse regression techniques such as the LASSO restrict the parameter space to produce sparse regression estimators (Tibshirani (1996)). Several authors have suggested parameter space restrictions for high-dimensional covariance estimation, e.g., Fan, Fan, and Lv (2008), Friedman, Hastie, and Tibshirani (2008), Ravikumar et al. (2011). Parameter space restrictions have also been applied in linear discriminant analysis (Tibshirani et al. (2002)). In graph inference, previous authors have explored ways of incorporating statistical regularization into eigenvector computations (Chaudhuri, Chung, and Tsiatas (2012); Amini et al. (2012); Mahoney and Orecchia (2010); Perry and Mahoney (2011); Mahoney (2012)).

We propose restricting the parameter space for the Stochastic Blockmodel, which results in a statistically regularized estimator. We show that the RMLE is suitable in the highest dimensional asymptotic setting.

2. Preliminaries

2.1. Highest dimensional asymptotic setting

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 (Holland and Leinhardt (1983)).

For a node set $\{1, \ldots, N\}$, let P_{ij} denote the probability of including an edge linking node i and j. Let $\tilde{z} : \{1, \ldots, N\} \to \{1, \ldots, K\}$ partition the N nodes into K blocks, with$\tilde{z}_i$ 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, and $P_{ij} = \theta_{\tilde{z}_i \tilde{z}_j}$ for any $i, j = 1, \ldots, n$. Under the SBM, the probability of observing an adjacency matrix A is

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

As we only consider undirected graphs without self-loops, the product here is over $i < j$.

The highest dimensional asymptotic setting is an SBM with the following asymptotic restrictions.

(R1) For s equal to the population of the smallest block. $s = \omega(\log^\beta N), \beta > 4$, where $x_n = \omega(y_n) \Leftrightarrow y_n/x_n = o(1)$.

(R2) Let (c, d) be the interval between c and d and let Q contain a subset of the indices for $\boldsymbol{\theta}$. For constants C and $f(N) = o(s/\log N)$,

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

Assumption (R1) requires that the population of the smallest block $s = \omega(\log^\beta N), \beta > 4$. This includes the scenario where each block size is very small (e.g., $o(\log^5 N)$). In this case, the expected degree for each node is $o(\log^5 N)$. Assumption (R2) ensures that the numbers of out-of-block edges and in-block edges do not grow too fast. This setting allows a set Q to prevent this restriction from becoming too stringent; if $(a, b) \in Q$, then θ_{ab} is not required to shrink as the network grows, allowing blocks a and b to have a tight connection.

2.2. Regularized maximum likelihood estimator

Under the highest dimensional asymptotic setting, 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.

Let z denote an arbitrary partition. The log-likelihood for an observed adjacency matrix A under the SBM w.r.t. z is

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

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, $n_{ab} = N_a N_b$ if $a \neq b$ and $n_{aa} = \binom{N_a}{2}$. For an arbitrary partition z, the MLE of $\boldsymbol{\theta}$ is

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

And it is straightforward to show that

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

By substituting $\hat{\boldsymbol{\theta}}^{(\boldsymbol{z})}$ into $L(A; z, \boldsymbol{\theta})$, we get the profiled log-likelihood (Bickel and Chen (2009)) $L(A; z) = L(A; z, \hat{\boldsymbol{\theta}}^{(\boldsymbol{z})})$. Define $\hat{z} = \arg\max_z L(A; z)$ as the MLE of $\tilde{z}$. To get RMLE, let the restricted parameter space be

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

We call the new estimator "regularized" because Θ^R has only $K+1$ free parameters.

The RMLE $\boldsymbol{\theta}^{R,(z)}$ is given by

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

This optimization problem can be treated as an unconstrained optimization problem since we force the off-diagonal elements of $\boldsymbol{\theta}$ to be equal to some number r. It has the closed form solution

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

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 the same as the ordinary MLE, while the Regularized MLE for $\theta_{ab}, a \neq b$ is set to the total off-diagonal average. By substituting $\hat{\boldsymbol{\theta}}^{R,(z)}$ into $L(A; z, \boldsymbol{\theta})$, we define the regularized profile log-likelihood as

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

and denote the RMLE of the true partition $\tilde{z}$ as $\hat{z}^R = \arg\max_z L^R(A; z)$.

