PHASE TRANSITION FOR DETECTING A SMALL COMMUNITY IN A LARGE NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

How to detect a small community in a large network is an interesting problem, including clique detection as a special case, where a naive degree-based χ^2-test was shown to be powerful in the presence of an Erdös-Renyi background. Using Sinkhorn's theorem, we show that the signal captured by the χ^2-test may be a modeling artifact, and it may disappear once we replace the Erdös-Renyi model by a broader network model. We show that the recent SgnQ test is more appropriate for such a setting. The test is optimal in detecting communities with sizes comparable to the whole network, but has never been studied for our setting, which is substantially different and more challenging. Using a degree-corrected block model (DCBM), we establish phase transitions of this testing problem concerning the size of the small community and the edge densities in small and large communities. When the size of the small community is larger than $\sqrt{n}$, the SgnQ test is optimal for it attains the computational lower bound (CLB), the information lower bound for methods allowing polynomial computation time. When the size of the small community is smaller than $\sqrt{n}$, we establish the parameter regime where the SgnQ test has full power and make some conjectures of the CLB. We also study the classical information lower bound (LB) and show that there is always a gap between the CLB and LB in our range of interest.

1 INTRODUCTION

Consider an undirected network with n nodes and K communities. We assume n is large and the network is connected for convenience. We are interested in testing whether $K = 1$ or $K > 1$ and the sizes of some of the communities are much smaller than n (communities are scientifically meaningful but mathematically hard to define; intuitively, they are clusters of nodes that have more edges "within" than "across" (Jin, 2015; Zhao et al., 2012)). The problem is a special case of network global testing, a topic that has received a lot of attention (e.g., Jin et al. (2018; 2021b)). However, existing works focused on the so-called *balanced case*, where the sizes of communities are at the same order. Our case is *severely unbalanced*, where the sizes of some communities are much smaller than n (e.g., n^ε).

The problem also includes clique detection (a problem of primary interest in graph learning (Alon et al., 1998; Ron & Feige, 2010)) as a special case. Along this line, Arias-Castro & Verzelen (2014); Verzelen & Arias-Castro (2015) have made remarkable progress. In detail, they considered the problem of testing whether a graph is generated from a one-parameter Erdös-Renyi model or a two-parameter model: for any nodes $1 \leq i, j \leq n$, the probability that they have an edge equals b if i, j both are in a small planted subset and equals a otherwise. A remarkable conclusion of these papers is: a naive degree-based χ^2-test is optimal, provided that the clique size is in a certain range. Therefore, at first glance, it seems that the problem has been elegantly solved, at least to some extent.

Unfortunately, recent progress in network testing tells a very different story: the signal captured by the χ^2-test may be a modeling artifact. It may disappear once we replace the models in Arias-Castro & Verzelen (2014); Verzelen & Arias-Castro (2015) by a properly broader model. When this happens, the χ^2-test will be asymptotically powerless in the whole range of parameter space.

We explain the idea with the popular *Degree-Corrected Block Model (DCBM)* (Karrer & Newman, 2011), though it is valid in broader settings. Let $A \in \mathbb{R}^{n,n}$ be the network adjacency matrix, where $A(i, j) \in \{0, 1\}$ indicates whether there is an edge between nodes i and j, $1 \leq i, j \leq n$. By convention, we do not allow for self-edges, so the diagonals of A are always 0. Suppose there

are K communities, $\mathcal{C}_1, \ldots, \mathcal{C}_K$. For each node i, $1 \leq i \leq n$, we use a parameter θ_i to model the degree heterogeneity and π_i to model the membership: when $i \in \mathcal{C}_k$, $\pi_i(\ell) = 1$ if $\ell = k$ and $\pi_i(\ell) = 0$ otherwise. For a $K \times K$ symmetric and irreducible non-negative matrix P that models the community structure, DCBM assumes that the upper triangle of A contains independent Bernoulli random variables satisfying

$$\mathbb{P}(A(i,j) = 1) = \theta_i \theta_j \pi_i' P \pi_j, \qquad 1 \leq i, j \leq n. \tag{1.1}$$

In practice, we interpret $P(k, \ell)$ as the baseline connecting probability between communities k and ℓ. Write $\theta = (\theta_1, \theta_2, \ldots, \theta_n)'$, $\Pi = [\pi_1, \pi_2, \ldots, \pi_n]'$, and $\Theta = \mathrm{diag}(\theta) \equiv \mathrm{diag}(\theta_1, \theta_2, \ldots, \theta_n)$. Introduce $n \times n$ matrices Ω and W by $\Omega = \Theta \Pi P \Pi' \Theta$ and $W = A - \mathbb{E}[A]$. We can re-write (1.1) as

$$A = \Omega - \mathrm{diag}(\Omega) + W. \tag{1.2}$$

We call Ω the *Bernoulli probability matrix* and W the noise matrix. When θ_i in the same community are equal, DCBM reduces to the Stochastic Block Model (SBM) (Holland et al., 1983). When $K = 1$, the SBM reduces to the Erdös-Renyi model, where $\Omega(i,j)$ take the same value for all $1 \leq i, j \leq n$.

To see why the signal captured by the χ^2-test in Arias-Castro & Verzelen (2014); Verzelen & Arias-Castro (2015) may be a modeling artifact, we use Sinkhorn's matrix scaling theorem (Sinkhorn, 1974). Fix (θ, P) and suppose π_i are iid from a distribution F. Let $h = \mathbb{E}_F[\pi_i]$. By the main result of Sinkhorn (1974), there is a unique positive diagonal matrix $D = \mathrm{diag}(d_1, \ldots, d_K)$ such that $DPDh = \mathbf{1}_K$. Consider a pair of two DCBM, a null with $K = 1$ and an alternative with $K > 1$, where the Bernoulli probability matrices are Ω and Ω^*, respectively. Here, $\Omega = \Theta \mathbf{1}_n \mathbf{1}_n' \Theta \equiv \theta\theta'$ and $\Omega^*(i,j) = \theta_i^* \theta_j^* \pi_i' P \pi_j$ with $\theta_i^* = d_k \theta_i$ if $i \in \mathcal{C}_k$, $1 \leq k \leq K$. It is seen that $\theta_i^* \pi_i = \theta_i D \pi_i$ for all $1 \leq i \leq n$. Fix i. In the null, the expected degree of node i is $\sum_{j=1}^n \theta_i \theta_j = \theta_i \|\theta\|_1$. In the alternative, conditioning on π_i, the expected degree of node i is $\theta_i^* \pi_i' \mathbb{E}[\sum_{j=1}^n \theta_j^* P \pi_j] = \theta_i \|\theta\|_1 \pi_i'(DPD)h = \theta_i \|\theta\| \pi_i' \mathbf{1}_K$, which is also equal to $\theta_i \|\theta\|_1$. Here, we have neglected the matrix $\mathrm{diag}(\Omega)$ in (1.2), the effect of which is secondary. Also, note that $\Omega^* = \Theta \Pi (DPD) \Pi' \Theta$ and $(DPD)h = \mathbf{1}_K$.

