The Annals of Statistics
2023, Vol. 51, No. 6, 2366–2385
https://doi.org/10.1214/23-AOS2329

ADJUSTED CHI-SQUARE TEST FOR DEGREE-CORRECTED BLOCK MODELS

BY LINFAN ZHANG[a] AND ARASH A. AMINI[b]

Department of Statistics, University of California, Los Angeles, [a]*linfanz@g.ucla.edu,* [b]*aaamini@ucla.edu*

We propose a goodness-of-fit test for degree-corrected stochastic block models (DCSBM). The test is based on an adjusted chi-square statistic for measuring equality of means among groups of n multinomial distributions with $d_1, \dots, d_n$ observations. In the context of network models, the number of multinomials, n, grows much faster than the number of observations, d_i, corresponding to the degree of node i, hence the setting deviates from classical asymptotics. We show that a simple adjustment allows the statistic to converge in distribution, under null, as long as the harmonic mean of $\{d_i\}$ grows to infinity. When applied sequentially, the test can also be used to determine the number of communities. The test operates on a compressed version of the adjacency matrix, conditional on the degrees, and as a result is highly scalable to large sparse networks. We incorporate a novel idea of compressing the rows based on a $(K+1)$-community assignment when testing for K communities. This approach increases the power in sequential applications without sacrificing computational efficiency, and we prove its consistency in recovering the number of communities. Since the test statistic does not rely on a specific alternative, its utility goes beyond sequential testing and can be used to simultaneously test against a wide range of alternatives outside the DCSBM family. In particular, we prove that the test is consistent against a general family of latent-variable network models with community structure. We show the effectiveness of the approach by extensive numerical experiments with simulated and real data. In particular, applying the test to the Facebook-100 data set, a collection of one hundred social networks, we find that a DCSBM with a small number of communities (say < 25) is far from a good fit in almost all cases.

1. Introduction. Network analysis has become an increasingly prominent part of data analysis as the developments in the age of the internet and in various sciences, especially life and social sciences, have produced a substantial collection of network data. Given a network, it is of interest to understand its structure, which is often done by finding communities or clusters. Probabilistic network models such as the Stochastic Block Model (SBM) [15] and its variant the Degree-Corrected Stochastic Block Model (DCSBM) [18] are commonly used to recover the community structure from network data. Both models use a latent variable, the node label, to categorize nodes in a network into different communities. In the SBM, the probability of an edge formation between two nodes depends on the communities they belong to. The DCSBM incorporates an additional propensity parameter to determine the edge probability, allowing heterogeneous node degrees within a community.

The SBM and its degree-corrected variant have been the subject of intense study in recent years and numerous methods have been developed for fitting them (see Appendix B.2 [46]). Many of these methods are based on the assumption that the number of communities K is given and most come with consistency guarantees, when the data is generated from the corresponding model with K communities. On the other hand, how well these network models fit

Received January 2021; revised September 2023.
MSC2020 subject classifications. 62G20, 62E17, 62H99.
Key words and phrases. Degree-corrected stochastic block model, goodness-of-fit test, community detection, adjusted chi-square statistic, nonasymptotic.

the data, the so-called goodness-of-fit question, is studied comparatively much less. Prominent work in this area include the graphical approach of [17] for general network models, and the recent work of Bickel and Sarkar [6] and its extension by Lei [23], on a spectral goodness-of-fit test for the SBM. Developing goodness-of-fit tests specifically for the DCSBM is more challenging and to the best of our knowledge has not been considered so far, except for the work of Karwa et al. [19] on the related β-SBM. A related problem is model selection, that is, determining the number of communities assuming that the network is generated from some SBM (or DCSBM). An application of model selection is designing the stopping rule in hierarchical clustering [25]. Model selection has been studied more extensively, with a literature overview provided in Section 1.2.

Compared to model selection, goodness-of-fit testing is a more general problem. When applied sequentially, such tests can also be used for model selection. However, their utility goes beyond model selection and they can be used to test against a wide range of alternatives. They also provide a quantitative and baseline-normalized measure of how well the model fits in various situations. On the other hand, the ability to simultaneously test against many alternatives can be considered a weakness. To quote L. Breiman [7]:

> "Work by Bickel, Ritov and Stoker (2001) [5] shows that goodness-of-fit tests have very little power unless the direction of the alternative is precisely specified. The implication is that omnibus goodness-of-fit tests, which test in many directions simultaneously, have little power and will not reject until the lack of fit is extreme."

In our experiments, we have found the opposite to be true for current network models. It is possible to construct powerful tests, without specifying the direction of the alternative, for one of the most established families of network models. For example, we demonstrate both theoretically and empirically that the tests we develop for DCSBM are extremely powerful against a latent-variable community-structured model outside the DCSBM family (cf. Section 5.2 and Appendix A.2.2). Moreover, for the majority of the real networks that we tested, the null hypothesis of a DCSBM with a small number of communities is strongly rejected (cf. Section 6). This is all the more surprising given that the DCSBM is considered the state-of-the-art in modeling real community-structured networks.

1.1. *Our contributions.* In this paper, we propose the adjusted chi-square test for measuring the goodness-of-fit of a DCSBM. The idea is as follows: Given a set of column labels, we compress the adjacency matrix by summing each row over the communities specified by the labels, a process we will refer to as column aggregation. Under a DCSBM, the rows of the compressed matrix will have a multinomial distribution, conditional on the node degrees d_i (i.e., the row sums). Rows in the same (row) community will have the same multinomial parameter. Thus, the problem reduces to testing whether groups of multinomials have equal means. The challenge is that the number of multinomials in each group is proportional to n, the total number of nodes, which grows to infinity fast, while the number of observations in each multinomial, d_i, grows much slower. We study this general multigroup testing problem in Section 2 and show that under mild conditions, as long as the harmonic mean $h(d_1, \ldots, d_n)$ goes to infinity, a modified version of the classical chi-square statistic, which we refer to as Adjusted Chi-square (*AC*), has the standard normal distribution under the null hypothesis.

We then extend these ideas to the analysis of networks, leading to the Network Adjusted Chi-square (*NAC*) family of tests. The family includes many variants depending on which subsets of the adjacency matrix are used and how the columns are aggregated. Assume that we want to test a K-community DCSBM. One variant of the test uses a subsampling scheme and aggregates using K communities for the columns. We refer to this version as *SNAC*, for subsampled NAC. We show that given a consistent set of labels, *SNAC* has the standard normal distribution under null. Another variant of the test uses subsampling but aggregates using

$(K+1)$-community column labels, while still using K-community row labels when testing for the equality of multinomials. We refer to this variant as *SNAC+*. We show that *SNAC+* has the same null distribution as *SNAC*, but is more powerful against DCSBM alternatives in sequential applications (Section 5.1).

We also develop bootstrapped versions of the tests, which are more robust in practice and can be applied even when the null distribution of the test statistic is difficult to compute. Moreover, we introduce a smoothing idea that can further increase the robustness of sequential model selection.

Our theoretical results are nonasymptotic, controlling the Kolomogrov distance of the distribution of the test statistic to the target, with explicit constants. The results are valid in the regime where the expected average degree of the network, λ, scales as $\gtrsim \log n$, hence applicable in the same sparsity regime where strong consistency (i.e., exact label recovery) is possible for DCSBMs. Our results, however, only require weakly consistent labels subject to bounds on the rate of convergence that are more relaxed than that of strong consistency. From a computational standpoint, evaluating the statistic is highly scalable, with an expected computational overhead of $O(n(\lambda+K))$ over the cost of applying the community detection algorithm. To test a sequence of DCSBMs with $K = K_1, \ldots, K_2$, the test requires an application of a community detection algorithm at most $K_2 - K_1 + 2$ times.

We show the effectiveness of these ideas with extensive experiments on simulated and real networks. The code for these experiments is available at [45]. In particular, we apply the test to the Facebook-100 data set [35, 36], a collection of one hundred social networks, and find that a DCSBM (or SBM) with a small number of communities (say < 25) is far from a good fit in almost all cases.

1.2. *Related work.* Various methods have been developed to address the model selection problem in the SBM and DCSBM. The popular Bayesian information criterion (BIC) has been adapted to the network setting in [16, 38, 40]. Likelihood ratio tests have been developed for comparing two block models in [29, 38, 41, 42]. Bayesian approaches, though computationally intensive, can estimate the structure and the number of communities simultaneously. Ideas include the use of Dirichlet process prior [2] and mixture of priors [14, 30, 33]. Cross-validation, another widely used idea for model selection, has too been adapted to network settings [8, 20, 26]. A leave-one-out scheme has been used in [20] with the posterior predictive density of an edge, under the SBM, as the loss function. Chen and Lei [8] use a node-pair splitting idea, while [26] uses edge sampling followed by low-rank matrix completion, an approach that can be applied to any low-rank network model. Spectral approaches exploring the eigenstructures of the Bethe Hessian matrix and its variants are shown to be useful in determining the number of communities in the SBM [21] and DCSBM [11]. The approach of [21] can be extended to other low-rank structured models such as DCSBM. Semidefinite programming have been shown in [39] to be capable of performing label recovery and model selection in one shot. Modularity maximization can also perform the two tasks simultaneously [31].

Comparatively, the goodness-of-fit problem has been explored much less. The pioneering work of [17] graphically compares certain network statistics (such as degree distribution) between the observed network and a collection of networks simulated from the fitted model. In later work, Bickel et al. [4] establish general asymptotic properties of network statistics like the empirical graph moments, which could potentially be used for goodness-of-fit testing.The Monte Carlo simulation procedures in [17] have also been further exploited in other works [27, 32] to test the goodness-of-fit of graph models. Among them, we note that Karwa et al. [19] develops a chi-square test for SBM and uses Markov Chain Monte Carlo sampling to approximate its exact p-value. For SBMs, a spectral goodness-of-fit test was developed in [6]

for the case of $K = 2$ communities and subsequently extended to general K in [23]. The test is based on the largest eigenvalue of a standardized residual adjacency matrix (cf. Appendix A.1 for more details). Using results from random matrix theory [12, 22], this eigenvalue has an asymptotic Tracy–Widom distribution under the null, a result that can be used to set the critical threshold. Although, we can apply the same ideas in the DCSBM setting, the null distribution result does not hold, due to the uncertainty in estimating the node connection propensity parameters. Whether a rigorous spectral goodness-of-fit test of this form exists for DCSBM is not clear. We refer to Appendix B.1 for a more detailed comparison with the existing literature.