3. Performance of the RMLE in the Highest Dimensional Asymptotic Setting

Our main result shows that most nodes are correctly clustered by the RMLE under the highest dimensional asymptotic setting, where for any estimated class assignment z, $N_e(z)$ is 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. (Choi, Wolfe, and Airoldi (2012).)

Our result uses the KL divergence between two Bernoulli distributions,

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

and we take

$$|Q| = \sum_{\{a,b\}\in Q} n_{ab}. \tag{3.1}$$

Theorem 1. *Under the highest dimensional asymptotic setting, assume that* $|Q| = o(Ns)$*, and that for any distinct class pairs* (a,b)*, there exists a class* c *such that*

$$D\Big(\theta_{ac}\|\frac{\theta_{ac}+\theta_{bc}}{2}\Big) + D\Big(\theta_{bc}\|\frac{\theta_{ac}+\theta_{bc}}{2}\Big) \geq C\frac{MK}{N^2}, \tag{3.2}$$

then $N_e(\hat{z}^R)/N = o_p(1)$.

With $|Q| = o(Ns)$, the expected number of edges $M = \sum_{i<j} EA_{ij}$ grows slowly, specifically $M = \omega(N(\log N)^{3+\delta})$, where $\delta > 0$. The other assumption here relates to the identifiability of $\tilde{z}$.

4. Simulations

This section compares the RMLE's and the MLE's ability to estimate the block memberships in the Stochastic Blockmodel. In our simulations, the RMLE outperformed the MLE in a wide range of scenarios, particularly when there are several blocks and when the out-of-block probabilities were not too heterogeneous.

4.1. Implementation

Computing the exact RMLE and MLE is not tractable owing to the combinatorial nature of the parameter space. We fit the MLE with the pseudo-likelihood algorithm proposed in Amini et al. (2012). A slight change to the pseudo-likelihood algorithm can fit the RMLE as well: immediately after the pseudo-likelihood algorithm updates $\theta^{(z)}$, we replace the off-diagonal elements with the average of the off-diagonal elements. This often returns an estimated partition that contains empty sets. When the implementation discards a block, we reseeds a new block. This is done in an algorithm motivated by follow-up work to the current paper (see Rohe and Qin (2013)):

1. Find the block in the current iteration of the partition with the smallest empirical in-block probability.
2. For each node in this block, take its neighborhood and remove any nodes that do not connect to any other nodes in the neighborhood. Call this the transitive neighborhood.

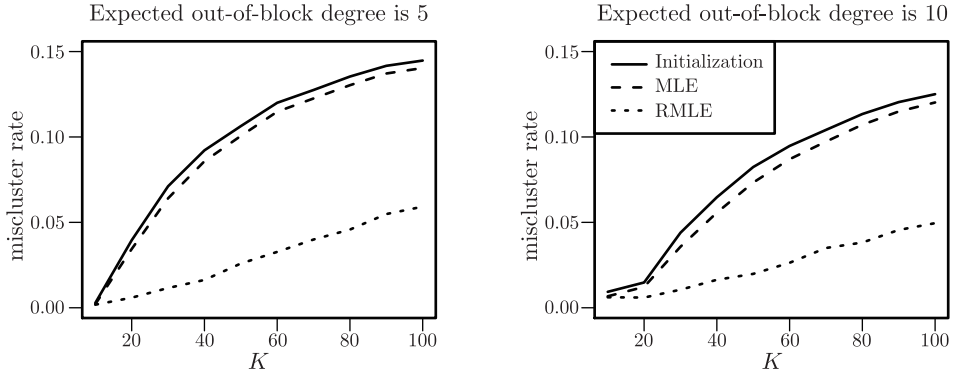

Figure 1. In the simulations, every block contains 20 nodes. Both algorithms were initialized with regularized spectral clustering and the results for this initialization are displayed by the solid line. Each point represents the average of 300 simulations. All methods were run on the same simulated adjacency matrices.

3. Combine into a new block the node with the most nodes in its transitive neighborhood with its transitive neighborhood.

We found it beneficial to do this reseeding not only when blocks disappear, but also whenever they are smaller than two nodes. Section 7 demonstrates how this reseeding provides consistently better values of the restricted likelihood.

