Optimal Estimation of the Number of Communities

Jiashun Jin*
Department of Statistics, Carnegie Mellon University
and
Zheng Tracy Ke
Department of Statistics, Harvard University
and
Shengming Luo
Department of Statistics, Carnegie Mellon University
and
Minzhe Wang
Department of Statistics, University of Chicago

January 25, 2022

Abstract

In network analysis, how to estimate the number of communities K is a fundamental problem. We consider a broad setting where we allow severe degree heterogeneity and a wide range of sparsity levels, and propose Stepwise Goodness-of-Fit (StGoF) as a new approach. This is a stepwise algorithm, where for $m = 1, 2, \ldots$, we alternately use a community detection step and a goodness-of-fit (GoF) step. We adapt SCORE [19] for community detection, and propose a new GoF metric. We show that at step m, the GoF metric diverges to ∞ in probability for all $m < K$ and converges to $N(0, 1)$ if $m = K$. This gives rise to a consistent estimate for K. Also, we discover the right way to define the signal-to-noise ratio (SNR) for our problem and show that consistent estimates for K do not exist if SNR $\to 0$, and StGoF is uniformly consistent for K if SNR $\to \infty$. Therefore, StGoF achieves the optimal phase transition.

Similar stepwise methods (e.g., [38, 34]) are known to face analytical challenges. We overcome the challenges by using a different stepwise scheme in StGoF and by deriving sharp results that are not available before. The key to our analysis is to show that SCORE has the *Non-Splitting Property (NSP)*. Primarily due to a non-tractable rotation of eigenvectors dictated by the Davis-Kahan $\sin(\theta)$ theorem, the NSP is non-trivial to prove and requires new techniques we develop.

*JJ and SL gratefully acknowledge the support of the NSF grant *DMS-2015469*. ZK gratefully acknowledges the support of the NSF CAREER grant *DMS-1943902*.

Keywords: Community detection, k-means, lower bound, Non-Splitting Property (NSP), over-fitting, under-fitting

1 Introduction

Suppose A is the adjacency matrix for a symmetric and connected network with n nodes:

$$A_{ij} = \begin{cases} 1, & \text{if node } i \text{ and node } j \text{ have an edge,} \\ 0, & \text{otherwise,} \end{cases} \qquad 1 \le i \ne j \le n. \tag{1.1}$$

As a convention, self-edges are not allowed so all the diagonal entries of A are 0. As usual, we assume the network has K (unknown) communities $\mathcal{N}_1$, $\mathcal{N}_2$, $\ldots, \mathcal{N}_K$. Similar to that of a cluster in multivariate analysis, the precise meaning a community is hard to formalize, but frequently and intuitively, communities in a network are groups of nodes that have more edges within than between (e.g., [42]).

Our primary goal is to estimate K. This is a fundamental problem in network analysis: In many recent approaches, K is assumed as known a priori (e.g., [40, 39, 16] on community detection, [22, 9] on mixed-membership estimation, [31, 18, 41] on dynamic networks, and [43, 3, 15] on network regression analysis). Unfortunately, K is rarely known in applications, so the performance of these approaches hinges on how well we can estimate K.

Real world networks have several noteworthy features. First, a network may have severe degree heterogeneity. Take the Polblog network in Table 1 for example. The maximum degree is 351 and the minimum degree is 1. Second, the network sparsity (e.g., measured by the average degree) may range significantly from one network to another. Last, frequently, the desired community structure is masked by strong noise, and the signal-to-noise ratio (SNR) is usually relatively small. Motivated by these features, we adopt the widely-used degree-corrected block model (DCBM) [25]. Recall that the network has K communities $\mathcal{N}_1$, $\mathcal{N}_2$, $\ldots, \mathcal{N}_K$. For each $1 \le i \le n$, we encode the community label of node i by a vector $\pi_i \in \mathbb{R}^K$ where for $i \in \mathcal{N}_k$, $\pi_i(k) = 1$ and $\pi_i(m) = 0$ for $m \ne k$. Moreover, for a $K \times K$ symmetric nonnegative matrix P which models the community structure and positive parameters $\theta_1, \theta_2, \ldots, \theta_n$ which model the degree heterogeneity, we assume the upper triangular entries of A are independent Bernoulli variables satisfying

$$\mathbb{P}(A_{ij} = 1) = \theta_i \theta_j \cdot \pi_i' P \pi_j \equiv \Omega_{ij}, \qquad 1 \le i < j \le n, \tag{1.2}$$

where Ω denotes the matrix $\Theta \Pi P \Pi' \Theta$, with Θ being the $n \times n$ diagonal matrix $\mathrm{diag}(\theta_1, \ldots, \theta_n)$

and Π being the $n \times K$ matrix $[\pi_1, \pi_2, \ldots, \pi_n]'$. For identifiability, we assume

$$P \text{ is non-singular and all diagonal entries of } P \text{ are 1.} \tag{1.3}$$

Write for short $\mathrm{diag}(\Omega) = \mathrm{diag}(\Omega_{11}, \Omega_{22}, \ldots, \Omega_{nn})$, and let W be the matrix where for $1 \leq i, j \leq n$, $W_{ij} = A_{ij} - \Omega_{ij}$ if $i \neq j$ and $W_{ij} = 0$ otherwise. In matrix form, we have

$$A = \Omega - \mathrm{diag}(\Omega) + W, \qquad \text{where we recall } \Omega = \Theta\Pi P \Pi' \Theta. \tag{1.4}$$

When $\theta_1 = \theta_2 = \ldots = \theta_n$, DCBM reduces to the stochastic block model (SBM).

