Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Using Maximum Entry-Wise Deviation to Test the Goodness of Fit for Stochastic Block Models

Jianwei Hu, Jingfei Zhang, Hong Qin, Ting Yan & Ji Zhu

To cite this article: Jianwei Hu, Jingfei Zhang, Hong Qin, Ting Yan & Ji Zhu (2021) Using Maximum Entry-Wise Deviation to Test the Goodness of Fit for Stochastic Block Models, Journal of the American Statistical Association, 116:535, 1373-1382, DOI: 10.1080/01621459.2020.1722676

To link to this article: https://doi.org/10.1080/01621459.2020.1722676

View supplementary material

Published online: 08 Sep 2020.

Submit your article to this journal

Article views: 595

View related articles

View Crossmark data

Citing articles: 5 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uasa20

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 535, 1373–1382: Theory and Methods
https://doi.org/10.1080/01621459.2020.1722676

Taylor & Francis
Taylor & Francis Group

Using Maximum Entry-Wise Deviation to Test the Goodness of Fit for Stochastic Block Models

Jianwei Hu*[a], Jingfei Zhang*[b], Hong Qin[a,c], Ting Yan[a], and Ji Zhu[d]

[a]Department of Statistics, Central China Normal University, Wuhan, China; [b]Department of Management Science, University of Miami, Coral Gables, FL; [c]Department of Statistics, Zhongnan University of Economics and Law, Wuhan, China; [d]Department of Statistics, University of Michigan, Ann Arbor, MI

ABSTRACT

The stochastic block model is widely used for detecting community structures in network data. How to test the goodness of fit of the model is one of the fundamental problems and has gained growing interests in recent years. In this article, we propose a novel goodness-of-fit test based on the maximum entry of the centered and rescaled adjacency matrix for the stochastic block model. One noticeable advantage of the proposed test is that the number of communities can be allowed to grow linearly with the number of nodes ignoring a logarithmic factor. We prove that the null distribution of the test statistic converges in distribution to a Gumbel distribution, and we show that both the number of communities and the membership vector can be tested via the proposed method. Furthermore, we show that the proposed test has asymptotic power guarantee against a class of alternatives. We also demonstrate that the proposed method can be extended to the degree-corrected stochastic block model. Both simulation studies and real-world data examples indicate that the proposed method works well. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2018
Accepted November 2019

KEYWORDS
Community detection; Degree-corrected stochastic block model; Goodness-of-fit test; Network data; Stochastic block model.

1. Introduction

One of the fundamental problems in network data analysis is community detection that aims to divide nodes into communities such that the links are dense within communities and relatively sparse between communities. The stochastic block model proposed by Holland, Laskey, and Leinhardt (1983) is probably the most studied network model for this purpose; see Snijders and Nowicki (1997), Nowicki and Snijders (2001), Bickel and Chen (2009), Rohe, Chatterjee, and Yu (2011), Choi, Wolfe, and Airoldi (2012), Jin (2015), and Zhang and Zhou (2016) for some of the representative work.

In a stochastic block model with k communities, n nodes are clustered into k blocks, that is, there exists a mapping σ of community membership: $[n] \to [k]^n$, where $[n] = \{1, \ldots, n\}$. Given the community membership σ, the entries A_{ij} $(i > j)$ of the symmetric adjacency matrix $A \in \{0, 1\}^{n\times n}$ of an undirected random graph $\mathcal{G}$ are then assumed to be mutually independent Bernoulli random variables with the occurrence probabilities $P_{ij} = B_{\sigma(i)\sigma(j)}$ for certain symmetric probability matrix $B \in [0, 1]^{k\times k}$. A large number of methods for recovering the community membership have been proposed, including modularity (Newman 2006), profile-likelihood maximization (Bickel and Chen 2009), pseudo-likelihood maximization (Amini et al. 2013), variational methods (Daudin, Picard, and Robin 2008), and spectral clustering (Rohe, Chatterjee, and Yu 2011; Jin 2015). Asymptotic properties of the estimators of the community membership have also been established; see Choi, Wolfe, and Airoldi (2012), Rohe, Chatterjee, and Yu (2011), Zhao, Levina, and Zhu (2012), Sarkar and Bickel (2015), Jin (2015), Lei and Rinaldo (2015), and Zhang and Zhou (2016). For a review of the subject, we refer to Bhattacharyya and Bickel (2016). However, how to validate the stochastic block model is a challenging problem and has not been addressed only until recently. Specifically, Wang and Bickel (2017) developed a likelihood-based approach to test the model and derived the asymptotic distribution of the log-likelihood ratio statistic under model misspecification when the number of communities k is fixed. Bickel and Sarkar (2015) used the largest eigenvalue of the centered and scaled adjacency matrix to test the Erdős–Rényi model and derived the asymptotic null distribution. By extending their arguments, Lei (2016) developed a goodness-of-fit test for stochastic block models using the largest singular value of the centered and rescaled adjacency matrix and derived its asymptotic null distribution when the condition $k = o(n^{1/6})$ holds. It was also acknowledged that it is difficult to extend these results to the more flexible degree-corrected block model. Karwa et al. (2016) developed a finite-sample Monte Carlo goodness-of-fit test for the stochastic block model. The proposed test calculates goodness-of-fit statistics of graphs sampled from a conditional distribution given sufficient statistics of the stochastic block model; then the sample statistics are compared to the one calculated from the observed network, from which a naive p-value estimator is obtained. The proposed procedure is computationally expensive and there is no theoretical guarantee for the null distribution and asymptotic power of such finite-sample Monte Carlo tests.

CONTACT Ji Zhu jizhu@umich.edu Department of Statistics, University of Michigan, 1085 South University, Ann Arbor, MI 48109-1107.

*The first two authors contributed equally to this work.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

© 2020 American Statistical Association

In this article, we propose a novel goodness-of-fit test based on the maximum entry-wise deviation of the centered and rescaled adjacency matrix. We show that the asymptotic null distribution of the test statistic is a Gumbel distribution when $k = o(n/\log^2 n)$. This condition implies that k is allowed to grow linearly with n ignoring a logarithmic factor. This kind of scenario was referred to by Rohe, Qin, and Fan (2014) as the highest dimensional stochastic block model as the number of communities must be smaller than the number of nodes, and no reasonable model would allow k to grow faster than that. As a result, the proposed test significantly relaxes the condition that $k = o(n^{1/6})$ in Lei (2016). Moreover, we show that the proposed test is asymptotically powerful against a class of alternatives. We also propose an augmented test statistic that improves the power of the goodness-of-fit test, while having the same asymptotic null distribution as the original test statistic. The maximum entry-wise deviation approach was first introduced by Jiang (2004) for testing the hypothesis $H_0 : R = I$ versus $H_1 : R \neq I$, where R is a correlation matrix; therefore, the setting is quite different from ours.

The remainder of the article is organized as follows. In Section 2, we introduce the new test statistic, and state its asymptotic null distribution and asymptotic power. Further, we propose an augmented test statistic to improve the power of the test. We extend our results to the degree-corrected stochastic block model in Section 3. Simulation studies and real-world data examples are given in Sections 4 and 5, respectively.

2. A New Goodness-of-Fit Test for the Stochastic Block Model

Consider a stochastic block model on n nodes with the membership vector σ and probability matrix B. For any fixed (B, σ), the probability mass function for the adjacency matrix A is

$$P(A) = \prod_{1 \le i < j \le n} B_{\sigma(i)\sigma(j)}^{A_{ij}} (1 - B_{\sigma(i)\sigma(j)})^{(1-A_{ij})},$$

and the corresponding log-likelihood under the stochastic block model can be written as

$$\ell(A|B, \sigma) = \frac{1}{2} \sum_{u,v=1}^{k} (m_{uv} \log B_{uv} + (n_{uv} - m_{uv}) \log(1 - B_{uv})),$$

where

$$n_{uv} = \sum_{i=1}^{n} \sum_{j \neq i} \mathbf{1}\{\sigma(i) = u, \sigma(j) = v\} \quad \text{and}$$

$$m_{uv} = \sum_{i=1}^{n} \sum_{j \neq i} A_{ij} \mathbf{1}\{\sigma(i) = u, \sigma(j) = v\}.$$

It is not difficult to see that given a number of communities k_0 and a membership vector σ_0, the maximum likelihood estimate of B is given by

$$\widehat{B}_{uv}^{\sigma_0} = \begin{cases} \frac{\sum_{i \in \sigma_0^{-1}(u), j \in \sigma_0^{-1}(v)} A_{ij}}{|\sigma_0^{-1}(u)| \cdot |\sigma_0^{-1}(v)|}, & u \neq v, \\ \frac{\sum_{i \neq j \in \sigma_0^{-1}(u)} A_{ij}}{|\sigma_0^{-1}(u)| \cdot (|\sigma_0^{-1}(u)| - 1)}, & u = v, \end{cases} \tag{1}$$

where $\sigma_0^{-1}(u) = \{i : 1 \le i \le n, \sigma_0(i) = u\}$ and $|\sigma_0^{-1}(u)|$ is the number of nodes in block u.