As with the suggestion in Amini et al. (2012), we initialize the pseudo-likelihood algorithm with spectral clustering using the regularized graph Laplacian (Chaudhuri, Chung, and Tsiatas (2012)). This runs k-means on the top K eigenvectors of the matrix $D_\tau^{-1/2} A D_\tau^{-1/2}$, where $D_\tau^{-1/2}$ is a diagonal matrix whose i,ith element is $1/\sqrt{D_{ii}+\tau}$. Here $D_{ii} = \sum_j A_{ij}$ is the degree of node i, and the tuning parameter τ is set to the average degree of all nodes, as proposed in Qin and Rohe (2013).

4.2. Numerical results

This section reports on two sets of simulations. In the first, K grows while everything else remains fixed. The second investigates the sensitivity of the algorithms to heterogeneous values in the off-diagonal elements of θ.

The results in Figure 1 compare the RMLE and MLE under an asymptotic regime that keeps the population of each block fixed at twenty nodes and simply adds blocks. In both the left and right panels, the probability of a connection between two nodes in the same block was $8/20$. In the left panel, the probability of a connection between two nodes in separate blocks was $5/N$; in the right panel, it was $10/N$. In these asymptotics, the expected number of "signal" edges

connected to each node was eight, while the expected number of "noisy" edges was either five or ten.

The results in Figure 2 examine the sensitivity of the algorithms to deviations from the model in Figure 1 that makes the off-diagonal elements of θ equal to one another. In all simulations, the expected number of "signal edges" per node was eight, the expected number of "noisy edges" per node was 5, $s = 20$, and $K = 40$. On the left side of Figure 2, the off-diagonal elements of θ come from the Gamma distribution. In the top left figure, the shape parameter in the Gamma distribution (α) varies along the horizontal axis. While the shape parameter varies, the rate parameter changes to ensure that each node has an expected out-of-block degree equal to five. Under our scaling of the rate parameter, the variance of the Gamma distribution is proportional to $1/\alpha$. As such, the small values of α make the out-of-block probabilities more heterogeneous, deviating further from the implicit model. For values of α greater than 0.18, the RMLE outperforms the MLE. The bottom left plot shows the top left 400×400 submatrix of the adjacency matrix for a simulated example when $\alpha = 0.18$; the block pattern is clearly recognizable at this level of α, suggesting that the RMLE is surprisingly robust to deviations from the implicit model.

The plots on the right side of Figure 2 are similar, except the off-diagonal elements of θ are scaled Bernoulli(p) random variables. When $p = 1$, this simulation is identical to a setting in Figure 1. The scaling ensures that the expected out-of-block degree is always five. Here, the break-even point is around $p = 0.14$ and the bottom right figure shows the top left 400×400 submatrix of the adjacency matrix for a sample when $p = 0.14$; the block pattern is clearly recognizable for this level of p. In both cases, the RMLE appears robust to deviations from the implied model. At the same time, for small levels of p and α, the MLE misclusters fewer nodes than the RMLE.

5. Discussion

This paper examines the theoretical properties of the regularized maximum likelihood estimator (RMLE) under the highest dimensional asymptotic setting, showing that under relevant asymptotic regime, regularization allows for weakly consistent estimation of the block memberships.

Under the highest dimensional asymptotic setting, the size of the communities grows at a poly-logarithmic rate, not at a polynomial rate, aligning with several empirical observations (Dunbar (1992); Leskovec et al. (2008)). There are two natural implications of the block populations growing this slowly. Under any Stochastic Blockmodel, to ensure the sampled graph has sparse edges, the probability of an out-of-block connection must decay. In previous "low-dimensional"

Since θ is now random, this expectation is taken over both A and θ.