We let n be the driving asymptotic parameter, and allow (Θ, Π, P) to depend on n, so DCBM is broad enough to capture the three features aforementioned. In detail, let $\theta = (\theta_1, \theta_2, \ldots, \theta_n)'$, $\theta_{max} = \max\{\theta_1, \ldots, \theta_n\}$, and $\theta_{min} = \min\{\theta_1, \ldots, \theta_n\}$. First, a reasonable metric for the degree heterogeneity is $\theta_{max}/\theta_{min}$, so to allow severe degree heterogeneity, we prefer not to put an artificial upper bound on $\theta_{max}/\theta_{min}$. Second, a reasonable metric for network sparsity is $\|\theta\|$ (e.g., see [23, 19]). [1] To cover all sparsity levels of interest, and especially the very sparse case (e.g., $\theta_i = O(\sqrt{\log(n)/n})$ for all $1 \leq i \leq n$) and the very dense case (e.g., $\theta_1 = O(1)$ for all $1 \leq i \leq n$), we assume ($C > 0$ is a constant)

$$C\sqrt{\log(n)} \leq \|\theta\| \leq C\sqrt{n}. \tag{1.5}$$

Last, let $\lambda_1, \lambda_2, \ldots, \lambda_K$ be the K nonzero eigenvalues of Ω, arranged in the descending order of magnitudes. We will soon see that the signal strength and noise level in our setting are captured by $|\lambda_K|$ and $\|W\|$, respectively, where under mild conditions,

$$\|W\| = \text{a multi-log}(n) \text{ term} \cdot \sqrt{\lambda_1} \text{ with high probability, where } \lambda_1 \asymp \|\theta\|^2. \tag{1.6}$$

Therefore, a reasonable metric for the signal to noise ratio (SNR) is $|\lambda_K|/\sqrt{\lambda_1}$ (see Section 3 for more discussion). We consider two extreme cases (assuming $n \to \infty$).

- *Strong signal case.* $|\lambda_1|, |\lambda_2|, \ldots, |\lambda_K|$ are at the same magnitude, and so SNR $\asymp \sqrt{\lambda_1}$.
- *Weak signal case.* $|\lambda_K|/\sqrt{\lambda_1}$ is much smaller than $\sqrt{\lambda_1}$ and grows to ∞ slowly.

[1] An appropriate measure for sparsity is $\|\Omega\|$ (e.g.,[23]). In (1.3), we assume all diagonal entries of P are 1, so if K is finite and some regularity conditions hold, $\|\Omega\| \asymp \|\theta\|^2$. Also, d_i (degree of node i) is at the order of $\theta_i\|\theta\|_1$, which is $O(n\theta_i^2)$ if all θ_i are at the same order. Therefore, the range of interest for θ_i is between $1/\sqrt{n}$ and 1, up to some logarithmic factors (e.g., $\log(n)$).

For example, in a weak signal case, we may have $\lambda_1 = O(\sqrt{n})$ and SNR $= \log\log(n)$ and $\lambda_1 = \sqrt{n}$. Section 3.3 suggests that when SNR $= o(1)$, consistent estimate for K does not exist, so the weak signal case is very challenging. Motivated by the above observations, it is desirable to find a consistent estimate for K that satisfies the following requirements.

- (R1) Allow severe degree heterogeneity (i.e., no artificial bound on $\theta_{max}/\theta_{min}$).
- (R2) Optimally adaptive to all sparsity levels of interest (e.g., see (1.5)).
- (R3) *Attain the information lower bound*. Consistent for both the strong signal case where SNR is large and the weak signal case where SNR may be as small as $\log\log(n)$.

Example 1. A frequently considered DCBM is to assume $P = P_0$ and $\theta_i \asymp \sqrt{\alpha_n}$ for all $1 \leq i \leq n$, where $\alpha_n > 0$ is a scaling parameter and P_0 is a fixed matrix. It is seen that $\lambda_1, \ldots, \lambda_K$ are at the same order, so the model only considers the strong signal case.

Example 2. Let $e_1, \ldots, e_K$ be the standard basis vectors of $\mathbb{R}^K$. Fix a positive vector $\theta \in \mathbb{R}^n$ and $b_n \in (0, 1)$. Consider a DCBM where each community has n/K nodes, and $P = (1-b_n)I_K + b_n 1_K 1_K'$. Here, $(1-b_n)$ measures the "dis-similarity" of different communities. By basic algebra, $\lambda_1 \asymp \|\theta\|^2$, $\lambda_2 = ... = \lambda_K \asymp \|\theta\|^2(1-b_n)$, and SNR $\asymp \|\theta\|(1-b_n)$; moreover, $\|\theta\| = O(\sqrt{\log(n)})$ in the very sparse case, and $\|\theta\| = O(\sqrt{n})$ in the dense case. When $b_n \leq c_0$ for a constant $c_0 < 1$, $|\lambda_K| \geq C|\lambda_1|$ and SNR $\asymp \|\theta\|$; we are in the strong signal case if $\|\theta\| \geq n^a$ for a constant $a > 0$. When $b_n = 1 + o(1)$ and $\|\theta\|(1-b_n) = \log\log(n)$ (say), SNR $\asymp \log\log(n)$ and we are in the weak signal case.

Example 3. An SBM can be identifiable even if P is singular (e.g., [37]). However, a DCBM can be non-identifiable if P is singular. For example, consider an SBM with parameters $(\widetilde{\Pi}, \widetilde{P})$ where $\widetilde{P} \in \mathbb{R}^{2,2}$, $\widetilde{P}_{11} = a$, $\widetilde{P}_{22} = c$, $\widetilde{P}_{12} = \widetilde{P}_{21} = b$, and $ac = b^2$ (so the rank of $\widetilde{P}$ is 1). The model is an identifiable SBM with two communities. But if we treat it with a DCBM with parameters (K, Θ, Π, P), then we can either take $(K, \Theta, \Pi, P) = (2, I_n, \widetilde{\Pi}, \widetilde{P})$, or take $(K, \Theta, \Pi, P) = (1, \Theta, \widetilde{\Pi}, 1)$, so it is not identifiable. Here $\Theta = \mathrm{diag}(\theta_1, \ldots, \theta_n)$ and $\theta_i = \sqrt{a}$ if i is in community 1 and $\theta_i = \sqrt{c}$ if i is in community 2, $1 \leq i \leq n$.

