The Annals of Statistics
2016, Vol. 44, No. 1, 401–424
DOI: 10.1214/15-AOS1370

A GOODNESS-OF-FIT TEST FOR STOCHASTIC BLOCK MODELS

By Jing Lei

Carnegie Mellon University

The stochastic block model is a popular tool for studying community structures in network data. We develop a goodness-of-fit test for the stochastic block model. The test statistic is based on the largest singular value of a residual matrix obtained by subtracting the estimated block mean effect from the adjacency matrix. Asymptotic null distribution is obtained using recent advances in random matrix theory. The test is proved to have full power against alternative models with finer structures. These results naturally lead to a consistent sequential testing estimate of the number of communities.

1. Introduction. Large-scale network data with community structures have been the focus of much research efforts in the past decade [see, e.g., Newman and Girvan (2004), Newman (2006)]. In the statistics and machine learning literature, the stochastic block model (Holland, Laskey and Leinhardt, 1983) is a very popular model for community structures in network data. In a stochastic block model, the observed network is often recorded in the form of an $n \times n$ adjacency matrix A, representing the presence/absence of pairwise interactions among n individuals in a population of interest. The model assumes that (i) the individuals are partitioned into K disjoint communities, and (ii) given the memberships, the upper diagonal entries of A are independent Bernoulli random variables, where the parameter $E(A_{ij})$ depends only on the memberships of nodes i and j. Such a model naturally captures the community structures commonly observed in complex networks, and has close connection to nonparametric exchangeable random graphs (Bickel and Chen, 2009). The stochastic block model can be made more realistic by incorporating additional parameters to better approximate real world network data. For example, Karrer and Newman (2011) incorporated individual node activeness into the stochastic block model to allow

Received December 2014; revised August 2015.
AMS 2000 subject classifications. 62H15.
Key words and phrases. Network data, stochastic block model, goodness-of-fit test, consistency, Tracy–Widom distribution.

for arbitrary degree distributions. In the mixed membership model (Airoldi et al., 2008), each individual may belong to more than one community.

In this paper, we develop a goodness-of-fit test for stochastic block models. Given an adjacency matrix A and a positive integer K_0, we test whether A can be adequately fitted by a stochastic block model with K_0 communities. Our test statistic is the largest singular value of a residual matrix obtained by removing the estimated block mean effect from the observed adjacency matrix. Intuitively, if A is generated by a stochastic block model and the block mean effect is estimated appropriately, the residual matrix will approximate a generalized Wigner matrix: a symmetric random matrix with independent mean zero upper diagonal entries. Our first contribution is the asymptotic null distribution of the test statistic (Theorem 3.1). The proof uses some recent advances in random matrix theory, such as the local semicircle law of generalized Wigner matrices and its consequences (Erdős, Yau and Yin, 2012, Erdős et al., 2013a, Bloemendal et al., 2014). Our second contribution is asymptotic power guarantee of the test against models with finer structures (Theorems 3.3 and 3.5). In particular, we establish the growth rate of the test statistic under alternatives that correspond to stochastic block models with more communities or with individual node degree variations. It is of particular interest to consider alternative stochastic block models with more communities because any exchangeable random graph can be approximated by a stochastic block model (Bickel and Chen, 2009). In our simulation study, we observe that the proposed test is powerful against not only stochastic block models with more communities, but also other network models with finer structures such as the degree corrected block model and the mixed membership block model.

A related test statistic using the largest eigenvalue of the centered and scaled adjacency matrix has been studied in Bickel and Sarkar (2013). They derive asymptotic null distribution for Erdős–Rényi models, which corresponds to a stochastic block model with one community. We generalize their argument to prove the asymptotic null distribution result for stochastic block models with more than one community. The key step is to bound the fluctuation in the leading eigenvalue of a random matrix under perturbation of a block-wise constant noise matrix. Moreover, their asymptotic power analysis requires the alternative model to be diagonal dominant. Our test statistic uses the largest singular value of the residual matrix, so we are able to capture signals affecting either the largest or the smallest eigenvalues, and our asymptotic power guarantee holds for a much wider class of alternative models.

Our goodness-of-fit test can also serve as a main building block to estimate the number of communities. As a key inference problem in stochastic block models and its variants, the community recovery problem concerns

estimating the hidden communities from a single observed adjacency matrix [see McSherry (2001), Bickel and Chen (2009), Decelle et al. (2011), Zhao, Levina and Zhu (2012), Jin (2012), Fishkind et al. (2013), Lei and Rinaldo (2013), Chen, Sanghavi and Xu (2012), Chaudhuri, Chung and Tsiatas (2012), Krzakala et al. (2013), Massoulie (2013), Mossel, Neeman and Sly (2013), Abbe, Bandeira and Hall (2014), Anandkumar et al. (2014), e.g.]. A common assumption made in all these methods is that K, the total number of communities, is known. Therefore, estimating the number of communities in a stochastic block model is of great practical and theoretical importance. Some methods have been proposed to estimate the number of communities in stochastic block models (Zhao, Levina and Zhu, 2011, Bickel and Sarkar, 2013, Chen and Lei, 2014, Saldana, Yi and Feng, 2014), but without consistency guarantee.

To estimate the number of communities, we consider hypothesis test

$$H_{0,K_0} : K = K_0, \quad \text{against} \quad H_{\mathrm{a},K_0} : K > K_0 \tag{1}$$

sequentially for each $K_0 \geq 1$ until the null hypothesis is not rejected. We prove the consistency of this sequential testing estimator in Corollary 3.4 of Section 3. Throughout this paper, we use K to denote the true number of communities in a stochastic block model and use K_0 to denote a hypothetical number of communities.

Recently, Chatterjee (2015) studied a general method for matrix denoising using singular value thresholding, which covers the stochastic block model as a special case. Following the ideas developed there, one may use the number of significant singular values, for example, those greater than $\sqrt{n}$, of the adjacency matrix as an estimate of K. But this method only works when the community-wise connectivity matrix has full rank. Empirically, we also find that the sequential testing estimator developed in this paper performs better than singular value thresholding for sparse networks.

Glossary. For a square matrix M, $\operatorname{diag}(M)$ denotes the diagonal matrix induced by M. For any $n \times n$ symmetric matrix M, $\lambda_j(M)$ denotes its jth largest eigenvalue value, ordered as $\lambda_1(M) \geq \lambda_2(M) \geq \cdots \geq \lambda_n(M)$, and $\sigma_1(M)$ is the largest singular value. Denote $\mathcal{B}_K$ the set of all $K \times K$ symmetric matrices with entries in $(0, 1)$ and all rows being distinct.

2. Stochastic block models and a goodness-of-fit test. A stochastic block model on n nodes with K communities is parameterized by a membership vector $g \in \{1, \ldots, K\}^n$ and a symmetric community-wise edge probability matrix $B \in [0, 1]^{K \times K}$. The observed adjacency matrix A is a symmetric binary matrix with diagonal entries being 0. Given (g, B), the probability mass function for the adjacency matrix A is

$$P_{g,B}(A) = \prod_{1 \leq i < j \leq n} B_{g_i g_j}^{A_{ij}} (1 - B_{g_i g_j})^{(1 - A_{ij})}. \tag{2}$$

In other words, given (g, B), the edges are independent Bernoulli random variables with parameters determined by the node memberships.

To avoid triviality, we say that a stochastic block model parameterized by (g, B) has K communities if (i) g contains all K distinct values in $\{1, \ldots, K\}$, and (ii) any two rows of B are distinct. A stochastic block model is identifiable up to a label permutation on g and a corresponding row/column permutation on B.

Given an observed adjacency matrix A, and a positive integer K_0, we would like to know if A can be well fitted by a stochastic block model with K_0 communities. If we assume that A is generated by a stochastic block model with K communities, this leads to a goodness-of-fit test for stochastic block models with a composite null hypothesis

$$H_{0,K_0} : K = K_0. \tag{3}$$

To derive a goodness-of-fit test for stochastic block models, a natural idea is to estimate the model parameters and remove the signal from the observed adjacency matrix, and test whether the residual matrix looks like a noise matrix. To this end, consider the $n \times n$ matrix P given by

$$P_{ij} = B_{g_i g_j},$$

so that $E(A) = P - \operatorname{diag}(P)$. Let $\tilde{A}^*$ be

$$\tilde{A}^*_{ij} = \frac{A_{ij} - P_{ij}}{\sqrt{(n-1)P_{ij}(1-P_{ij})}}, \qquad i \neq j \quad \text{and} \quad \tilde{A}^*_{ii} = 0, \forall i. \tag{4}$$

Then $\tilde{A}^*$ is a generalized Wigner matrix, satisfying $E(\tilde{A}^*_{ij}) = 0$ for all (i, j) and $\sum_j \operatorname{var}(\tilde{A}^*_{ij}) = 1$ for all i. The asymptotic distribution of the extreme eigenvalues of $\tilde{A}^*$ has been well studied in random matrix theory. In particular, combining results in Erdős, Yau and Yin (2012) and Lee and Yin (2014) we have

$$n^{2/3}[\lambda_1(\tilde{A}^*) - 2] \rightsquigarrow TW_1 \quad \text{and} \quad n^{2/3}[-\lambda_n(\tilde{A}^*) - 2] \rightsquigarrow TW_1, \tag{5}$$

where TW_1 denotes the Tracy–Widom distribution with index 1 and "$\rightsquigarrow$" denotes convergence in distribution. We remark that (5) cannot be obtained using results for standard Wigner matrices as the diagonal entries of $\tilde{A}^*$ are fixed to be 0. We formally state and prove this result as Lemma A.1 in Section A.1.