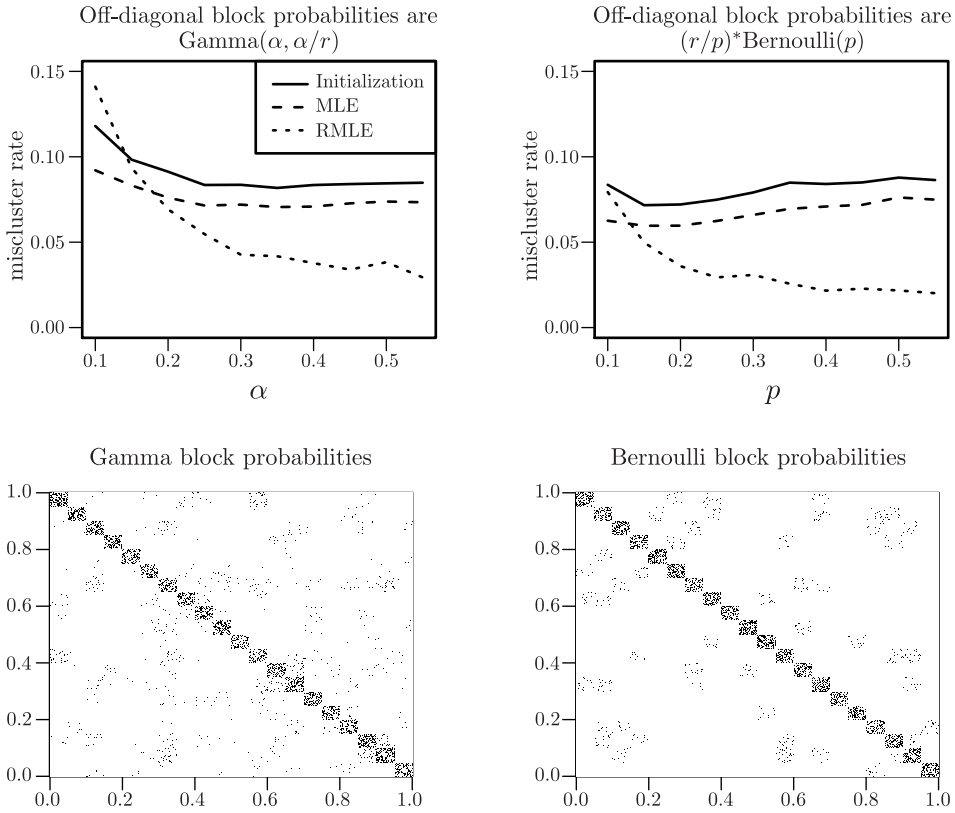

Figure 2. The top left figure displays results when elements of θ come from the Gamma distribution with varying shape parameter; the top right figure displays results when elements of θ come from the Bernoulli distribution with varying probability p. In both cases, adjustments are made so that each node has five expected out-of-block neighbors. The bottom plots illustrate the how these heterogenous probabilities manifest in the adjacency matrix; in both cases, A is sampled with the parameterization that corresponds to the break-even point between the MLE and the RMLE. Each point represents an average over 200 simulations.

analyses, it was also necessary for the probability of an in-block connection to decay. The first implication of small blocks is that the probability of an in-block connection must stay bounded away from zero. The second implication of small block sizes is that the number of off diagonal elements in Θ grows nearly quadratically with N, while the number of in-block parameters (diagonal elements of Θ) grows linearly with N.

The proposed estimator, restricts the parameter space of the SBM in a way that leverages these implications. Since the out-of-block edge probabilities decay to zero, we maximize the likelihood over a parameter space that estimates

the probabilities as equal. Theorem 1 shows that under the highest dimensional asymptotic setting and certain conditions that are similar to identifiability conditions, the RMLE can estimate the correct block for most nodes. Correspondingly, the simulation section demonstrates the advantages of the RMLE over the MLE. This paper represents a first step in applying statistically regularized estimators to high dimensional network analysis in a parametric setting. Because of the computational issues involved in computing both the MLE and the RMLE, future work will propose a "local estimator" that incorporates the insights gained from the current analysis and is computationally straight-forward.