1.1 Literature review and our contributions

Exiting approaches for estimating K can be roughly divided into the spectral approaches, cross validation approaches, penalization approaches, and likelihood ratio approaches.

For spectral approaches, Le and Levina [28] proposed to estimate K using the eigenvalues of the non-backtracking matrix or Bethe Hessian matrix. The approach uses interesting ideas from graph theory, but unfortunately, it requires relatively strong conditions for consistency. For example, their Theorem 4.1 only considers the very sparse SBM model where $\theta_1 = \theta_2 = \ldots = \theta_n = 1/\sqrt{n}$ and $P = P_0$ for a fixed matrix P_0. Liu *et al.* [33] proposed to estimate K using the scree plot with careful theoretical justification, but the approach is unsatisfactory for networks with severe degree heterogeneity, for it is hard to derive a sharp bound for the spectral norm of the noise matrix W (e.g., [19]). Therefore, their approach requires the condition of $\theta_{max} \leq C\theta_{min}$. The paper also assumed $\|\theta\| = O(\sqrt{n})$ so it did not address the settings of sparse networks (e.g., see (1.5)). For cross-validation approaches, we have [4, 30], and among the penalization approaches, we have [36, 5, 27], where K is estimated by the integer that optimizes some objective functions. For example, Salda *et al.* [36] used a BIC-type objective function and [5, 27] used an objective function of the Bayesian model selection flavor. However, these methods did not provide explicit theoretical guarantee on consistency (though a partial result was established in [30], which stated that under SBM, the proposed estimator $\widehat{K}$ is no greater than K with high probability).

For likelihood ratio approaches, Wang and Bickel [38] proposed to estimate K by solving a BIC type optimization problem, where the objective function is the sum of log-likelihood and model complexity. The major challenge is that the likelihood is the sum of exponentially many terms and is hard to compute. In a remarkable paper, Ma *et al.* [34] extended the idea of [38] by proposing a new approach that is computationally more feasible.

On a high level, we can recast their methods as a *stepwise testing or sequential testing* algorithm. Consider a stepwise testing scheme where for $m = 1, 2, \ldots$, they construct a test statistic $\ell_n^{(m)}$ (e.g. log-likelihood) assuming m is the correct number of communities. They estimate K as the smallest m such that the pairwise log-likelihood ratio $(\ell_n^{(m+1)} - \ell_n^{(m)})$ falls below a *threshold*. Call $m < K$, $m = K$, and $m > K$ the *under-fitting*, *null*, and *over-fitting* cases, respectively. As mentioned in [38, 34], such an approach faces a two-fold challenge. First, one has to analyze $\ell_n^{(m)}$ for both the under-fitting case and the over-fitting case, but there are no efficient technical tools to address either case. Second, it is hard to derive sharp results on the limiting distribution of $\ell_n^{(m+1)} - \ell_n^{(m)}$ in the null case, and so it is unclear how to pin down the threshold. Ma *et al.* [34] (see also [38]) made interesting progress but

unfortunately the problems are not resolved satisfactorily. For example, they require hard-to-check conditions on both the under-fitting and over-fitting cases. Also, it is unclear whether their results are sharp in the over-fitting case and how to standardize $\ell_n^{(m+1)} - \ell_n^{(m)}$ in the under-fitting case as the variance term is unknown (so it unclear how to pin down the threshold). Most importantly, both papers focus on the setting in Example 1 (see above), where severe degree heterogeneity is not allowed and they only consider the strong signal case.

We propose *Stepwise Goodness-of-Fit (StGoF)* as a new approach to estimating K. Our idea follows a different vein, and is different both in the statistics we developed and in the stepwise scheme we use. In detail, for $m = 1, 2, \ldots$, StGoF alternately uses a community detection sub-step (where we apply SCORE [19] assuming m is the correct number of communities) and a Goodness-of-Fit (GoF) sub-step. We propose a new GoF approach and let $\psi_n^{(m)}$ be the GoF test statistic in step m. Assuming SNR $\to \infty$, we show that

$$\psi_n^{(m)} \begin{cases} \to N(0,1), & \text{when } m = K \text{ (null case)}, \\ \to \infty \text{ in probability}, & \text{when } 1 \le m < K \text{ (under-fitting case)}. \end{cases} \tag{1.7}$$

For a properly chosen threshold t, define the StGoF estimate by $\widehat{K} = \min_m\{\psi_n^{(m)} \le t\}$. By (1.7), $\widehat{K}$ is consistent. Now, first, (1.7) shows that $N(0,1)$ is the limiting null. Such an explicit limiting null is crucial in pinning down the threshold t. Second, a noteworthy advantage of StGoF is that, we do not need to analyze the over-fitting case to prove the consistency of $\widehat{K}$. In comparison, if we follow the approaches by [34, 38] and similarly define $\widehat{K}$ by $\min_m\{\ell_n^{(m+1)} - \ell_n^{(m)} \le t\}$, then we have to derive the limiting distribution for $\ell_n^{(m+1)} - \ell_n^{(m)}$ with $m = K$, which is an over-fitting case. In this case, how to derive tight bounds is an open problem (even if the limiting distribution of $\ell_n^{(m+1)} - \ell_n^{(m)}$ can be derived theoretically, it contains unknown parameters, so it is hard to pin down the threshold t). For these reasons, it is unclear how to derive sharp results with these approaches.

Fortunately, sharp results are possible if we use the StGoF approach. In Section 3.3, we show that when SNR $\to 0$, consistent estimates for K do not exist. Therefore, our consistency result above is sharp in terms of the rate of SNR, so StGoF achieves the optimal phase transition, in a broad setting (where we allow degree heterogeneity, flexible sparsity levels, and weak signals). The phase transition is a well-known optimality framework. It is related to the minimax framework but can be frequently more informative [7].