Now given an observed adjacency matrix A, one may be interested in knowing whether A can be well fitted by a stochastic block model with k_0 communities and/or a membership vector σ_0. This leads to the following two hypothesis tests for fitness of the stochastic block model:

(1) $H_0 : k = k_0$ versus $H_1 : k > k_0$, and
(2) $H_0 : \sigma = \sigma_0$ versus $H_1 : \sigma \neq \sigma_0$,

where we use k and σ to denote the true number of communities and the true membership vector, respectively, and use k_0 and σ_0 to denote a hypothetical number of communities and a hypothetical membership vector, respectively. Note in hypothesis test (1), we consider the one-sided alternative in which nodes are partitioned into less than k communities (i.e., $k_0 < k$). For $k_0 > k$, nodes are partitioned into more than k communities. In this case, goodness-of-fit tests may not have theoretical guarantee, as a stochastic block model with k communities can also be reformulated as one with $k_0 > k$ communities by artificially splitting one or more true communities. As a result, we focus on the one-sided alternative $H_1 : k > k_0$, similar to what has been considered in Lei (2016), Chen and Lei (2018), and Wang and Bickel (2017).

Let the centered and rescaled adjacency matrix $\widetilde{A}$ be

$$\widetilde{A}_{ij} = \begin{cases} \frac{A_{ij} - \widehat{P}_{ij}^{\sigma_0}}{\sqrt{(n-1)\widehat{P}_{ij}^{\sigma_0}(1 - \widehat{P}_{ij}^{\sigma_0})}}, & i \neq j \\ 0, & i = j, \end{cases}$$

where $\widehat{P}_{ij}^{\sigma_0} = \widehat{B}_{\sigma_0(i)\sigma_0(j)}^{\sigma_0}$, as defined in (1). Under the null hypothesis $H_0 : k = k_0, \sigma = \sigma_0$, if $k = o(n^{1/6})$, Lei (2016) showed that

$$n^{2/3}(\lambda_1(\widetilde{A}) - 2) \xrightarrow{d} \text{TW}_1 \quad \text{and} \quad n^{2/3}(-\lambda_n(\widetilde{A}) - 2) \xrightarrow{d} \text{TW}_1,$$

where TW_1 denotes the Tracy–Widom distribution with index 1 and $\lambda_i(A)$ denotes the ith largest eigenvalue of the matrix A. Further, to test (1), Lei (2016) proposed to obtain $\widehat{\sigma}$ using spectral clustering (under $k = k_0$) and developed the following test statistic:

$$T_{n,k_0} = \max[n^{2/3}(\lambda_1(\widetilde{A}) - 2), n^{2/3}(-\lambda_n(\widetilde{A}) - 2)],$$

where σ_0 in $\widetilde{A}$ has been replaced by $\widehat{\sigma}$. Note T_{n,k_0} is a Bonferroni correction, and the corresponding level-α rejection rule is then

$$\text{reject } H_0 : k = k_0 \text{ if } T_{n,k_0} \ge t_{1-\alpha/2},$$

where t_α is the αth quantile of the TW_1 distribution for $\alpha \in (0, 1)$. As an improvement to many previous methods, the number of communities k in Lei (2016) is allowed to grow as n increases, but at the rate of $k = o(n^{1/6})$, which suggests that the test may not perform well when k is large.

We aim to develop a new test statistic that allows k to grow, up to a logarithm factor, linearly with n, and is able to test the goodness of fit of stochastic block models in both hypothesis tests (1) and (2). Most existing work in the literature have only considered the hypothesis test (1), while as we will see, as a natural by-product of our result, we are also able to consider the hypothesis test (2), which is often of practical interest as well. Moreover, the proposed test statistic can be extended to the degree-corrected stochastic block model. Specifically, we propose a new test statistic based on the maximum entry-wise deviation:

$$L_n(k_0, \sigma_0) \quad \triangleq \quad \max_{1 \le i \le n, 1 \le v \le k_0} \mid \widehat{\rho}_{iv} \mid,$$

where $\widehat{\rho}_{iv} \;=\; \frac{1}{\sqrt{|\sigma_0^{-1}(v)/\{i\}|}} \sum_{j \in \sigma_0^{-1}(v)/\{i\}} \frac{A_{ij} - \widehat{B}^{\sigma_0}_{\sigma_0(i)\sigma_0(j)}}{\sqrt{\widehat{B}^{\sigma_0}_{\sigma_0(i)\sigma_0(j)}(1 - \widehat{B}^{\sigma_0}_{\sigma_0(i)\sigma_0(j)})}}$, and $\sigma_0^{-1}(v)/\{i\}$ denotes the set of nodes, excluding node i, that belong to community v in σ_0.

2.1. The Asymptotic Null Distribution

To derive the asymptotic distribution for $L_n(k_0, \sigma_0)$, we make the following assumptions:

(A1) The entries of B are uniformly bounded away from 0 and 1, and B has no identical rows.

(A2) There exist $C_1 > 0$ and $C_2 > 0$ such that

$$C_1 n/k \le \min_{1 \le u \le k} \mid \sigma^{-1}(u) \mid \le \max_{1 \le u \le k} \mid \sigma^{-1}(u) \mid \le C_2 n^2/(k^2 \log^2 n)$$

for all n.

Condition (A1) requires that the entries in the probability matrix B are bounded away from 0 and 1, which was also considered in Lei (2016). At the same time, Condition (A1) requires that B is identifiable. Such a condition was considered in Wang and Bickel (2017) as well. Condition (A2) requires the size of the smallest community is at least proportional to n/k. This is a reasonable and mild condition; for example, it is satisfied almost surely if the membership vector σ is generated from a multinomial distribution with n trials and probability $\pi = (\pi_1, \dots, \pi_k)$ such that $\min_{1 \le u \le k} \pi_u \ge C_1/k$. Condition (A2) also places an upper bound on the largest community size. This is a reasonable condition as well and similar conditions have been considered by Zhang and Zhou (2016) and Gao et al. (2018). The upper bound on the largest community size is used to control the maximum grouped bias between $\widehat{B}_{\sigma(i)\sigma(j)}$ and its population version $B_{\sigma(i)\sigma(j)}$, that is,

$$\max_{1 \le i \le n, 1 \le v \le k} \times \left| \frac{1}{\sqrt{\mid \sigma^{-1}(v)/\{i\} \mid}} \sum_{j \in \sigma^{-1}(v)/\{i\}} \frac{B_{\sigma(i)\sigma(j)} - \widehat{B}_{\sigma(i)\sigma(j)}}{\sqrt{B_{\sigma(i)\sigma(j)}(1 - B_{\sigma(i)\sigma(j)})}} \right|$$

such that it converges in probability to 0.

We now state the asymptotic properties of $L_n(k_0, \sigma_0)$ and delay the proof to the supplementary materials.

Theorem 1. Suppose that Conditions (A1) and (A2) hold. Then under the null hypothesis $H_0 : k = k_0, \sigma = \sigma_0$, as $n \to \infty$, if $k = o(n/\log^2 n)$, we have

$$\frac{L_n(k_0, \sigma_0)}{\sqrt{\log(2k_0 n)}} \xrightarrow{P} \sqrt{2} \quad \text{and}$$

$$\lim_{n \to \infty} P(L_n^2(k_0, \sigma_0) - 2\log(2k_0 n) + \log\log(2k_0 n) \le y) = \exp\left\{ -\frac{1}{2\sqrt{\pi}} e^{-y/2} \right\}, \qquad (2)$$

where the right hand-side of (2) is the cumulative distribution function of the Gumbel distribution with $\mu = -2\log(2\sqrt{\pi})$ and $\beta = 2$.

Using the above theorem, we can carry out both hypothesis tests (1) and (2). To carry out hypothesis test (1), we need to first estimate the community membership $\widehat{\sigma}$ under $H_0 : k = k_0$, and then compute

$$T_n = L_n^2(k_0, \widehat{\sigma}) - 2\log(2k_0 n) + \log\log(2k_0 n).$$

Assume that $\widehat{\sigma}$ is strongly consistent (i.e., $P(\widehat{\sigma} = \sigma_0) \to 1$). Following Theorem 1, we have that T_n follows a Gumbel distribution with $\mu = -2\log(2\sqrt{\pi})$ and $\beta = 2$. To carry out the test, we reject $H_0 : k = k_0$, if $T_n > t_{(1-\alpha)}$, where t_α is the αth quantile of the Gumbel distribution with $\mu = -2\log(2\sqrt{\pi})$ and $\beta = 2$.