6. Proof of the Main Result

The proof requires some additional definitions. Let the expectations of $\hat{\boldsymbol{\theta}}^{(z)}$ and $\hat{\boldsymbol{\theta}}^{R,(z)}$ be $\bar{\boldsymbol{\theta}}^{(z)}$ and $\bar{\boldsymbol{\theta}}^{R,(z)}$, and take 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 $\bar{L}_P(z)$ and $\bar{L}_P^R(z)$ be

$$\bar{L}_P(z) = \bar{L}_P(z, \bar{\boldsymbol{\theta}}^{(z)}) = \sup_{\boldsymbol{\theta} \in \Theta} \bar{L}_P(z, \boldsymbol{\theta}), \tag{6.1}$$

$$\bar{L}_P^R(z) = \bar{L}_P(z, \bar{\boldsymbol{\theta}}^{R,(z)}) = \sup_{\boldsymbol{\theta} \in \Theta^R} \bar{L}_P(z, \boldsymbol{\theta}). \tag{6.2}$$

The proof of the theorem is divided into lemmas. We bound the difference between $\bar{L}_P(\tilde{z})$ and $\bar{L}_P^R(\hat{z}^R)$ in Lemma 3. Lemma 3 divides $\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R)$ into parts as a bias-variance tradeoff; we sacrifice some bias $\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\tilde{z})$ to decrease the variance $\max_z |L^R(A; z) - \bar{L}_P^R(z)|$. it is necessary to develop the concept of regularized refinement, an extension of the refinement idea proposed in Choi, Wolfe, and Airoldi (2012). Using a 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 and Lemma 5 use a new regularized refinement to connect the bounds on the log-likelihood with the error rate $N_e(\hat{z}^R)/N$. We write $\hat{\theta}$ and $\bar{\theta}$ for $\hat{\theta}^{(z)}$ and $\bar{\theta}^{(z)}$ when the choice of z is understood.

Lemma 1. *If $M = \sum_{i<j} EA_{ij}$, then*

$$\max_z |L^R(A; z) - \bar{L}_P^R(z)| = o_p(M). \tag{6.3}$$

This proof follows a similar argument made in Choi, Wolfe, and Airoldi (2012).

Proof. Let $H(p) = -p \log p - (1-p) \log(1-p)$ and take $X = \sum_{i<j} A_{ij} \log\{\bar{\theta}_{z_i z_j} / (1 - \bar{\theta}_{z_i z_j})\}$. If n_{ab} denotes the maximum number of possible edges between all different blocks,

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

For the first terms, by a similar argument as in Choi, Wolfe, and Airoldi (2012), we have that, for every regularized estimator $\hat{\theta}^R$,

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

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 $|\hat{\Theta}| = (n_{out}+1) \cdot \Pi_{a=1}^{K}(n_{aa}+1)$. As $\sum_{a=1}(n_{aa}+1) + (n_{out}+1) = N(N-1)/2 + K + 1$, we have $|\hat{\Theta}| \leq ((N(N-1)/2(K-1)) + 1)^{K+1} \leq (N^2/2K)^{(K+1)}$. Then $\forall \epsilon > 0$,

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

As for $X - E(X)$, each $X_{ij} = A_{ij} \log\{\bar{\theta}_{z_i z_j}/(1-\bar{\theta}_{z_i z_j})\}$ is bounded in magnitude by $C = 2\log N$, we have

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

with $\sum_{i<j} E(X_{ij}^2) \leq 4M\log^2 N$. By a union bound inequality over all partitions z, we have

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

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

Lemma 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,

$$\begin{aligned}\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\tilde{z}) &= \sum_{a<b} n_{ab} D(\theta_{ab} \| \bar{r}) \\ &= \sum_{a<b, \{a,b\} \in Q} n_{ab} D(\theta_{ab} \| \bar{r}) + \sum_{a<b, \{a,b\} \notin Q} n_{ab} D(\theta_{ab} \| \bar{r}) \\ &\leq |Q| C_1 + \frac{N(N-1)}{2} - \sum_{a=1}^{K} n_{aa} - |Q|) \frac{Cf(N)}{N} (\log(CNf(N))) \\ &\leq |Q| C_1 + N^2 \frac{Cf(N)}{N} (\log N + \log Cf(N)) = o(M).\end{aligned}$$

Here $C_1 > 0$ is some constant. The last equality is due to the fact that $M = \Omega(Ns)$.

Lemma 3. *For 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. The difference is nonnegative since $\tilde{z}$ maximizes $\bar{L}_P(\cdot)$ and $\bar{L}_P(\hat{z}^R) \geq \bar{L}_P^R(\hat{z}^R)$. Adding another positive term, and using Lemma 1 and Lemma 2, we have

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

To make $N_e(z)$ mathematically tractable, Choi, Wolfe, and Airoldi (2012) introduced the concept of block refinements. The next paragraphs extend this definition to the regularized block refinement.