The matrix $\tilde{A}^*$ involves unknown model parameters and cannot be used as a test statistic. Now we describe a natural estimate of $\tilde{A}^*$ by plugging in an estimated stochastic block model.

Let $\hat{g}$ be an estimated community membership vector with target number of communities being K_0. Define $\hat{\mathcal{N}}_k = \{i : 1 \leq i \leq n, \hat{g}_i = k\}$, and $\hat{n}_k = |\hat{\mathcal{N}}_k|$

for all $1 \le k \le K_0$. We consider the plug-in estimator of B:

$$\hat{B}_{kl} = \begin{cases} \dfrac{\sum_{i \in \hat{\mathcal{N}}_k, j \in \hat{\mathcal{N}}_l} A_{ij}}{\hat{n}_k \hat{n}_l}, & k \neq l, \\ \dfrac{\sum_{i,j \in \hat{\mathcal{N}}_k, i<j} A_{ij}}{\hat{n}_k(\hat{n}_k - 1)/2}, & k = l. \end{cases} \tag{6}$$

The estimates $(\hat{g}, \hat{B})$ leads to the empirically centered and re-scaled adjacency matrix $\tilde{A}$:

$$\tilde{A}_{ij} = \begin{cases} \dfrac{A_{ij} - \hat{P}_{ij}}{\sqrt{(n-1)\hat{P}_{ij}(1 - \hat{P}_{ij})}}, & i \neq j, \\ 0, & i = j, \end{cases} \tag{7}$$

where

$$\hat{P}_{ij} = \hat{B}_{\hat{g}_i \hat{g}_j}. \tag{8}$$

It is natural to conjecture that under the null hypothesis $K = K_0$ and when the estimates $(\hat{g}, \hat{B})$ are accurate enough, the convergence in (5) will carry over to the corresponding eigenvalues of $\tilde{A}$. Therefore, we can use the largest singular value of $\tilde{A}$, after centering and scaling, as our test statistic:

$$T_{n,K_0} = n^{2/3}[\sigma_1(\tilde{A}) - 2]. \tag{9}$$

The corresponding level α rejection rule for testing problem (3) is

$$\text{Reject} \quad H_{0,K_0}, \quad \text{if } T_{n,K_0} \ge t(\alpha/2), \tag{10}$$

where $t(\alpha/2)$ is the $\alpha/2$ upper quantile of the TW_1 distribution for $\alpha \in (0,1)$. We use $t(\alpha/2)$ instead of $t(\alpha)$ for Bonferroni correction because

$$\sigma_1(\tilde{A}) = \max(\lambda_1(\tilde{A}), -\lambda_n(\tilde{A})),$$

and hence

$$T_{n,K_0} = \max[n^{2/3}(\lambda_1(\tilde{A}) - 2), n^{2/3}(-\lambda_n(\tilde{A}) - 2)].$$

A similar result concerning the largest eigenvalue of $\tilde{A}$ in the simple case $K_0 = 1$ has been obtained in Bickel and Sarkar (2013). In Section 3 below, we formally state and prove the validity of our test statistic T_{n,K_0} in Theorem 3.1 by establishing the asymptotic null distributions of both the largest and smallest eigenvalues and for general values of K_0.

The use of $\sigma_1(\tilde{A})$ instead of $\lambda_1(\tilde{A})$ as our test statistic in (9) is crucial for power guarantee. Under some alternative hypotheses, the signal may be carried solely by $\lambda_n(\tilde{A})$. For example, consider a model with two equal-sized

communities and $B_{11} = B_{22} = 1/4$, $B_{12} = B_{21} = 1/2$. Suppose we would like to test $H_0 : K = K_0 = 1$. In this case, $A - \hat{P}$ has block-wise mean

$$\begin{pmatrix} -1/8 & 1/8 \\ 1/8 & -1/8 \end{pmatrix},$$

which has no positive eigenvalues. Therefore, the test using only $\lambda_1(\tilde{A})$ has no power.

Given the rejection rule (10) for testing problem (3), we have the following sequential testing estimator of K:

$$\hat{K} = \inf\{K_0 \geq 1 : T_{n,K_0} < t_n\}. \tag{11}$$

In other words, we perform the goodness-of-fit test for $K_0 = 1, 2, \ldots,$ until failing to reject H_{0,K_0}. We prove consistency of $\hat{K}$ for appropriate choices of t_n in Corollary 3.4 below, as a consequence of (i) a large deviation inequality of the extreme eigenvalues of $\tilde{A}$ under the null hypothesis $K = K_0$, and (ii) the growth rate of T_{n,K_0} under the alternative hypothesis $K > K_0$.

3. Asymptotic null distribution and power. The asymptotic distribution of the test statistic T_{n,K_0} under the null hypothesis depends on the accuracy of the estimated community membership $\hat{g}$. In order to consider the asymptotic behavior of community recovery, we consider a sequence of stochastic block models $\{(g^{(n)}, B^{(n)}) : n \geq 1\}$ where $g^{(n)} \in \{1, \ldots, K^{(n)}\}^n$ for each n, and $B^{(n)} \in \mathcal{B}_{K^{(n)}}$. Here, the number of communities $K = K^{(n)}$ and the community-wise edge probability matrix $B = B^{(n)}$ are allowed to change with n.

We will focus on relatively balanced communities.

(A1) There exists $c_0 > 0$ such that $\min_{1 \leq k \leq K^{(n)}} |\{i : g_i^{(n)} = k\}| \geq c_0 n / K^{(n)}$ for all n.

Assumption (A1) assumes each community has size at least proportional to $n/K^{(n)}$. For example, it is satisfied almost surely if the membership vector $g^{(n)}$ is generated from a multinomial distribution with n trials and probability $\pi = (\pi_1, \ldots, \pi_{K^{(n)}})$ such that $\min_{1 \leq k \leq K} \pi_k > c_0 / K^{(n)}$ and $K^{(n)}$ grows slowly.

Definition (Consistency of community recovery). For a sequence of stochastic block models $\{(g^{(n)}, B^{(n)}) : n \geq 1\}$ with $K^{(n)}$ communities and $B^{(n)} \in \mathcal{B}_{K^{(n)}}$, we say a community membership estimator $\hat{g} = \hat{g}(A, K^{(n)})$ is consistent if

$$P_{A \sim (g^{(n)}, B^{(n)})}(\hat{g} = g^{(n)}) \to 1.$$

REMARK. The notion "$\hat{g} = g$" shall be interpreted as being equal up to a label permutation. Such a label permutation does not affect our methodological and theoretical development so we will assume that the label permutation is identity for simplicity. The definition of consistent community recovery can be satisfied by several methods. For example, in the case of fixed finite $K^{(n)} = K$ and $B^{(n)} = B$, the profile likelihood method (Bickel and Chen, 2009) is consistent for all $(g^{(n)} : n \geq 1)$ satisfying (A1) and all $B \in \mathcal{B}_K$; the spectral clustering method can be made consistent, with slight modification, for all $(g^{(n)} : n \geq 1)$ satisfying (A1) and $B \in \mathcal{B}_K$ with full rank (McSherry, 2001, Vu, 2014, Lei and Zhu, 2014). In the case of slowly growing $K^{(n)}$, consistent community recovery can be achieved in some special cases such as the planted partition model (Chaudhuri, Chung and Tsiatas, 2012, Amini and Levina, 2014).

3.1. *The asymptotic null distribution.*

THEOREM 3.1 (Asymptotic null distribution). *Let A be an adjacency matrix generated from stochastic block model $(g^{(n)}, B^{(n)})$, where $B^{(n)} \in \mathcal{B}_{K^{(n)}}$ and $(g^{(n)} : n \geq 1)$ satisfies condition* (A1). *Let $\tilde{A}$ be given as in (7) using a consistent community estimate $\hat{g}$ and corresponding plug-in estimate of $B^{(n)}$ as in (6). Assume in addition that $K^{(n)} = O(n^{1/6-\tau})$ for some $\tau > 0$, and the entries of $B^{(n)}$ are uniformly bounded away from 0 and 1. The following holds under the null hypothesis $K^{(n)} = K_0$:*

$$n^{2/3}(\lambda_1(\tilde{A}) - 2) \rightsquigarrow TW_1, \qquad n^{2/3}(-\lambda_n(\tilde{A}) - 2) \rightsquigarrow TW_1. \tag{12}$$

Theorem 3.1 is proved in Section A.2. The main challenge is that, assuming $\hat{g} = g$, the entry-wise estimation error in $\hat{B}$ is of order $K^{(n)}/n$. The simple upper bound of $\tilde{A} - \tilde{A}^*$ in Frobenius norm is of order $K^{(n)}n^{-1/2}$ which exceeds the $n^{-2/3}$ scaling required in (12). Our proof establishes (12) using a more delicate analysis that exploits the block-wise constant structure in $\tilde{A} - \tilde{A}^*$, combined with random matrix theory results which ensure that (i) the leading eigenvectors of $\tilde{A}^*$ are delocalized, in the sense that the chance these eigenvectors being close to any fixed vector is small, and (ii) the number of large eigenvalues of $\tilde{A}^*$ in an interval of length $K^{(n)}/\sqrt{n}$ can be accurately approximated. This result is a nontrivial generalization of Theorem 2.1 in Bickel and Sarkar (2013).