Compared with the approaches in [34, 38], (a) they focused on more restricted settings, with either strong signals, or strong eigen-gap conditions, or the more specific SBM model, (b) they did not have an explicit limiting null, and (c) they have to analyze the over-fitting case but it remains an open problem to derive sharp bounds. For these reasons, it is unclear whether they are able to achieve the optimal phase transition.

To prove (1.7), the key is to show that when $m \leq K$, SCORE has the so-called *Non-Splitting Property (NSP)*, meaning that with high probability all nodes in each (true) community are always clustered together. The proof of NSP is non-trivial. It depends on the row-wise distances of the matrix Ξ consisting of the first m columns of $[\xi_1, \ldots, \xi_K]\Gamma$, where ξ_k is the k-th eigenvector of Ω and Γ is an orthogonal matrix dictated by the Davis-Kahan $\sin(\theta)$ theorem [6]. Γ is data dependent and hard to track, and when it ranges, the row-wise distances of Ξ are the same if $m = K$ but may vary significantly if $m < K$. This is why SCORE is much harder to study in the under-fitting case than in the null case. To overcome the challenge, we need new and non-trivial proof ideas; see Section 4.

While our paper uses SCORE, it is very different from [19]. The goal of [19] is community detection where K is known, focusing on the null case ($m = K$). Here, the goal is to estimate K: SCORE is only used as part of our stepwise algorithm, and the focus is on the under-fitting case ($m < K$), where the property of SCORE is largely unknown, and our results on the NSP of SCORE are new. Our contributions are two fold. First, we propose StGoF as a new approach to estimating K. We show that StGoF has $N(0, 1)$ as the limiting null, achieves the optimal phase transition, and is uniformly consistent in broad settings (so it satisfies all requirements (R1)-(R3) as desired). Second, we overcome the technical challenges for stepwise algorithms of this kind by (a) developing a new stepwise scheme as in StGoF, (b) deriving sharp results as in (1.7), and (c) developing new techniques to prove the NSP of SCORE.

1.2 Content

Section 2 introduces the StGoF algorithm, and Section 3 shows that StGoF is consistent for K uniformly in a broad setting, and achieves the optimal phase transition. Section 4 shows that SCORE has the Non-Splitting Property (NSP) for $1 \leq m \leq K$, which is one of the keys to our study in Section 2. Section 5 presents simulation results, and Section 6 contains real data analysis. The supplementary material contains the proofs of all theorems and lemmas.

2 The stepwise Goodness-of-Fit (StGoF) algorithm

StGoF is a stepwise algorithm where for $m = 1, 2, \ldots$, we alternately use a community detection step and a Goodness-of-Fit (GoF) step. We may view StGoF as a general framework, where for either step, we can use a different algorithm. However, for most existing community detection algorithms (e.g., [25, 35]), it is unclear whether they have the desired theoretical properties (especially the NSP), so we may face analytical challenges. For this reason, we choose to use SCORE [19], which we prove to have the NSP. For GoF, existing algorithms (e.g., [14, 29]) do not apply to the current setting, so we propose Refitted Quadrilateral (RQ) as a new GoF metric (a quadrilateral in a graph is a length-4 cycle [1]; see details below).

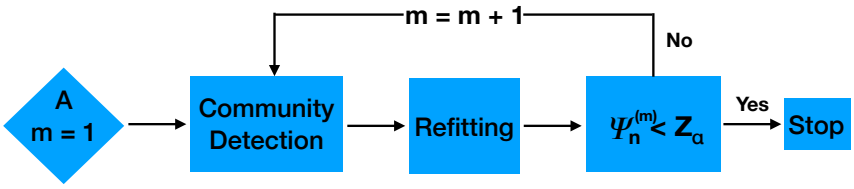

Figure 1: The flow chart of StGoF.

In detail, fix $0 < \alpha < 1$ (e.g., $\alpha = 1\%$ or 5%). Let z_α be the α upper-quantile of $N(0,1)$, StGoF runs as follows. Input: adjacency matrix A (initialize with $m = 1$; see Figure 1).

- *(a). Community detection.* If $m = 1$, let $\widehat{\Pi}^{(m)}$ be the n-dimensional vector of 1's. If $m > 1$, apply SCORE to A assuming m is the correct number of communities and obtain an $n \times m$ matrix $\widehat{\Pi}^{(m)}$ for the estimated community labels.
- *(b). Goodness-of-Fit.* Pretending $\widehat{\Pi}^{(m)}$ is the matrix of true community labels, we obtain an estimate $\widehat{\Omega}^{(m)}$ for Ω by refitting the DCBM, following (2.2)-(2.3) below. Obtain the Refitted Quadrilateral test score $\psi_n^{(m)}$ as in (2.5)-(2.8).
- *(c). Termination.* If $\psi_n^{(m)} \geq z_\alpha$, repeat (a)-(b) with $m = m + 1$. Otherwise, output m as the estimate for K. Denote the final estimate by $\hat{K}_\alpha^*$.

We now fill in the details for steps (a)-(b). Consider (a) first. The case of $m = 1$ is trivial so we only consider the case of $m > 1$. Let $\hat{\lambda}_k$ be the k-th largest (in magnitude) eigenvalue of A, and let $\hat{\xi}_k$ be the corresponding eigenvector. For each $m > 1$, we apply SCORE as follows. Input: A and m. Output: estimated community label matrix $\widehat{\Pi}^{(m)} \in \mathbb{R}^{n,m}$.

- Obtain the first m eigenvectors $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_m$ of A. Define the $n \times (m-1)$ matrix of entry-wise ratios $\widehat{R}^{(m)}$ by $\widehat{R}^{(m)}(i,k) = \hat{\xi}_{k+1}(i)/\hat{\xi}_1(i)$, $1 \le i \le n, 1 \le k \le m-1$. [2]
- Cluster the rows of $\widehat{R}^{(m)}$ by the k-means assuming we have m clusters. Output $\widehat{\Pi}^{(m)} = [\hat{\pi}_1^{(m)}, \ldots, \hat{\pi}_n^{(m)}]'$ ($\hat{\pi}_i^{(m)}(k) = 1$ if node i is clustered to cluster k and 0 otherwise).