To obtain the asymptotic null distribution of T_n calculated with $\widehat{\sigma}$, the estimated $\widehat{\sigma}$ is required to be strongly consistent. This assumption is analogous to the strong consistency condition on $\widehat{\sigma}$ in Lei (2016) and the global optimum condition on the maximum likelihood estimation in Wang and Bickel (2017). Under Conditions (A1) and (A2), strong consistency (or exact recovery) is achievable when $k = o(n/\log^2 n)$ by Theorem 1.1 in Gao et al. (2018). To achieve strong consistency, we consider the majority voting algorithm in Gao et al. (2017), initialized by spectral clustering (Lei and Rinaldo 2015). Based on Theorem 4 in Gao et al. (2017), this procedure can achieve strong consistency when $k = o(n/\log^2 n)$. Alternatively, to obtain $\widehat{\sigma}$, one may consider spectral clustering combined with the sample splitting method in Lei and Zhu (2017), or the variational EM method in Daudin, Picard, and Robin (2008). While the latter two methods perform well empirically, they do not have theoretical guarantee on strong consistency when k diverges.

As for hypothesis test (2), since σ_0 gives a corresponding k_0, we can compute

$$T_n = L_n^2(k_0, \sigma_0) - 2\log(2k_0 n) + \log\log(2k_0 n),$$

and we reject $H_0 : \sigma = \sigma_0$, if $T_n > t_{(1-\alpha)}$, where t_α is again the αth quantile of the Gumbel distribution with $\mu = -2\log(2\sqrt{\pi})$ and $\beta = 2$. In Section 4, we carry out extensive simulation studies to investigate the finite sample performance of the two proposed tests of hypothesis.

2.2. The Asymptotic Power

In this section, we study the asymptotic power of the proposed tests. To do so, we first define a class of alternatives. For a stochastic block model with true membership vector σ and true probability matrix B, define probability matrix B^{σ_0} with respect to a given membership vector σ_0 as

$$B^{\sigma_0}_{uv} = \begin{cases} \frac{\sum_{i \in \sigma_0^{-1}(u), j \in \sigma_0^{-1}(v)} B_{\sigma(i)\sigma(j)}}{|\sigma_0^{-1}(u)| \cdot |\sigma_0^{-1}(v)|}, & u \ne v, \\ \frac{\sum_{i \ne j \in \sigma_0^{-1}(u)} B_{\sigma(i)\sigma(j)}}{|\sigma_0^{-1}(u)| \cdot (|\sigma_0^{-1}(u)| - 1)}, & u = v. \end{cases}$$

From the above definition, we can see that $B^{\sigma} = B$. We introduce the following condition on k_0 and σ_0:

(A2′) There exist $C_1 > 0$ and $C_2 > 0$ such that

$$C_1 n/k_0 \le \min_{1\le u\le k_0} \mid \sigma_0^{-1}(u) \mid \le \max_{1\le u\le k_0} \mid \sigma_0^{-1}(u) \mid \le C_2 n^2/(k_0^2 \log^2 n),$$

for all n.

This condition is analogous to (A2) and as we have argued, is a reasonably mild condition on community sizes.

Define the maximum grouped difference between B and B^{σ_0} as

$$\ell(k_0, \sigma_0) = \max_{1\le i\le n, 1\le v\le k_0} \times \left| \frac{1}{\sqrt{\mid \sigma_0^{-1}(v) \mid}} \sum_{j\in\sigma_0^{-1}(v)} (B_{\sigma(i)\sigma(j)} - B^{\sigma_0}_{\sigma_0(i)\sigma_0(j)}) \right|.$$

Consider the following alternative class of number of communities and membership vectors:

$$\mathcal{F}(k, \sigma, B) = \{(k_0, \sigma_0) : k_0 \le k,\ \ell(k_0, \sigma_0)/\sqrt{\log\, n} \longrightarrow \infty\}.$$

The set $\mathcal{F}(k, \sigma, B)$ specifies that under the alternative, the maximum grouped difference between B and B^{σ_0} diverges faster than $\sqrt{\log n}$. It can be seen that when $\sum_{j\in\sigma_0^{-1}(v)}(B_{\sigma(i)\sigma(j)} - B^{\sigma_0}_{\sigma_0(i)\sigma_0(j)}) = O(|\sigma_0^{-1}(v)|)$ for some i and v, under Condition (A2′) and $k_0 = o(n/\log^2 n)$, we have that $\ell(k_0, \sigma_0)/\sqrt{\log\, n} \longrightarrow \infty$. For example, when $k_0 = k$, for an alternative σ_0 such that $B^{\sigma_0} \neq B$ (up to row/column permutations), we have $\sum_{j\in\sigma_0^{-1}(v)}(B_{\sigma(i)\sigma(j)} - B^{\sigma_0}_{\sigma_0(i)\sigma_0(j)}) = O(|\sigma_0^{-1}(v)|)$ for some i and v, and consequently $(k_0, \sigma_0) \in \mathcal{F}(k, \sigma, B)$.

Given k, σ, and B, it is straightforward to calculate $\ell(k_0, \sigma_0)$ for an alternative (k_0, σ_0) and verify if it belongs to $\mathcal{F}(k, \sigma, B)$. We next provide some sufficient conditions for $(k_0, \sigma_0) \in \mathcal{F}(k, \sigma, B)$ when $k_0 < k$.

Corollary 1. Suppose that Conditions (A1) and (A2) hold. Consider the stochastic block model with B and σ from multinomial $(\pi_1, \ldots, \pi_k)$. Let B^- denote B after removing the diagonal entries, that is, $B^-_{u,\cdot} = (B_{uv})_{1\le v\le k, v\neq u}$, where $B^-_{u,\cdot}$ denotes the uth row of matrix B^-. For any $k_0 < k$ and σ_0 satisfying $\sigma_0(i) = \sigma_0(j)$ if $\sigma(i) = \sigma(j)$, we have $(k_0, \sigma_0) \in \mathcal{F}(k, \sigma, B)$, if at least one of the following conditions holds for some $c_0 > 0$:

(i) $\min_{u\neq v} |B_{uu} - B_{vv}| > c_0$,
(ii) $\min_{u\neq v} \|B^-_{u,\cdot} - B^-_{v,\cdot}\|_\infty > c_0$, where $\|\cdot\|_\infty$ denotes the vector infinity norm,
(iii) $\min_{u\neq v} |\pi_u/\pi_v - 1| > c_0$.

The proof is collected in the supplementary materials. In Corollary 1, we focus on the merged alternatives (i.e., communities in σ are merged to form communities in σ_0) to reduce the number of possible alternatives in developing the theoretical result, similar to that in Wang and Bickel (2017). Condition (i) specifies that the absolute differences between diagonal entries in B are lower bounded, Condition (ii) specifies that the differences, in terms of the infinity norm, between rows in B^- are lower bounded, and Condition (iii) specifies that the differences between elements in $(\pi_1, \ldots, \pi_k)$ are lower bounded. These conditions cover a large class of stochastic block models. Next, we discuss the asymptotic power of our proposed test against alternatives in $\mathcal{F}(k, \sigma, B)$. The following theorem provides a lower bound on the growth rate of the test statistic under an alternative $(k_0, \sigma_0) \in \mathcal{F}(k, \sigma, B)$.

Theorem 2. Suppose that Conditions (A1) and (A2′) hold. For any alternative $(k_0, \sigma_0) \in \mathcal{F}(k, \sigma, B)$, let $T_n = L_n^2(k_0, \sigma_0) - 2\log(2k_0 n) + \log\log(2k_0 n)$. If $k_0 = o(n/\log^2 n)$, we have

$$P(T_n \ge c_1 \log(n)) \to 1, \tag{3}$$

for some positive constant c_1.

The proof is collected in the supplementary materials. Theorem 2 shows that the growth rate of T_n under the alternative is at least $\log(n)$. The asymptotic null distribution in Theorem 1 and the growth rate under the alternative suggest that the null and the alternative hypotheses are well separated, and our proposed test is asymptotically powerful against $(k_0, \sigma_0) \in \mathcal{F}(k, \sigma, B)$. Specifically, our proposed test is asymptotically powerful when $k_0 < k$, if at least one of the conditions in Corollary 1 holds.