An immediate consequence of Theorem 3.1 is an asymptotic type I error bound for the rejection rule (10):

$$\begin{aligned} &P[T_{n,K_0} \geq t(\alpha/2)] \\ &\quad \leq P[n^{2/3}(\lambda_1(\tilde{A}) - 2) \geq t(\alpha/2)] + P[n^{2/3}(-\lambda_n(\tilde{A}) - 2) \geq t(\alpha/2)] \\ &\quad = \alpha/2 + o(1) + \alpha/2 + o(1) = \alpha + o(1). \end{aligned}$$

Formally, we have the following corollary.

COROLLARY 3.2 (Asymptotic type I error control). *Under the assumptions of Theorem 3.1, the rejection rule in (10) has asymptotic level* α.

3.2. *Asymptotic power against* $K > K_0$ *and consistent estimation of* K. Now we consider the power of the test against finer stochastic block models. The following theorem provides a lower bound of the growth rate of the test statistic T_{n,K_0} under the alternative model $K^{(n)} > K_0$.

THEOREM 3.3 (Asymptotic power guarantee). *Let* A *be an adjacency matrix generated from stochastic block model* $(g^{(n)}, B^{(n)})$ *with* $B^{(n)} \in \mathcal{B}_{K^{(n)}}$ *and* $(g^{(n)} : n \geq 1)$ *satisfying condition* (A1). *Let* δ_n *be the smallest* ℓ_∞ *distance among all pairs of distinct rows of* $B^{(n)}$. *For any* $K_0 < K^{(n)}$ *and any community estimator* $\hat{g}$, *we have*

$$\sigma_1(\tilde{A}) \geq \tfrac{1}{2}\delta_n c_0 [K^{(n)}]^{-2} n^{1/2} + O_P(1).$$

Theorem 3.3 is powerful in that it puts no structural condition on the connectivity matrix $B^{(n)}$, nor does it make any assumption about the particular method used to estimate the membership. Theorem 3.3 is proved in Section A.3. The main idea is that if the nodes are partitioned into less than $K^{(n)}$ groups, the corresponding block partition of the expected adjacency matrix cannot be block-wise constant, and hence it is impossible to remove the mean effect by subtracting a constant from each estimated block submatrix of A.

When $B^{(n)} = B$ and $K^{(n)} = K$ are fixed and do not change with n, the community separation parameter δ_n is constant and Theorem 3.3 gives a growth rate at least $n^{1/2}$. When $K^{(n)}$ is allowed to grow with n, consistent community recovery can be achieved for the planted partition model where $B^{(n)}_{kk} = p$ and $B^{(n)}_{kk'} = q$ $(k \neq k')$ for some $0 \leq q < p \leq 1$. If p and q are constants independent of n, then $\delta_n = p - q$ is also a constant. Therefore, in this case Theorem 3.3 says that the growth rate of T_{n,K_0} is at least $[K^{(n)}]^{-2} n^{1/2}$.

The asymptotic null distribution and growth rate under alternative $K^{(n)} > K_0$ suggest that the null and alternative hypotheses are well separated. Therefore, if in the sequential testing estimator (11) we choose the rejection threshold t_n to increase with the network size n, we shall expect to have a consistent estimate of $K^{(n)}$.

THEOREM 3.4 (Consistency of estimating K). *Under the assumptions of Theorem 3.1 and Theorem 3.3, assume in addition that* $\liminf_{n\to\infty} \delta_n > 0$. *Let* $\hat{K}$ *be the sequential testing estimator given in (11) with threshold* t_n *satisfying* $t_n \asymp n^{\varepsilon}$ *for some* $\varepsilon \in (0, 5/6)$, *then*

$$P(\hat{K} = K^{(n)}) \to 1.$$

Corollary 3.4 is proved in Section A.3. We note that the asymptotic null distribution given in Theorem 3.1 cannot be directly used to bound the probability of $P(T_{n,K^{(n)}} \geq t_n)$ because t_n changes with n. Instead, we need to use a tail probability bound on the largest singular value of $\tilde{A}$ (Lemma A.4). The condition that δ_n is bounded away from zero is satisfied both when $B^{(n)}$ is fixed or when $B^{(n)}$ is given by a planted partition model with constant diagonal and off-diagonal edge probabilities. This condition can be relaxed to requiring δ_n to decay no faster than $n^{-1/6}$ and having $t_n \ll n^{5/6}\delta_n$.

3.3. *Asymptotic power against degree corrected block models.* The goodness-of-fit test (10) is also powerful against certain degree corrected block models. A degree corrected block model is parameterized by a triplet (g, B, ψ), where $\psi \in [0,1]^n$ is the node activeness parameter and the edge probability between nodes i and j is $\psi_i\psi_j B_{g_i g_j}$. The probability mass function of the observed adjacency matrix is

$$P_{g,B,\psi}(A) = \prod_{1\leq i<j\leq n} (\psi_i\psi_j B_{g_i g_j})^{A_{ij}}(1-\psi_i\psi_j B_{g_i g_j})^{1-A_{ij}}.$$

Let ϕ_k be the subvector of ψ corresponding to the entries in community k, and $\tilde{\phi}_k = \phi_k/\|\phi_k\|$.

The condition we will need on ψ is that there exists a community whose node activeness parameters cannot be approximated by block-wise constant vectors. Formally, for any vector v and positive integer L let

$$\mathcal{E}(v, L) = \min\{\|v-u\|_2^2 : \text{the entries of } u \text{ take at most } L \text{ distinct values}\}.$$

THEOREM 3.5. *Let A be generated by a degree corrected block model (g, B, ψ) on n nodes and K communities. If there exists $1 \leq k^* \leq K$ such that $\mathcal{E}(\tilde{\phi}_{k^*}, K_0) > 0$, then for any estimator $(\hat{g}, \hat{B})$ of a K_0-stochastic block model, we have*

$$\|\tilde{A}\| \geq \frac{\mathcal{E}(\tilde{\phi}_{k^*}, K_0)}{2K_0^{3/2}}\|B_{k^*,\cdot}\|_\infty \kappa_n n^{1/2} + O_P(1),$$

where $\kappa_n = \min_{1\leq k\leq K}\|\phi_k\|^2/n$ and $\|B_{k^*,\cdot}\|_\infty = \max_{1\leq k\leq K} B_{k^*,k}$.

Theorem 3.5 is proved in Section A.4. The quantity $\mathcal{E}(\tilde{\psi}_{k^*}, K_0)$ reflects the idea that there exists at least one community whose node activeness cannot be approximated by a simple K_0-block structure. Recall that for each k, $\tilde{\phi}_k$ is a vector with unit ℓ_2 norm. If the entries of ϕ_k are sampled from a compact interval with strictly positive density, then $\mathcal{E}(\tilde{\phi}_k, K_0) \asymp K_0^{-2}$ when K_0 is small compared to the length of ϕ_k. When K_0 increases, $\mathcal{E}(\tilde{\phi}_k, K_0)$ decreases for all k and the test will be less powerful. This agrees with the

fact that any degree corrected block model can be approximated by regular stochastic block models with a large number of communities.

Consider an opposite case, where A is generated by a degree corrected block model with one community and degree parameter vector ψ containing only K_0 distinct values. Here, the model can also be viewed as a regular stochastic block model with K_0 communities. Then $\mathcal{E}(\tilde{\psi}_{k^*}, K_0) = 0$ and the test will not tend to reject the null hypothesis, provided that a consistent community recovery method is used.

The quantity κ_n acts as a lower bound on the overall node activeness. A larger value of κ_n usually leads to better power because there are more observed edges for inference. Under the balanced community assumption (A1), $\kappa_n \asymp K^{-1}$ if the entries of ψ are uniformly bounded away from zero, or are sampled from a common distribution independent of n.

Applying Theorem 3.5 in the simple special case where $B \in \mathcal{B}_K$ (and hence K) is fixed and ψ_i's are sampled from a compact interval with strictly positive density, under Assumption (A1) we have, for any given K_0,

$$\|\tilde{A}\| \geq Cn^{1/2} + O_P(1).$$

Therefore, with probability tending to one, the test will reject the null hypothesis that A is generated from a regular stochastic block model with K_0 blocks. If K grows with n and the entries of B scale at rate ρ_n, the test is still powerful as long as $n^{1/2}\rho_n/(K_0^{7/2}K) \to \infty$.

4. Numerical experiments. Now we illustrate the performance of the proposed test and the estimator of K in various simulations. In our simulation, we use simple spectral clustering for community recovery. Given an adjacency matrix A and a hypothetical number of communities K_0, this algorithm estimates the community membership by applying k-means clustering to the rows of the matrix formed by the K_0 leading singular vectors of A.

4.1. *Simulation 1: The null distribution and bootstrap correction.* In the first simulation, we consider the finite sample null distribution of the scaled and centered extreme eigenvalues of $\tilde{A}$ and empirically verify Theorem 3.1 for a simple stochastic block model. Following the observation in Bickel and Sarkar (2013), the speed of convergence to the limit distribution may be slow. A practical solution to this issue using a fused bootstrap correction has been proposed in Bickel and Sarkar (2013) for the special case of $K_0 = 1$. Here, we extend this idea to the more general case considered in this paper.