Consider (b). The idea is to *pretend* that the SCORE estimate $\widehat{\Pi}^{(m)}$ is accurate. We then use it to estimate Ω by re-fitting, and check how well the estimated Ω fits with the adjacency matrix A. In detail, let d_i be the degree of node i, $1 \le i \le n$, and let $\widehat{\mathcal{N}}_k^{(m)}$ be the set of nodes that SCORE assigns to group k, $1 \le k \le m$. We decompose $\mathbf{1}_n$ as follows

$$\mathbf{1}_n = \sum_{k=1}^{m} \hat{\mathbf{1}}_k^{(m)}, \qquad \text{where } \hat{\mathbf{1}}_k^{(m)}(j) = 1 \text{ if } j \in \widehat{\mathcal{N}}_k^{(m)} \text{ and 0 otherwise.} \tag{2.1}$$

For most quantities that have superscript (m), we may only include the superscript when introducing these quantities for the first time, and omit it later for notational simplicity when there is no confusion. Introduce a vector $\hat{\theta}^{(m)} = (\hat{\theta}_1^{(m)}, \hat{\theta}_2^{(m)}, \ldots, \hat{\theta}_n^{(m)})' \in \mathbb{R}^n$ and a matrix $\widehat{P}^{(m)} \in \mathbb{R}^{m,m}$ where for all $1 \le i \le n$ and $1 \le k, \ell \le m$,

$$\hat{\theta}_i^{(m)} = [d_i/(\hat{\mathbf{1}}_k' A \mathbf{1}_n)] \cdot \sqrt{\hat{\mathbf{1}}_k' A \hat{\mathbf{1}}_k}, \qquad \widehat{P}_{k\ell}^{(m)} = (\hat{\mathbf{1}}_k' A \hat{\mathbf{1}}_\ell)/\sqrt{(\hat{\mathbf{1}}_k' A \hat{\mathbf{1}}_k)(\hat{\mathbf{1}}_\ell' A \hat{\mathbf{1}}_\ell)}. \tag{2.2}$$

Let $\widehat{\Theta}^{(m)} = \mathrm{diag}(\hat{\theta})$. We refit Ω by

$$\widehat{\Omega}^{(m)} = \widehat{\Theta}^{(m)} \widehat{\Pi}^{(m)} \widehat{P}^{(m)} (\widehat{\Pi}^{(m)})' \widehat{\Theta}^{(m)}. \tag{2.3}$$

Recall that $\Omega = \Theta \Pi P \Pi' \Theta$ and P has unit diagonal entries. In the ideal case where $m = K$, $\widehat{\Pi}^{(m)} = \Pi$, and $A = \Omega$, we have $(\widehat{\Theta}^{(m)}, \widehat{P}^{(m)}, \widehat{\Omega}^{(m)}) = (\Theta, P, \Omega)$. This suggests that the refitting in (2.3) is reasonable. The Refitted Quadrilateral (RQ) test statistic is then

$$Q_n^{(m)} = \sum_{i_1, i_2, i_3, i_4 (dist)} (A_{i_1 i_2} - \widehat{\Omega}_{i_1 i_2}^{(m)})(A_{i_2 i_3} - \widehat{\Omega}_{i_2 i_3}^{(m)})(A_{i_3 i_4} - \widehat{\Omega}_{i_3 i_4}^{(m)})(A_{i_4 i_1} - \widehat{\Omega}_{i_4 i_1}^{(m)}), \tag{2.4}$$

("dist" means the indices are distinct). Without the refitted matrix $\widehat{\Omega}^{(m)}$, $Q_n^{(m)}$ reduces to

$$C_n = \sum_{i_1, i_2, i_3, i_4 (dist)} A_{i_1 i_2} A_{i_2 i_3} A_{i_3 i_4} A_{i_4 i_1} = \text{total number of quadrilaterals.} \tag{2.5}$$

In the null case of $m = K$, first, $\mathrm{Var}(Q_n^{(m)})$ can be well-approximated by $8C_n$. Second, while the mean of $Q_n^{(K)}$ is 0 in the ideal case of $\widehat{\Omega}^{(K)} = \Omega$, in the real case, it is comparable to

[2] As the network is connected, $\hat{\xi}_1$ is uniquely defined with all positive entries, by Perron's theorem [19].

$[\mathrm{Var}(Q_n^{(K)})]^{1/2}$ and is not negligible, so we need bias correction. Motivated by these, for any $m \geq 1$, we introduce two vectors $\hat{g}^{(m)}, \hat{h}^{(m)} \in \mathbb{R}^m$ where

$$\hat{g}_k^{(m)} = (\hat{\mathbf{1}}_k'\hat{\theta})/\|\hat{\theta}\|_1, \qquad \hat{h}_k^{(m)} = (\hat{\mathbf{1}}_k'\widehat{\Theta}^2\hat{\mathbf{1}}_k)^{1/2}/\|\hat{\theta}\|, \qquad 1 \leq k \leq m. \tag{2.6}$$