Notably, however, under the planted partition model (i.e., $B_{uu} = p$ and $B_{uv} = q, u \neq v$ for some $0 \le q < p \le 1$) with equal sized communities, some straightforward algebra shows that $\ell(k_0, \sigma_0) = 0$ for any (k_0, σ_0) satisfying $k_0 < k$ and $\sigma_0(i) = \sigma_0(j)$ if $\sigma(i) = \sigma(j)$. Consequently, such alternatives do not belong to $\mathcal{F}(k, \sigma, B)$. Additionally, one can verify that our test is not powerful against such alternatives under the planted partition model with equal sized communities. For example, consider a simple case with $k = 2$ and $\pi_1 = \pi_2$. Under $k_0 = 1$, we have $B^{\sigma_0} = (p + q)/2$, and the entry-wise deviation is calculated as

$$\rho_{i1} = \frac{1}{\sqrt{n}} \left| \sum_{j\in\sigma_0^{-1}(1)} \frac{A_{ij} - \frac{p+q}{2}}{\sqrt{\frac{p+q}{2}(1 - \frac{p+q}{2})}} \right| = \frac{1}{\sqrt{n}} \frac{\left|\sum_{j\in\sigma^{-1}(1)}(A_{ij} - p) + \sum_{j\in\sigma^{-1}(2)}(A_{ij} - q)\right|}{\sqrt{\frac{p+q}{2}(1 - \frac{p+q}{2})}}.$$

It can be seen that this entry-wise deviation is not well separated from those calculated under the null, that is, $\frac{1}{\sqrt{n/2}} \frac{\left|\sum_{j\in\sigma^{-1}(1)}(A_{ij}-p)\right|}{\sqrt{p(1-p)}}$ and $\frac{1}{\sqrt{n/2}} \frac{\left|\sum_{j\in\sigma^{-1}(2)}(A_{ij}-q)\right|}{\sqrt{q(1-q)}}$, and our proposed test is not powerful. The above calculation can be generalized to the cases where $k \ge 2$ and $k_0 < k$. Returning to the case of $k = 2$ and $k_0 = 1$, it is straightforward to show that when $|\pi_1/\pi_2 - 1| > c_0$ for some $c_0 > 0$, the growth rate of $\max_{i,v} \rho_{i1}$ is $\sqrt{n}$. Consequently, the null and the alternative are well separated, and our proposed test is powerful. More generally, for $k \ge 2$ and $k_0 < k$, if $\min_{u\neq v} |\pi_u/\pi_v - 1| > c_0$ for some $c_0 > 0$, our proposed test is powerful (see Corollary 1).

2.3. An Augmented Test Statistic

In this section, we discuss a practical solution to improve the power of the proposed test for hypothesis test (1) under the planted partition model with equal sized communities. Consider a planted partition model with n nodes, k equal sized communities, and within and between community connecting probabilities p and q, respectively. We propose adding a community of size $n/(2k)$ to the model. For the added community, we let the within and between community connecting probabilities be p' and q', respectively. Note that (p', q') can be the same as or different from (p, q). For this new model with $k+1$ communities, by Theorem 1, the asymptotic null distribution of the test statistic still follows a Gumbel distribution. Moreover, under an alternative $k_0 < k+1$, if the added small community is merged with others to form a new community in σ_0, we have $\ell(k_0, \sigma_0) \in \mathcal{F}(k, \sigma, B)$ and our proposed test is powerful. For spectral clustering based algorithms, such as that in Gao et al. (2017), it is reasonable to assume that the added small community is merged with others in σ_0 when $k_0 < k+1$, as this tends to lead to smaller within-cluster sum of squares.

The discussion above implies that when carrying out hypothesis test (1), an additional community can be added to the observed network to improve the power of our proposed test. We refer to the test statistic calculated with the added community as the *augmented test statistic*. Denote $k_0^+ = k_0 + 1$. For a given adjacency matrix A and null hypothesis $H_0 : k = k_0, \sigma = \sigma_0$, the augmented test statistic is calculated through the following steps:

1. Calculate $\widehat{B}$ using (1).
2. Add a k_0^+th community of size $n_{k_0^+} = \min_{1\le u\le k_0} |\sigma_0^{-1}(u)|/2$ to the observed network. For the added community, let the within and between community connecting probabilities be $\max_{1\le u\le k_0} \widehat{B}_{uu}$ and $\min_{u\neq v} \widehat{B}_{uv}/2$, respectively.
3. Calculate the size n^+ and the adjacency matrix A^+ of the network from step (2). With the membership vector $\sigma_0^+ = (\sigma_0, \underbrace{k_0^+, \dots, k_0^+}_{n_{k_0^+}})$, calculate the probability matrix $\widehat{B}^+$.
4. The augmented test statistic is calculated as

$$T_n^+ = L_n^2(k_0^+, \sigma_0^+) - 2\log(2k_0^+ n^+) + \log\log(2k_0^+ n^+).$$

To carry out hypothesis test (1), we reject the null hypothesis if $T_n^+ > t_{(1-\alpha)}$, where t_α is the αth quantile of the Gumbel distribution with $\mu = -2\log(2\sqrt{\pi})$ and $\beta = 2$.

Note that the asymptotic null distribution of the augmented test statistic T_n^+ is the same as T_n, provided that the added community satisfies Conditions (A1) and (A2). Under Conditions (A1) and (A2), other procedures (e.g., different size or connecting probabilities) for adding the additional community are also feasible. We adopt the above procedure as it is easy to implement and shows good empirical performance in our numerical studies.

3. Extension to the Degree-Corrected Stochastic Block Model

It has been observed that a typical real-world network often contains a few high-degree "hub" nodes which have many edges and many low-degree nodes that have few edges. The stochastic block model, however, does not accommodate such heterogeneity. To incorporate the degree heterogeneity of nodes for community detection, Karrer and Newman (2011) proposed the degree-corrected stochastic block model. Specifically, the degree-corrected stochastic block model assumes that $P(A_{ij} = 1 \mid \sigma(i) = u, \sigma(j) = v) = \omega_i\omega_j B_{uv}$, where $\omega = (\omega_i)_{1\le i\le n}$ are a set of node degree parameters measuring the degree variation. For identifiability of the model, we use the following constraint for the degree-corrected stochastic block model:

(A3) $\sum_i \omega_i \mathbf{1}\{\sigma(i) = u\} = |\sigma^{-1}(u)|$ for $1 \le u \le k$.

To develop a goodness-of-fit test for the degree-corrected stochastic block model, we consider two cases: (1) ω is known, and (2) ω is unknown. We first consider the case where ω is known. In this case, we propose the following test statistic:

$$L_{n1}(k_0, \sigma_0) \quad \triangleq \quad \max_{1\le i\le n, 1\le v\le k_0} \mid \widehat{\tau}_{iv} \mid,$$

where $\widehat{\tau}_{iv} = \frac{1}{\sqrt{|\sigma_0^{-1}(v)/\{i\}|}} \sum_{j\in\sigma_0^{-1}(v)/\{i\}} \frac{A_{ij} - \omega_i\omega_j \widehat{B}^{\sigma_0}_{\sigma_0(i)\sigma_0(j)}}{\sqrt{\omega_i\omega_j \widehat{B}^{\sigma_0}_{\sigma_0(i)\sigma_0(j)}(1 - \omega_i\omega_j \widehat{B}^{\sigma_0}_{\sigma_0(i)\sigma_0(j)})}}$
.

To derive the asymptotic distribution of $L_{n1}(k_0, \sigma_0)$, we make the following additional assumption:

(A4) The entries of $(\omega_i\omega_j B_{\sigma(i)\sigma(j)})_{1\le i\le n, 1\le j\le n}$ are uniformly bounded away from 0 and 1.

We now state the asymptotic properties of $L_{n1}(k_0, \sigma_0)$.

Theorem 3. Suppose that Conditions (A2)–(A4) hold. Then under the null hypothesis $H_0 : k = k_0, \sigma = \sigma_0$, as $n \to \infty$, if $k = o(n/\log^2 n)$, we have

$$\frac{L_{n1}(k_0, \sigma_0)}{\sqrt{\log(2k_0 n)}} \xrightarrow{P} \sqrt{2} \quad \text{and}$$

$$\lim_{n\to\infty} P(L_{n1}^2(k_0, \sigma_0) - 2\log(2k_0 n) + \log\log(2k_0 n) \le y)$$
$$= \exp\left\{-\frac{1}{2\sqrt{\pi}} e^{-y/2}\right\}.$$

Note that $E(\frac{A_{ij} - \omega_i\omega_j P_{ij}}{\sqrt{\omega_i\omega_j P_{ij}(1 - \omega_i\omega_j P_{ij})}}) = 0$ and $E(\frac{A_{ij} - \omega_i\omega_j P_{ij}}{\sqrt{\omega_i\omega_j P_{ij}(1 - \omega_i\omega_j P_{ij})}})^2 = 1$, which is analogous to the result under the stochastic block model, in which $E(\frac{A_{ij} - B_{\sigma(i)\sigma(j)}}{\sqrt{B_{\sigma(i)\sigma(j)}(1 - B_{\sigma(i)\sigma(j)})}}) = 0$ and $E(\frac{A_{ij} - B_{\sigma(i)\sigma(j)}}{\sqrt{B_{\sigma(i)\sigma(j)}(1 - B_{\sigma(i)\sigma(j)})}})^2 = 1$. Henceforth, the proof of Theorem 3 is very similar to that of Theorem 1, and we omit the details in the article. Using the result in the above theorem, we can carry out hypothesis tests (1) and (2) using the test Φ_α defined as

$$\Phi_\alpha = I(T_{n1} > t_{(1-\alpha)}),$$

where $T_{n1} = L_{n1}^2(k_0, \sigma_0) - 2\log(2k_0 n) + \log\log(2k_0 n)$ and t_α is the αth quantile of the Gumbel distribution with $\mu = -2\log(2\sqrt{\pi})$ and $\beta = 2$. To estimate $\widehat{\sigma}$, we adopt the regularized spherical spectral clustering algorithm in Lei and Rinaldo (2015). Other methods such as the SCORE algorithm in Jin (2015) and the normalized neighbor voting procedure in Gao et al. (2018) can also be considered. Following similar arguments as in the case of stochastic block model, it can also be shown that the test Φ_α is powerful against a class of alternatives defined similarly as in (2.2). Following similar arguments as in Section 2.2, we can show that when the communities are equal sized and $B_{uu} = p$ and $B_{uv} = q, u \neq v$ for some $0 \le q < p \le 1$, our proposed test is not powerful when $k_0 < k$. To overcome this challenge, we propose an augmented test statistic for the degree-corrected stochastic block model. The calculation of this augmented test statistic is similar to that under the stochastic block model, and we include the computational details in the supplementary materials.