For a given adjacency matrix A on n nodes and null hypothesis $K = K_0$, the goodness-of-fit test statistic with fused bootstrap correction is given as follows:

1. Let $\hat{g}$ be an estimated community membership vector with K_0 communities, and $(\hat{B}, \hat{P})$ be the corresponding estimates in (6) and (8).
2. Calculate $\tilde{A}$ as in (7) and its extreme eigenvalues $\lambda_1(\tilde{A})$, $\lambda_n(\tilde{A})$.
3. For $m = 1, \ldots, M$:

   (a) Let $A^{(m)}$ be an adjacency matrix independently generated from stochastic block model $(\hat{g}, \hat{B})$.

   (b) Let $\tilde{A}^{(m)} = (\tilde{A}_{ij}^{(m)})_{i,j=1}^{n}$ be such that

$$\tilde{A}_{ii}^{(m)} = 0 \quad \text{and} \quad \tilde{A}_{ij}^{(m)} = \frac{A_{ij}^{(m)} - \hat{P}_{ij}}{\sqrt{(n-1)\hat{P}_{ij}(1-\hat{P}_{ij})}}, \qquad 1 \le i < j \le n.$$

   (c) Let $\lambda_1^{(m)}$ and $\lambda_n^{(m)}$ be the largest and smallest eigenvalues of $\tilde{A}^{(m)}$, respectively.

4. Let $(\hat{\mu}_1, \hat{s}_1^2)$ and $(\hat{\mu}_n, \hat{s}_n^2)$ be the sample mean and variance of $(\lambda_1^{(m)} : 1 \le m \le M)$ and $(\lambda_n^{(m)} : 1 \le m \le M)$, respectively.
5. The bootstrap corrected test statistic is

$$T_{n,K_0}^{(\mathrm{boot})} = \mu_{\mathrm{tw}} + s_{\mathrm{tw}} \max\left(\frac{\lambda_1(\tilde{A}) - \hat{\mu}_1}{\hat{s}_1}, -\frac{\lambda_n(\tilde{A}) - \hat{\mu}_n}{\hat{s}_n} \right), \tag{13}$$

   where μ_{tw} and s_{tw} are the mean and standard deviation of the Tracy–Widom distribution.

The fused bootstrap correction is computationally appealing as the bootstrap sample size M can be chosen as small as 50. All of our simulations use $M = 50$.

REMARK. The bootstrap correction is based on the empirical observation that although the finite sample null distribution is different from the theoretical limit, it has a similar shape, with different location and spread. Instead of using the theoretical centering and scaling as in (5) and (12), the corresponding bootstrap corrected extreme eigenvalues are

$$\mu_{\mathrm{tw}} + s_{\mathrm{tw}} \frac{\lambda_1(\tilde{A}) - \hat{\mu}_1}{\hat{s}_1} \quad \text{and} \quad \mu_{\mathrm{tw}} + s_{\mathrm{tw}} \frac{-\lambda_n(\tilde{A}) + \hat{\mu}_n}{\hat{s}_n}. \tag{14}$$

The largest and smallest eigenvalues of $\tilde{A}$ are individually corrected using the bootstrap population, because they are individually shown to have asymptotic Tracy–Widom distribution in Theorem 3.1.

In Figure 1, we plot the estimated density of the scaled and centered extreme eigenvalues of $\tilde{A}$ calculated from 1000 independent realizations, with

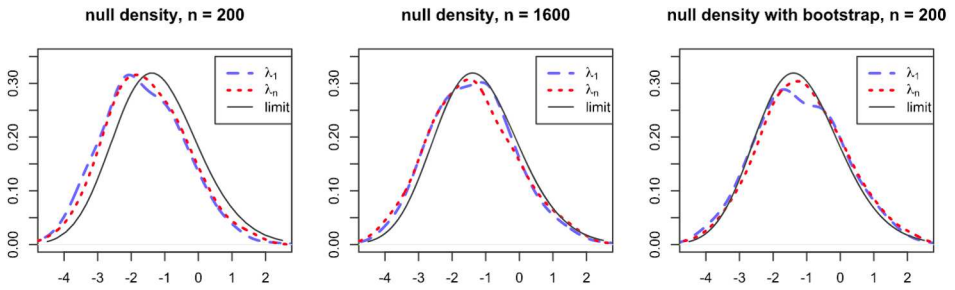

FIG. 1. *The empirical null distributions of scaled and centered extreme eigenvalues of* $\tilde{A}$ *over 1000 repetitions. Dashed line: largest eigenvalue; dotted line: smallest eigenvalue; solid line: theoretical limit distribution. Left: centered and scaled extreme eigenvalues as in (12) for* $n = 200$*; middle: centered and scaled extreme eigenvalues as in (12) for* $n = 1600$*; right: bootstrap corrected extreme eigenvalues as in (14). The stochastic block model used has two equal-sized communities, and* $B_{11} = B_{22} = 0.7$*,* $B_{12} = B_{21} = 0.3$*.*

and without bootstrap correction. The stochastic block model used here has two equal-sized communities, with $B_{11} = B_{22} = 0.7$ and $B_{12} = B_{21} = 0.3$. It is clear that the finite sample null distribution is systematically different from the limiting distribution when $n = 200$, and the difference is reduced but still visible when $n = 1600$. When bootstrap correction is used, the finite sample null distributions for both the largest and smallest eigenvalues are close to the limit even when $n = 200$.

4.2. *Simulation 2: Type* I *and type* II *errors.* Now we investigate the type I error of the proposed test under the null hypothesis and the power against various alternative distributions. For each $K_0 \in \{2, 3, 4\}$, we investigate four different models: (i) the null model, which is a stochastic block model with $K = K_0$ communities; (ii) a finer stochastic block model (finer SBM) with $K = K_0 + 1$ communities; (iii) a degree corrected block model [DCBM, Karrer and Newman (2011)] with $K = K_0$ communities; and (iv) a mixed membership block model [MMBM, Airoldi et al. (2008)] with $K = K_0$ communities. For any value of K, the community-wise edge probability matrix B is chosen such that $B_{kl} = 0.2 + 0.4 \times \mathbf{1}(k = l)$, for all $1 \le k, l \le K$. For the stochastic block model, the membership vector g is generated by sampling each entry independently from $\{1, \ldots, K\}$ with equal probability. For the degree corrected model, the membership vector is generated the same way as for the stochastic block model, with additional node activeness parameter ψ_i independently sampled from $\mathrm{Unif}(0, 1)$. In the degree corrected block model, the edge probability between nodes i and j is $\psi_i \psi_j B_{g_i g_j}$. For the mixed membership block model, the community mixing probability ϕ_i for each node i is an independent sample from a Dirichlet distribution with parameter $0.5 \times \mathbf{e}_K$ where $\mathbf{e}_K$ is a vector of ones with length K. With such

TABLE 1

Simulation 2: proportion of rejection at nominal level 0.05 *over* 200 *independent samples. The models considered are* (i) *Null: the stochastic block model with* $K = K_0$ *communities;* (ii) *Finer SBM: the stochastic block model with* $K = K_0 + 1$ *communities;* (iii) *DCBM: degree corrected block models with* $K = K_0$ *communities; and* (iv) *MMBM: mixed membership block model with* $K = K_0$ *communities. The edge probability between communities* k *and* l *is* $B_{kl} = 0.2 + 0.4 \times \mathbf{1}(k = l)$

	K_0	Null	Finer SBM	DCBM	MMBM
Without bootstrap	2	0.02	1	1	1
	3	0.04	1	1	1
	4	0.03	1	1	0.92
With bootstrap	2	0.02	1	1	1
	3	0.05	1	1	1
	4	0.06	1	1	0.93

a parameter, each node will tend to favor one or two communities so there is a weak community structure. The edge probability between nodes i and j in the mixed membership block model is $\phi_i^T B \phi_j$. For each model, we generate 200 independent adjacency matrices with $n = 1000$ nodes and perform the proposed hypothesis test, with or without bootstrap correction. The proportion of rejection at nominal level 0.05 is summarized in Table 1. We observe that the type I error is correctly kept at the nominal level. The type I error of bootstrap correction method is slightly closer to the nominal level. Also we observe that the test can successfully detect all three types of alternative hypotheses.