Write for short $\widehat{V}^{(m)} = \mathrm{diag}(\widehat{P}\hat{g})$ and $\widehat{H}^{(m)} = \mathrm{diag}(\hat{h})$. We estimate the mean of $Q_n^{(m)}$ by

$$B_n^{(m)} = 2\|\hat{\theta}\|^4 \cdot [\hat{g}'\widehat{V}^{-1}(\widehat{P}\widehat{H}^2\widehat{P} \circ \widehat{P}\widehat{H}^2\widehat{P})\widehat{V}^{-1}\hat{g}], \tag{2.7}$$

where for matrixes A and B, $A \circ B$ is their Hadamard product [13]. Here, in the null case, $B_n^{(m)}$ is a good estimate for $\mathbb{E}[Q_n^{(m)}]$, and in the under-fitting case, it is much smaller than the leading term of $Q_n^{(m)}$ and so is negligible. Finally, the StGoF statistic is defined by

$$\psi_n^{(m)} = [Q_n^{(m)} - B_n^{(m)}]/\sqrt{8C_n}. \tag{2.8}$$

For each m, StGoF has a SCORE step (consisting of a PCA step and a k-means step) [19] and a GoF step. The complexity of PCA step is $O(n^2m)$ if we use the power method, and the complexity of the GoF step is $O(n^2\bar{d})$, where $\bar{d}$ is the average node degree. In Section 3, we show that under mild conditions, StGoF terminates in K steps with high probability. So aside from running K times of k-means, the complexity of StGoF is $O(n^2K^2 + n^2K\bar{d})$. Note that many real networks are sparse, where the factor $\bar{d}$ is relatively small. Similarly, [34] iterates for $m = 1, 2, \ldots, k_{max}$ (k_{max} is a prescribed upper bound for K), where for each m, it runs PCA once, k-means for $(m+1)$ times, and then computes a quantity with a cost of $O(n^2m)$. Therefore, aside from running k-means for $O(k_{max}^2)$ times, the cost is $O(n^2k_{max}^2)$. The approach by [38] also iterates for $m = 1, 2, \ldots, k_{max}$, where for each m, they need an exhaustive search step which is NP hard. To overcome the challenge, they use a spectral clustering approach to approximate the solution, where the cost (aside from running k-means for k_{max} times) is $O(n^2k_{max}^2)$. In theory, as the complexity of k-means is relatively high, the main costs of three algorithms come from the k-means part, and StGoF is less expensive (the times it runs for k-means is fewer than those of the others). In practice, we usually implement the k-means with the (relatively fast) Lloyd's algorithm [12], so all three algorithms are reasonably fast. For example, for a typical setting in Experiment 5a of Section 6 with $(n, K) = (600, 6)$, the computing time of three methods for 100 repetitions are $1, 10, 8$ minutes, respectively, and for a typical setting in Experiment 4b of Section 6 with $(n, K) = (1200, 3)$, the computing time of three methods for 100 repetitions are $2, 30, 40$ minutes, respectively.

Lemma 2.1. *Suppose $K = O(\bar{d})$, where $\bar{d}$ is the average degree of the network. For each $m = 1, 2, \ldots, K$, the complexity for computing $\psi_n^{(m)}$ by (2.2)-(2.5) is $O(n^2\bar{d})$.*

Our StGoF procedure is new. Existing stepwise algorithms (e.g., those in [38, 34]) iterate by comparing $\ell_n^{(m+1)} - \ell_n^{(m)}$ with a benchmark (which unfortunately has unknown parameters) for $m = 1, 2, \ldots, K$, and can not avoid the over-fitting case. StGoF iterates by comparing $\psi_n^{(m)}$ with $N(0,1)$ for $m = 1, 2, \ldots, K$, and successfully avoids the over-fitting case. Such a difference is crucial for obtaining sharp theoretical result; see Section 3.

Comparing with [19], though we use SCORE in the clustering step, but is for a different purpose: The orthodox SCORE is for community detection in the null case of $m = K$. We use SCORE to construct a low-rank matrix $\widehat{\Omega}^{(m)}$ in the under-fitting case of $m < K$, where the analysis is quite different and requires new technical tools; see Section 4.

The RQ test $\psi_n^{(m)}$ is connected to the SgnQ test [23] (a recent idea for global testing, which can be viewed as an improved version of the GC test by [21] and the EZ test by [10]), but there are major differences. First, the SgnQ test is for global testing where we test $K = 1$ v.s. $K > 1$, and it is unclear how to use it for goodness-of-fit in each step of StGoF. Second, SgnQ is not a stepwise algorithm and does not depend on any intermediate clustering results. The RQ critically depends on the intermediate clustering results by SCORE, where the NSP of SCORE plays a key role. Third, SgnQ does not need re-fitting, but RQ requires a re-fitting step. The re-fitting errors cause a non-negligible bias in $Q_n^{(m)}$. To obtain a tractable limiting null (where $m = K$), we need to figure out the right bias correction as in (2.7), with long and careful calculations. At the same time, by similar proofs as in our main theorems, we can show that $\psi_n^{(1)} \to N(0,1)$ if $K = 1$ and $\psi_n^{(1)} \to \infty$ in probability if $K > 1$ and $|\lambda_2|/\sqrt{\lambda_1} \to \infty$, where λ_k is the k-th largest (in magnitude) eigenvalue of Ω. Comparing with the lower bound in [23], $\psi_n^{(1)}$ is optimal for global testing.

Remark 1. Existing GoF algorithms include [14, 29], but they only address narrower settings (e.g., dense networks that follow SBM and have strong signals). As mentioned in [14], it remains unclear how to generalize these approaches to the DCBM setting here. In principle, a GoF approach only focuses on the null case, and can not be used for estimating K without sharp results in the under-fitting case, or the over-fitting case, or both.

Remark 2. For SBM settings where P is singular (see Example 3), $r < K$ ($r = \mathrm{rank}(\Omega)$). In this case, StGoF can consistently estimate r. To estimate K, we may revise StGoF by

replacing the SgnQ test in the GoF step by a degree-based χ^2-test (the success of which was shown for global testing with SBM; e.g. [2, 20]). By the NSP of SCORE, we can show that the new estimator is consistent under similar regularity conditions. Though the χ^2-tests may be powerful in some SBM settings, they usually lose power in more general DCBM settings, as suggested by the following result on *degree matching*. Consider a DCBM setting where we test $K = 1$ vs. $K > 1$ (i.e., global testing). It was shown in [23, 20] that for any alternative (i.e., $K > 1$), we can pair it with a null such that for each node, the expected degrees under the two models in the pair match with each other. Therefore, a naive degree-based test may lose power in separating the two models in the pair.