If ω is unknown, we can plug in its estimate for $L_{n1}(k_0, \sigma_0)$. Similar to Karrer and Newman (2011), we replace the Bernoulli distribution of A_{ij} by the Poisson distribution with the mean $\omega_i\omega_j B_{uv}$. As discussed in Zhao, Levina, and Zhu (2012), there is no practical difference in performance between the log-likelihood and its slightly more elaborate version based on the Bernoulli observations. The reason is that the Bernoulli distribution with a small mean can be well approximated by a Poisson distribution. One advantage of using the Poisson distribution is that it greatly simplifies the calculation. Another advantage is that it admits networks containing both multi-edges and self-edges. Specifically, for any fixed (B, ω, σ), the log-likelihood of observing the adjacency matrix A under the degree-corrected stochastic block model can be written as

$$\ell(A|B, \omega, \sigma) = \sum_{1\le i\le n} d_i \log \omega_i + \frac{1}{2}\sum_{u,v=1}^{k}(m_{uv}\log B_{uv} - n_{uv}B_{uv}),$$

where $d_i = \sum_{1\le j\le n} A_{ij}$, and m_{uv} and n_{uv} are defined the same as before. It is not difficult to show that given σ_0, the maximum likelihood estimate of the parameter ω is given by $\widehat{\omega}_i = |\sigma_0^{-1}(u)|d_i / \sum_{j:\sigma_0(j)=\sigma_0(i)} d_j$. Then the proposed plug-in test statistic is given by

$$L_{n2}(k_0, \sigma_0) \quad \triangleq \quad \max\nolimits_{1\le i\le n, 1\le v\le k_0} \mid \widehat{\tau}_{iv} \mid, \tag{4}$$

where $\widehat{\tau}_{iv} = \frac{1}{\sqrt{|\sigma_0^{-1}(v)/\{i\}|}} \sum_{j\in\sigma_0^{-1}(v)/\{i\}} \frac{A_{ij} - \widehat{\omega}_i\widehat{\omega}_j\widehat{B}^{\sigma_0}_{\sigma_0(i)\sigma_0(j)}}{\sqrt{\widehat{\omega}_i\widehat{\omega}_j\widehat{B}^{\sigma_0}_{\sigma_0(i)\sigma_0(j)}(1-\widehat{\omega}_i\widehat{\omega}_j\widehat{B}^{\sigma_0}_{\sigma_0(i)\sigma_0(j)})}}$.

When ω is unknown, it is very challenging to derive the asymptotic distribution of $L_{n2}(k_0, \sigma_0)$, due to the complex dependency between the centered and rescaled entries in $\widehat{\tau}_{iv}$. We perform simulation studies and find that the empirical distribution of $L_{n2}^2(k_0, \sigma_0) - 2\log(2k_0 n) + \log\log(2k_0 n)$ deviates from the Gumbel distribution by a location and scale shift (see Figure 2). This shift is especially large when the number of communities k is small. As a practical solution, in Section S2.2, we describe a bootstrap correction procedure. With the bootstrap corrected test statistic, both hypothesis tests (1) and (2) can be carried out, similar to what have been done in Section 2.

4. Simulation Studies

In this section, we carry out extensive simulation studies to evaluate the performance of the proposed test statistic. We consider both the stochastic block model and the degree-corrected stochastic block model. In the stochastic block model setting, the majority voting algorithm in Gao et al. (2017), initialized by spectral clustering is used to obtain the community membership, whereas in the degree-corrected stochastic block model setting, the regularized spherical spectral clustering algorithm in Lei and Rinaldo (2015) is employed. For the stochastic block model, we consider the test statistic $T_n = L_n^2(k_0, \sigma_0) - 2\log(2k_0 n) + \log\log(2k_0 n)$, and the augmented test statistic T_n^+ proposed in Section 2.3. For the degree-corrected stochastic block model, we consider $T_{n2} = L_{n2}^2(k_0, \sigma_0) - 2\log(2k_0 n) + \log\log(2k_0 n)$, and the augmented test statistic T_{n2}^+ proposed in Section S2.1. In our comparative simulation studies, Lei (2016), Karwa et al. (2016) and our method are all implemented in R.[1]

4.1. Simulation 1: The Null Distribution Under the Stochastic Block Model and a Bootstrap Correction

In this simulation, we examine the finite sample null distribution of the test statistic T_n and verify the result in Theorem 1. As the speed of convergence to a Gumbel distribution may be slow, one may consider a finite sample bootstrap correction. Such a finite-sample correction was first proposed in Bickel and Sarkar (2015) and later considered in Lei (2016). Here, we extend their ideas to our setting.

For an adjacency matrix A and null hypothesis $k = k_0, \sigma = \sigma_0$, the bootstrap corrected goodness-of-fit test statistic is calculated as the following:

1. Calculate $\widehat{B}$ using (1). Calculate T_n using A and $(\widehat{B}, \sigma_0)$.
2. For $m = 1, \dots, M$, generate $A^{(m)}$ from the stochastic block model $(\widehat{B}, \sigma_0)$, and calculate $T_n^{(m)}$ using $A^{(m)}$ and $(\widehat{B}, \sigma_0)$.
3. Using $(T_n^{(m)} : 1 \le m \le M)$, estimate the location and scale parameters $\widehat{\mu}$ and $\widehat{\beta}$ of the Gumbel distribution using maximum likelihood estimation.
4. The bootstrap corrected test statistic is calculated as

$$T_{n,\text{boot}} = \mu + \beta\left(\frac{T_n - \widehat{\mu}}{\widehat{\beta}}\right),$$

where $\mu = -2\log(2\sqrt{\pi})$ and $\beta = 2$.

Since the limiting distribution of the test statistic is provably Gumbel, finite sample corrections can be made inexpensively by generating a small number of bootstrap samples to estimate the location and scale parameters. In all of our simulations, we use $M = 100$.

In Figure 1, we plot the distribution of T_n with and without bootstrap corrections from 1000 data replications. In this simulation, we set $k = k_0 = 3$ with $\pi_1 = \pi_2 = \pi_2 = 1/3$. The edge probability between communities u and v is $B_{uv} = 0.1(1 + 2 \times \mathbf{1}(u = v))$. We consider sample sizes $n = 300$ and $n = 1500$. It can be seen that the finite sample null distribution of T_n deviates slightly from the limiting distribution when $n = 300$, and the difference is much reduced when $n = 1500$. When bootstrap correction is considered, the finite sample null distribution is close to the limit even when $n = 300$.

4.2. Simulation 2: Hypothesis Test (1) Under the Stochastic Block Model

In the stochastic block model setting, we consider the hypothesis test

$$H_0 : k = k_0 \;\text{ versus }\; H_1 : k > k_0.$$

[1] We obtained the code for Lei (2016) from the author's website and implemented the code for Karwa et al. (2016) by ourselves following the algorithm proposed in the article.

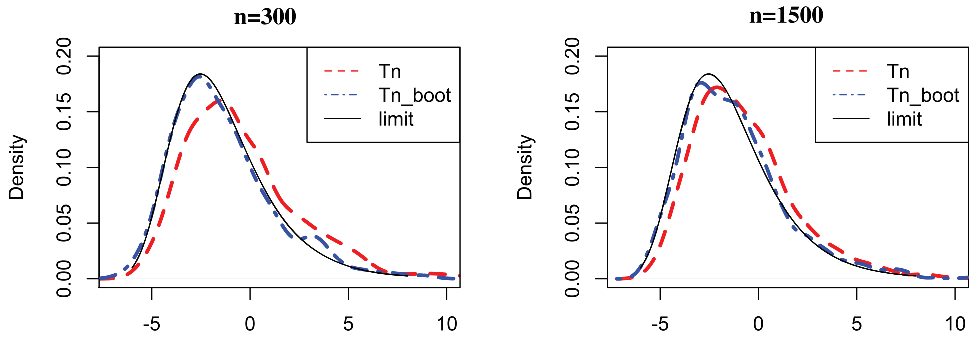

Figure 1. Null densities under the stochastic block model in Simulation 1 with $n = 300$ (left plot) and $n = 1500$ (right plot). The red dashed lines, blue dash-dotted lines, and black solid lines show the densities of the test statistic T_n, the bootstrap corrected test statistic $T_{n,\text{boot}}$, and the theoretical limit, respectively.