4.3. *Simulation 3: Estimating K using sequential testing.* Our third simulation examines the performance of the sequential testing estimator of K given in (11). We use two settings for this simulation. The first setting concerns different levels of network sparsity, where the community-wise connectivity matrices B is given by $B_{kl} = r(1 + 2 \times \mathbf{1}(k = l))$. That is, the edge probability is $3r$ within community and r between communities. We consider $r \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$ for different levels of network sparsity, and values of K between 2 and 8. For each combination of K and r, we generate 200 independent adjacency matrices A with $n = 1000$ nodes and K equal-sized communities. The number of communities is estimated for each observation as in (11) using threshold t_n corresponding to nominal type I error bound 10^{-4}. The proportion of correct estimates is summarized in Table 2. The sequential testing estimator with bootstrap correction works well for $K = 2, 3$ at all sparsity levels. When K gets larger, both methods require denser models to consistently estimate K. When the model is moderately dense, both methods work well for all values of K. For very sparse models,

TABLE 2
Simulation 3: proportion of correct estimates of K over 200 simulations under different sparsity levels indexed by r. The edge probability between communities k and l is $r(1+2\times\mathbf{1}(k=l))$. The network size is $n=1000$ with equal sized communities

	With bootstrap					**Without bootstrap**				
$\boldsymbol{r}$	**0.01**	**0.02**	**0.05**	**0.1**	**0.2**	**0.01**	**0.02**	**0.05**	**0.1**	**0.2**
$K=2$	1	1	1	1	1	0.30	0.98	1	1	1
$K=3$	0.99	1	1	1	1	0.11	0.91	1	1	1
$K=4$	0	1	1	1	1	0.24	0.89	1	1	1
$K=5$	0	0.5	1	1	1	0.25	0.93	1	1	1
$K=6$	0	0	1	1	1	0.16	0.09	1	1	1
$K=7$	0	0	1	1	1	0.04	0	1	1	1
$K=8$	0	0	0.71	1	1	0.03	0	0.9	1	1

the null distribution without bootstrap is biased and the sequential testing method tends to pick larger values of K.

In the second setting, the focus is on different types of block structures. To this end, for each $K\in\{2,3,4\}$ we generate matrices B whose diagonal and upper diagonal entries are independently drawn from a uniform distribution between 0 and 0.5. The success of spectral clustering requires the smallest singular value of B to be bounded away from zero, so we only use those B matrices whose smallest singular values are at least 0.1. The membership vector g is generated by sampling each entry independently from $\{1,\dots,K\}$ with equal probability. For each K and network size $n=500$ and $n=1000$, we generate 200 independent adjacency matrices using random B and g described above. Similarly, K is estimated as in (11) using threshold t_n corresponding to nominal type I error bound 10^{-4}. In Table 3, we summarize the proportion of correct estimates. The proposed test can correctly estimate the number of communities in a very large proportion of these randomly generated models. In general, the bootstrap correction helps improve the estimation accuracy.

TABLE 3
Simulation 3: proportion of correct estimates of K over 200 simulations with randomly generated matrices B and membership vectors g

	With bootstrap			**Without bootstrap**		
$\boldsymbol{K}$	**2**	**3**	**4**	**2**	**3**	**4**
$n=500$	0.99	0.90	0.76	0.91	0.84	0.74
$n=1000$	1	1	0.93	0.98	0.93	0.90

4.4. *The political blog data.* The political blog data (Adamic and Glance, 2005) records hyperlinks between web blogs shortly before the 2004 US presidential election. It has been used widely in the network community detection literature as an example of significant within-community node degree variation [see Karrer and Newman (2011), Zhao, Levina and Zhu (2012), Jin (2012), e.g.]. It is widely believed that a degree corrected block model is more suitable for this data, rather than a regular stochastic block model. Yan et al. (2014) used a likelihood ratio method to choose the degree corrected model over the regular stochastic block model. Theoretical justification of the χ^2 approximation used in this method is still an open problem, and maximizing the likelihood is computationally demanding. Following common practice, we consider the largest connected component of the political blog data. There are 1222 nodes with community sizes 586 and 636. We set $\hat{g}$ to be the true labeling given in the data—the results are similar for $\hat{g}$ estimated from the data. Under the null hypothesis that the data is generated from a stochastic block model of two communities, the test statistic is 1172.3 for the original test and 491.5 for the bootstrap corrected test, both indicating strong evidence to reject the null hypothesis. In addition, we apply the sequential testing procedure at type I error level 10^{-5}, with block model parameters estimated by spectral clustering using two leading eigenvectors of the adjacency matrix. The procedure partitions the nodes into 17 groups. Sixteen of these estimated groups mostly contain nodes from one true community, with 8 groups for each community and stratified by degrees. The additional estimated group contains nodes with very small degrees, whose community memberships are very hard to recover.

5. Discussion. The goodness-of-fit test developed in this paper is an attempt to perform principled statistical inference for stochastic block models. The test statistic reflects a fundamental difference between network models and traditional statistical models on independent individuals. In traditional independent and identically distributed data samples, the goodness-of-fit is usually assessed by the sum of residuals or squared residuals. For stochastic block models, the residual is a matrix, where the signal is not carried in the sum of individual residuals but is determined by how these residuals align across the rows and columns. For example, suppose A is generated from a stochastic block model with two communities and we want to test if $K = 1$. If we simply treat the upper diagonal entries of A as independent Bernoulli variables, the goodness-of-fit test reduces to testing whether the $n(n-1)/2$ upper diagonal entries look like an independent sample of a Bernoulli variable. Such tests have little power in detecting the block structure. On the other hand, the extreme singular value of the residual matrix accurately captures the block structure. This is an example of detecting low-rank mean effect from a noisy random matrix using its extreme eigenvalues.

Other examples using the similar idea include Kargin (2014) for reduced rank multivariate regression and Montanari, Reichman and Zeitouni (2014) for the Gaussian hidden clique problem. It would be interesting to further develop goodness-of-fit testing methods for more realistic null hypotheses, such as the degree corrected block model or even the nonparametric graphon model (Wolfe and Olhede, 2013).

It is possible to extend the method and theory developed in this paper to certain sparse stochastic block models. Consider sparse stochastic block models with $B = \rho_n B_0$ where the entries of B_0 are of order 1 and $\rho_n \downarrow 0$ controls the overall network sparsity. Most random matrix theory used in this paper (namely, Lemmas A.1, A.3, A.4) has been developed for moderately sparse stochastic block models with $\rho_n \gg n^{-1/3}$ in Erdős et al. (2013b, 2012). However, existing arguments do not guarantee isotropic delocalization of eigenvectors (Lemma A.2) due to the heavy tail of the normalized adjacency matrix entries $(A_{ij} - P_{ij})/[P_{ij}(1-P_{ij})]$. The possibility of proving such a result using modified techniques has been mentioned in Erdős et al. (2013a).

APPENDIX: PROOFS

Additional notation. Let $(\lambda_j^*, u_j^*)_{j=1}^n$ be the eigenvalue-eigenvector pairs of $\tilde{A}^*$ such that $\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_n^*$. For a pair of random sequences (a_n) and (b_n), we write $a_n = \tilde{O}_P(b_n)$ if for any $\varepsilon > 0$ and $D > 0$ there exists $n_0 = n_0(\varepsilon, D)$ such that

$$P(a_n \geq n^{\varepsilon} b_n) \leq n^{-D} \qquad \text{for all } n \geq n_0.$$

For any matrix M with singular value decomposition $M = \sum_j \sigma_j u_j v_j^T$, define $|M| = \sum_j |\sigma_j| u_j v_j^T$. We will use c and C to denote positive constants independent of n, which may vary from line to line.

A.1. Results from random matrix theory. We first collect some useful results from random matrix theory regarding the distributions of the eigenvalues and eigenvectors of $\tilde{A}^*$.

LEMMA A.1 [Asymptotic distributions of $\lambda_1(\tilde{A}^*)$ and $\lambda_n(\tilde{A}^*)$]. *For $\tilde{A}^*$ defined in (4) we have*

$$n^{2/3}(\lambda_1(\tilde{A}^*) - 2) \rightsquigarrow TW_1, \qquad n^{2/3}(-\lambda_n(\tilde{A}^*) - 2) \rightsquigarrow TW_1.$$

PROOF. Let G^* be an $n \times n$ symmetric matrix whose upper diagonal entries are independent normal with mean zero and variance $1/(n-1)$, and diagonal entries are zero. Then $\tilde{A}^*$ and G^* have the same first and second moments. According to Theorem 2.4 of Erdős, Yau and Yin (2012), we

know that $n^{2/3}(\lambda_1(\tilde{A}^*) - 2)$ and $n^{2/3}(\lambda_1(G^*) - 2)$ have the same limiting distribution. But $n^{2/3}(\lambda_1(G^*) - 2) \rightsquigarrow TW_1$ according to Lee and Yin (2014). The same argument applies to $\lambda_n(\tilde{A}^*)$. □

LEMMA A.2 (Eigenvector delocalization). *For each deterministic unit vector u and each $1 \le j \le n$, for any $\varepsilon > 0$ and $D > 0$ there exists $n_0 = n_0(\varepsilon, D)$ such that*

$$P[(u^T u_j^*)^2 \ge n^{-1+\varepsilon}] \le n^{-D} \qquad \text{for all } n \ge n_0.$$

It is worth noting that the above result is uniform over j and u in the sense that $n_0(\varepsilon, D)$ does not depend on u or j. Lemma A.2 can be equivalently stated as $(u^T u_j^*)^2 = \tilde{O}_P(n^{-1})$ uniformly over all u_j^* $(1 \le j \le n)$ and all deterministic unit vector u.

Lemma A.2 is Theorem 2.16 of Bloemendal et al. (2014). Although Bloemendal et al. (2014) requires the diagonal entries of $\tilde{A}^*$ to have positive variance, their Theorem 2.16 is a consequence of the local semicircle law [Theorem 2.12 of Bloemendal et al. (2014)], which can be established for matrices with zero diagonals using the result of Erdős et al. (2013a). See also the discussion in Bickel and Sarkar (2013).