3 The consistency and optimality of StGoF

In this section, we discuss the consistency and optimality of StGoF. The NSP of SCORE (one of the key components in our proofs and a second part of our main results) is deferred to Section 4. Consider a DCBM with K communities as in (1.4). We assume

$$\|P\| \leq C, \qquad \|\theta\| \to \infty, \qquad \text{and} \qquad \theta_{\max}\sqrt{\log(n)} \to 0. \tag{3.1}$$

The first one is a mild regularity condition on the $K \times K$ community structure matrix P. The other two are mild conditions on sparsity. See (1.5) for the interesting range of $\|\theta\|$. We exclude the case where $\theta_i = O(1)$ for all $1 \leq i \leq n$ for convenience, but our results continue to hold in this case provided that we make some small changes in our proofs. Moreover, for $1 \leq k \leq K$, let $\mathcal{N}_k$ be the set of nodes belonging to community k, let n_k be the cardinality of $\mathcal{N}_k$, and let $\theta^{(k)}$ be the n-dimensional vector where $\theta_i^{(k)} = \theta_i$ if $i \in \mathcal{N}_k$ and $\theta_i^{(k)} = 0$ otherwise. We assume the K communities are balanced in the sense that

$$\min_{\{1\leq k\leq K\}}\{n_k/n,\ \|\theta^{(k)}\|_1/\|\theta\|_1,\ \|\theta^{(k)}\|^2/\|\theta\|^2\} \geq C. \tag{3.2}$$

In the presence of severe degree heterogeneity, the valid SNR for SCORE is

$$s_n = a_0(\theta)(|\lambda_K|/\sqrt{\lambda_1}), \qquad \text{where } \ a_0(\theta) = (\theta_{min}/\theta_{max}) \cdot (\|\theta\|/\sqrt{\theta_{max}\|\theta\|_1}) \leq 1.$$

In the special case of $\theta_{max} \leq C\theta_{min}$, it is true that $a_0(\theta) \asymp 1$ and $s_n \asymp |\lambda_K|/\sqrt{\lambda_1}$. In this case, s_n is the SNR introduced in (1.6). We assume

$$s_n \geq C_0\sqrt{\log(n)}, \qquad \text{for a sufficiently large constant } C_0 > 0. \tag{3.3}$$

In the special case $\theta_{max} \leq C\theta_{min}$, (3.3) is equivalent to $|\lambda_K|/\sqrt{\lambda_1} \geq C\sqrt{\log(n)}$, which is mild. Define a diagonal matrix $H \in \mathbb{R}^{K,K}$ by $H_{kk} = \|\theta^{(k)}\|/\|\theta\|$, $1 \leq k \leq K$. For the matrix HPH and $1 \leq k \leq K$, let μ_k be the k-th largest eigenvalue (in magnitude) and η_k be the corresponding eigenvector. By Perron's theorem [13], if P is irreducible, then the multiplicity of μ_1 is 1, and all entries of η_1 are strictly positive. Note also the size of the matrix P is small. It is therefore only a mild condition to assume that for a constant $0 < c_0 < 1$,

$$\min_{2\leq k\leq K} |\mu_1 - \mu_k| \geq c_0|\mu_1|, \qquad \text{and} \qquad \frac{\max_{1\leq k\leq K}\{\eta_1(k)\}}{\min_{1\leq k\leq K}\{\eta_1(k)\}} \leq C. \tag{3.4}$$

In fact, (3.4) holds if all entries of P are lower bounded by a positive constant or $P \to P_0$ for a fixed irreducible matrix P_0. We also note that the most challenging case for network analysis is when P is close to the matrix of 1's (where it is hard to distinguish one community from another), and (3.4) always holds in such a case. In this paper, we implicitly assume K is fixed. Our method can be extended to the case where K diverges with n at a speed not too fast, but the right hand side of (3.2) needs to be replaced by C/K. See Section 7 for discussions.

3.1 The null case and a confidence lower bound for K

In the null case, $m = K$, so if we apply SCORE to the rows of $\widehat{R}^{(m)}$ assuming m clusters, then we have perfect community recovery with overwhelming probability, and StGoF provides a confidence lower bound for K. The next theorem is proved in the supplement.

Theorem 3.1. *Fix $0 < \alpha < 1$. Suppose we apply StGoF to a DCBM model where (3.1)-(3.4) hold. As $n \to \infty$, up to a permutation of the columns of $\widehat{\Pi}^{(K)}$, $\mathbb{P}(\widehat{\Pi}^{(K)} \neq \Pi) \leq Cn^{-3}$, $\psi_n^{(K)} \to N(0,1)$ in law, and $\mathbb{P}(\widehat{K}^*_\alpha \leq K) \geq (1-\alpha) + o(1)$.*

Theorem 3.1 allows for severe degree heterogeneity. If the degree heterogeneity is moderate, $s_n \asymp |\lambda_K|/\sqrt{\lambda_1}$, and we have the following corollary.

Corollary 3.1. *Fix $0 < \alpha < 1$. Suppose we apply StGoF to a DCBM model where (3.1)-(3.2) and (3.4) hold. Suppose $\theta_{max} \leq C\theta_{min}$ and $|\lambda_K|/\sqrt{\lambda_1} \geq C_0\sqrt{\log(n)}$ for a sufficiently large constant $C_0 > 0$. As $n \to \infty$, up to a permutation of the columns of $\widehat{\Pi}^{(K)}$, $\mathbb{P}(\widehat{\Pi}^{(K)} \neq \Pi) \leq Cn^{-3}$, $\psi_n^{(K)} \to N(0,1)$ in law, and $\mathbb{P}(\widehat{K}^*_\alpha \leq K) \geq (1-\alpha) + o(1)$.*

It follows that $\widehat{K}^*_\alpha$ is a level-$(1-\alpha)$ confidence lower bound for K. If α depends on n and tends to 0 slowly enough, these results continue to hold. In this case, $\mathbb{P}(\widehat{K}^*_\alpha \leq K) = 1 - o(1)$.