Table 1. Proportion of rejection at nominal level $\alpha = 0.05$ for hypothesis test $H_0 : k = k_0$ versus $H_1 : k > k_0$.

	T_n					T_n^+					$T_{n,\text{boot}}^+$				
k	2	4	6	8	10	2	4	6	8	10	2	4	6	8	10
$k_0 = 2$	**0.03**	0.09	0.43	0.66	0.82	**0.06**	1.00	1.00	1.00	1.00	**0.05**	1.00	1.00	1.00	1.00
$k_0 = 4$	*	**0.04**	0.08	0.42	0.88	*	**0.08**	1.00	1.00	1.00	*	**0.05**	1.00	1.00	1.00
$k_0 = 6$	*	*	**0.08**	0.14	0.28	*	*	**0.08**	1.00	1.00	*	*	**0.04**	1.00	1.00
$k_0 = 8$	*	*	*	**0.08**	0.14	*	*	*	**0.09**	1.00	*	*	*	**0.05**	1.00
$k_0 = 10$	*	*	*	*	**0.08**	*	*	*	*	**0.10**	*	*	*	*	**0.04**

NOTE: Each community has 200 nodes and $B_{uv} = 0.1(1 + 4 \times \mathbf{1}(u = v))$. * indicates alternatives that are not considered (since we only consider a one-sided test with the alternative $H_1 : k > k_0$).

Table 2. Proportion of rejection at nominal level $\alpha = 0.05$ for $H_0 : k = k_0$ versus $H_1 : k > k_0$.

	$T_{n,\text{boot}}^+$							Lei$_{\text{boot}}$						
k	1	2	3	5	10	20	30	1	2	3	5	10	20	30
$k_0 = 1$	**0.03**	1.00	1.00	1.00	1.00	1.00	1.00	**0.04**	1.00	1.00	1.00	1.00	1.00	1.00
$k_0 = 2$	*	**0.06**	1.00	1.00	1.00	1.00	1.00	*	**0.04**	1.00	1.00	1.00	1.00	1.00
$k_0 = 3$	*	*	**0.06**	1.00	1.00	1.00	1.00	*	*	**0.02**	1.00	1.00	1.00	1.00
$k_0 = 5$	*	*	*	**0.05**	1.00	1.00	1.00	*	*	*	**0.03**	1.00	1.00	1.00
$k_0 = 10$	*	*	*	*	**0.06**	1.00	1.00	*	*	*	*	**0.46**	1.00	1.00
$k_0 = 20$	*	*	*	*	*	**0.04**	1.00	*	*	*	*	*	**0.82**	1.00
$k_0 = 30$	*	*	*	*	*	*	**0.04**	*	*	*	*	*	*	**0.98**

NOTE: The network size is $n = 3000$ with equal sized communities, and $B_{uv} = 0.1(1 + 4 \times \mathbf{1}(u = v))$. * indicates alternatives that are not considered (since we only consider a one-sided test with the alternative $H_1 : k > k_0$).

We first compare the performance of the test statistic T_n, the augmented test statistic T_n^+ and the bootstrap corrected augmented test statistic $T_{n,\text{boot}}^+$ under varying k and k_0. We set the edge probability between communities u and v as $0.1(1 + 4 \times \mathbf{1}(u = v))$, and let the size of each block be 200. Table 1 reports the result from 100 data replications. While the Type I errors from all three test statistics are close to the nominal level, $T_{n,\text{boot}}^+$'s Type I errors are closer to the nominal level when k is large. As this simulation setting considers a planted partition model with equal sized communities, T_n does not have good power. This agrees with our theoretical results in Section 2.2. It is seen that, with the augmentation, the power from both T_n^+ and $T_{n,\text{boot}}^+$ improve significantly.

Next, we compare our method with Lei (2016). For their test statistic, we also use the bootstrap correction procedure suggested in their work, and this test statistic is referred to as Lei$_{\text{boot}}$. We fix the network size at $n = 3000$ and let both k and k_0 vary. Table 2 reports the results from $T_{n,\text{boot}}^+$ and Lei$_{\text{boot}}$ from 200 data replications. It can be seen from Table 2 that the two tests have comparable Type I errors when k is small (i.e., $k \leq 5$). However, when k is large (i.e., $k \geq 10$), the Type I errors from $T_{n,\text{boot}}^+$ are much closer to the nominal level. This agrees with our theoretical finding that our proposed test allows k to grow at a much faster rate than that of Lei (2016). Moreover, it is seen that both tests have good power. Specifically, both tests are powerful against the Erdős–Rényi model alternative (i.e., $k_0 = 1$) when $k \geq 2$.

We have also run comparative simulations with sparser networks, networks with unbalanced community sizes and networks with randomly generated B. In the interest of space, we report these additional results in the supplementary materials.

4.3. Simulation 3: Hypothesis Test (2) Under the Stochastic Block Model

In the stochastic block model setting, we also consider the hypothesis test

$$H_0 : \sigma = \sigma_0 \quad \text{versus} \quad H_1 : \sigma \neq \sigma_0.$$

We use the true number of communities k when we obtain the membership vector σ_0. We investigate the probability of Type I error of the test statistic T_n and $T_{n,\text{boot}}$. The network size n is the same as in Simulation 2. The edge probability between communities u and v is $0.1(1 + 2 \times \mathbf{1}(u = v))$. Each simulation is repeated 200 times. The simulation results are given in Table 3. It can be seen from this table that the probabilities of Type I error

Table 3. Proportion of rejection at nominal level $\alpha = 0.05$ for hypothesis test $H_0 : \sigma = \sigma_0$ versus $H_1 : \sigma \neq \sigma_0$ under settings in Simulation 3.

$k = k_0$	2	3	4	5	6	7	8
T_n	0.05	0.05	0.07	0.07	0.09	0.07	0.10
$T_{n,\text{boot}}$	0.04	0.08	0.05	0.03	0.06	0.03	0.04
Karwa et al. (2016)	0.19	0.10	0.14	0.15	0.15	0.15	0.20

of both T_n and $T_{n,\text{boot}}$ are close to the nominal level, with $T_{n,\text{boot}}$ having a slightly smaller Type I error when k is large. We also compare our method with Karwa et al. (2016). We can see that the estimated Type I errors from our tests are much closer to the nominal level than that of Karwa et al. (2016).

We have also run simulations to examine the power of the test when a proportion of the labels in σ are corrupted. We find that our test is powerful under this setting. In the interest of space, we report these additional results in the supplementary materials.

4.4. Simulation 4: The Null Distribution Under the Degree-Corrected Stochastic Block Model and a Bootstrap Correction

In this simulation, we examine the finite sample null distribution of the test statistic T_{n2} under the degree-corrected stochastic block model. Similar to Simulation 1, we also consider a finite sample bootstrap correction. The calculation for the bootstrap corrected test statistic $T_{n2,\text{boot}}$ is similar to that in Simulation 1 and we include the computational details in the supplementary materials.

To generate the degree parameters ω, we follow the approach in Zhao, Levina, and Zhu (2012). The identifiability constraint $\sum_i \omega_i \mathbf{1}\{\sigma(i) = u\} = |\sigma^{-1}(u)|$ for each community $1 \leq u \leq k$ is replaced by the requirement that the ω_i be independently generated from a distribution with unit expectation, that is,

$$\omega_i = \begin{cases} \eta_i, & \text{w.p. } 0.8, \\ 9/11, & \text{w.p. } 0.1, \\ 13/11, & \text{w.p. } 0.1, \end{cases}$$

where η_i is uniformly distributed on the interval $[\frac{4}{5}, \frac{6}{5}]$. In this simulation, we consider $k = k_0 = 3$ with $\pi_u = 1/3$, $u = 1, \ldots, 3$, and $k = k_0 = 5$ with $\pi_u = 1/5$, $u = 1, \ldots, 5$. The edge probability between communities u and v is $B_{uv} = 0.1(1 + 2 \times \mathbf{1}(u = v))$. We consider sample sizes $n = 300, 500$, and 1500.

In Figure 2, we plot the distribution of T_{n2} with and without bootstrap corrections from 1000 data replications. It can be seen that when $k = 3$, the null distribution of T_{n2} deviates from the Gumbel distribution by a location and scale shift. Such a deviation does not decrease even when the sample size is increased to $n = 1500$. However, when k is increased to 5, the distribution of T_{n2} is much less deviated from the Gumbel distribution in Theorem 3. Note that when the bootstrap correction is considered, the sample null distribution is close to the limit even when $k = 3$ and $n = 300$.