In summary, fix an alternative DCBM ($K > 1$) with a Bernoulli probability matrix Ω^*, we can always use the Sinkhorn's theorem to redefine (Θ, P) (while Π is unchanged) so that for the new (Θ, P), $\Omega^* = \Theta \Pi P \Pi' \Theta$ and $Ph = \mathbf{1}_K$. This motivates an interesting *identifiability condition for DCBM*; see Section 2.1. Moreover, if we pair the alternative with the null with $\Omega = \Theta \mathbf{1}_n \mathbf{1}_n' \Theta \equiv \theta\theta'$, then for each node, the expected degrees in two models match with each other, and a naive degree-based χ^2-test as in Arias-Castro & Verzelen (2014) is asymptotically powerless.

In the Erdös-Renyi setting in Arias-Castro & Verzelen (2014), the null has one parameter and the alternative has two parameters. In such a setting, we cannot have degree-matching. In these cases, a naive degree-based χ^2-test may have good power, but it is due to the very specific models they choose. For clique detection in more realistic settings, we prefer to use a broader model such as the DCBM, where by the degree-matching argument above, the χ^2-test is asymptotically powerless.

This motivates us to look for a different test. One candidate is the scan statistic Bogerd et al. (2021). However, a scan statistic is only computationally feasible when each time we scan a very small subset of nodes. For example, if each time we only scan a finite number of nodes, then the computational cost is polynomial; we call the test the *Economic Scan Test (EST)*. Another candidate may come from the Signed-Polygon test family (Jin et al., 2021b), including the Signed-Quadrilateral (SgnQ) as a special case. Let $\hat{\eta} = (\mathbf{1}_n A \mathbf{1}_n)^{-1/2} A \mathbf{1}_n$ and $\widehat{A} = A - \hat{\eta}\hat{\eta}$. Define $Q_n = \sum_{i_1, i_2, i_3, i_4 (dist)} \widehat{A}_{i_1 i_2} \widehat{A}_{i_2 i_3} \widehat{A}_{i_3 i_4} \widehat{A}_{i_4 i_1}$. The SgnQ test statistic is

$$\psi_n = \big[Q_n - 2(\|\hat{\eta}\|^2 - 1)^2\big] / \sqrt{8(\|\hat{\eta}\|^2 - 1)^4}. \tag{1.3}$$

SgnQ is computationally attractive because it can be evaluated in time $O(n^2 \bar{d})$, where $\bar{d}$ is the average degree of the network (Jin et al., 2021b).

Moreover, it was shown in Jin et al. (2021b) that (a) when $K = 1$ (the null case), $\psi_n \to N(0,1)$, and (b) when $K > 1$ and all communities are at the same order (i.e., a balanced alternative case), the SgnQ test achieves the classical information lower bound (LB) for global testing and so is optimal. Unfortunately, our case is much more delicate: the signal of interest is contained in a community with a size that is much smaller than n (e.g., n^ε), so the signal can be easily overshadowed by the noise term of Q_n. Even in the simple alternative case where we only have two communities (with sizes N

and $(n - N)$), it is unclear (a) how the lower bounds vary as $N/n \to 0$, and especially whether there is a gap between the computation lower bound (CLB) and classical information lower bound (LB), and (b) to what extent the SgnQ test attains the CLB and so is optimal.

1.1 RESULTS AND CONTRIBUTIONS

Fix $K > m \geq 1$ and consider a DCBM with K communities as in (1.2). We assume for $N \ll n$, the sizes of the m smallest communities are $O(N)$, and the sizes of other communities are $O(n)$. Let $\widetilde{\Omega} = \Omega - (\mathbf{1}_n'\Omega\mathbf{1}_n)^{-1}\Omega\mathbf{1}_n\mathbf{1}_n'\Omega$, and let λ_1 and $\tilde{\lambda}_1$ be the largest (in magnitude) eigenvalues of Ω and $\widetilde{\Omega}$, respectively. Consider the aforementioned global testing problem where we test for the presence of a small community. We show that if $|\tilde{\lambda}_1|/\sqrt{\lambda_1} \to \infty$, then SgnQ has full power asymptotically. This extends the results by Jin et al. (2021b) in the balanced case case (i.e., $N \geq Cn$) to the severely unbalanced case where N/n may tend to 0 arbitrarily fast (as long as $|\tilde{\lambda}_1|/\sqrt{\lambda_1} \to \infty$). In this work, we provide a complete analysis of SgnQ for the case $(K, m) = (2, 1)$. We also expect that our result holds for general (K, m) and can be proven with the same techniques (but more tedious calculations).

For optimality, we obtain a computational lower bound (CLB), relying on the low-degree polynomial conjecture, which is a standard approach in studying CLB (e.g., Kunisky et al. (2019)). Consider a case where $K = 2$ and we have a small community with size N. Suppose the edge probability within the community and outside the community are a and c, where $a > c$. The quantity $(a - c)/\sqrt{c}$ acts as the *Node-wise Signal-to-Noise Ratio (SNR)* for the detection problem.[1] When $N \gg \sqrt{n}$, we find that the CLB is completely determined by N and node-wise SNR; moreover, SgnQ matches with the CLB and is optimal. When $N \ll \sqrt{n}$, the situation is more subtle: if the node-wise SNR $(a - c)/\sqrt{c} \to 0$ (weak signal case), we show the problem is computationally hard and the LB depends on N and the node-wise SNR. If $(a - c)/\sqrt{c} \gg n^{1/2}$ (strong signal case), then SgnQ solves the detection problem. In the range $1 \ll (a - c)/\sqrt{c} \ll n^{1/2}$ (moderate signal case), the CLB depends on not only N and the node-wise SNR but also the background edge density c. In this regime, we make conjectures of the CLB, from the study of the aforementioned economic scan test (EST).

We also obtain the classical information lower bound (LB), and discover that as $N/n \to 0$, there is big gap between CLB and LB. Notably the LB is achieved by an (inefficient) signed scan test. In the balanced case in Jin et al. (2021b), the SgnQ test is optimal among all tests (even those that are allowed unbounded computation time), and such a gap does not exist.

We also show that, for any alternative, we can pair it with a null, so that the naive degree-based χ^2-test is asymptotically powerless in separating the pair. This verifies our degree-matching argument above.