The rest of the paper is organized as follows: Section 2 introduces the adjusted chi-square test and its multigroup extension and establishes its null limiting distribution. Section 3 introduces NAC family of tests. Section 4 establishes the null limiting distribution of SNAC and SNAC+ and Section 5 shows their consistency against underfitted DCSBM and a latent-variable community-structured network model. In Section 6, we illustrate how SNAC+ can be used to assess the goodness-of-fit for an ensemble of real networks, namely the Facebook-100 data set.

2. Adjusted chi-square test. We start by developing a general test for the equality of the parameters among groups of multinomial observations. To set the ideas, we first consider the case of a single group and show how the classical chi-square test can be adjusted to accommodate a growing number of multinomials. We then discuss the multigroup extension and provide quantitative bounds for the null distribution of the test statistic in this general setting.

2.1. *Single-group case.* Let $\mathcal{P}_L$ be the probability simplex in $\mathbb{R}^L$, and consider the following problem: We have

$$X_i \sim \mathrm{Mult}(d_i, \boldsymbol{p}^{(i)}), \quad i = 1, \dots, n, \tag{1}$$

independently, where $X_i = (X_{i\ell}) \in \mathbb{N}_0^L$ and $\boldsymbol{p}^{(i)} \in \mathcal{P}_L$, and we would like to test the null hypothesis

$$H_0 : \boldsymbol{p}^{(1)} = \boldsymbol{p}^{(2)} = \cdots = \boldsymbol{p}^{(n)} = \boldsymbol{p}. \tag{2}$$

Let $\psi(x, y) := (x - y)^2/y$. The chi-square statistic for testing this hypothesis is

$$\widetilde{Y}^*_{(n,d)} := \sum_{i=1}^{n} \sum_{\ell=1}^{L} \psi(X_{i\ell}, d_i \widetilde{p}_\ell), \quad \text{where } \widetilde{p}_\ell = \frac{\sum_{i=1}^n X_{i\ell}}{\sum_{i=1}^n d_i}, \ell \in [L].$$

Here, $\widetilde{\boldsymbol{p}} = (\widetilde{p}_\ell) \in \mathcal{P}_L$ is the pooled estimate of $\boldsymbol{p}$ under the null, and $\boldsymbol{d} = (d_1, \dots, d_n)$. We are also using the shorthand notation $[L] := \{1, \dots, L\}$.

Standard asymptotic theory gives the following (cf. [37], Chapter 17): If n is fixed and $d_{\min} := \min_i d_i \to \infty$, then

$$\widetilde{Y}^*_{(n,d)} \rightsquigarrow \chi^2_{(n-1)(L-1)}, \quad \text{under } H_0. \tag{3}$$

A heuristic for the degrees of freedom of the limiting χ^2 distribution can be given by counting parameters. In the unrestricted model, we have a total of $n(L - 1)$ free parameters among $p^{(1)}, \dots, p^{(n)}$, while under the restricted null model, we only have $L - 1$ free parameters. The difference gives the degrees of freedom of the limit.

The setting we are interested in, however, is the opposite of the classical setting. We would like to use the statistic when $n \to \infty$, while $d_{\min}$ is fixed or grows slowly with n. Assuming

that n is large enough so that $(n-1)(L-1) \approx n(L-1)$, (3) suggests that we can approximate $\widetilde{Y}^*_{(n,d)}$ in distribution by the sum of n independent χ^2_{L-1} variables, that is,

$$\widetilde{Y}^*_{(n,d)} \overset{d}{\approx} \sum_{i=1}^{n} \xi_i$$

for some i.i.d. random variables $\xi_i \sim \chi^2_{L-1}$. The approximate inequality above is only in distribution and $\{\xi_i\}$ are not necessarily related to $\widetilde{Y}^*_{(n,d)}$. Moreover, the central limit theorem suggests that the standardized version of $\sum_i \xi_i$ has a distribution close to a standard normal.

Based on the above heuristic argument, we propose the following adjusted test statistic:

$$\widetilde{T}^*_n = \frac{1}{\sqrt{2}}\Big(\frac{\widetilde{Y}^*_{(n,d)}}{\gamma_n} - \gamma_n\Big), \quad \text{where } \gamma_n = \sqrt{n(L-1)}. \tag{4}$$

Note that γ_n^2 is the expectation of $\sum_i \xi_i$ and $\sqrt{2}\gamma_n$ is its standard deviation. We refer to (4) as the *adjusted chi-square* (*AC*) statistic.

2.2. *Multigroup extension*. Before proceeding, let us introduce an extension of the testing problem (2) to groups of observations. This extension is needed for the network applications. Consider model (1) and assume that each observation is assigned to one of the K known groups, denoted as $[K] = \{1, \dots, K\}$. Let $g_i \in [K]$ be the group assignment of observation i and let $\mathcal{G}_k = \{i \in [n] : \ g_i = k\}$ be the kth group. We would like to test the null hypothesis that all the observations in the same group have the same parameter vector, that is,

$$H_0 : \boldsymbol{p}^{(i)} = \boldsymbol{p}_{k*}, \quad \forall i \in \mathcal{G}_k, k \in [K], \tag{5}$$

where for each $k \in [K]$, $\boldsymbol{p}_{k*} = (p_{k\ell})_{\ell \in [L]} \in \mathcal{P}_L$.

In some problems, it is reasonable to assume that the groups $\mathcal{G}_k$ are known. However, in our network applications, the groups themselves are not known. In such settings, we first estimate the label vector g from data, to obtain $\hat{g}$, and then form the test statistic based on the estimated groups $\widehat{\mathcal{G}}_k = \{i : \hat{g}_i = k\}$. The resulting test is based on the extended chi-square statistic

$$\widehat{Y}_{(n,d)} = \sum_{k=1}^{K} \sum_{i \in \widehat{\mathcal{G}}_k} \sum_{\ell=1}^{L} \psi(X_{i\ell}, d_i \widehat{p}_{k\ell}), \quad \text{where } \widehat{p}_{k\ell} = \frac{\sum_{i \in \widehat{\mathcal{G}}_k} X_{i\ell}}{\sum_{i \in \widehat{\mathcal{G}}_k} d_i}, \ \ell \in [L]. \tag{6}$$

Alternatively, we have $\widehat{Y}_{(n,d)} = \sum_{i=1}^{n} \sum_{\ell=1}^{L} \psi(X_{i\ell}, d_i \widehat{p}_{\hat{g}_i \ell})$. We also let $Y_{(n,d)}$ be the idealized version of $\widehat{Y}_{(n,d)}$ with $\widehat{p}_{k\ell}$ replaced with $p_{k\ell}$ and $\widehat{\mathcal{G}}_k$ replaced with $\mathcal{G}_k$. Let $\widehat{T}_n$ and T_n be the adjusted chi-square statistics based on $\widehat{Y}_{(n,d)}$ and $Y_{(n,d)}$, respectively, that is,

$$\widehat{T}_n = \frac{1}{\sqrt{2}}\Big(\frac{\widehat{Y}_{(n,d)}}{\gamma_n} - \gamma_n\Big), \qquad T_n = \frac{1}{\sqrt{2}}\Big(\frac{Y_{(n,d)}}{\gamma_n} - \gamma_n\Big). \tag{7}$$

We are interested in understanding under what conditions $\widehat{T}_n$ has an approximately normal null distribution. This question is nontrivial, since we would like to allow $\{d_i\}$ as well as groups sizes $|\mathcal{G}_k|, k \in [K]$ to vary with n. Moreover, we would like to allow the groups to be estimated based on the same data we use for testing, in which case, $\hat{g}$ and $\widehat{T}_n$ are most likely statistically dependent.

We give a precise answer to the above question by quantifying the Kolomogorv distance between the distribution of $\widehat{T}_n$ and that of a standard normal variable Z, for any choice of $\{d_i\}$ and $\{|\mathcal{G}_k|\}$ that satisfy a mild set of conditions, and for consistent label estimates of a certain quality. We measure the quality of label estimation in terms of misclassification rate.

DEFINITION 1. The misclassification rate between two label vectors $g \in [K]^n$ and $\hat{g} \in [K]^n$ is

$$\mathrm{Mis}(g, \hat{g}) = \min_{\omega} \frac{1}{n} \sum_{i=1}^{n} 1\{g_i \neq \omega(\hat{g}_i)\},$$

where the minimization ranges over all bijective maps $\omega : [K] \to [K]$.

Recall that for two random variables X and Y, the Kolomogrov distance between their distributions is defined as

$$d_{\mathrm{K}}(X, Y) := \sup_{t \in \mathbb{R}} \big|\mathbb{P}(X \leq t) - \mathbb{P}(Y \leq t)\big|. \tag{8}$$

For a vector $\boldsymbol{d} = (d_1, \ldots, d_n)$, we write $h(\boldsymbol{d}) = (n^{-1} \sum_{i=1}^n d_i^{-1})^{-1}$ for the harmonic mean of its elements, and $d_{\mathrm{av}} = n^{-1} \sum_{i=1}^n d_i$ for the arithmetic mean. Since $\boldsymbol{d}$ has positive elements, $d_{\mathrm{av}} \geq h(\boldsymbol{d}) \geq d_{\min} := \min_i d_i$. Let $\pi_k = |\mathcal{G}_k|/n$ and write $d_{\mathrm{av}}^{(k)} = \frac{1}{|\mathcal{G}_k|} \sum_{i \in \mathcal{G}_k} d_i$ for the arithmatic average of $\{d_i\}$ within group $\mathcal{G}_k$, and define

$$\omega_n := \min_k \pi_k d_{\mathrm{av}}^{(k)}, \qquad d_{\max} := \max_i d_i, \qquad \tau_d := \omega_n / d_{\max} \tag{9}$$

The following result formalizes the heuristic argument of Section 2.1, by providing a quantitative finite-sample bound on the Kolomogrov distances of T_n and $\widehat{T}_n$ to a standard normal variable:

THEOREM 1. *Let $X_i \sim \mathrm{Mult}(d_i, \boldsymbol{p}_{k*})$, $i \in \mathcal{G}_k$, $k \in [K]$ be n independent L-dimensional multinomial variables, with probability vectors $\boldsymbol{p}_{k*} = (p_{k\ell})$ and group labels $g = (g_i) \in [K]^n$ so that $\mathcal{G}_k = \{i : g_i = k\}$. Let $\hat{g}$ be some (estimated) group labels, potentially dependent on $\{X_i\}$ and consider $\widehat{T}_n$, based on $\hat{g}$, and T_n as in (7). Let $Z \sim N(0,1)$ and $\underline{p} = \min_{k,\ell} p_{k\ell}$. Assume that $\min\{h(\boldsymbol{d}), L\} \geq 2$.*

(a) *Then, under the null hypothesis (5), for all $n \geq 1$,*

$$d_K(T_n, Z) \leq \frac{C_{1,p}}{\sqrt{Ln}} + \frac{C_{2,p}}{h(\boldsymbol{d})}, \tag{10}$$

where $C_{1,p} = 55/\underline{p}^4$ and $C_{2,p} = (\pi e)^{-1/2} \max\{1, \underline{p}^{-1} - L - 1\}$.