6.1. Partitions and refinements

Refinement is the key concept to connect $\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R)$ with the error rate $N_e(\hat{z}^R)/N$. We review the concept of partition and refinement, then give its regularized version.

For positive integer N, take $[N]$ 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, $\Pi : \{(i,j)\}_{i \in [N], j \in [N], i<j} \to (1, \ldots, L)$, as

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

The partition log-likelihood is defined as

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

Any class assignment z induces a corresponding partition Π^z, $\Pi^z(i,j) = \ell$, where $\ell = z_i + (z_j - 1) \cdot K$, and 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, so $\Pi'(i_1, j_1) = \Pi'(i_2, j_2) \Longrightarrow \Pi(i_1, j_1) = \Pi(i_2, j_2)$, for any $i_1 < j_1$ and $i_2 < j_2$. From Lemma A2 in Choi, Wolfe, and Airoldi (2012),

$$\bar{L}_P^*(\Pi) \leq \bar{L}_P^*(\Pi'). \tag{6.4}$$

To define Π^*, a specific refinement of partition Π^z, we first need to define a set of triples T. The following construction comes directly from Choi, Wolfe, and Airoldi (2012):

> "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 $\tilde{z}_i \neq \tilde{z}_j$."

Take C_1 as the number of (i, j) pairs selected by the above routine. Here at least one of i or j is misclustered, and $N_e(z)/2 \leq C_1 \leq N_e(z)$. 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 (3.2), 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. 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.

6.2. Regularized partition and regularized refinement

To extend the analysis to the RMLE, we 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} as follows. 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. The resulting partition is Π^{*R}. Take

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

Here R is a group in Π^{*R}, make a refinement Π' by subdividing R into $\binom{K}{2}$ new groups:

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

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

Lemma 4 (Choi, Wolfe, and Airoldi (2012))**.** *For any partition z and Π^* being its refinement, if the size of the smallest block $s = \Omega(MK/N^2)$, and for any distinct class pairs (a,b), there exists a class c such that Equation* (3.2) *holds, then*

$$\bar{L}_P(\tilde{z}) - \bar{L}_P^*(\Pi^*) = \frac{N_e(z)}{N}\Omega(M). \tag{6.5}$$

Lemma 5. *Let $\Pi^{\hat{z}^R}$ be the partition corresponding to $\hat{z}^R$. If Π' is the refinement of $\Pi^{\hat{z}^R}$, and Π'^R is the regularized refinement of $\Pi^{\hat{z}^R}$,*

$$\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R) \geq \bar{L}_P(\tilde{z}) - \bar{L}_P^*(\Pi'^R) \geq \bar{L}_P(\tilde{z}) - \bar{L}_P^*(\Pi'). \tag{6.6}$$

Proof. The first inequality holds since Π'^R is a refinement of the partition $\Pi^{\hat{z}^R}$, the second since Π' is a refinement of Π'^R.

Proof of main theorem. The conditions in Lemma 4 are satisfied by the highest dimensional asymptotic setting assumption. By Lemmas 3, 4 and 5, we have

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

7. Reseeding

Section 4 describes a reseeding technique that ensures that the pseudo-likelihood implementation of the RMLE returns an estimated partition with the desired number of non-empty sets. We compareed the implementation with reseeding (`reseed`) to the implementation without reseeding (`no.reseed`). Overall, `reseed` never attains a smaller likelihood score and often attains a larger likelihood scores. Moreover, `reseed` is more stable over different initializations.

In the following simulation, $K = 30$, $n = 600$, $\theta_{ii} = 8/20$ for all i, and $\theta_{ij} = 10/580$ for all $i \neq j$. So, in expectation, each node connects to 8 nodes in the same block and 10 nodes in other blocks. For each simulated adjacency matrix A, both `reseed` and `no.reseed` were initialized 50 times with spectral clustering. The simulations in Section 4 reseeded whenever a block contains either zero nodes or a single node. In this simulation, blocks were reseeded whenever they had fewer than five nodes.