Our statistical lower bound, computational lower bound, and the powerlessness of χ^2 based on degree-matching are valid for all (K, m) since any model with $K \geq 2, m \geq 1$ contains $(K, m) = (2, 1)$ as a special case. We also expect that our lower bounds are tight for these broader models and that our lower bound constructions for $(K, m) = (2, 1)$ represent the least favorable cases.

Compared to Verzelen & Arias-Castro (2015); Arias-Castro & Verzelen (2014), we consider network global testing in a more realistic setting, and show that optimal tests there (i.e., a naive degree-based χ^2 test) may be asymptotically powerless here. Compared with Bogerd et al. (2021), our setting is very different (they considered a setting where both the null and alternative are DCBM with $K = 1$). Compared to the study in the balanced case (e.g., Jin et al. (2018; 2021b); Gao & Lafferty (2017)), our study is more challenging for two reasons. First, in the balanced case, there is no gap between the UB (the upper bound provided by the SgnQ test) and LB, so there is no need to derive the CLB, which is usually technical demanding. Second, the size of the smaller community can get as small as n^{ε}, where $\varepsilon > 0$ is any constant. In comparison, the standard deviation of the SgnQ test statistics is at a much larger order. Therefore, to prove our results, we have to analyze each term carefully, so the signal would not be overwhelmed by the noise.

Our study of the CLB is connected to the study of CLB by Hajek et al. (2015) in the Erdös-Renyi setting of Arias-Castro & Verzelen (2014); Verzelen & Arias-Castro (2015). In this simpler model, degree matching is not possible, and Hajek et al. (2015) prove that the naive χ^2-test is the optimal polynomial-time test (conditionally on the planted-clique hypothesis). However, we consider the

[1] Note that the node-wise SNR captures the ratio of the mean difference and standard deviation of Bernoulli(a) versus Bernoulli(c), which motivates our terminology.

DCBM model, where due to the subtle phenomenon of degree matching between the null and alternative hypotheses, both CLB and LB are different from those obtained by Hajek et al. (2015).

Our work is also connected to property testing in theoretical computer science (Goldreich et al., 1998; Goldreich & Ron, 2011). Goldreich & Ron (2011) studies testing if a given graph is ϵ-close (two graphs are said to be ϵ-close if one can be obtained from the other by removing or deleting an ϵ-fraction of its edges) to one that has community structure, in the sense of its normalized adjacency matrix having a large second eigenvalue $\hat{\lambda}_2^*$. Although our notion of communities is also related to spectral properties, the test statistic $\hat{\lambda}_2^*$ can be very noisy due to the severe degree heterogeneity in DCBM, and moreover it does not have a tractable null. For these reasons, the algorithms of property testing are not directly applicable in our setting.

Notations: We use $\mathbf{1}_n$ to denote a n-dimensional vector of ones. For a vector $\theta = (\theta_1, \ldots, \theta_n)$, $\mathrm{diag}(\theta)$ is the diagonal matrix where the i-th diagonal entry is θ_i. For a matrix $\Omega \in \mathbb{R}^{n\times n}$, $\mathrm{diag}(\Omega)$ is the diagonal matrix where the i-th diagonal entry is $\Omega(i,i)$. For a vector $\theta \in \mathbb{R}^n$, $\theta_{max} = \max\{\theta_1, \ldots, \theta_n\}$ and $\theta_{min} = \min\{\theta_1, \ldots, \theta_n\}$. For two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \asymp b_n$ if $c_1 \le a_n/b_n \le c_2$ for constants $c_2 > c_1 > 0$. We say $a_n \sim b_n$ if $(a_n/b_n) = 1 + o(1)$.

2 Main results

In Section 2.1, following our discussion on Sinkhorn's theorem in Section 1, we introduce calibrations (including conditions on identifiability and balance) that are appropriate for severely unbalanced DCBM and illustrate with some examples. In Sections 2.2-2.3, we analyze the power of the SgnQ test and compare it with the χ^2-test. In Sections 2.4-2.5, we discuss the information lower bounds (both the LB and CLB) and show that SgnQ test is optimal among polynomial time tests, when $N \gg \sqrt{n}$. In Section 2.6, we study the EST and make some conjectures of the CLB when $N \ll \sqrt{n}$. In Section 2.7, we summarize our results and present the phase transitions.

2.1 DCBM for severely unbalanced networks: identifiability, balance metrics, and global testing

In the DCBM (1.1)-(1.2), $\Omega = \Theta\Pi P\Pi'\Theta$. It is known that the matrices (Θ, Π, P) are not identifiable. One issue is that (Π, P) are only unique up to a permutation: for a $K \times K$ permutation matrix Q, $\Pi P \Pi = (\Pi Q)(Q'PQ)(\Pi Q)'$. This issue is easily fixable in applications so is usually neglected. A bigger issue is that, (Θ, P) are not uniquely defined. For example, fixing a positive diagonal matrix $D \in \mathbb{R}^{K\times K}$, let $P^* = DPD$ and $\Theta^* = \mathrm{diag}(\theta_1^*, \theta_2^*, \ldots, \theta_n^*)$ where $\theta_i^* = \theta_i/\sqrt{D(k,k)}$ if $i \in \mathcal{C}_k$, $1 \le k \le K$. It is seen that $\Theta\Pi P\Pi'\Theta = \Theta^*\Pi P^*\Pi'\Theta^*$, so (Θ, P) are not uniquely defined.

There are many ways to resolve the issue. For example, in the balanced case (e.g., Jin et al. (2021b; 2022)), we can resolve it by requiring that P has unit diagonals. However, for our case, this is inappropriate. Recall that, in practice, $P(k,\ell)$ represents as the baseline connecting probability between community k and ℓ. If we forcefully rescale P to have a unit diagonal here, both (P, Θ) lose their practical meanings.