(b) *Let $C_{3,p} = 6/(\underline{p}\, \tau_d)$ and pick a sequence $\{\alpha_n\}$ such that*

$$\alpha_n \leq \min\left\{ \frac{\underline{p}}{8C_{3,p}}, \frac{2}{C_{3,p}^2 L} \right\}, \quad \text{for all } n \geq 1. \tag{11}$$

Assume that $\sqrt{2} d_{\max} \geq L C_{3,p}$, $\omega_n \geq L$ and $\log(K\omega_n)/\omega_n \leq (\underline{p}/8)^2 n$. Then, under the null hypothesis (5), for all $n \geq 1$,

$$\begin{aligned} d_K(\widehat{T}_n, Z) \leq d_K(T_n, Z) + 12 \frac{\sqrt{L}}{\underline{p}} \Bigg(& \sqrt{\frac{\log(K\omega_n)}{\omega_n}} + \frac{K \log(K\omega_n)}{\sqrt{n}} \\ & + \frac{C_{3,p}}{3L} d_{\max} \sqrt{Kn}\, \alpha_n \Bigg) + 2\mathbb{P}\big(\mathrm{Mis}(\hat{g}, g) > \alpha_n\big). \end{aligned} \tag{12}$$

Note that we always have $\underline{p}^{-1} \geq L$ since the elements of $\boldsymbol{p}_{k*}$ are nonnegative and sum to one. In the proof of Theorem 1, we will show that $\mathbb{E}[Y_{(n,d)}] = \gamma_n^2$. But the standard deviation $v_n(\boldsymbol{p}) := \sqrt{\mathrm{var}[Y_{(n,d)}]}$ has a more complicated form and is not equal to $\sqrt{2}\gamma_n$ in general. The proof gives an explicit expression for this variance, and we could have alternatively defined

$\widehat{T}_n$ by dividing by $v_n(\widehat{\boldsymbol{p}})$ instead of $\sqrt{2}\gamma_n$. Nevertheless, Theorem 1 shows that we do not lose much by using the simpler standardization by $\sqrt{2}\gamma_n$.

In general, for T_n to converge in distribution to the standard normal, we need $n \to \infty$ and $h(\boldsymbol{d}) \to \infty$. For $\widehat{T}_n$ to converge to the normal distribution, we further need $\omega_n \to \infty$, $K \log(K\omega_n) = o(\sqrt{n})$,

$$\alpha_n = o\big((d_{\max}\sqrt{n})^{-1}\big) \quad \text{and} \quad \mathbb{P}\big(\mathrm{Mis}(\hat{g}, g) > \alpha_n\big) = o(1). \tag{13}$$

Note that $\log(K\omega_n)/\omega_n \le (\underline{p}/8)^2 n$ and (11) are satisfied for large n, as long as $\underline{p}$ is bounded away from zero. The assumption $\sqrt{2}d_{\max} \ge C_{3,p}L$ also holds since $d_{\max} \ge h(\boldsymbol{d})$ and we require $h(\boldsymbol{d}) \to \infty$.

As we will see, in network applications, typically K, L and $\underline{p}$ are of constant order. Then the requirements reduce to (13), $h(\boldsymbol{d}) \to \infty$, $\omega_n \to \infty$ and $\log(\omega_n) = o(\sqrt{n})$. The condition $h(\boldsymbol{d}) \to \infty$ is fairly mild in network applications, since d_i will be the degree of node i, and one often assumes that the network degrees grow to infinity as $n \to \infty$ (a necessary condition for weak label consistency). See also the empirical evidence in Appendix B.3. Even if one does not want to assume $h(\boldsymbol{d}) \to \infty$ over the whole network, the condition can still be reasonably achieved by manually filtering out nodes with small d_i, as will be discussed in detail in Section 3.

Since, in networks, $d_{\max}$ grows much slower than $\sqrt{n}$ (closer to $\log n$ in fact), Condition (13) on the misclassification rate α_n is, in general, much milder than strong consistency which is equivalent to $\alpha_n = o(n^{-1})$. In the network setting, it is typical to assume that all the degrees grow at the same rate, in which case, $h(\boldsymbol{d}) \asymp d_{\max} \asymp \omega_n$. Under these assumptions, we obtain the following simplified bound.

COROLLARY 1. *Suppose* $1 \lesssim h(\boldsymbol{d}) \asymp d_{\max} \asymp \omega_n \lesssim \sqrt{n}$ *and that* K, L *and* $\underline{p}$ *are of constant order. Then*

$$d_K(\widehat{T}_n, Z) \lesssim \sqrt{\frac{\log \omega_n}{\omega_n}} + \alpha_n \omega_n \sqrt{n} + \mathbb{P}\big(\mathrm{Mis}(\hat{g}, g) > \alpha_n\big).$$

3. Network AC test. We are now ready to apply the AC test to DCSBMs. Let $A_{n\times n}$ be the adjacency matrix of a random network on n nodes. A DCSBM with connectivity matrix $B \in [0,1]^{K\times K}$, node label vector $z = (z_i) \in [K]^n$ and connection propensity vector $\theta = (\theta_i) \in \mathbb{R}_+^n$, assumes the following structure for the mean of A:

$$\mathbb{E}[A_{ij} \mid z] = \theta_i \theta_j B_{z_i z_j}, \quad \forall\, i \ne j. \tag{14}$$

One further assumes that A is symmetric and the entries $A_{ij}, i < j$ are drawn independently, while $A_{ii} = 0$ for all i. Common choices for the distribution of each element, A_{ij}, are Bernoulli and Poisson. In this paper, unless otherwise stated, we assume the Poisson distribution for derivations, following the original DCSBM paper [18]. The Poisson assumption simplifies the arguments and provides computational advantages. We show in simulations that the tests so-derived work well in the Bernoulli case when the network is sparse. The SBM is a special case of (14) with $\theta_i = 1$ for all i.

3.1. *NAC family of tests.* The network AC test can be performed on a general submatrix $A_{S_2S_1} = (A_{ij} : i \in S_2, j \in S_1)$ of the adjacency matrix, for $S_1, S_2 \subseteq [n]$. We first present this general form, though one can assume $S_1 = S_2 = [n]$ on the first reading. Consider another label vector on S_1, say $\widehat{y} = (\widehat{y}_j)_{j\in S_1} \in [L]^{S_1}$—for some L that can be different from K. Let $R = (R_{k\ell}) \in \mathbb{R}_+^{K\times L}$ be the weighted confusion matrix between z_{S_1} and $\widehat{y}$, given by

$$R_{k\ell} = \frac{1}{|S_1|} \sum_{j\in S_1} \theta_j 1\{z_j = k, \widehat{y}_j = \ell\}. \tag{15}$$

Consider the column aggregation of $A_{S_2 S_1}$ w.r.t. $\widehat{y}$, defined as $X = (X_{i\ell}) \in \mathbb{R}_+^{|S_2| \times L}$, with

$$X_{i\ell}(\widehat{y}) = \sum_{j \in S_1} A_{ij} 1\{\widehat{y}_j = \ell\}. \tag{16}$$

Assuming that $\widehat{y}$ is deterministic, we have

$$\begin{aligned}\mathbb{E}\big[X_{i\ell}(\widehat{y})\big] &= \sum_{j \in S_1} B_{z_i z_j} \theta_i \theta_j 1\{\widehat{y}_j = \ell\} = \theta_i \sum_{k=1}^{K} B_{z_i k} \sum_{j \in S_1} \theta_j 1\{z_j = k, \widehat{y}_j = \ell\} \\ &= |S_1| \, \theta_i (BR)_{z_i \ell}.\end{aligned}$$

Let $d_i = \sum_{j \in S_1} A_{ij}$ be the degree of node i in S_2. Under the Poisson model, $(A_{ij}, j \in S_1)$ is a vector of independent Poisson cooridnates. It is well known that such a vector has a multinomial distribution conditional on the sum of its entries. That is,

$$X_{i*}(\widehat{y}) \mid d_i \sim \text{Mult}(d_i, \rho_{z_i *}), \tag{17}$$

where $\rho_{z_i *}$ denotes the z_ith row of $\rho = (\rho_{k\ell}) \in [0,1]^{K \times L}$, defined as

$$\rho_{k\ell} = \frac{(BR)_{k\ell}}{\sum_{\ell'} (BR)_{k\ell'}}. \tag{18}$$

In other words, conditioned on the degree sequence $\boldsymbol{d} = (d_i, i \in S_2)$, all the rows of X corresponding to z-community k, have multinomial distributions with probability vector ρ_{k*}. This observation allows us to apply the AC test developed in Section 2.2, to test whether all the rows with $z_i = k$, have the same multinomial distribution.