LEMMA A.3 (Counting large eigenvalues). *Let c_n be a possibly random number of order $o_P(1)$ and $m(c_n)$ be the number of eigenvalues of $\tilde{A}^*$ larger than $\lambda_1^* - c_n$. Then $m(c_n) = O_P(nc_n^{3/2}) + \tilde{O}_P(1)$.*

Lemma A.3 extends equation (26) of Bickel and Sarkar (2013).

PROOF OF LEMMA A.3. For any $a < b < 5$, let $N^*(a, b)$ be the number of eigenvalues of $\tilde{A}^*$ in the interval $(a, b]$, and $N(a, b) = n\int_a^b \rho_{\mathrm{sc}}(x)\,dx$ where $\rho_{\mathrm{sc}}(x) = (1/2\pi)((4 - x^2)_+)^{1/2}$ is the density of the semicircle law. Let $\delta(a, b) = N^*(a, b) - N(a, b)$ then according to Theorem 2.2 of Erdős, Yau and Yin (2012) we have $\sup_{a,b<5} |\delta(a, b)| = \tilde{O}_P(1)$. Then, conditioning on the event that $\{|2 - \lambda_1^*| + c_n \le 1\}$, we have

$$\begin{aligned} m(c_n) &= N^*(\lambda_1^* - c_n, \lambda_1^*) \\ &= N(\lambda_1^* - c_n, \lambda_1^*) + \sup_{a,b<5} |\delta(a, b)| \\ &\le n \int_{2-(2-\lambda_1^*)-c_n}^{2} ((4 - x^2)_+)^{1/2}\,dx + \tilde{O}_P(1) \\ &\le 2n(c_n + |2 - \lambda_1^*|)^{3/2} + \tilde{O}_P(1) \\ &\le O(nc_n^{3/2}) + \tilde{O}_P(1). \end{aligned}$$

□

The claimed result follows by observing that the event $\{|2-\lambda_1^*|+c_n \le 1\}$ has probability $1-o(1)$.

LEMMA A.4 (Deviation of largest singular value). *There exists absolute positive constants a, b, c, C, such that*

$$P[n^{2/3}(\sigma_1(\tilde{A}^*)-2) \ge (\log n)^{a\log\log n}] \le C\exp[-b(\log n)^{c\log\log n}].$$

Lemma A.4 is a direct consequence of equation (2.22) in Erdős, Yau and Yin (2012). We can simplify the statement so that there exists an absolute constant $b>0$ such that for any $\varepsilon>0$

$$P[n^{2/3}(\sigma_1(\tilde{A}^*)-2) \ge n^{\varepsilon}] = O(n^{-b}). \tag{15}$$

A.2. Proof of asymptotic null distribution. Now we provide proofs for theoretical results in Section 3. Here, we omit the dependence on n in g, B and K for simplicity.

PROOF OF THEOREM 3.1. The consistency of $\hat{g}$ allows us to focus on the event $\hat{g}=g$.

We will prove the claim for $\lambda_1(\tilde{A})$. The other claim can be proved by applying the same argument on $-\tilde{A}$.

Let $\tilde{A}' \in \mathbb{R}^{n\times n}$ be such that

$$\tilde{A}'_{ij} = \begin{cases} \dfrac{A_{ij}-\hat{P}_{ij}}{\sqrt{(n-1)P_{ij}(1-P_{ij})}}, & i\neq j, \\ \dfrac{P_{ii}-\hat{P}_{ii}}{\sqrt{(n-1)P_{ii}(1-P_{ii})}}, & i=j. \end{cases} \tag{16}$$

Thus, $\tilde{A}'=\tilde{A}^*+\Delta'$, where $\Delta'_{ij}=(P_{ij}-\hat{P}_{ij})/\sqrt{(n-1)P_{ij}(1-P_{ij})}$. Because Δ' is a $K\times K$ block-wise constant symmetric matrix, its rank is at most K, and the corresponding principal subspace is spanned by $(\theta_1,\dots,\theta_K)$, where $\theta_k\in\mathbb{R}^n$ is the unit norm indicator of the kth community in g. That is, the ith entry of θ_k is $n_k^{-1/2}$ if $g_i=k$ and zero otherwise, where n_k is the size of the kth community.

The consistency of $\hat{g}$ implies that with probability tending to one, for each $1\le k,k'\le K$, $\hat{B}_{k,k'}$ is the sample mean of independent Bernoulli random variables with parameter $B_{k,k'}$ and sample size of order $(n/K)^2$. Thus, standard large deviation inequalities such as Bernstein's inequality or Hoeffding's inequality suggest that $\sup_{k,k'}|\hat{B}_{k,k'}-B_{k,k'}| = o_P(K\log n/n)$, which implies that $\sup_{i,j}|\hat{P}_{ij}-P_{ij}| = o_P(K\log n/n)$. Note here the o_P statement goes through a union bound over K^2 terms, which is valid since the tail probability bound for $\hat{P}_{ij}-P_{ij}$ can be made exponentially small in n. Let

$\Delta' = \Theta\Gamma\Theta^T$, where $\Theta = (\theta_1, \ldots, \theta_K)$ and Γ is a $K \times K$ symmetric matrix. Then each entry of Γ is $o_P(n^{-1/2}\log n)$, and hence $\|\Gamma\| = o_P(Kn^{-1/2}\log n)$.

We will show that

$$\lambda_1(\tilde{A}') = \lambda_1(\tilde{A}^*) + o_P(n^{-2/3}), \tag{17}$$

by establishing a lower and upper bound on $\lambda_1(\tilde{A}')$. Both parts uses the eigenvector delocalization result (Lemma A.2) as follows. Let $\Theta = (\theta_1, \ldots, \theta_K)$, then, uniformly over j we have

$$\|\Theta^T u_j^*\|_2^2 = \sum_{k=1}^{K} (\theta_k^T u_j^*)^2 = \tilde{O}_P(Kn^{-1}), \tag{18}$$

and hence

$$\begin{aligned} |(u_j^*)^T \Delta' u_j^*| &\le |(\Theta^T u_j^*)^T \Gamma (\Theta^T u_j^*)| \le \|\Theta^T u_j^*\|_2^2 \|\Gamma\| \\ &= \tilde{O}_P(K^2 n^{-3/2}\log n). \end{aligned} \tag{19}$$

Here, the $\tilde{O}_P$ statement in (18) holds when taking union bound over K terms by choosing D large enough in Lemma A.2.

First, we provide a lower bound on $\lambda_1(\tilde{A}')$:

$$\begin{aligned} \lambda_1(\tilde{A}') &\ge (u_1^*)^T \tilde{A}' u_1^* = \lambda_1^* + (u_1^*)^T \Delta' u_1^* \\ &\ge \lambda_1^* - \tilde{O}_P(K^2 n^{-3/2}\log n) \\ &\ge \lambda_1^* - o_P(n^{-2/3}), \end{aligned} \tag{20}$$

where the last inequality uses the assumed upper bound on the rate at which K grows with n, and the second last inequality uses (19).

Next, we provide an upper bound of $\lambda_1(\tilde{A}')$. For any unit vector $u \in \mathbb{R}^n$, let $(a_1, \ldots, a_n)$ be a unit vector in $\mathbb{R}^n$ such that

$$u = \sum_{j=1}^{n} a_j u_j^*.$$

Let m be the number of λ_j^*'s in the interval $(\lambda_1^* - 2\|\Delta'\|, \lambda_1^*]$, and $u_1 = \sum_{j=1}^{m} a_j u_j^*$, $u_2 = \sum_{j=m+1}^{n} a_j u_j^*$. Then

$$\begin{aligned} u^T \tilde{A}' u &= u^T \tilde{A}^* u + u^T \Delta' u \\ &\le \lambda_1^* \sum_{j=1}^{m} a_j^2 + (\lambda_1^* - 2\|\Delta'\|) \sum_{j=m+1}^{n} a_j^2 + 2u_1^T |\Delta'| u_1 + 2u_2^T |\Delta'| u_2 \\ &\le \lambda_1^* \sum_{j=1}^{m} a_j^2 + (\lambda_1^* - 2\|\Delta'\|) \sum_{j=m+1}^{n} a_j^2 \end{aligned}$$

$$
\begin{aligned}
&+ 2m\sum_{j=1}^{m} a_j^2 (u_j^*)^T |\Delta'| u_j^* + 2u_2^T |\Delta'| u_2 \\
&\le \lambda_1^* \sum_{j=1}^{m} a_j^2 + (\lambda_1^* - 2\|\Delta'\|) \sum_{j=m+1}^{n} a_j^2 \qquad (21)\\
&\quad + 2m\tilde{O}_P(K^2 n^{-3/2}\log n)\sum_{j=1}^{m} a_j^2 + 2\|\Delta'\| \sum_{j=m+1}^{n} a_j^2 \\
&\le \lambda_1^* + m\tilde{O}_P(K^2 n^{-3/2}\log n) \\
&\le \lambda_1^* + (O(n\|\Delta'\|^{3/2}) + \tilde{O}_P(1))\tilde{O}_P(K^2 n^{-3/2}\log n) \\
&= \lambda_1^* + \tilde{O}_P(K^{7/2}(\log n)^{5/2} n^{-5/4}),
\end{aligned}
$$

where the third inequality uses (19) and uniformity over j, and the second last line uses Lemma A.3 together with $\|\Delta'\| = o_P(Kn^{-1/2}\log n)$.