Motivated by our earlier discussion on Sinkhorn's theorem in Section 1, we propose an identifiability condition that is more appropriate for the severely unbalanced DCBM. By our discussion in Section 1, for any DCBM with a Bernoulli probability matrix Ω, we can always use Sinkhorn's theorem to define (Θ, P) (while Π is unchanged) such that for the new (Θ, P), $\Theta = \Theta\Pi P\Pi'\Theta$ and $Ph \propto \mathbf{1}_K$, where $h = (h_1, \ldots, h_K)'$ and $h_k > 0$ is the fraction of nodes in community k, $1 \le k \le K$. This motivates the following identifiability condition (which is more appropriate for our case):

$$\|\theta\|_1 = n, \qquad Ph \propto \mathbf{1}_K, \quad \text{where } h_k \text{ is fraction of nodes in } \mathcal{C}_k,\ 1 \le k \le K. \tag{2.1}$$

Lemma 2.1. *For any Ω that satisfies the DCBM (1.2) and has positive diagonal elements, we can always find (Θ, Π, P) such that $\Omega = \Theta\Pi P\Pi'\Theta$ and (2.1) holds. Also, any (Θ, P) that satisfy $\Omega = \Theta\Pi P\Pi'\Theta$ and (2.1) are unique.*

Moreover, for network balance, the following two vectors in $\mathbb{R}^K$ are natural metrics:

$$d = (\|\theta\|_1)^{-1}\Pi'\Theta\mathbf{1}_n, \qquad g = (\|\theta\|)^{-2}\Pi'\Theta^2\Pi\mathbf{1}_K, \tag{2.2}$$

In the balanced case (e.g., Jin et al. (2021b; 2022)), we usually assume the entries of d and g are at the same order. For our setting, this is not the case.

Next we introduce the null and alternative hypotheses that we consider. Under each hypothesis, we impose the identifiability condition (2.1).

General null model for the DCBM. When $K = 1$ and $h = 1$, P is scalar (say, $P = \alpha$), and $\Omega = \alpha\theta\theta'$ satisfies $\|\theta\|_1 = n$ by (2.1). The expected total degree is $\alpha(\|\theta\|_1^2 - \|\theta\|^2) \sim \alpha\|\theta\|_1^2 = n^2\alpha$ under mild conditions, so we view α as the parameter for network sparsity. In this model, $d = g = 1$.

Alternative model for the DCBM . We assume $K = 2$ and that the sizes of the two communities, $\mathcal{C}_0$ and $\mathcal{C}_1$, are $(n - N)$ and N, respectively. For some positive numbers a, b, c, we have

$$P = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \qquad \text{and} \qquad \Omega(i,j) = \begin{cases} \theta_i\theta_j \cdot a, & \text{if } i,j \in \mathcal{C}_1, \\ \theta_i\theta_j \cdot c, & \text{if } i,j \in \mathcal{C}_0, \\ \theta_i\theta_j \cdot b, & \text{otherwise.} \end{cases} \tag{2.3}$$

In the classical clique detection problem (e.g., Bogerd et al. (2021)), a and c are the baseline probability where two nodes have an edge when both of them are *in* the clique and *outside* the clique, respectively. By (2.1), $a\epsilon + b(1-\epsilon) = b\epsilon + c(1-\epsilon)$ if we write $\epsilon = N/n$. Therefore,

$$b = (c(n-N) - aN)/(n - 2N). \tag{2.4}$$

Note that this is the *direct result* of Sinkhorn's theorem and the parameter calibration we choose, not a condition we choose for technical convenience. Write $d = (d_0, d_1)'$ and $g = (g_0, g_1)'$. It is seen that $d_0 = 1 - d_1$, $g_0 = 1 - g_0$, $d_1 = \|\theta\|_1^{-1}\sum_{i\in\mathcal{C}_1}\theta_i$, and $g_1 = \|\theta\|^{-2}\sum_{i\in\mathcal{C}_1}\theta_i^2$. If all θ_i are at the same order, then $d_1 \asymp g_1 \asymp (N/n)$ and $d_0 \sim g_0 \sim 1$.

Extension . An extension of our alternative hypothesis is that, for the K communities, the sizes of m of them are at the order of N, for an $N \ll n$ and an integer m, $1 \le m < K$, and the sizes of remaining $(K - m)$ are at the order of n. In this case, m entries of d are $O(N/n)$ and other entries are $O(1)$; same for g.

2.2 THE SGNQ TEST: LIMITING NULL, P-VALUE, AND POWER

In the null case, $K = 1$ and we assume $\Omega = \alpha\theta\theta'$, where $\|\theta\|_1 = n$. As $n \to \infty$, both (α, θ) may vary with n. Write $\theta_{\max} = \|\theta\|_\infty$. We assume

$$n\alpha \to \infty, \qquad \text{and} \qquad \alpha\theta_{\max}^2\log(n^2\alpha) \to 0. \tag{2.5}$$

The following theorem is adapted from Jin et al. (2021b) and the proof is omitted.

Theorem 2.1 (Limiting null of the SgnQ statistic)**.** *Suppose the null hypothesis is true and the regularity conditions (2.1) and (2.5) hold. As $n \to \infty$, $\psi_n \to N(0,1)$ in law.*

We have two comments. First, since the DCBM has many parameters (even in the null case), it is not an easy task to find a test statistic with a limiting null that is completely parameter free. For example, if we use the largest eigenvalue of A as the test statistic, it is unclear how to normalize it so to have such a limiting null. Second, since the limiting null is completely explicit, we can approximate the (one-sided) p-value of ψ_n by $\mathbb{P}(N(0,1) \ge \psi_n)$. The p-values are useful in practice, as we show in our numerical experiments.. For example, using a recent data set on the statisticians' publication (Ji et al., 2022), for each author, we can construct an ego network and apply the SgnQ test. We can then use the p-value to measure the co-authorship diversity of the author. Also, in many hierarchical community detection algorithms (which are presumably recursive, aiming to estimate the tree structure of communities), we can use the p-values to determine whether we should further divide a sub-community in each stage of the algorithm (e.g. Ji et al. (2022)).

The power of the SgnQ test hinges on the matrix $\widetilde{\Omega} = \Omega - (\mathbf{1}_n'\Omega\mathbf{1}_n)^{-1}\Omega\mathbf{1}_n\mathbf{1}_n'\Omega$. For d and g in (2.2), let $G \in \mathbb{R}^{K,K}$ be the diagonal matrix $G = \mathrm{diag}(g) = \mathrm{diag}(g_1, g_2, \ldots, g_K)$. By basic algebra,

$$\widetilde{\Omega} = \Theta\Pi\widetilde{P}\Pi'\Theta, \qquad \text{where} \quad \widetilde{P} = P - (d'Pd)^{-1}Pdd'P. \tag{2.6}$$

Let $\tilde{\lambda}_1$ be the largest (in magnitude) eigenvalue of $\widetilde{\Omega}$. Lemma 2.2 is proved in the supplement.