Now, consider two estimated label vectors $\widehat{z} = (\widehat{z}_i) \in [K]^n$ and $\widehat{y} = (\widehat{y}_i) \in [L]^{S_1}$. Let $\hat{\mathcal{C}}_k = \{i \in [n] : \widehat{z}_i = k\}$, $\widehat{\mathcal{G}}_k = \hat{\mathcal{C}}_k \cap S_2$ and $\tilde{n} = |S_2|$. Consider the multigroup version of the AC statistic based on $\widehat{z}$ and $\widehat{y}$:

$$\widehat{T}_n = \frac{1}{\sqrt{2}} \left(\frac{1}{\gamma_{\tilde{n}}} \sum_{k=1}^{K} \sum_{i \in \widehat{\mathcal{G}}_k} \sum_{\ell=1}^{L} \psi\big(X_{i\ell}(\widehat{y}), d_i \widehat{\rho}_{k\ell}\big) - \gamma_{\tilde{n}} \right), \tag{19}$$

where $\gamma_{\tilde{n}} = \sqrt{\tilde{n}(L-1)}$ and

$$\widehat{\rho}_{k\ell} = \frac{\sum_{i \in \widehat{\mathcal{G}}_k} X_{i\ell}(\widehat{y})}{\sum_{i \in \widehat{\mathcal{G}}_k} d_i}, \quad k \in [K], \ell \in [L]. \tag{20}$$

The above construction specifies a family of test statistics, depending on the choices of label vectors $\widehat{z}$ and $\widehat{y}$, and subsets S_1 and S_2. We refer to this family, as the NAC family of tests. The acronym NAC stands for Network Adjusted Chi-square, since the test is the natural extension of the adjusted chi-square test, introduced earlier, to networks. As we see in Section 4, since the degrees are fixed by conditioning and act as multinomial sample sizes, the asymptotic null distribution of the test statistic is invariant to the degree distribution. Furthermore, NAC family of tests are easily applicable to nonsquare and nonsymmetric adjacency matrices, as we discuss in Appendix B.4.

3.2. *Full version.* We now single out two specifc members of the NAC family. Let $S_1 = S_2 = [n]$ and consider the following choices for $\widehat{z}$ and $\widehat{y}$:

1. *FNAC*: $\widehat{y} = \widehat{z}$ and $\widehat{z}$ is an estimated label vector with K communities,
2. *FNAC+*: $\widehat{z}$ and $\widehat{y}$ are estimated label vectors with K and $L = K + 1$ communities.

The acronym FNAC stands for Full NAC, where "full" refers to the choice $S_1 = S_2 = [n]$. There are two main reasons for introducing the FNAC+ version with $L = K + 1$ column communities. First, FNAC only works when $K \geq 2$; when $K = L = 1$, (17) leads to a noninformative statistic for FNAC, because then $X_{i*} = d_i$ almost surely, conditioned on d_i. FNAC+ on the other hand still produces an informative statistic when $K = 1$. Second, the choice $L = K + 1$ makes FNAC+ especially powerful in determining the number of communities by sequential testing from below, as we discuss extensively in Section 5.1.

3.3. *Subsampled version.* The asymptotic null distribution of the full version statistics, FNAC and FNAC+, can be complicated. There are two main obstacles in applying Theorem 1 to these statistics. First, although the theorem allows for the dependence of $\widehat{z}$ on the entire adjacency matrix A as long as it converges to the true label vector z, it cannot directly handle the dependence of $\widehat{y}$ on the entire A. Because then, $X_{i*}(\widehat{y})$ will be formed by summing elements of A_{i*} (the ith row of A) over subsets of the columns that depend on A_{i*} itself. This dependence between $\widehat{y}$ and A is algorithm-specific, that is, itself depends on the particular community detection algorithm used, leading to an unknown deviation of the distribution of individual $X_{i\ell}(\widehat{y})$ from a Poisson. Moreover, the joint dependence of $\widehat{y}$ and A induces an algorithm-specific joint distribution on $(X_{i*}(\widehat{y}), i \in [n])$, which is hard to characterize for interesting algorithms such as spectral clustering.

The above issues are resolved if we assume $\widehat{y} = z$ w.h.p., which holds if the algorithm is strongly consistent, but this can only happen for FNAC; in the case of FNAC+, we always estimate with one more community relative to the truth, and the breaking of at least one true community causes an unknown skewness in the distribution of the resulting partitions; imagine bisecting an Erdős–Rényi (ER) network, resulting in two subnetworks that are more clustered than a typical ER network. The second obstacle to using the full version statistics is the symmetry of A, which makes $X_{i*}(\widehat{y})$ and $X_{j*}(\widehat{y})$ (mildly) dependent through the shared element $A_{ij} = A_{ji}$, even when $\widehat{y} = z$, and hence applies to both FNAC and FNAC+.

To circumvent the above obstacles, we introduce a particular subsampling scheme, which provides several advantages. It takes care of the dependence issues, making the results independent of the community detection algorithm used. It also allows us to state unified results that apply regardless of the choice of L, hence the same results will be applicable to both SNAC and SNAC+. Moreover, as we will show, by using the scheme, we avoid the assumption $\widehat{y} = z$. In fact, we no longer even need $\widehat{y}$ to be consistent for z for the results to go through. Finally, it allows us to implement a further degree filtering step, which potentially improves the growth rate of the harmonic mean of the remaining degrees, $h(\boldsymbol{d})$, making the assumption $h(\boldsymbol{d}) \to \infty$ easier to satisfy in practice.

The scheme is detailed in Algorithm 1. It involves a sampling step so that: (a) $\widehat{y}$ no longer depends on the entries of A needed to be summed; (b) the symmetry is broken. It also has a filtering step to leave out nodes with small degrees, so that $h(\boldsymbol{d})$ is large. A practical recipe for selecting the quantile filtering threshold σ is discussed in Appendix B.5.

We refer to Algorithm 1 as subsampled NAC, or SNAC for short, when $L = K$ and as SNAC+ when $L = K + 1$. Note that step 5, the quantile filtering, can be skipped if the degrees are mostly large or if the normal approximation to the null distribution is not required. In the latter case, we can use the bootstrap debiasing of Section 3.4 to determine the critical region. In such cases, we set $S_2' = S_2$ (equivalently $\sigma = 0$) and perform the test on $A_{S_2 S_1}$.

REMARK 1 (On notation). In the sequel, we often state results that apply to either of SNAC or SNAC+. We will use the notation *SNAC(+)* to mean the statement holds for either version. Similarly, *FNAC(+)* refers to either of FNAC or FNAC+.

Algorithm 1 SNAC(+)

Input: Adjacency matrix A, number of row clusters K, number of column clusters $L \in \{K, K+1\}$, degree-filtering threshold $\sigma \in [0, 1)$. Critical threshold $\tau > 0$.

Output: Test statistic $\widehat{T}_n$ and whether null is rejected.

1: Fit K clusters to the whole network to get labels $\widehat{z} \in [K]^n$ and clusters $\hat{\mathcal{C}}_k = \{i : \widehat{z}_i = k\}$.
2: (*Sampling*) Choose a subset $S_1 \subset [n]$ by including each index $i \in [n]$, independently, with probability $1/2$. Let $S_2 = [n] \setminus S_1$ be the complement of S_1.
3: Fit L clusters to $A_{S_1 S_1} = (A_{ij} : i, j \in S_1)$ to learn the label vector $\widehat{y}$ on S_1.
4: Form partial degrees $d_i := \sum_{j \in S_1} A_{ij}$ for all $i \in S_2$.
5: (*Quantile filtering*) Within each $\widehat{\mathcal{G}}_k = \hat{\mathcal{C}}_k \cap S_2$, keep nodes with d_i at least the σth quantile of all d_i in $\widehat{\mathcal{G}}_k$ to form $\widehat{\mathcal{G}}'_k$. Let $S'_2 = \bigcup_{k=1}^{K} \widehat{\mathcal{G}}'_k$.
6: Perform the test on $A_{S'_2 S_1}$ using row labels $\widehat{z}_{S'_2}$ and column labels $\widehat{y}$ from Step 3 to form $\widehat{T}_n$ as in (19) and reject the null if $\widehat{T}_n > \tau$.

In Section 4, we show that, under the null model, the distributions of the test statistics of SNAC(+) are close to a standard normal. Furthermore, we show that they are large when the model is underfitted, that is, the presumed number of communities is smaller than that of the true model, with SNAC+ often being much larger than SNAC. We also show that under DCLVM, a latent variable network model with clusters, SNAC(+) values are large. Such properties allow us to use SNAC+ for assessing the goodness-of-fit of DCSBM or SBM to an observed network and to determine the number of clusters in community detection.

3.4. *Bootstrap debiasing*. Per our discussion above, without subsampling, the full version statistics, FNAC(+), do not have a standard normal null distribution in general. However, they are expected to produce more powerful tests since they utilize all the information in the network. As a result, they are great choices in practice if we can approximate their null distribution. The remedy is to use bootstrap simulation to determine their critical regions. In addition, bootstrap can correct deviations of the null distribution of SNAC(+) from the standard normal when some of the underlying assumptions fail to hold; see Remark 2.

Given adjacency matrix A, the null hypothesis that the number of communities is K, and the test statistic $\hat{T} = \hat{T}(A)$, the bootstrap debiasing is performed as follows:

1. Fit a K-community SBM to A and get label estimates $\hat{z}$ and connectivity matrix $\hat{B}$.
2. For $j = 1, \ldots, J$, sample $A^{(j)} \sim \mathrm{SBM}(\hat{z}, \hat{B})$ and compute test statistic $\hat{T}^{(j)}$ based on $A^{(j)}$.
3. Construct the debiased statistic $\hat{T}^{(\mathrm{boot})} = (\hat{T} - \hat{\mu})/\hat{\sigma}$ where $\hat{\mu}$ and $\hat{\sigma}$ are the sample mean and the standard deviation of $\{\hat{T}^{(j)}\}_{j=1}^{J}$.

Note that we sample from SBM instead of DCSBM. To simulate from DCSBM, one has to estimate $(\theta_i)_{i=1}^n$, which cannot be done consistently, and whose estimates are highly variable. As a result, generating from an estimated DCSBM adds extra variance and produces samples that are actually further from the original network than those produced by the SBM fit. We also note that the distribution of our statistics are invariant to degrees, making SBM generation further justified.