4.5. Simulation 5: Hypothesis Test Under the Degree-Corrected Stochastic Block Model

In the degree-corrected stochastic block model setting, we then consider the hypothesis test

$$H_0 : k = k_0 \;\text{ versus }\; H_1 : k > k_0.$$

We investigate the probability of Type I error and the power of the test statistic $T^{+}_{n2,\text{boot}}$, which is the augmented test statistic (calculated as in Section S2.1) with bootstrap correction. The connecting probabilities are generated the same as in Simulation 4. We consider equal sized communities with 200 nodes each. Each simulation is repeated 100 times. The simulation results are given in Table 4. It is seen that the probability of Type I error is close to the nominal level and the test also shows good power.

Table 4. Proportion of rejection at nominal level $\alpha = 0.05$ for hypothesis test $H_0 : k = k_0$ versus $H_1 : k > k_0$ under the setting in Simulation 5.

	$T^{+}_{n,\text{boot}}$				
k	2	4	6	8	10
$k_0 = 2$	**0.05**	1.00	1.00	1.00	1.00
$k_0 = 4$	*	**0.06**	1.00	1.00	1.00
$k_0 = 6$	*	*	**0.04**	1.00	1.00
$k_0 = 8$	*	*	*	**0.04**	1.00
$k_0 = 10$	*	*	*	*	**0.07**

NOTE: * indicates alternatives that are not considered (since we only consider a one sided test with the alternative $H_1 : k > k_0$).

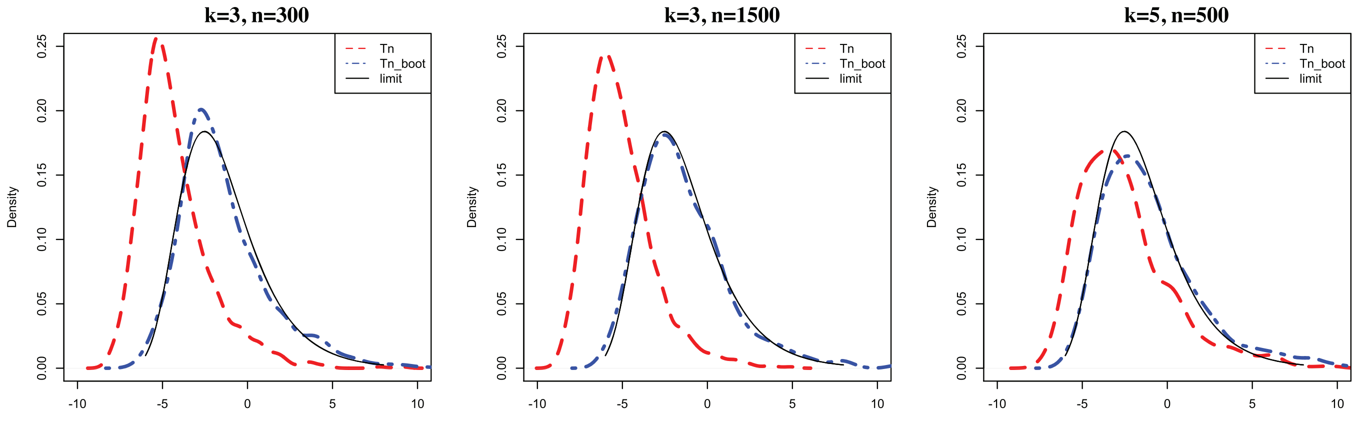

Figure 2. Null densities under the degree-corrected stochastic block model setting in Simulation 4 with $k = 3, n = 300$ (left plot), $k = 3, n = 1500$ (middle plot), and $k = 5, n = 500$ (right plot). The red dashed lines, blue dash-dotted lines, and black solid lines show the densities of the test statistic T_{n2}, the bootstrap corrected test statistic $T_{n2,\text{boot}}$, and the theoretical limit, respectively.

5. Data Examples

5.1. International Trade Data

In this subsection, we apply the proposed method to the international trade dataset that was studied in Westveld and Hoff (2011). The dataset contains yearly international trade information among $n = 58$ countries from 1981 to 2000. The original network is directed and weighted, in which each node corresponds to a country and for a given year, Trade_{ij} indicates the amount of exports from country i to country j; see Westveld and Hoff (2011) for details. Saldana, Yu, and Feng (2017) revisited the dataset for the purpose of estimating the number of communities. Following Saldana, Yu, and Feng (2017), we focus on the international trade network in 1995 and transform the directed and weighted adjacency matrix to an undirected binary matrix. Specifically, let $W_{ij} = \text{Trade}_{ij} + \text{Trade}_{ji}$, and we set $A_{ij} = 1$ if $W_{ij} \geq W_{0.5}$, and $A_{ij} = 0$ otherwise, where $W_{0.5}$ denotes the 50th percentile of $\{W_{ij}\}_{1 \leq i < j \leq n}$. Saldana, Yu, and Feng (2017) used three different methods and identified three different numbers of communities, specifically, 3, 7, and 10, respectively. Note due to the limited size of the network, some communities can be very small (e.g., less than 5 nodes). In that case, the augmentation procedure may not work well since the added community may have less than or equal to 2 nodes. We thus use $T_{n,\text{boot}}$ as the test statistic for the stochastic block model and obtain $T_{n,\text{boot}} = 44.28, 3.33$, and 13.44 for $k_0 = 3$, 7, and 10, respectively. Since $t_{0.95} = 3.41$ for the Gumbel distribution, we do not reject $H_0 : k = 7$ at the level of 0.05. As discussed in Saldana, Yu, and Feng (2017), $k = 7$ seems to be a reasonable choice for the number of communities, corresponding to countries with highest GDPs, industrialized European and Asian countries with medium-level GDPs, and developing countries in South America with the lowest GDPs.

5.2. Political Blog Data

In this subsection, we use the political blog network (Adamic and Glance 2005) to demonstrate the proposed methods for both the stochastic block model and the degree-corrected stochastic block model. The dataset consists of political blogs, with edges representing web links. Each node is labeled either as "conservative" or "liberal" based on the blogger's political stance. We only consider the largest connected component of this network which consists of 1222 nodes as is commonly done in the literature. Chen and Lei (2018) applied a network cross-validation method to the political blog data to select the number of communities, and they identified $k = 10$ and $k = 2$, respectively, for the stochastic block model and the degree-corrected stochastic block model. Here, we reanalyze the data using the test statistics that we have developed to test the significance of the number of communities identified by Chen and Lei (2018). Specifically, we use $T^{+}_{n,\text{boot}}$ and $T^{+}_{n2,\text{boot}}$ as the test statistic for the stochastic block model and the degree-corrected stochastic block model, respectively. We obtain $T^{+}_{n,\text{boot}} = 23.59$ under $k = 10$, and $T^{+}_{n2,\text{boot}} = 2.06$ under $k = 2$. Since $t_{0.95} = 3.41$ for the Gumbel distribution, at the level of 0.05, we would reject $H_0 : k = 10$ under the stochastic block model and not reject $H_0 : k = 2$ under the degree-corrected stochastic block model. The result of our analysis agrees with that of Chen and Lei (2018) for the degree-corrected stochastic block model but not so for the stochastic block model. It is possible that the stochastic block model is not an appropriate model for this particular dataset as it was observed that there is a big variation among node degrees.

6. Discussion

In this article, we have developed a novel goodness-of-fit test based on the maximum entry-wise deviation of the centered and rescaled observed adjacency matrix and demonstrated that its asymptotic null distribution is the Gumbel distribution when $k = o(n/\log^2 n)$, which significantly relaxes the condition in Lei (2016). The test is different from those used in traditional methods based on independent random variables, in which the goodness of fit is assessed by the sum of residual squares. For stochastic block models, the residual is a matrix, and the proposed test incorporates the signal change among different blocks nested in the residual matrix to test the goodness of fit of the model. In the case of degree-corrected stochastic block model with unknown degree parameters, through simulation studies, we show that the distribution of T_{n2} under the null deviates from the Gumbel distribution with $\mu = -2\log(2\sqrt{\pi})$ and $\beta = 2$. Finding the asymptotic null distribution of T_{n2} under the degree-corrected stochastic block model is a challenging task, as the estimated degree parameters $\widehat{\omega}_i$, $i = 1, \ldots, n$, introduce complex dependencies between the entries of the rescaled adjacency matrix. As such, the theoretical arguments used in the current article can no longer be directly applied or extended to obtain the asymptotic null distribution of T_{n2}.