Lemma 2.2. *The rank and trace of the matrix $\widetilde{\Omega}$ are $(K-1)$ and $\|\theta\|^2\mathrm{diag}(\tilde{P})'g$, respectively. When $K = 2$, $\tilde{\lambda}_1 = \mathrm{trace}(\widetilde{\Omega}) = \|\theta\|^2(ac - b^2)(d_0^2g_1 + d_1^2g_0)/(ad_1^2 + 2bd_0d_1 + cd_0^2)$.*

We now discuss the power of the SgnQ test. We focus on the alternative hypothesis in Section 2.1. Let $d = (d_1, d_0)'$ and $g = (g_1, g_0)'$ be as in (2.2), and let $\theta_{\max,0} = \max_{i\in\mathcal{C}_0}\theta_i$ and $\theta_{\max,1} = \max_{i\in\mathcal{C}_1}\theta_i$. Suppose

$$d_1 \asymp g_1 \asymp N/n, \qquad a\theta^2_{\max,1} = O(1), \qquad cn \to \infty, \qquad c\theta^2_{\max,0}\log(n^2c) \to 0. \tag{2.7}$$

These conditions are mild. For example, when θ_i's are at the same order, the first inequality in (2.7) automatically holds, and the other inequalities in (2.7) hold if $a \le C$ for an absolute constant $C > 0$, $cn \to \infty$, and $c\log(n) \to 0$.

Fixing $0 < \kappa < 1$, let $z_\kappa > 0$ be the value such that $\mathbb{P}(N(0,1) \ge z_\kappa) = \kappa$. The level-$\kappa$ SgnQ test rejects the null if and only if $\psi_n \ge z_\kappa$, where ψ_n is as in (1.3). Theorem 2.2 and Corollary 2.1 are proved in the supplement. Recall that our alternative hypothesis is defined in Section 2.1.

Theorem 2.2 (Power of the SgnQ test)**.** *Suppose that* (2.7) *holds, and let* $\kappa \in (0,1)$. *Under the alternative hypothesis, if* $|\tilde{\lambda}_1|/\sqrt{\lambda_1} \to \infty$, *the power of the level-*κ *SgnQ test tends to* 1.

Corollary 2.1. *Suppose the same conditions of Theorem 2.2 hold, and additionally* $\theta_{\max} \le C\theta_{\min}$ *so all* θ_i *are at the same order. In this case,* $\lambda_1 \asymp cn$ *and* $|\tilde{\lambda}_1| \asymp N(a-c)$, *and the power of the level-*κ *SgnQ test tends to* 1 *if* $N(a-c)/\sqrt{cn} \to \infty$.

In Theorem 2.2 and Corollary 2.1, if $\kappa = \kappa_n$ and $\kappa_n \to 0$ slowly enough, then the results continues to hold, and the sum of Type I and Type II errors of the SgnQ test at level-$\kappa_n \to 0$.

The power of the SgnQ test was only studied in the balanced case (Jin et al., 2021b), but our setting is a severely unbalanced case, where the community sizes are at different orders, and the entries of d and g are also at different orders. In the balanced case, the signal-to-noise ratio of SgnQ is governed by $|\lambda_2|/\sqrt{\lambda_1}$, but in our setting, the signal-to-noise ratio is governed by $|\tilde{\lambda}_1|/\sqrt{\lambda_1}$. The proof is also subtly different. Since the entries of P are at different orders, many terms deemed negligible in the power analysis of the balanced case may become non-negligible in the unbalanced case and require careful analysis.

2.3 Comparison with the naive degree-based χ^2-test

Consider a setting where $\Omega = \alpha\Theta\mathbf{1}_n\mathbf{1}_n'\Theta \equiv \alpha\theta\theta'$ under the null and $\Omega = \Theta\Pi P\Pi'\Theta$ under the alternative, and (2.1) holds. When θ is unknown, it is unclear how to apply the χ^2-test: the null case has n unknown parameters $\theta_1, \ldots, \theta_n$, and we need to use the degrees to estimate θ_i first. As a result, the resultant χ^2-statistic may be trivially 0. Therefore, we consider a simpler SBM case where $\theta = \mathbf{1}_n$. In this case, $\Omega = \alpha\mathbf{1}_n\mathbf{1}_n$, and $\Omega = \Pi P\Pi'$ and the null case only has one unknown parameter α. Let y_i be the degree of node i, and let $\hat{\alpha} = [n(n-1)]^{-1}\mathbf{1}_n'A\mathbf{1}_n$. The χ^2-statistic is

$$X_n = \sum_{i=1}^{n}(y_i - n\hat{\alpha})^2/[(n-1)\hat{\alpha}(1-\hat{\alpha})]. \tag{2.8}$$

It is seen that as $n\alpha \to \infty$ and $\alpha \to 0$, $(X_n - n)/\sqrt{2n} \to N(0,1)$ in law. For a fixed level $\kappa \in (0,1)$, consider the χ^2-test that rejects the null if and only if $(X_n - n)/\sqrt{2n} > z_\kappa$. Let $\alpha_0 = n^{-2}(\mathbf{1}_n'\Omega\mathbf{1}_n)$. The power of the χ^2-test hinges on the quantity $(n\alpha_0)^{-1}\|(\Omega\mathbf{1}_n - n\alpha_0)\|^2 = (n\alpha_0)^{-1}\|\Pi Ph - (h'Ph)^{-1}\mathbf{1}_n\|^2 = 0$, if $Ph \propto \mathbf{1}_K$. The next theorem is proved in the supplement.

Theorem 2.3. *Suppose* $\theta = \mathbf{1}_n$ *and* (2.7) *holds. If* $|\tilde{\lambda}_1|/\sqrt{\lambda_1} \to \infty$ *under the alternative hypothesis, the power of the level-*κ *SgnQ test goes to* 1, *while the power of the level-*κ χ^2*-test goes to* κ.

2.4 The statistical lower bound and the optimality of the scan test

For lower bounds, it is standard to consider a random-membership DCBM (Jin et al., 2021b), where $\|\theta\|_1 = n$, P is as in (2.3)-(2.4) and for a number $N \ll n$, $\Pi = [\pi_1, \pi_2, \ldots, \pi_n]'$ satisfies

$$\pi_i = (X_i, 1 - X_i), \qquad \text{where } X_i \text{ are iid Bernoulli}(\varepsilon) \text{ with } \varepsilon = N/n. \tag{2.9}$$

Theorem 2.4 (Statistical lower bound)**.** *Consider the null and alternative hypotheses of Section 2.1, and assume that* (2.9) *is satisfied,* $\theta_{\max} \le C\theta_{\min}$ *and* $Nc/\log n \to \infty$. *If* $\sqrt{N}(a-c)/\sqrt{c} \to 0$, *then for any test, the sum of the type-I and type-II errors tends to* 1.