The test rejects for large values of $\hat{T}^{(\mathrm{boot})}$ (or $|\hat{T}^{(\mathrm{boot})}|$), with the threshold set assuming that $\hat{T}^{(\mathrm{boot})}$ has (approximately) a standard normal distribution under null. A similar idea is used in [23] for the spectral test. An alternative to debiasing is to use the empirical quantiles of $\{T^{(j}\}$ to set the critical threshold. We, however, found that the debiasing approach performs better in practice. See Appendix B.6 for a detailed discussion and comparison of all the bootstrap methods in a simulation setting.

3.5. *Model selection.* A goodness-of-fit test can also be used as a model selection method, through a process of sequential testing. In particular, we can use FNAC(+) (with bootstrap debiasing) and SNAC(+) statistics to determine the number of communities when fitting DCSBMs.

The idea is to test the null hypothesis of K communities, starting with $K = K_{\min}$, which is usually taken to be 1, and increasing K to $K + 1$ if the null is rejected. The process is repeated until we can no longer reject the null or a preset maximum number of communities, $K_{\max}$, is reached. The value of K on which we stop is selected as the optimal number of communities. We refer to this procedure as *sequential testing from below*. There is also the possibility of starting at $K = K_{\max}$ and working backwards. Testing from below is, however, more advantageous, especially if one expects a small number of communities a priori.

The rejection thresholds for SNAC(+) can be determined based on the standard normal distribution. For FNAC(+), we need to apply the bootstrap debiasing of Section 3.4 before comparing the statistic with the threshold. Theorem 3 provides a theoretical guarantee for the consistency of the sequential testing from below when SNAC(+) is used. An empirical comparison of the model selection performance of this approach with existing methods is provided in Appendix A.2.1.

4. Null distribution. We now derive the null distribution of SNAC(+). We consider a DCSBM with K_0 true community, and the edge probability matrix $B = (\nu_n/n)B^0$ where ν_n is a scaling factor and B^0 satisfies

$$\min_{k,\ell} B^0_{k\ell} \geq \tau_B \cdot \max_{k,\ell} B^0_{k\ell}. \tag{21}$$

Let $\mathcal{C}_k = \{i \in [n] : z_i = k\}$ be the true community k. We assume that

$$n_k := |\mathcal{C}_k| \geq \tau_{\mathcal{C}}\, n, \quad \theta_i \geq \tau_\theta \cdot \max_i \theta_i \tag{22}$$

for all $k \in [K_0]$ and $i \in [n]$. Here, τ_B, $\tau_{\mathcal{C}}$ and τ_θ are in $(0, 1]$ and measure the deviation of the corresponding parameters from being balanced. To make ν_n identifiable, we further assume without loss of generality that $\|B^0\|_\infty := \max_{k,\ell} B^0_{k\ell} = 1$ and $\|\theta\|_\infty := \max_i \theta_i = 1$. We require the following on the community detection algorithm.

ASSUMPTION 1. The community detection algorithm applied with K communities to the DCSBM described above, producing labels $\{\widehat{z}_i\}$, satisfies:

(a) Weak consistency: When $K = K_0$, there is a sequence $\alpha_n = o(1)$ such that

$$\mathbb{P}\big(\mathrm{Mis}(\widehat{z}, z) \leq \alpha_n\big) = 1 - o(1).$$

(b) Stability: For $K \in [K_0 + 1]$, we have $|\{i : \widehat{z}_i = k\}| \geq \tau_0 n$ for all $k \in [K]$.

Assumption 1(a), known as the weak consistency or partial recovery, allows us to focus on the event where $\widehat{z}$ is close to z, the true label vector. As long as $\nu_n \to \infty$, there are algorithms that can achieve this [1]. We, in fact, need α_n in Assumption 1(a) to go down faster than $o(1)$, but still much slower than what is needed for exact recovery (or strong consistency); see the discussion after Theorem 2. The growth rate of ν_n is roughly that of the expected average degree (EAD) of the network, assuming that B^0, $\{n_k/n\}_k$ and the distribution of $\{\theta_i\}$ are roughly constant.

Assumption 1(b) is even milder, and ensures that the algorithm does not produce extremely small communities when applied with $K \neq K_0$. It can be guaranteed by explicitly enforcing it in the algorithm: If the size of a recovered community is too small relative to n, we merge

it with another community. Whether a specific community detection algorithm satisfies this condition automatically without explicit enforcement is an interesting research question.

Recall σ, the threshold in step 5 of Algorithm 1, and let $\bar{\sigma} := 1 - \sigma$. To state further assumptions, we define the following constants:

$$c_1 := \frac{\bar{\sigma}\tau_{\mathcal{C}}}{5K_0}, \qquad C_1 := \tau_\theta^2 \tau_{\mathcal{C}} \min_h \|B^0_{h*}\|_1, \tag{23}$$

$$\tau_a := \tau_\theta \tau_B \tau_{\mathcal{C}}, \qquad \tau_\rho := \tau_\theta \tau_B \tau_0, \tag{24}$$

where τ_0 is the constant in Assumption 1(b). Let $\beta_n = \log[(3/4)K_0^2 \nu_n]$. We make the following assumptions:

$$\frac{\log n}{n} \le \frac{C_1}{300}, \qquad L \ge 2, \tag{25}$$

$$\nu_n \ge \frac{1}{C_1} \max\left\{2\sqrt{2}C_2 L, 10^3 \log n, \frac{154}{\tau_\rho^2 c_1 K_0}\frac{\beta_n}{n}\right\}, \tag{26}$$

$$\alpha_n \le \min\left\{\frac{2}{LC_2^2}, \frac{\tau_{\mathcal{C}}}{5}\frac{1-\sigma}{1+\sigma}\right\}, \tag{27}$$

where C_1 is as defined in (23) and $C_2 = 11/(c_1 C_1 \tau_\rho)$.

THEOREM 2 (Null distribution). *Consider an $n \times n$ adjacency matrix A that is generated from a Poisson DCSBM with K_0 blocks, satisfying (21) and (22). Let $\widehat{z} \in [K_0]^n$ be an estimated label vector based on A and $\widehat{y} \in [L]^{|S_1|}$ an estimated label vector based on $A_{S_1 S_1}$ satisfying Assumption 1(b). Let $\widehat{T}_n$ be the test statistic of SNAC(+). Assume that (25)–(27) hold. Then*

$$\begin{aligned} d_K(\widehat{T}_n, Z) \le{}& \frac{C_3}{\sqrt{\bar{\sigma} Ln}} + \frac{C_4}{C_1 \nu_n} \\ &+ \frac{19\sqrt{L}}{\tau_\rho}\left(\frac{1}{\sqrt{c_1 C_1}}\sqrt{\frac{\beta_n}{K_0 \nu_n}} + \frac{K_0 \beta_n}{\sqrt{\bar{\sigma} n}} + C_2 \frac{K_0^{3/2}}{\bar{\sigma} L}\nu_n \sqrt{n}\,\alpha_n\right) \\ &+ 3\mathbb{P}\big(\mathrm{Mis}(\widehat{z}, z) > \alpha_n\big), \end{aligned} \tag{28}$$

where $C_3 = 94\tau_\rho^{-4}$ and $C_4 = 4(\pi e)^{-1/2}\max\{1, \tau_\rho^{-1} - L - 1\}$.

The bound in Theorem 2 applies to both SNAC and SNAC+. Assuming the common scaling $\log n \lesssim \nu_n \lesssim \sqrt{n}$ and $\alpha_n = o(1)$, the conditions on ν_n and α_n are satisfied as $n \to \infty$ and the bound simplifies to

$$d_{\mathrm{K}}(\widehat{T}_n, Z) \lesssim \sqrt{\frac{\log \nu_n}{\nu_n}} + \nu_n \sqrt{n}\,\alpha_n + \mathbb{P}\big(\mathrm{Mis}(\widehat{z}, z) > \alpha_n\big).$$

To have a null distribution close to the standard normal, we need to have

$$\alpha_n = o\big((\nu_n \sqrt{n})^{-1}\big) \quad \text{with } \mathbb{P}\big(\mathrm{Mis}(\widehat{z}, z) > \alpha_n\big) = o(1). \tag{29}$$

There are community detection algorithms that can achieve this as long as $\nu_n \gtrsim \log n$ [10, 24, 34]. In fact, if $\nu_n \ge C \log n$ for a sufficiently large constant C, there are algorithms that achieve exact recovery, that is, we can take $\alpha_n = 0$ and still have $\mathbb{P}(\mathrm{Mis}(\widehat{z}, z) > \alpha_n) = o(1)$. It is also possible to satisfy (29) below the $\log n$ threshold on ν_n; see, for example, [13, 43, 47]. However, for the distribution to converge we still need $\nu_n \gtrsim \log n$ from (26). This is needed to guarantee the concentration of degrees d_i uniformly over all nodes $i \in S_2'$. Whether this requirement can be lifted and still achieve convergence in distribution is open.

REMARK 2 (Bernoulli vs. Poisson). Theorem 2 assumes Poisson generation for the DCSBM, and it is not clear if the result holds under the Bernoulli version. The main challenge is the conditional distribution of $X_{i*}(\widehat{y})$, which is no longer a multinomial—that is, (17) no longer holds—under the Bernoulli model. To prove Theorem 1, we use the Esseen's bound and control the moments of the conditional distribution of $X_{i*}(\widehat{y})$. Under the Bernoulli model, these moments do not have a closed form [9] and are also hard to approximate. Another approach is to show that the conditional distribution is close to a multinomial. For example, using results in [28], one can show that, for any i, the Kolmogorov distance between the distribution of $X_{i*}(\widehat{y})$ and a multinomial is of the order ν_n^2/n, which goes to zero fast under the typical sparse scaling of $\nu_n \sim \log n$. However, since SNAC(+) are roughly sums of n chi-square statistics divided by $\sqrt{n}$, the small distances of their individual terms to the desired distribution may not carry over to the distribution of their sum. In general, it is not clear if the Kolmogorov distance for sums of this form can be controlled based solely on the distances of their individual terms. Despite the above theoretical challenges, the null distribution under the Bernoulli setting is close enough to a standard normal in practice to make these results useful, especially if the bootstrap debiasing is also applied. As we show in the simulations, which are all based on Bernoulli DCSBM, SNAC+ can consistently select the correct number of communities when applied sequentially, and the performances are similar with or without bootstrap debiasing.