Thus, (17) is established by combining (20) and (21), provided that $K = O(n^{1/6-\tau})$ for some small positive τ.

Next, we show that $\lambda_1(\tilde{A}) = \lambda_1(\tilde{A}') + o_P(n^{-2/3})$. Let $\tilde{A}'' = \tilde{A}' - \mathrm{diag}(\tilde{A}')$. Consider the block representation of $\tilde{A}$:

$$\tilde{A} = (\tilde{A}_{(k,l)})_{k,l=1}^{K},$$

where $\tilde{A}_{(k,l)}$ is the submatrix corresponding to the rows in community k and columns in community l. Similar block representations can be defined for $\tilde{A}''$. It is obvious that

$$\tilde{A}_{(k,l)} = \tilde{A}''_{(k,l)} \frac{\sqrt{B_{kl}(1-B_{kl})}}{\sqrt{\hat{B}_{kl}(1-\hat{B}_{kl})}} = \tilde{A}''_{(k,l)}(1 + o_P(Kn^{-1}\log n)).$$

Therefore,

$$
\begin{aligned}
\|\tilde{A} - \tilde{A}''\| &\le K \max_{k,l} \|\tilde{A}_{(k,l)} - \tilde{A}''_{(k,l)}\| \le o_P(Kn^{-1}\log n) K \sum_{k,l} \|\tilde{A}''_{(k,l)}\| \\
&\le o_P(K^2 n^{-1}\log n)\|\tilde{A}''\| \le o_P(K^2 n^{-1}\log n)(\|\tilde{A}'\| + \|\mathrm{diag}(\tilde{A}')\|) \\
&\le o_P(K^2 n^{-1}\log n)(O_P(1) + O_P(Kn^{-3/2}\log n)) \\
&= o_P(K^2 n^{-1}\log n) = o_P(n^{-2/3}).
\end{aligned}
$$

Then

$$\|\tilde{A} - \tilde{A}'\| \le \|\tilde{A} - \tilde{A}''\| + \|\mathrm{diag}(\tilde{A}')\| = o_P(n^{-2/3}). \qquad (22)$$

Combining (17) and (22), we have

$$\lambda_1(\tilde{A}) = \lambda_1(\tilde{A}^*) + o_P(n^{-2/3}). \qquad (23)$$

Now applying Lemma A.1 and combining with (23) we have

$$n^{2/3}(\lambda_1(\tilde{A}) - 2) \rightsquigarrow TW_1. \qquad \square$$

A.3. Proof of power and consistency.

PROOF OF THEOREM 3.3. For all $1 \le l \le K$, $1 \le k \le K_0$, let $\mathcal{N}_l = \{i : g_i = l\}$, $\hat{\mathcal{N}}_k = \{i : \hat{g}_i = k\}$ and $\hat{\mathcal{N}}_{k,l} = \{i : \hat{g}_i = k, g_i = l\}$. For each $1 \le l \le K$, $\mathcal{N}_l$ is partitioned into $\{\hat{\mathcal{N}}_{k,l} : 1 \le k \le K_0\}$. Thus, for each $1 \le l \le K$ there exists a k_l such that $1 \le k_l \le K_0$ and $|\hat{\mathcal{N}}_{k_l,l}| \ge |\mathcal{N}_l|/K_0 \ge c_0 n/(K \times K_0) \ge c_0 n K^{-2}$. Because $K_0 < K$, there exist l_1 and l_2 such that $k_{l_1} = k_{l_2} = k$. Since $B \in \mathcal{B}_K$, there exists an l_3 such that $B_{l_1,l_3} \ne B_{l_2,l_3}$. Let $k' = k_{l_3}$ and we have $|\hat{\mathcal{N}}_{k',l_3}| \ge c_0 n K^{-2}$.

Let $\tilde{A}^{(0)}$ be the submatrix of $\tilde{A}$ consisting the rows in $\hat{\mathcal{N}}_{k,l_1} \cup \hat{\mathcal{N}}_{k,l_2}$, and the columns in $\hat{\mathcal{N}}_{k',l_3}$. Define $A^{(0)}$, $\hat{P}^{(0)}$, and $P^{(0)}$ correspondingly.

When $k \ne k'$, or $k = k'$ but $l_3 \notin \{l_1, l_2\}$, the submatrix $A^{(0)}$ contains only off-diagonal entries of A. Therefore, $\hat{P}^{(0)}$ is a constant matrix in that all of its entries are equal. We have

$$\begin{aligned}
\|\tilde{A}\| &\ge \|\tilde{A}^{(0)}\| \ge n^{-1/2}\|A^{(0)} - \hat{P}^{(0)}\| \\
&\ge n^{-1/2}(\|P^{(0)} - \hat{P}^{(0)}\| - \|A^{(0)} - P^{(0)}\|) \\
&\ge n^{-1/2}(\|P^{(0)} - \hat{P}^{(0)}\| - O_P(n^{1/2})) \\
&\ge n^{-1/2}(\delta_B c_0 n K^{-2}/2 - O_P(n^{1/2})).
\end{aligned} \tag{24}$$

To obtain the last inequality, first note that $P^{(0)}$ has two distinct blocks each with at least $c_0 n K^{-2}$ rows and at least $c_0 n K^{-2}$ columns. Each of these two blocks has constant entries and at least one of them has absolute entry value at least $\delta_B/2$. Thus, $\|P^{(0)} - \hat{P}^{(0)}\| \ge \delta_B c_0 n K^{-2}/2$.

When $k = k'$ and $l_3 \in \{l_1, l_2\}$, the submatrix $A^{(0)}$ defined above contains diagonal entries of A. The corresponding entries of $\hat{P}^{(0)}$ are zero. These zero entries causes an additional $O(1)$ term in $\|\hat{P}^{(0)} - P^{(0)}\|$ and (24) still goes through. $\square$

PROOF OF COROLLARY 3.4. Following the notation in the proof of Theorem 3.3, for any $K_0 < K$ we have, in view of (24) and letting $C = \inf_n \delta_B c_0/2$,

$$\begin{aligned}
P(T_{n,K_0} < t_n) &= P[n^{2/3}(\|\tilde{A}\| - 2) < t_n] = P[\|\tilde{A}\| < n^{-2/3}t_n + 2] \\
&\le P[n^{-1/2}(\|P^{(0)} - \hat{P}^{(0)}\| - \|A^{(0)} - P^{(0)}\|) \le n^{-2/3}t_n + 2] \\
&\le P[n^{-1/2}\|A^{(0)} - P^{(0)}\| \ge Cn^{1/2}K^{-2} - n^{-2/3}t_n - 2] \\
&\le n^{-1},
\end{aligned}$$

the last inequality is obtained by first using the assumption $K = O(n^{1/6-\tau})$, and $t_n = O(n^{5/6})$, so that $Cn^{1/2}K^{-2} + n^{-2/3}t_n + 2 \geq n^{1/6}$ for large n, and then applying operator norm deviation bound results such as Theorem 5.2 of Lei and Rinaldo (2013) [see also Theorem 3.4 of Chatterjee (2015)].

Therefore,

$$P(\hat{K} < K) \leq \sum_{K_0=1}^{K-1} P(T_{n,K_0} < t_n) \leq n^{-1}(K-1) = o(1).$$

On the other hand,

$$\begin{aligned} P(\hat{K} > K) &\leq P(T_{n,K} \geq t_n) = P(n^{2/3}(\sigma_1(\tilde{A}) - 2) \geq t_n) \\ &\leq P(n^{2/3}(\sigma_1(\tilde{A}^*) - 2) \geq t_n/2) + P(n^{2/3}|\sigma_1(\tilde{A}^*) - \sigma_1(\tilde{A})| \geq t_n/2) \\ &= o(1), \end{aligned}$$

where the first probability is controlled using Lemma A.4 and the second probability is controlled using (23) and its analogous result for $\lambda_n(\tilde{A}) - \lambda_n(\tilde{A}^*)$. $\square$

A.4. Asymptotic power against degree corrected block models.

PROOF OF THEOREM 3.5. Recall that $\hat{\mathcal{N}}_{l,k} = \{i : g_i = k, \hat{g}_i = l\}$ $(1 \leq l \leq K_0,\ 1 \leq k \leq K)$. Let $\tilde{\phi}_{k,\hat{\mathcal{N}}_{l,k}}$ be the subvector of $\tilde{\phi}_k$ on the entries in $\hat{\mathcal{N}}_{l,k}$.

Let $\eta_l = \tilde{\phi}_{k^*,\hat{\mathcal{N}}_{l,k^*}}$ for each $1 \leq l \leq K_0$. By definition of $\mathcal{E}$, $\sum_{l=1}^{K_0} \mathcal{E}(\eta_l, 1) \geq \mathcal{E}(\tilde{\psi}_{k^*}, K_0)$, and hence there exists an l^* such that $\mathcal{E}(\eta_{l^*}, 1) \geq \mathcal{E}(\tilde{\phi}_{k^*}, K_0)/K_0$.