To show the tightness of this lower bound, we introduce the signed scan test, by adapting the idea in Arias-Castro & Verzelen (2014) from the SBM case to the DCBM case. Unlike the SgnQ test and the χ^2-test, signed scan test is not a polynomial time test, but it provides sharper upper bounds. Let $\hat{\eta}$ be the same as in (1.3). For any subset $S \subset \{1, 2, \ldots, n\}$, let $\mathbf{1}_S \in \mathbb{R}^n$ be the vector whose ith coordinate is $1\{i \in S\}$. Define the signed scan statistic

$$\phi_{sc} = \max_{S \subset \{1,2,\ldots,n\}: |S|=N} \mathbf{1}_S'(A - \hat{\eta}\hat{\eta}')\mathbf{1}_S. \tag{2.10}$$

Theorem 2.5 (Tightness of the statistical lower bound)**.** *Consider the signed scan test* (2.10) *that rejects the null hypothesis if $\phi_{sc} > t_n$. Under the assumptions of Theorem 2.4, if $\sqrt{N}(a-c)/\sqrt{c\log(n)} \to \infty$, then there exists a sequence t_n such that the sum of type I and type II errors of the signed scan test tends to* 0.

By Theorems 2.4-2.5 and Corollary 2.1, the two hypotheses are asymptotically indistinguishable if $\sqrt{N}(a-c)/\sqrt{c} \to 0$, and are asymptotically distinguishable by the SgnQ test if $N(a-c)/\sqrt{cn} \to \infty$. Therefore, the lower bound is sharp, up to log-factors, and the signed scan test is nearly optimal. Unfortunately, the signed scan test is not polynomial-time computable. Does there exist a polynomial-time computable test that is optimal? We address this in the next section.

2.5 The computational lower bound

Consider the same hypothesis pair as in Section 2.4, where $K = 2$, P is as in (2.3)-(2.4), and Π is as in (2.9). For simplicity, we only consider SBM, i.e., $\theta_i \equiv 1$. The low-degree polynomials argument emerges recently as a major tool to predicting the average-case computational barriers in a wide range of high-dimensional problems (Hopkins & Steurer, 2017; Hopkins et al., 2017). Many powerful methods, such as spectral algorithms and approximate message passing, can be formulated as functions of the input data, where the functions are polynomials with degree at most logarithm of the problem dimension. In comparison to many other schemes of developing computational lower barriers, the low-degree polynomial method yields the same threshold for various average-case hardness problems, such as community detection in the SBM (Hopkins & Steurer, 2017) and (hyper)-planted clique detection (Hopkins, 2018; Luo & Zhang, 2022). The foundation of the low-degree polynomial argument is the following *low-degree polynomial conjecture* (Hopkins et al., 2017) :

Conjecture 2.1 (Adapted from Kunisky et al. (2019))**.** *Let $\mathbb{P}_n$ and $\mathbb{Q}_n$ denote a sequence of probability measures with sample space $\mathbb{R}^{n^k}$ where $k = O(1)$. Suppose that every polynomial f of degree $O(\log n)$ with $\mathbb{E}_{\mathbb{Q}_n} f^2 = 1$ is bounded under $\mathbb{P}_n$ with high probability as $n \to \infty$ and that some further regularity conditions hold. Then there is no polynomial-time test distinguishing $\mathbb{P}_n$ from $\mathbb{Q}_n$ with type I and type II error tending to* 0 *as $n \to \infty$.*

We refer to Hopkins (2018) for a precise statement of this conjecture's required regularity conditions. The low-degree polynomial computational lower bound for our testing problem is as follows.

Theorem 2.6 (Computational lower bound)**.** *Consider the null and alternative hypotheses in Section 2.1, and assume $\theta_i \equiv 1$ and* (2.9) *holds. As $n \to \infty$, assume $c < a$, $c < 1-\delta$ for constant $\delta > 0$, $N < n/3$, $D = O(\log n)$, and $\limsup_{n\to\infty} \left\{ \left(\log_n \frac{N}{\sqrt{n}} + \log_n \frac{a-c}{\sqrt{c}} \right) \vee \left(\sqrt{D/2 - 1} \log_n \frac{a-c}{\sqrt{c}} \right) \right\} < 0$. For any series of degree-D polynomials $\phi_n : A \to \mathbb{R}$, whenever $\mathbb{E}_{H_0}\phi_n(A) = 0$, $Var_{H_0}(\phi_n(A)) = 1$, we must have $\mathbb{E}_{H_1}\phi_n(A) = o(1)$. This implies if Conjecture 2.1 is true, there is no consistent polynomial-time test for this problem.*

By Theorem 2.6, if both $(a-c)/\sqrt{c} \lesssim 1$ and $N(a-c)/\sqrt{cn} \to 0$, the testing problem is computationally infeasible. The region where the testing problem is statistically possible but the SgnQ test loses power corresponds to $N(a-c)/\sqrt{cn} \to 0$. If $N \gtrsim \sqrt{n}$, Theorem 2.6 already implies that this is the computationally infeasible region; in other words, SgnQ achieves the CLB and is optimal. If $N = o(\sqrt{n})$, SgnQ solves the detection problem only when $(a-c)/\sqrt{c} \gg n^{1/2}$, i.e. when the node-wise SNR is strong. We discuss the case of moderate node-wise SNR in the next subsection.

2.6 The power of EST, and discussions of the tightness of CLB

When $N = o(\sqrt{n})$ and $(a-c)/\sqrt{c} \to \infty$ both hold, the upper bound by SgnQ does not match with the CLB. It is unclear whether the CLB is tight. To investigate the CLB in this regime, we consider

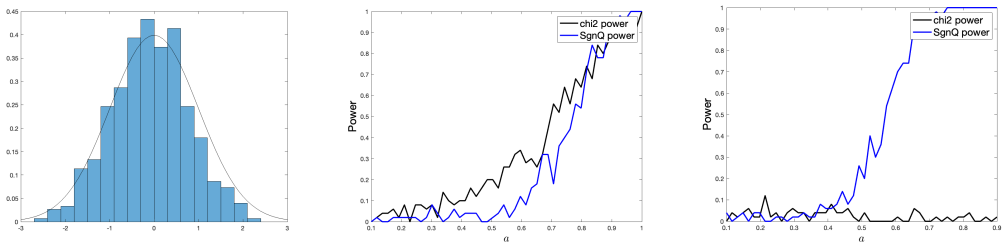

Figure 1: Left: Null distribution of SgnQ ($n = 500$). Middle and right: Power comparison of SgnQ and χ^2 ($n = 100$, $N = 10$, 50 repetitions). We consider a 2-community SBM with $P_{11} = a$, $P_{22} = 0.1$, $P_{12} = 0.1$ (middle plot) and $P_{12} = \frac{an-(a+0.1)N}{n}$ (right plot, the case of degree matching).

other possible polynomial-time tests. The economic scan test (EST) is one candidate. Given fixed positive integers v and e, the EST statistic is defined to be $\phi_{EST}^{(v)} \equiv \sup_{|S|\le v} \sum_{i,j\in S} A_{ij}$, and the EST is defined to reject if and only if $\phi_{EST}^{(v)} \ge e$. EST can be computed in time $O(n^v)$, which is polynomial time. For simplicity, we consider the SBM, i.e. where $\theta = \mathbf{1}_n$, and a specific setting of parameters for the null and alternative hypotheses.