5. Consistency. We show the consistency of SNAC(+) against alternative models by deriving lower bounds on the statistic that go to infinity, under the alternatives, as $n \to \infty$. We consider two alternative models: (1) DCSBM with the number of communities less than that of the null; (2) DCLVM, a general class of degree-corrected latent variable models discussed in more details in Section 5.2. Combined with the null distribution in Theorem 2, the first case above shows that SNAC(+) can be applied in sequential testing from below to determine the number of communities consistently. In addition, its power against DCLVM shows its utility as a very general goodness-of-fit test beyond the DCSBM family.

5.1. *Consistency against underfitted DCSBM.* We analyze the power of SNAC(+) in distinguishing the null hypothesis $H_0 : K = K_0$ from the alternative $H_1 : K < K_0$. Theorem 3 provides a lower bound on the growth rate of the test statistic $\widehat{T}_n$ under the alternative. Recall that $\widehat{y}$ are labels derived for nodes S_1 based on $A_{S_1 S_1}$. Let parameters $\rho_{k\ell}$ be defined as in (18), and let

$$\omega_2 = \frac{1}{18}\tau_\theta^2 \tau_a^2 c_1^2 \min_{k,h\in[K_0]:\, k\neq h} \frac{1}{L}\|\rho_{k*} - \rho_{h*}\|_2^2. \tag{30}$$

See (23) and (24) for the definitions of c_1 and τ_a.

THEOREM 3. *Let A be an $n \times n$ adjacency matrix generated from a Poisson DCSBM with $K_0 \geq 2$ blocks that satisfies (21) and (22). Let $\widehat{T}_n$ be the SNAC(+) test statistic (19) formed as detailed in Algorithm 1, with $K < K_0$, estimated by a community detection algorithm satisfying stability Assumption 1(b). Let $C_5 := c_1 C_1/9$, assume that $(\log n)/\nu_n \leq C_1 \tau_\rho^2/64$ and consider the event*

$$\Omega_n := \left\{ \max\left(\frac{1}{C_5 \nu_n}, \frac{768}{\tau_\rho^3}\sqrt{\frac{\log n}{C_1 \nu_n}} \right) \leq \omega_2 \right\}. \tag{31}$$

Then, with probability at least $1 - 9Ln^{-1} - \mathbb{P}(\Omega_n^c) - \mathbb{P}(\mathrm{Mis}(\widehat{z}, z) > \alpha_n)$,

$$\widehat{T}_n \geq C_5\, \omega_2\, \nu_n \sqrt{Ln}.$$

Quantity ω_2 that appears in Theorem 3 is random (via $\{\rho_{k\ell}\}$), due to the randomness in $\widehat{y}$, and depends on the specific community detection algorithm used to form the test statistic. As discussed below, for any reasonable algorithm, under mild conditions on the connectivity matrix, we expect ω_2 to be of constant order as $n \to \infty$, that is, $\omega_2 \asymp 1$. In particular, we expect to have $\mathbb{P}(\omega_2 \geq c_0) \to 1$ for some constant $c_0 > 0$, as $n \to \infty$. Then we have $\mathbb{P}(\Omega_n^c) \to 0$, as long as $(\log n)/\nu_n \lesssim c_0$.

Under these assumptions, Theorem 2 shows that for a given significance level $\alpha > 0$, SNAC(+) statistic $\widehat{T}_n \asymp 1$ with probability approaching $1 - \alpha$ when $K = K_0$, while Theorem 3 guarantees that $\widehat{T}_n \gtrsim \nu_n\sqrt{n}$, w.h.p., when $K < K_0$. This shows that SNAC(+) with a constant threshold or one that grows slower than $\nu_n\sqrt{n}$, leads to consistent model selection when applied sequentially from below. In short, model selection consistency of SNAC(+) only requires two assumptions: (a) $(\log n)/\nu_n = O(1)$, that is, the expected degree should grow no slower than $\log n$, and (b) ω_2 should remain bounded below in probability.

In addition to consistency, Theorem 3 suggests that SNAC+ is more powerful than SNAC in sequential testing from below, due to using $L = K + 1$ clusters for column compression. The difference between the two algorithms is manifested in their corresponding values of ω_2. Let us consider the hardest case in Theorem 3, that is, testing the null hypothesis $K = K_0 - 1$ against the alternative $K = K_0$. To simplify the discussion, assume that $\nu_n \gtrsim \log n$ and the community detection algorithm is strongly consistent (achieves exact recovery). First, consider the SNAC+. Since $L = K + 1 = K_0$ in this case, the estimated column labels $\widehat{y}$ match the true labels z when computing the SNAC+ statistic. Recalling the definition of the confusion matrix from (15), we obtain $R = \operatorname{diag}(\widetilde{\pi}_k)$, where $\widetilde{\pi}_k = \frac{1}{|S_1|}\sum_{j\in S_1}\theta_j 1\{z_j = k\}$ for all $k \in [K_0]$. Then $\rho_{k\ell} = B^0_{k\ell}\widetilde{\pi}_\ell/(\sum_{\ell'} B^0_{k\ell'}\widetilde{\pi}_{\ell'})$. Note that both B^0 and $\{\widetilde{\pi}_k\}$ are stable as $n \to \infty$. In particular, although the entries of B vanish under the scaling $\nu_n/n \to 0$, the entries of $(\rho_{k\ell})$ do not. To guarantee that $\omega_2 > 0$, it suffices that the $K_0 \times K_0$ matrix $(B^0_{k\ell}\widetilde{\pi}_\ell)$ has no two colinear rows, a mild identifiability condition.

On the other hand, for SNAC we have $L = K_0 - 1$, causing the multinomial parameter matrix $\rho \in \mathbb{R}^{K_0\times(K_0-1)}$ to have rows that are weighted averages of its counterpart when $L = K_0$. We refer to [38] for an example of how the weighted mixture of the rows of the connectivity matrix B emerges in the underfitted case, and ρ is mixed in the same way. Due to this averaging, the pairwise distances among the rows of ρ will be smaller compared to when $L = K_0$, leading to smaller ω_2, hence lower power for SNAC compared to SNAC+.

The ρ-mixtures in the case of SNAC still lead to an ω_2 that is bounded away from zero—and hence preserve consistency—provided that the mixture weights do not converge to specific values that make some rows of ρ identical. This implausible situation, however, can occur in some corner cases. Consider the extreme case of the SBM with a planted partition pattern for B (with entries equal to p on the diagonal and q off the diagonal) and equal community sizes. If the community detection algorithm recovers a superset of the true communities when underfiting, as shown, for example, for the spectral clustering in [29], ρ will have identical rows in the limit, and thus $\omega_2 \to 0$ as $n \to \infty$, making SNAC powerless. More details on this example are included in Appendix B.7.

In sequential testing, one may want to know the growth rate of the test statistic $\widehat{T}_n$ in the overfitted case where $K > K_0$. The same argument as in Theorem 2 shows that under $K > K_0$, if the community detection algorithm is *refinement consistent*—that is, recovers a refinement of the true clusters—then $\widehat{T}_n$ has asymptotically a standard normal distribution, hence $\widehat{T}_n \sim 1$ as $n \to \infty$. More precisely, we have the following result.

PROPOSITION 1. *Suppose the community detection algorithm is refinement consistent, that is, there exists label vector z^* with $K > K_0$ communities, such that z^* is a refinement of the true labels, and (29) holds with z replaced with z^*. Then Theorem 2 holds with K_0 and z replaced with K and z^*.*

Some algorithms, such as spectral clustering, exhibit refinement consistency in practice; for an example, see Appendix B.7. Recent theoretical discussions of the phenomenon appear in [29, 44].

5.2. *Consistency against DCLVM.* We define a K-community DCLVM, with degree parameter θ, label vector $z^* \in [K^*]^n$, mixture components $\{\mathbb{Q}_k^*\}_{k=1}^K$ and latent variables $\{x_i\}_{i=1}^n \subset \mathcal{X}$, as follows: Given $\{x_i\}$, each (i, j) is drawn independently (of other edges) from a Poisson distribution with mean

$$p_{ij} := \mathbb{E}[\, A_{ij} \,|\, x_i,\, x_j] = \frac{\nu_n}{n}\theta_i\theta_j g(x_i, x_j) \tag{32}$$

and $x_i \sim \mathbb{Q}_{z_i^*}^*$ independently across i. The mixture components $\{\mathbb{Q}_k^*\}$ are distributions on the space $\mathcal{X}$, and when they are different, they impose some latent community structure. An example, with specific forms for $g(\cdot,\cdot)$ and $\{\mathbb{Q}_k^*\}$ is given in Appendix A.2.2. Here, we consider the general case, with minimal assumptions on $g(\cdot,\cdot)$ and $\{\mathbb{Q}_k^*\}$. We use similar assumptions on θ as in the DCSBM, namely

$$\max_i \theta_i = 1, \quad \theta_i \geq \tau_\theta,\ \forall i \in [n]. \tag{33}$$

By rescaling ν_n if need be, we assume that g has range $[0, 1]$.

Without strong assumptions on $\{\mathbb{Q}_k^*\}$, the distribution of x_i is a nonparametric mixture model which, in general, is not identifiable. One can shift mass from one of $\{\mathbb{Q}_k^*\}$ to the other ones or create a new component, and redefine the label vector to get the same distribution. For example, suppose that we start with a two-community model with components $\mathbb{Q}_1^*$ and $\mathbb{Q}_2^*$. We relabel each x_i by assigning it the new label $z_i \in [K]$ (rather than z_i^*). The same model for x_i can be stated as $x_i \sim \mathbb{Q}_{z_i}$ for new mixture components $\mathbb{Q}_k = \pi_{k1}\mathbb{Q}_1^* + \pi_{k2}\mathbb{Q}_2^*$, which are convex combinations of the original ones. We refer to $\{\mathbb{Q}_k\}$ as the mixture components induced by z. The result that we present here applies to any of these parameterizations.