Both our work and Lei (2016) consider dense networks, that is, entries in the probability matrix B are bounded away from 0. In practice, networks can be sparse. To admit the sparse case, the probability matrix B is often assumed to be of the form $B = \rho_n B_0$, where the entries of B_0 are of order 1, and ρ_n is a parameter controlling the sparsity of the network and allowed to decrease to zero when n increases (Bickel and Chen 2009). Some recent work that consider tests based on subgraph statistics have obtained closed-form asymptotic null distributions in the sparse regime (Gao and Lafferty 2017; Jin, Ke, and Luo 2019). However, these asymptotic null distributions only hold under $k = 1$, that is, the Erdős–Rényi model. Thus, these tests can only be used to test if there are communities in the network but not how many are there. In addition, Wang and Bickel (2017) considers a likelihood-based model selection approach and derives the asymptotic distribution of the log-likelihood ratio statistic in the sparse case. However, the number of communities k is fixed in their work. For our method, when the network is sparse, existing arguments do not guarantee the moderate deviation bound (see Lemma 2) due to the heavy tail of the centered and rescaled adjacency entry $(A_{ij} - B_{\sigma(i)\sigma(j)})/\sqrt{B_{\sigma(i)\sigma(j)}(1 - B_{\sigma(i)\sigma(j)})}$. It would be of interest to investigate whether the moderate deviation bound in Lemma 2 can be modified to consider such heavy-tailed scenarios in future work.

Supplementary Materials

The supplementary materials collect all technical proofs, additional computational details and simulation results.

Acknowledgments

We are very grateful to two anonymous referees, an associate editor, and the editor for their valuable comments that have greatly improved the article.

Funding

Hu is partially supported by the National Natural Science Foundation of China (nos. 11471136, 11571133) and the Key Laboratory of Applied Statistics of MOE (KLAS) grants (nos. 130026507, 130028612). Qin is partially supported by the National Natural Science Foundation of China (no. 11871237). Yan is partially supported by the National Natural Science Foundation of China (no. 11771171), the Fundamental Research Funds for the Central Universities, and the Key Laboratory of Applied Statistics of MOE (KLAS) grants (nos. 130026507, 130028612). Zhu is partially supported by the National Science Foundation (DMS 1407698 and DMS 1821243). Zhang is partially supported by the National Science Foundation (DMS 2015190).

References

Adamic, L. A., and Glance, N. (2005), "The Political Blogosphere and the 2004 US Election," in *Proceedings of the WWW—2005 Workshop on the Weblogging Ecosystem*, ACM, pp. 36–43. [1381]

Amini, A. A., Chen, A., Bickel, P. J., and Levina, E. (2013), "Pseudo-Likelihood Methods for Community Detection in Large Sparse Networks," *The Annals of Statistics*, 41, 2097–2122. [1373]

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), "Mixed Membership Stochastic Blockmodels," *The Journal of Machine Learning Research*, 9, 1981–2014.

Arritia, R, Goldstein, L., and Gordon, L. (1989), "Two Moments Suffice for Poisson Approximations: The Chen-Stein Method," *The Annals of Probability*, 17, 9–25.

Bhattacharyya, S., and Bickel, P. J. (2016), "Spectral Clustering and Block Models: A Review and a New Algorithm," in *Statistical Analysis for High-Dimensional Data*, eds. A. Frigessi, P. Bühlmann, I. Glad, M. Langaas, S. Richardson, and M. Vannucci, Cham: Springer, pp. 67–90. [1373]

Bickel, P. J., and Chen, A. (2009), "A Nonparametric View of Network Models and Newman-Girvan and Other Modularities," *Proceedings of the National Academy of Sciences of the United States of America*, 106, 21068–21073. [1373,1381]

Bickel, P. J., and Sarkar, P. (2015), "Hypothesis Testing for Automated Community Detection in Networks," *Journal of the Royal Statistical Society*, Series B, 78, 253–273. [1373,1378]

Cai, T. T., and Jiang, T. (2011), "Limiting Laws of Coherence of Random Matrices With Applications to Testing Covariance Structure and Construction of Compressed Sensing Matrices," *The Annals of Statistics*, 39, 1496–1525.

Chatterjee, S. (2015), "Matrix Estimation by Universal Singular Value Thresholding," *The Annals of Statistics*, 43, 177–214.

Chen, K., and Lei, J. (2018), "Network Cross-Validation for Determining the Number of Communities in Network Data," *Journal of the American Statistical Association*, 113, 241–251. [1374,1381]

Chen, X. (1990), "Probabilities of Moderate Deviations for B-Valued Independent Random Vectors," *Chinese Annals of Mathematics*, 11, 621–629.

Choi, D. S., Wolfe, P. J., and Airoldi, E. M. (2012), "Stochastic Blockmodels With a Growing Number of Classes," *Biometrika*, 99, 273–284. [1373]

Daudin, J. J., Picard, F., and Robin, S. (2008), "A Mixture Model for Random Graphs," *Statistics and Computing*, 18, 173–183. [1373,1375]

Gao, C., and Lafferty, J. (2017), "Testing for Global Network Structure Using Small Subgraph Statistics," arXiv no. 1710.00862. [1381]

Gao, C., Ma, Z., Zhang, Y., and Zhou, H. (2017), "Achieving Optimal Misclassification Proportion in Stochastic Block Models," *The Journal of Machine Learning Research*, 18, 1980–2024. [1375,1377,1378]

——— (2018), "Community Detection in Degree-Corrected Block Models," *The Annals of Statistics*, 46, 2153–2185. [1375,1377]

Hoeffding, W. (1963), "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association*, 58, 13–30.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983), "Stochastic Blockmodels: First Steps," *Social Networks*, 5, 109–137. [1373]

Jin, J. (2015), "Fast Community Detection by SCORE," *The Annals of Statistics*, 43, 57–89. [1373,1377]

Jin, J., Ke, Z. T., and Luo, S. (2019), "Optimal Adaptivity of Signed-Polygon Statistics for Network Testing," arXiv no. 1904.09532. [1381]

Jiang, T. (2016), "The Asymptotic Distributions of the Largest Entries of Sample Correlation Matrices," *The Annals of Applied Probability*, 14, 865–880. [1374]

Karrer, B., and Newman, M. E. J. (2011), "Stochastic Blockmodels and Community Structure in Networks," *Physical Review E*, 83, 016107. [1377,1378]

Karwa, V., Pati, D., Petrovic, S., Solus, L., Alexeev, N., Raic, M., Wilburne, D., Williams, R., and Yan, B. (2016), "Exact Tests for Stochastic Block Models," arXiv no. 1612.06040. [1373,1378,1380]

Lei, J. (2016), "A Goodness-of-Fit Test for Stochastic Block Models," *The Annals of Statistics*, 44, 401–424. [1373,1374,1375,1378,1379,1381]

Lei, J., and Rinaldo, A. (2015), "Consistency of Spectral Clustering in Stochastic Block Models," *The Annals of Statistics*, 43, 215–237. [1373,1375,1377,1378]

Lei, J., and Zhu, L. (2015), "Generic Sample Splitting for Refined Community Recovery in Degree-Corrected Stochastic Block Models," *Statistica Sinica*, 27, 1639–1659. [1375]

Newman, M. E. J. (2006), "Modularity and Community Structure in Networks," *Proceedings of the National Academy of Sciences of the United States of America*, 103, 8577–8582. [1373]

Nowicki, K., and Snijders, T. A. B. (2001), "Estimation and Prediction for Stochastic Block Structures," *Journal of the American Statistical Association*, 96, 1077–1087. [1373]

Rohe, K., Chatterjee, S., and Yu, B. (2011), "Spectral Clustering and the High-Dimensional Stochastic Blockmodel," *The Annals of Statistics*, 39, 1878–1915. [1373]

Rohe, K., Qin, T., and Fan, H. (2014), "The Highest Dimensional Stochastic Blockmodel With a Regularized Estimator," *Statistica Sinica*, 24, 1771–1786. [1374]

Saldana, D. F., Yu, Y., and Feng, Y. (2017), "How Many Communities Are There?," *Journal of Computational and Graphical Statistics*, 26, 171–181. [1381]

Sarkar, P., and Bickel, P. J. (2015), "Role of Normalization in Spectral Clustering for Stochastic Blockmodels," *The Annals of Statistics*, 43, 962–990. [1373]

Snijders, T. A. B., and Nowicki, K. (1997), "Estimation and Prediction for Stochastic Block Models for Graphs With Latent Block Structure," *Journal of Classification*, 14, 75–100. [1373]

Wang, Y. X. R., and Bickel, P. J. (2017), "Likelihood-Based Model Selection for Stochastic Block Models," *The Annals of Statistics*, 45, 500–528. [1373,1374,1375,1376,1381]

Westveld, A. H., and Hoff, P. D. (2011), "A Mixed Effects Model for Longitudinal Relational and Network Data With Applications to International Trade and Conflict," *The Annals of Applied Statistics*, 5, 843–872. [1381]

Zhang, A. Y., and Zhou, H. H. (2016), "Minimax Rates of Community Detection in Stochastic Block Models," *The Annals of Statistics*, 44, 2252–2280. [1373,1375]

Zhao, Y., Levina, E., and Zhu, J. (2012), "Consistency of Community Detection in Networks Under Degree-Corrected Stochastic Block Models," *The Annals of Statistics*, 40, 2266–2292. [1373,1378,1380]