For simplicity, denote $\eta = \eta_{l^*}$ and $\bar{\eta} = \eta/\|\eta\|$. Let $m = |\hat{\mathcal{N}}_{l^*,k^*}|$ and define $\mathbf{e}_m$ as the $1 \times m$ vector with $1/\sqrt{m}$ on each entry. Therefore, we have

$$\|\bar{\eta} - \mathbf{e}_m\|^2 \geq \mathcal{E}(\bar{\eta}, 1) = \|\eta\|^{-2}\mathcal{E}(\eta, 1) \geq \|\eta\|^{-2}\mathcal{E}(\tilde{\phi}_{k^*}, K_0)/K_0. \tag{25}$$

Because $\bar{\eta}$ and $\mathbf{e}_m$ both have unit ℓ_2 norm, (25) implies that

$$|\mathbf{e}_m^T\bar{\eta}| \leq 1 - \frac{\mathcal{E}(\tilde{\phi}_{k^*}, K_0)}{2K_0\|\eta\|^2}.$$

Let $u = (\bar{\eta} - \mathbf{e}_m\mathbf{e}_m^T\bar{\eta})/\|\bar{\eta} - \mathbf{e}_m\mathbf{e}_m^T\bar{\eta}\|$, then

$$u^T\eta = u^T\bar{\eta}\|\eta\| = \|\eta\|\frac{1 - (\mathbf{e}_m^T\bar{\eta})^2}{\|\bar{\eta} - \mathbf{e}_m\mathbf{e}_m^T\bar{\eta}\|} \geq \frac{\mathcal{E}(\tilde{\phi}_{k^*}, K_0)}{2K_0\|\eta\|} \geq \frac{\mathcal{E}(\tilde{\phi}_{k^*}, K_0)}{2K_0}. \tag{26}$$

Now let k' be such that $B_{k^*,k'} = \|B_{k^*,\cdot}\|_\infty$. There exists an l' such that $\|\tilde{\phi}_{k',\hat{\mathcal{N}}_{l',k'}}\| \geq K_0^{-1/2}$.

Let $A^{(0)}$ be the submatrix of A corresponding to the rows in $\hat{\mathcal{N}}_{l^*,k^*}$ and columns in $\hat{\mathcal{N}}_{l',k'}$, and define $\tilde{A}^{(0)}$, $P^{(0)}$, $\hat{P}^{(0)}$ similarly. Thus, by construction we have, letting $m' = |\hat{\mathcal{N}}_{l',k'}|$,

$$P^{(0)} = \|\phi_{k^*}\|\|\phi_{k'}\|B_{k^*k'}\eta\tilde{\phi}^T_{k',\hat{\mathcal{N}}_{l',k'}}, \qquad \hat{P}^{(0)} = \hat{B}_{l^*l'}\sqrt{mm'}\mathbf{e}_m\mathbf{e}^T_{m'}.$$

Observing that $u^T\mathbf{e}_m = 0$, we have

$$\begin{aligned}\|u^T(P^{(0)} - \hat{P}^{(0)})\| &= \|u^TP^{(0)}\| = \|\psi_{k^*}\|\|\psi_{k'}\|B_{k^*k'}|u^T\eta|\|\tilde{\phi}_{k',\hat{\mathcal{N}}_{k',l'}}\| \\ &\geq \|\psi_{k^*}\|\|\psi_{k'}\|B_{k^*k'}\frac{\mathcal{E}(\tilde{\phi}_{k^*}, K_0)}{2K_0^{3/2}} \\ &\geq \kappa_n n\|B_{k^*,\cdot}\|_\infty\frac{\mathcal{E}(\tilde{\phi}_{k^*}, K_0)}{2K_0^{3/2}}.\end{aligned}$$

The claimed result follows by observing that

$$\|\tilde{A}^{(0)}\| \geq n^{-1/2}(\|P^{(0)} - \hat{P}^{(0)}\| - \|A^{(0)} - P^{(0)}\|)$$

and $\|A^{(0)} - P^{(0)}\| = O_P(\sqrt{n})$. □

REFERENCES

Abbe, E., Bandeira, A. S. and Hall, G. (2014). Exact recovery in the stochastic block model. Preprint. Available at arXiv:1405.3267.

Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery* 36–43. ACM, New York.

Airoldi, E. M., Blei, D. M., Fienberg, S. E. and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

Amini, A. A. and Levina, E. (2014). On semidefinite relaxations for the block model. Preprint. Available at arXiv:1406.5647.

Anandkumar, A., Ge, R., Hsu, D. and Kakade, S. M. (2014). A tensor approach to learning mixed membership community models. *J. Mach. Learn. Res.* **15** 2239–2312. MR3231594

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

Bickel, P. J. and Sarkar, P. (2013). Hypothesis testing for automated community detection in networks. Preprint. Available at arXiv:1311.2694.

Bloemendal, A., Erdős, L., Knowles, A., Yau, H.-T. and Yin, J. (2014). Isotropic local laws for sample covariance and generalized Wigner matrices. *Electron. J. Probab.* **19** 1–53. MR3183577

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.* **43** 177–214. MR3285604

Chaudhuri, K., Chung, F. and Tsiatas, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. *J. Mach. Learn. Res. Workshop Conf. Proc.* **2012** 35.1–35.23.

Chen, K. and Lei, J. (2014). Network cross-validation for determining the number of communities in network data. Preprint. Available at arXiv:1411.1715.

Chen, Y., Sanghavi, S. and Xu, H. (2012). Clustering sparse graphs. In *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) 2204–2212. Curran Associates, Red Hook, NY.

Decelle, A., Krzakala, F., Moore, C. and Zdeborová, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E (3)* **84** 066106.

Erdős, L., Yau, H.-T. and Yin, J. (2012). Rigidity of eigenvalues of generalized Wigner matrices. *Adv. Math.* **229** 1435–1515. MR2871147

Erdős, L., Knowles, A., Yau, H.-T. and Yin, J. (2012). Spectral statistics of Erdős-Rényi Graphs II: Eigenvalue spacing and the extreme eigenvalues. *Comm. Math. Phys.* **314** 587–640. MR2964770

Erdős, L., Knowles, A., Yau, H.-T. and Yin, J. (2013a). The local semicircle law for a general class of random matrices. *Electron. J. Probab.* **18** no. 59, 58. MR3068390

Erdős, L., Knowles, A., Yau, H.-T. and Yin, J. (2013b). Spectral statistics of Erdős–Rényi graphs I: Local semicircle law. *Ann. Probab.* **41** 2279–2375. MR3098073

Fishkind, D. E., Sussman, D. L., Tang, M., Vogelstein, J. T. and Priebe, C. E. (2013). Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM J. Matrix Anal. Appl.* **34** 23–39. MR3032990

Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088

Jin, J. (2012). Fast community detection by SCORE. Available at arXiv:1211.5803.

Kargin, V. (2014). On the singular values of the reduced-rank multivariate response regression. Preprint. Available at arXiv:1409.6779.

Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E (3)* **83** 016107, 10. MR2788206

Krzakala, F., Moore, C., Mossel, E., Neeman, J., Sly, A., Zdeborová, L. and Zhang, P. (2013). Spectral redemption in clustering sparse networks. *Proc. Natl. Acad. Sci. USA* **110** 20935–20940. MR3174850

Lee, J. O. and Yin, J. (2014). A necessary and sufficient condition for edge universality of Wigner matrices. *Duke Math. J.* **163** 117–173. MR3161313

Lei, J. and Rinaldo, A. (2013). Consistency of spectral clustering in stochastic block models. Preprint. Available at arXiv:1312.2050.

Lei, J. and Zhu, L. (2014). A generic sample splitting approach for refined community recovery in stochastic block models. Preprint. Available at arXiv:1411.1469.

Massoulie, L. (2013). Community detection thresholds and the weak Ramanujan property. Preprint. Available at arXiv:1311.3085.

McSherry, F. (2001). Spectral partitioning of random graphs. In *42nd IEEE Symposium on Foundations of Computer Science (Las Vegas, NV, 2001)* 529–537. IEEE Computer Soc., Los Alamitos, CA. MR1948742

Montanari, A., Reichman, D. and Zeitouni, O. (2014). On the limitation of spectral methods: From the Gaussian hidden clique problem to rank one perturbations of Gaussian tensors. Preprint. Available at arXiv:1411.6149.

Mossel, E., Neeman, J. and Sly, A. (2013). A proof of the block model threshold conjecture. Preprint. Available at arXiv:1311.4115.

Newman, M. E. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103** 8577–8582.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E (3)* **69** 026113.

Saldana, D. F., Yi, Y. and Feng, Y. (2014). How many communities are there? Preprint. Available at arXiv:1412.1684.

Vu, V. (2014). A simple SVD algorithm for finding hidden partitions. Preprint. Available at arXiv:1404.3918.

Wolfe, P. J. and Olhede, S. C. (2013). Nonparametric graphon estimation. Preprint. Available at arXiv:1309.5936.

Yan, X., Shalizi, C., Jensen, J. E., Krzakala, F., Moore, C., Zdeborová, L., Zhang, P. and Zhu, Y. (2014). Model selection for degree-corrected block models. *J. Stat. Mech. Theory Exp.* **2014** P05007.

Zhao, Y., Levina, E. and Zhu, J. (2011). Community extraction for social networks. *Proc. Natl. Acad. Sci. USA* **108** 7321–7326.

Zhao, Y., Levina, E. and Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40** 2266–2292. MR3059083

Department of Statistics
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
USA
E-mail: jinglei@andrew.cmu.edu