Assume that we perform the SNAC(+) with K row communities and L column communities. Let $\widehat{z} \in [K]^n$ be the estimated label vector based on the entire adjacency matrix A and $\widehat{y} \in [L]^{|S_1|}$ the label vector estimated based on $A_{S_1 S_1}$. We assume that there are deterministic vectors $z \in [K]^n$ and $y \in [L]^n$, and sequences $\{\alpha_n\}$ and $\{\kappa_n\}$ such that the event

$$\mathcal{M}_n := \big\{\mathrm{Mis}(\widehat{z}, z) \leq \alpha_n \text{ and } \mathrm{Mis}(\widehat{y}, y_{S_1}) \leq \kappa_n\big\}, \tag{34}$$

has probability converging to 1, as $n \to \infty$. Here, $y_{S_1} = (y_i : i \in S_1)$ is the subvector of y on S_1. Note that we do not require z (or y) to be the original z^*. Letting $n_k = |\{i : z_i = k\}|$, we assume

$$n_k \geq \tau_{\mathcal{C}} n, \quad \forall k \in [K]. \tag{35}$$

Let $\{\mathbb{Q}_k, k \in [K]\}$ be the mixture components induced by label vector z that appears in (34). Define

$$h_k(x) := \mathbb{E}\big[g(x, \xi)\big], \quad \xi \sim \mathbb{Q}_k, k \in [K].$$

We assume that there is an almost sure event Γ with the following property: There exists a constant $\tau_h > 0$ and $r_1, r_2 \ldots, r_K \in [K]$ such that on Γ,

$$\forall k \in [K], \forall i \in \mathcal{C}_k, \quad h_{r_k}(x_i) \geq \tau_h, \tag{36}$$

where $\mathcal{C}_k = \{i : z_i = k\}$. Note that (36) can be equivalently stated as $h_{r_{z_i}}(x_i) \geq \tau_h$ for all i. Condition (36) is mild and is satisfied if for any $k \in [K]$, one of $h_r(\cdot), r \in [K]$ is uniformly bounded below over the support of $\mathbb{Q}_k$. We also define

$$H_\ell(x) := \frac{\sum_k h_k(x) R_{k\ell}}{\sum_{\ell'}\sum_k h_k(x) R_{k\ell'}}, \qquad R_{k\ell} := \frac{1}{2}\sum_{j=1}^n \theta_j 1\{z_j = k,\ y_j = \ell\}. \tag{37}$$

Note that there exists a sequence $\{\ell_k\}_{k=1}^K$ such that

$$R_{k\ell_k} \geq \frac{1}{L}\sum_{\ell=1}^{L} R_{k\ell}, \quad \forall k \in [K]. \tag{38}$$

Fix one such sequence and consider the following quantities:

$$\begin{aligned} \vartheta_{k\ell} &:= \operatorname{var}\big(H_\ell(x)\big), \quad \text{where } x \sim \mathbb{Q}_k, \\ \underline{\vartheta} &:= \min_k \vartheta_{k\ell_k}. \end{aligned} \tag{39}$$

Let $\zeta_n = \max\{1, L\sqrt{\nu_n/n}, L/\sqrt{\nu_n \log n}\}$ and $c_2 = \tau_C \tau_h \tau_\theta^2/100$ and $\tau_\rho = \tau_C \tau_h \tau_\theta/(2L)$. We need the following assumptions:

$$\sqrt{\frac{\log n}{n}} \leq \frac{2}{9}\frac{\tau_\rho^2}{K}, \quad n \geq 2, \tag{40}$$

$$\alpha_n \leq \sqrt{\frac{\log n}{\nu_n}} \leq \frac{21\tau_C^2 \tau_h c_2^2}{L^2}, \quad \frac{n\kappa_n}{\nu_n} \leq 4c_2\tau_\rho, \tag{41}$$

$$\underline{\vartheta} \geq \frac{L^3}{c_2^3 \tau_\rho^3} \max\left\{\frac{2\zeta_n}{\tau_\rho \tau_C}\sqrt{\frac{\log n}{\nu_n}}, \frac{1}{5c_2}\frac{n\kappa_n}{\nu_n}\right\}. \tag{42}$$

THEOREM 4. *Let A be an $n \times n$ adjacency matrix generated from a K-community Poisson DCLVM* (32) *satisfying* (33) *and* (35). *Let $\widehat{T}_n$ be the SNAC(+) statistic* (19) *formed as detailed in Algorithm* 1. *Moreover, assume* (36) *and* (40)–(42). *Then, with probability at least* $1 - 12KLn^{-1} - Kn^{-c} - \mathbb{P}(\mathcal{M}_n^c)$,

$$\widehat{T}_n \geq \frac{49\, c_2^3}{\sqrt{L}}\, \underline{\vartheta}\sqrt{n}\nu_n,$$

where $c > 0$ is a universal constant.

The theorem roughly states the following: As long as the community detection algorithm produces row and columns labels that converge to some deterministic labels z and y at the rates $\alpha_n \sim \sqrt{(\log n)/\nu_n}$ and $\kappa_n \sim \nu_n/n$, respectively, and the resulting induced mixture components $\{\mathbb{Q}_k\}$ lead to a positive minimum variance $\underline{\vartheta}$, as defined in (39), then SNAC(+) are consistent in rejecting the underlying DCLVM model, with $\widehat{T}_n \gtrsim \sqrt{n}\nu_n \to \infty$. Note that $\underline{\vartheta} > 0$, unless there exists a sequence of constants $a_1, \ldots, a_K$ such that $\sum_r a_r h_r(x) = 0$ for $\mathbb{Q}_k$-almost all x. That is, unless $\{h_r\}_{r=1}^K$ satisfy a nontrivial linear constraint under $\mathbb{Q}_k$, the condition $\underline{\vartheta} > 0$ is guaranteed. An example where the condition $\underline{\vartheta} > 0$ is violated is when all $h_r(\cdot)$ are constant functions, as is the case for a DCSBM, consistent with the fact that we should not be able to reject a DCSBM. See Appendix B.8 for further remarks.

6. Goodness-of-fit to FB-100 data. The main utility of a goodness-of-fit test is to assess how well real data fits the model. Let us investigate how well a DCSBM fits real networks from the Facebook-100 data set [35, 36], hereafter referred to as FB-100. This data set is a collection of 100 social networks, each the entire Facebook network within one university from a date in 2005. The networks vary considerably in size and degree characteristics; some statistics are provided in Table 1.

Figure 1 shows the violin plots of the SNAC+ statistic, with degree-filtering threshold $\sigma = 0.2$, versus the number of communities, for the entire FB-100 data. The variation at each K is due to the variability of SNAC+ over the 100 networks in the data set. For each

TABLE 1
Statistics on the FB-100 data set. Qu. is a shorthand for quartile

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
n	769	4444	9950	12,083	17,033	41,554
Mean deg.	39	65	77	77	88	116
3rd Qu. deg.	54	91	110	108	124	166
Max. deg.	248	673	1202	1787	2123	8246

FB network, we sample a twin network from a synthetic 3-cluster DCSBM that matches the original network in degree distribution. Violin plots are also shown for these twin networks for comparison.

Each synthetic DCSBM has its own θ parameter proportional to the corresponding FB network degree vector, but they all share the same connectivity matrix B, which is set to the corresponding MLE based on all the FB networks. To get the shared B, we first apply spectral clustering with $K = 3$ to each FB network $A^{(s)}$, $s = 1, \ldots, 100$ to get estimated labels $\widehat{z}^{(s)}$. Then, for each $\widehat{z}^{(s)}$, we compute the corresponding block sum and block size matrices, $N^{(s)}$ and $M^{(s)}$; see Appendix A.1 for details. Finally, we set $B = \sum_s N^{(s)} / \sum_s M^{(s)}$, where the summation and division are elementwise. The community sizes for the synthetic networks are taken to be balanced. Kolmogorov– Smirnov tests were performed between the degree distributions of each FB network and its twin, and 84 out of 100 such pairs resulted in p-values greater than 0.05, indicating close matches.

The results in Figure 1 show a marked deviation of FB-100 networks from a DCSBM. If the networks were generated from a DCSBM, one would expect the distribution of SNAC+ to drop to within a narrow band around zero once K surpasses the true number of communities. Only at $K = 25$ a small fraction of FB-100 networks have SNAC+ values within, say, the interval $[-5, 5]$, showing that a DCSBM with $K < 25$ is not a good model for any of these networks. Even at $K = 25$, the majority of FB-100 networks are still ill-fitted. On the other hand, we observe that for simulated DCSBM twins, SNAC+ is nearly normally distributed for $K = 3$, while remaining large for $K = 1$ and $K = 2$. This corroborates the results of both Theorem 2 and Theorem 3 that predict exactly this behavior. Note that this conclusion holds despite the variation in the sizes and average degrees of the simulated networks, showing

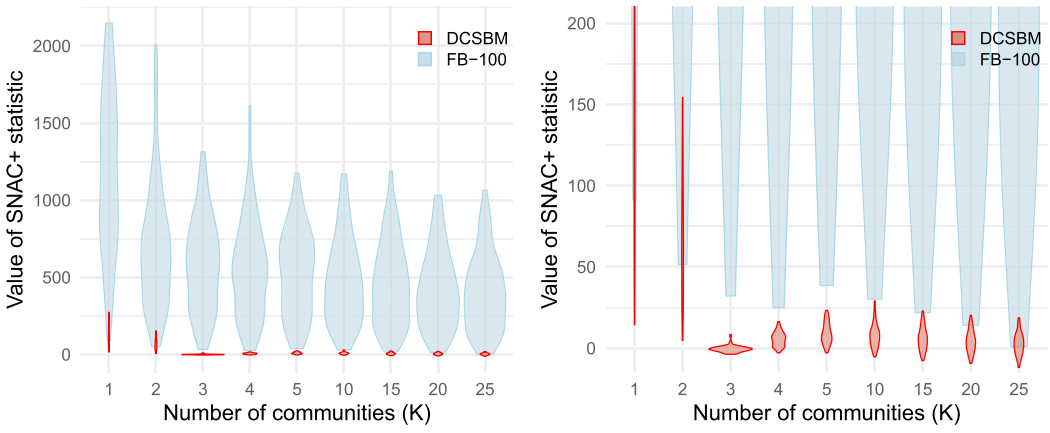

FIG. 1. *Comparing the goodness-of-fit of DCSBM to the FB-100 data set versus a data set simulated from twin DCSBMs with $K = 3$ communities, and having the same sizes and degree distributions as those of FB-100. The right plot is the zoomed-in version of the left.*

the insensitivity of the null distribution of SNAC+ to those parameters, as predicted by the theory.