Theorem 2.7 (Power of EST)**.** *Suppose $\beta \in [1/2, 1)$ and $0 < \omega < \delta < 1$ are fixed constants. Under the alternative, suppose $\theta = \mathbf{1}_n$, (2.9) holds, $N = n^{1-\beta}$, $a = n^{-\omega}$, and $c = n^{-\delta}$. Under the null, suppose $\theta = \mathbf{1}_n$ and $\alpha = a(N/n) + b(1 - N/n)$. If $\omega/(1-\beta) < \delta$, the sum of type I and type II errors of the EST with v and e satisfying $\omega/(1-\beta) < v/e < \delta$ tends to 0.*

Theorem 2.7 follows from standard results in probabilistic combinatorics (Alon & Spencer, 2016). It is conjectured in Bhaskara et al. (2010) that EST attains the CLB in the Erdös-Renyi setting considered by Arias-Castro & Verzelen (2014); Verzelen & Arias-Castro (2015). This suggests that the CLB in Theorem 2.6 is likely not tight when $N = o(\sqrt{n})$ and $(a-c)/\sqrt{c} \to \infty$. However, this is not because our inequalities in proving the CLB are loose. A possible reason is that the prediction from the low-degree polynomial conjecture does not provide a tight bound. It remains an open question whether other computational infeasibility frameworks provide a tight CLB in our problem.

2.7 The phase transition

We summarize our results in terms of phase transitions. Consider the null and alternative hypotheses from Section 2.1. For illustration purposes, we fix constants $\beta \in (0, 1)$ and $\gamma \in \mathbb{R}$ and assume that $N = n^{1-\beta}$ and $(a-c)/\sqrt{c} = n^{-\gamma}$. In the two-dimensional space of (γ, β), the region of $\beta > 1/2$ and $\beta < 1/2$ corresponds to that the size of the small community is $\gg \sqrt{n}$ and $o(\sqrt{n})$, respectively, and the regions of $\gamma > 0$, $-1/2 < \gamma < 0$ and $\gamma < -1/2$ correspond to 'weak node-wise signal', 'moderate node-wise signal,' and the 'strong node-wise signal', respectively. See Figure 2. By our results in Section 2.4, the testing problem is statistically impossible if $\beta + 2\gamma > 1$ (orange region). By our results in Section 2.2, SgnQ has a full power if $\beta + \gamma < 1/2$ (blue region). Our results in Section 2.5 state that the testing problem is computationally infeasible if both $\gamma > 0$ and $\beta + \gamma > 1/2$ (green and orange regions). Combining these results, when $\beta < 1/2$, we have a complete understanding of the LB and CLB.

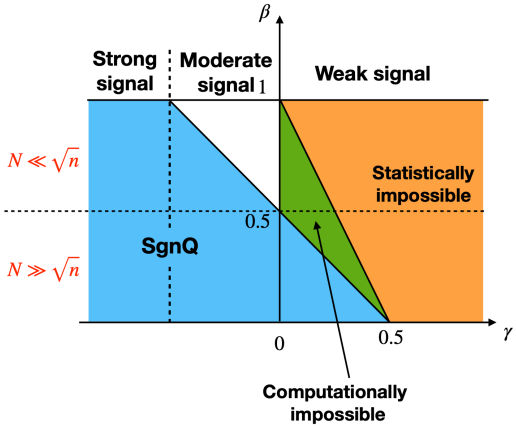

Figure 2: Phase diagram ($(a-c)/\sqrt{c} = n^{-\gamma}$ and $N = n^{1-\beta}$).

3 Numerical results

Simulations. First in Figure 1 (left panel) we demonstrate the asymptotic normality of SgnQ under a null of the form $\Omega = \theta\theta'$, where θ_i are i.i.d. generated from $\mathrm{Pareto}(4, 0.375)$. In this setting,

although the degree heterogeneity is severe, SgnQ when properly standardized is approximately standard normal under the null distribution. Second in Figure 1 we compare the power of SgnQ in an asymmetric and symmetric SBM model. As our theory predicts, both tests are powerful when degrees are not calibrated in each model, but only SgnQ is powerful in the symmetric case. We also compare the power of SgnQ with the scan test to provide numerical evidence of a statistical-to-computational gap. To save space, we relegate these experiments to the supplementary material.

Real data: In this section we demonstrate the effectiveness of SgnQ in detecting small communities in coauthorship networks studied in Ji et al. (2022). In Example 1, we consider the personalized network of Raymond Carroll, whose nodes consist of his coauthors for papers in a set of 36 statistics journals from the time period 1975 – 2015. An edge is placed between two coauthors if they wrote a paper in this set of journals during the specified time period. The SgnQ p-value for Carroll's personalized network G_{Carroll} is 0.02, which suggests the presence of more than one community. In Ji et al. (2022), the authors identify a small cluster of coauthors from a collaboration with the National Cancer Institute. We applied the SCORE community detection module with $K = 2$ (e.g. Ke & Jin (2022)) and obtained a larger community G^0_{Carroll} of size 218 and a smaller community G^1_{Carroll} of size 17. Precisely, we removed Carroll from his network, applied SCORE on the remaining giant component, and defined G^0_{Carroll} to be the complement of the smaller community. The SgnQ p-values in the table below suggest that both G^0_{Carroll} and G^1_{Carroll} are tightly clustered. Refer to the supplement for a visualization of Carroll's network and its smaller community labeled by author names. In Example 2, we consider three different coauthorship networks G_{old}, G_{recent}, and G_{new} corresponding to time periods (i) 1975-1997, (ii) 1995-2007, and (iii) 2005-2015 for the journals AoS, Bka, JASA, and JRSSB. Nodes are given by authors, and an edge is placed between two authors if they coauthored at least one paper in one of the aforementioned journals during the corresponding time period. For each network, we perform a similar procedure as in the first example. First we compute the SgnQ p-value, which turns out to be ≈ 0 (up to 16 digits of precision) for all networks. For each $i \in \{\mathsf{old}, \mathsf{recent}, \mathsf{new}\}$, we apply SCORE with $K = 2$ to G_i and compute the SgnQ p-value on both resulting communities, let us call them G^0_i and G^1_i. We refer to the table below for the results. We see that for G_{old} and G_{recent}, SCORE with $K = 2$ extracts a small community. The SgnQ p-value further supports the hypothesis that this small community is well-connected. In the third network, SCORE splits G_{new} into two roughly equally sized pieces whose SgnQ p-values suggests that they can be further split into smaller subcommunities.