There were 175 adjacency matrices A simulated from this model. For the ith simulated adjacency matrix, $\hat{z}^i_{reseed}$ was the partition that attained the largest likelihood over all 50 initializations of `reseed`. Similarly, $\hat{z}^i_{no.reseed}$ was the same partition for `no.reseed`. Over the 175 simulated adjacency matrices, $L^R(A; \hat{z}^i_{reseed}) > L^R(A; \hat{z}^i_{no.reseed})$ on 22% of the simulations. In the remaining simulations, they found the same maximum. Never did $\hat{z}^i_{no.reseed}$ attain a larger likelihood score. Moreover, on each initialization, the `reseed` was much more likely to find the maximum (72 % compared to 14 %).

Acknowledgements

Thanks to Sara Fernandes-Taylor for helpful comments. Research of KR is supported by DMS-1309998 and grants from the University of Wisconsin. Research of TQ is supported by NSF Grant DMS-0906818 and NIH Grant EY09946. Thank you to the anonymous referees and the associate editor for their thoughtful comments that have added to the quality of this research. Research supported in part by NIH Grant EY09946, NSF Grant DMS-0906818, and grants from WARF.

References

Amini, A. A., Chen, A., Bickel, P. J. and Levina, E. (2012). Pseudo-likelihood methods for community detection in large sparse networks. arXiv preprint arXiv:1207.2340.

Bickel, P.J. and Chen, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Nat. Acad. Sci.*, **106**, 21068-21073.

Bickel, P. J., Chen, A. and Levina, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.* **39**, 38-59.

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

Celisse, A., Daudin, J.-J. and Pierre, L. (2011). Consistency of maximum-likelihood and variational estimators in the stochastic block model. arXiv preprint arXiv:1105.3288.

Channarond, A., Daudin, J.-J. and Robin, S. (2011). title=Classification and estimation in the Stochastic Block Model based on the empirical degrees. arXiv preprint arXiv:1110.6517.

Chaudhuri, K., Chung, F. and Tsiatas, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. *J. Machine Learn. Res.* 1-23.

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

Dunbar, R. I. M. (1992). Neocortex size as a constraint on group size in primates. *J. Human Evolution* **22**, 469-493.

Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* **147**, 186-197.

Flynn, C. J. and Perry, P. O. (2013). Consistent Biclustering. Arxiv preprint arXiv:1206.6927.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432-441.

Holland, P.W. and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5**, 109-137.

Leskovec, J., Lang, K. J., Dasgupta, A. and Mahoney, M. W. (2008), Statistical properties of community structure in large social and information networks. Proceeding of the 17th international conference on World Wide Web, 695-704.

Mahoney, M. W. (2012). Approximate computation and implicit regularization for very large-scale data analysis. Arxiv preprint arXiv:1203.0786.

Mahoney, M. W. and Orecchia, L. (2010). Implementing regularization implicitly via approximate eigenvector computation, arXiv preprint arXiv:1010.0703.

McSherry, F. (2001). Spectral partitioning of random graphs. Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on, 529-537.

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

Perry, P. O. and Mahoney, M. W. (2011). Regularized Laplacian estimation and fast eigenvector approximation. arXiv preprint arXiv:1110.1757.

Qin, T. and Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. arXiv preprint arXiv:1309.4111.

Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1 penalized log-determinant divergence. *Electronic J. Statist.* **5**, 935-980.

Rohe, K., Chatterjee, S. and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878-1915.

Rohe, K. and Qin, T. (2013). The blessing of transitivity in sparse and stochastic networks. arXiv preprint arXiv:1307.2302.

Sussman, D. L., Tang, M., Fishkind, D. E. and Priebe, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc.* **2012 107**, 1119-1128.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* (Methodological), 267-288.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Nat. Acad. Sci.* **99**, 65-67.

Zhao, Y., Levina, E. and Zhu, J. (2011). On Consistency of Community Detection in Networks. Arxiv preprint arXiv:1110.3854.

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA.

E-mail: karlrohe@stat.wisc.edu

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA.

E-mail: qin@stat.wisc.edu

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA.

E-mail: haoyang@stat.wisc.edu

(Received March 2013; accepted December 2013)