We further explore the FB-100 data in Appendix A.3, where we show that removing high-degree nodes can result in a better overall DCSBM fit, though still far from ideal. Despite the lack of fit, in Appendix A.4, we propose a smoothed SNAC+ curve and show how it can be used to build *community profiles* of real networks, turning the statistic into an effective tool for exploring community structure. In addition to FB-100 data, the results of extensive simulation studies, comparing to competing methods, are reported in Appendix A. The R package `nett` implementing our proposed tests and the competing methods is available at [3]. The code for reproducing the experiments can be found at [45].

Acknowledgments. We thank Mason Porter who provided access to the Facebook-100 data set.

Funding. This work was supported by NSF Grant DMS-1945667.

SUPPLEMENTARY MATERIAL

Appendices (DOI: 10.1214/23-AOS2329SUPP; .pdf). This supplement contains further numerical experiments, discussion and proofs.

REFERENCES

[1] ABBE, E. (2017). Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18** Paper No. 177, 86. MR3827065

[2] AMINI, A., PAEZ, M. and LIN, L. (2022). Hierarchical stochastic block model for community detection in multiplex networks. *Bayesian Anal. Advance Publication* 1–27.

[3] AMINI, A. A. and ZHANG, L. (2020). *nett package*, https://aaamini.github.io/nett/index.html.

[4] BICKEL, P. J., CHEN, A. and LEVINA, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.* **39** 2280–2301. MR2906868 https://doi.org/10.1214/11-AOS904

[5] BICKEL, P. J., RITOV, Y. and STOKER, T. M. (2006). Tailor-made tests for goodness of fit to semiparametric hypotheses. *Ann. Statist.* **34** 721–741. MR2281882 https://doi.org/10.1214/009053606000000137

[6] BICKEL, P. J. and SARKAR, P. (2016). Hypothesis testing for automated community detection in networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 253–273. MR3453655 https://doi.org/10.1111/rssb.12117

[7] BREIMAN, L. (2001). Statistical modeling: The two cultures. *Statist. Sci.* **16** 199–231. MR1874152 https://doi.org/10.1214/ss/1009213726

[8] CHEN, K. and LEI, J. (2018). Network cross-validation for determining the number of communities in network data. *J. Amer. Statist. Assoc.* **113** 241–251. MR3803461 https://doi.org/10.1080/01621459.2016.1246365

[9] CHEN, S. X. (2000). General properties and estimation of conditional Bernoulli models. *J. Multivariate Anal.* **74** 69–87. MR1790614 https://doi.org/10.1006/jmva.1999.1872

[10] CHEN, Y., LI, X. and XU, J. (2018). Convexified modularity maximization for degree-corrected stochastic block models. *Ann. Statist.* **46** 1573–1602. MR3819110 https://doi.org/10.1214/17-AOS1595

[11] DALL'AMICO, L., COUILLET, R. and TREMBLAY, N. (2021). A unified framework for spectral clustering in sparse graphs. *J. Mach. Learn. Res.* **22** Paper No. 217, 56. MR4329796

[12] ERDŐS, L., YAU, H.-T. and YIN, J. (2012). Rigidity of eigenvalues of generalized Wigner matrices. *Adv. Math.* **229** 1435–1515. MR2871147 https://doi.org/10.1016/j.aim.2011.12.010

[13] GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.* **18** Paper No. 60, 45. MR3687603

[14] GENG, J., BHATTACHARYA, A. and PATI, D. (2019). Probabilistic community detection with unknown number of communities. *J. Amer. Statist. Assoc.* **114** 893–905. MR3963189 https://doi.org/10.1080/01621459.2018.1458618

[15] HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088 https://doi.org/10.1016/0378-8733(83)90021-7

[16] HU, J., QIN, H., YAN, T. and ZHAO, Y. (2020). Corrected Bayesian information criterion for stochastic block models. *J. Amer. Statist. Assoc.* **115** 1771–1783. MR4189756 https://doi.org/10.1080/01621459.2019.1637744

[17] HUNTER, D. R., GOODREAU, S. M. and HANDCOCK, M. S. (2008). Goodness of fit of social network models. *J. Amer. Statist. Assoc.* **103** 248–258. MR2394635 https://doi.org/10.1198/016214507000000446

[18] KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107, 10. MR2788206 https://doi.org/10.1103/PhysRevE.83.016107

[19] KARWA, V., PATI, D., PETROVIĆ, S., SOLUS, L., ALEXEEV, N., RAIČ, M., WILBURNE, D., WILLIAMS, R. and YAN, B. (2016). Exact tests for stochastic block models. Preprint. Available at arXiv:1612.06040.

[20] KAWAMOTO, T. and KABASHIMA, Y. (2017). Cross-validation estimate of the number of clusters in a network. *Sci. Rep.* **7**.

[21] LE, C. M. and LEVINA, E. (2022). Estimating the number of communities by spectral methods. *Electron. J. Stat.* **16** 3315–3342. MR4422967 https://doi.org/10.1214/21-ejs1971

[22] 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 https://doi.org/10.1215/00127094-2414767

[23] LEI, J. (2016). A goodness-of-fit test for stochastic block models. *Ann. Statist.* **44** 401–424. MR3449773 https://doi.org/10.1214/15-AOS1370

[24] LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605 https://doi.org/10.1214/14-AOS1274

[25] LI, T., LEI, L., BHATTACHARYYA, S., VAN DEN BERGE, K., SARKAR, P., BICKEL, P. J. and LEVINA, E. (2022). Hierarchical community detection by recursive partitioning. *J. Amer. Statist. Assoc.* **117** 951–968. MR4436325 https://doi.org/10.1080/01621459.2020.1833888

[26] LI, T., LEVINA, E. and ZHU, J. (2020). Network cross-validation by edge sampling. *Biometrika* **107** 257–276. MR4108931 https://doi.org/10.1093/biomet/asaa006

[27] LI, Y. and CHOUGH CARRIERE, K. (2013). Assessing goodness of fit of exponential random graph models. *Int. J. Stat. Probab.* **2** 64.

[28] LOH, W.-L. (1992). Stein's method and multinomial approximation. *Ann. Appl. Probab.* **2** 536–554. MR1177898

[29] MA, S., SU, L. and ZHANG, Y. (2021). Determining the number of communities in degree-corrected stochastic block models. *J. Mach. Learn. Res.* **22** Paper No. 69, 63. MR4253762

[30] NEWMAN, M. E. and REINERT, G. (2016). Estimating the number of communities in a network. *Phys. Rev. Lett.* **117** 078301.

[31] NEWMAN, M. E. J. and GIRVAN, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E, Stat. Nonlin. Soft Matter Phys.* **69** 03.

[32] OSPINA-FORERO, L., DEANE, C. M. and REINERT, G. (2019). Assessment of model fit via network comparison methods based on subgraph counts. *J. Complex Netw.* **7** 226–253.

[33] RIOLO, M. A., CANTWELL, G. T., REINERT, G. and NEWMAN, M. E. (2017). Efficient method for estimating the number of communities in a network. *Phys. Rev. E* **96** 032310.

[34] QIN, T. and ROHE, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems* 3120–3128.

[35] TRAUD, A. L., KELSIC, E. D., MUCHA, P. J. and PORTER, M. A. (2011). Comparing community structure to characteristics in online collegiate social networks. *SIAM Rev.* **53** 526–543. MR2834086 https://doi.org/10.1137/080734315

[36] TRAUD, A. L., MUCHA, P. J. and PORTER, M. A. (2012). Social structure of Facebook networks. *Phys. A, Stat. Mech. Appl.* **391** 4165–4180.

[37] VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[38] WANG, Y. X. R. and BICKEL, P. J. (2017). Likelihood-based model selection for stochastic block models. *Ann. Statist.* **45** 500–528. MR3650391 https://doi.org/10.1214/16-AOS1457

[39] YAN, B., SARKAR, P. and CHENG, X. Provable estimation of the number of blocks in block models. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, *Playa Blanca*, *Lanzarote*, *Canary Islands*, 09–11 *Apr* 2018 (A. Storkey and F. Perez-Cruz, eds.). *Proceedings of Machine Learning Research* **84** 1185–1194. PMLR.

[40] YAN, X. (2016). Bayesian model selection of stochastic block models. In 2016 *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (*ASONAM*) 323–328.

[41] 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.

[42] YUAN, M., FENG, Y. and SHANG, Z. (2022). A likelihood-ratio type test for stochastic block models with bounded degrees. *J. Statist. Plann. Inference* **219** 98–119. MR4355951 https://doi.org/10.1016/j.jspi.2021.12.005

[43] ZHANG, A. Y. and ZHOU, H. H. (2016). Minimax rates of community detection in stochastic block models. *Ann. Statist.* **44** 2252–2280. MR3546450 https://doi.org/10.1214/15-AOS1428

[44] ZHANG, L. and AMINI, A. (2021). Label consistency in overfitted generalized k-means. *Adv. Neural Inf. Process. Syst.* **34**.

[45] ZHANG, L. and AMINI, A. A. (2020). Adjusted chi-square test for degree-corrected block models: Experiments in R. https://github.com/linfanz/nac-test.

[46] ZHANG, L. and AMINI, A. A (2023). Supplement to "Adjusted chi-square test for degree-corrected block models." https://doi.org/10.1214/23-AOS2329SUPP

[47] ZHOU, Z. and AMINI, A. A. (2020). Optimal bipartite network clustering. *J. Mach. Learn. Res.* **21** Paper No. 40, 68. MR4073773