Example	Network	Size	SgnQ p-value	Communities	Sizes	SgnQ p-values
1	G_{Carroll}	235	0.02	$(G^0_{\mathsf{Carroll}}, G^1_{\mathsf{Carroll}})$	(218, 17)	(0.134, 0.682)
2	G_{old}	2647	0	$(G^0_{\mathsf{old}}, G^1_{\mathsf{old}})$	(2586, 61)	(0, 0.700)
	G_{recent}	2554	0	$(G^0_{\mathsf{recent}}, G^1_{\mathsf{recent}})$	(2540,14)	(0, 0.759)
	G_{new}	2920	0	$(G^0_{\mathsf{new}}, G^1_{\mathsf{new}})$	(1685,1235)	(0, 0)

Discussions: Global testing is a fundamental problem and often the starting point of a long line research. For example, in the literature of Gaussian models, certain methods started as a global testing tool, but later grew into tools for variable selection, classification, and clustering and motivated many researches (e.g., Donoho & Jin (2004; 2015)). The SgnQ test may also motivate tools for many other problems, such as estimating the locations of the clique and clustering more generally. For example, in Jin et al. (2022), the SgnQ test motivated a tool for estimating the number of communities (see also Ma et al. (2021)). SgnQ is also extendable to clique detection in a tensor (Yuan et al., 2021; Jin et al., 2021a) and for network change point detection. The LB and CLB we obtain in this paper are also useful for studying other problems, such as clique estimation. If you cannot tell whether there is a clique in the network, then it is impossible to estimate the clique. Therefore, the LB and CLB are also valid for the clique estimation problem (Alon et al., 1998; Ron & Feige, 2010).

The limiting distribution of SgnQ is $N(0, 1)$. This is not easy to achieve if we use other testing ideas, such as the leading eigenvalues of the adjacency matrix: the limiting distribution depends on many unknown parameters and it is hard to normalize (Liu et al., 2019). The p-value of the SgnQ test is easy to approximate and also useful in applications. For example, we can use it to measure the research diversity of a given author. Consider the ego sub-network of an author in a large co-authorship or citation network. A smaller p-value suggests that the ego network has more than 1 communities and so has more diverse interests. The p-values can also be useful as a stopping criterion in hierarchical community detection modules.

REFERENCES

Noga Alon and Joel H Spencer. *The probabilistic method*. John Wiley & Sons, 2016.

Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. *Random Structures & Algorithms*, 13(3-4):457–466, 1998.

Ery Arias-Castro and Nicolas Verzelen. Community detection in dense random networks. *The Annals of Statistics*, 42(3):940–969, 2014.

Aditya Bhaskara, Moses Charikar, Eden Chlamtac, Uriel Feige, and Aravindan Vijayaraghavan. Detecting high log-densities: an o (n $1/4$) approximation for densest k-subgraph. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 201–210, 2010.

Kay Bogerd, Rui M Castro, Remco van der Hofstad, and Nicolas Verzelen. Detecting a planted community in an inhomogeneous random graph. *Bernoulli*, 27(2):1159–1188, 2021.

David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

David Donoho and Jiashun Jin. Special invited paper: Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science*, pp. 1–25, 2015.

Chao Gao and John Lafferty. Testing for global network structure using small subgraph statistics. *arXiv:1710.00862*, 2017.

Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, pp. 68–75. Springer, 2011.

Oded Goldreich, Shari Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM (JACM)*, 45(4):653–750, 1998.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Computational lower bounds for community detection on random graphs. In *Conference on Learning Theory*, pp. 899–928. PMLR, 2015.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Soc. Netw.*, 5(2):109 – 137, 1983.

Samuel B Hopkins and David Steurer. Efficient bayesian estimation from few samples: community detection and related problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 379–390. IEEE, 2017.

Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 720–731. IEEE, 2017.

Samuel Brink Klevit Hopkins. Statistical inference and the sum of squares method. 2018.

Pengsheng Ji, Jiashun Jin, Zheng Tracy Ke, and Wanshan Li. Co-citation and co-authorship networks for statisticians (with discussions). *J. Bus. Econ. Statist.*, 40(2), 2022.

Jiashun Jin. Fast community detection by SCORE. *Ann. Statist.*, 43(1):57–89, 2015.

Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. Netwok global testing by counting graphlets. *35th International Conference on Machine Learning*, 2018.

Jiashun Jin, Zheng Tracy Ke, and Jiajun Liang. Sharp impossibility results for hyper-graph testing. *Advances in Neural Information Processing Systems*, 34:7220–7231, 2021a.

Jiashun Jin, Zheng Tracy Ke, and S. Luo. Optimal adaptivity of signed-polygon statistics for network testing. *Ann. Statist.*, 49(6):3408–3433, 2021b.

Jiashun Jin, Zheng Tracy Ke, Shengming Luo, and Minzhe Wang. Optimal estimation of the number of network communities. *J. Amer. Statist. Assoc.*, pp. 1–16, 2022.

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

Zheng Tracy Ke and Jiashun Jin. The score normalization, especially for highly heterogeneous network and text data. *arXiv preprint arXiv:2204.11097*, 2022.

Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. *arXiv preprint arXiv:1907.11636*, 2019.

Yan Liu, Zhiqiang Hou, Zhigang Yao, Zhidong Bai, Jiang Hu, and Shurong Zheng. Community detection based on the ℓ_∞ convergence of eigenvectors in dcbm. *arXiv preprint arXiv:1906.06713*, 2019.

Yuetian Luo and Anru R Zhang. Tensor clustering with planted structures: Statistical optimality and computational limits. *The Annals of Statistics*, 50(1):584–613, 2022.

Shujie Ma, Liangjun Su, and Yichong Zhang. Determining the number of communities in degree-corrected stochastic block models. *Journal of Machine Learning Research*, 22(69):1–63, 2021.

Dorit Ron and Uriel Feige. Finding hidden cliques in linear time. *Discrete Mathematics & Theoretical Computer Science*, 2010.

Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *Proc. Amer. Math. Soc.*, (2):195–198, 1974.

Nicolas Verzelen and Ery Arias-Castro. Community detection in sparse random networks. *The Annals of Applied Probability*, 25(6):3465–3510, 2015.

Mingao Yuan, Ruiqi Liu, Yang Feng, and Zuofeng Shang. Testing community structures for hypergraphs. *Ann. Statist. (to appear)*, 2021.

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.*, 40(4):2266–2292, 2012.