When perfect community recovery is impossible but the faction of misclassified nodes is small with high probability (e.g., for a slightly smaller SNR), the asymptotic normality continues to hold. Similar comments apply to Theorem 3.3 and Corollary 3.2. As far as we know, this is the first time in the literature that we have derived a (completely) explicit limiting null. The result can be used to derive p-values in settings such as Goodness-of-Fit [14, 29]. If the assumed model in GoF is DCBM with K communities, then by Theorem 3.1, we can apply StGoF with $m = K$ and derive an approximate p-value as $\mathbb{P}(N(0,1) \geq \psi_n^{(K)})$. The proof of Theorem 3.1 is non-trivial and tedious. The main reason is that Ω is unknown and we must estimate it with refitting (see $\widehat{\Omega}^{(K)}$ in (2.3)). The refitting errors are non-negligible even when Π is given: we must choose a bias correction term as in (2.4) and analyze $Q_n^{(K)}$ carefully.

3.2 The under-fitting case of $m < K$ and consistency of StGoF

Fixing an m such that $1 < m < K$ (the case of $m = 1$ is trivial), suppose we apply SCORE to the rows of $\widehat{R}^{(m)}$ assuming m is the correct number of communities. Let $\widehat{\Pi}^{(m)}$ be the matrix of estimated community labels. In this case, we underestimate the number of clusters, so perfect community recovery is impossible. Fortunately, SCORE satisfies the *Non-Splitting Property (NSP)*. Recall that Π is the matrix of true community labels.

Definition 3.1. *Fix $K > 1$ and $m \leq K$. We say that a realization of the $n \times m$ matrix of estimated labels $\widehat{\Pi}^{(m)}$ satisfies the NSP if for any pair of nodes in the same (true) community, the estimated community labels are the same (i.e., each community in Π is contained in a community in the realization of $\widehat{\Pi}^{(m)}$). When this happens, we write $\Pi \preceq \widehat{\Pi}^{(m)}$.*

Theorem 3.2. *Consider a DCBM where (3.1)-(3.4) hold. With probability at least $1 - O(n^{-3})$, for each $1 < m \leq K$, $\Pi \preceq \widehat{\Pi}^{(m)}$ up to a permutation in the columns.*

By Theorem 3.2, SCORE has the NSP (with high probability). Theorem 3.2 is the key to our upper bound study below. In Section 4, we explain the main technical challenges for proving Theorem 3.2, and present the key theorems and lemmas required for the proof.

Theorem 3.3. *Fix $0 < \alpha < 1$. Suppose we apply StGoF to a DCBM model where (3.1)-(3.4) hold. As $n \to \infty$, $\min_{1 \leq m < K}\{\psi_n^{(m)}\} \to \infty$ in probability and $\mathbb{P}(\widehat{K}^*_\alpha \neq K) \leq \alpha + o(1)$.*

Theorem 3.3 allows for severe degree heterogeneity. When the degree heterogeneity is moderate, SNR $\asymp |\lambda_K|/\sqrt{\lambda_1}$ and we have the following corollary.

Corollary 3.2. *Fix $0 < \alpha < 1$. Suppose we apply StGoF to a DCBM model where (3.1)-(3.2) and (3.4) hold, $\theta_{max} \leq C\theta_{min}$, and $|\lambda_K|/\sqrt{\lambda_1} \geq C_0\sqrt{\log(n)}$ for a sufficiently large constant $C_0 > 0$. As $n \to \infty$, $\min_{1\leq m<K}\{\psi_n^{(m)}\} \to \infty$ in probability and $\mathbb{P}(\hat{K}_\alpha^* \neq K) \leq \alpha + o(1)$.*

Now in Theorem 3.3 and Corollary 3.2, if we let α depend on n and tend to 0 slowly enough, then we have $\mathbb{P}(\hat{K}_\alpha^* = K) \to 1$. Theorem 3.3 is proved in the supplement. The proof is non-trivial and long, so for instruction, we explain (a) what are the technical challenges and especially why the NSP is critical, and (b) why StGoF provides a consistent estimate.

Consider (a) first. The main technical challenge is how to analyze $\psi_n^{(m)}$ where we not only need sharp row-wise large deviation bounds for the matrix $\widehat{R}^{(m)}$, but also need to establish the NSP of SCORE, where we note $m \leq K$. To see why NSP is important, note that $Q_n^{(m)}$ depends on $\widehat{\Omega}^{(m)}$ (see (2.4)), where $\widehat{\Omega}^{(m)}$ is obtained by refitting using the SCORE estimate $\widehat{\Pi}^{(m)}$, and depends on A in a complicate way. The dependence poses challenges for analyzing $Q_n^{(m)}$, to overcome which, a conventional approach is to use concentrations. However, $\widehat{\Pi}^{(m)}$ has $\exp(O(n))$ possible realizations, and how to characterize the concentration of $\widehat{\Pi}^{(m)}$ is a challenging problem (e.g., [38, 34]). [3] Fortunately, if SCORE has the NSP, then $\widehat{\Pi}^{(m)}$ only has $\binom{K}{m}$ possible realizations. In fact, $\widehat{\Pi}^{(m)}$ may have even fewer possible realizations if we impose some mild conditions. Therefore, for each $1 \leq m \leq K$, $\widehat{\Omega}^{(m)}$ only concentrates on a few non-stochastic matrices. Using this and union bound, we can therefore remove the technical hurdle for analyzing $\psi_n^{(m)}$ in the under-fitting case.

The proof of NSP is non-trivial, partially due to the intractable rotation of eigenvectors dictated by the Davis-Kahan $\sin(\theta)$ theorem. See Section 4 for detailed explanations.

Consider (b). Fix $1 \leq m \leq K$. By the NSP of SCORE, except for a small probability, the estimated membership matrix $\widehat{\Pi}^{(m)} \in \mathbb{R}^{n,m}$ only has finitely many realizations. Fixing a realization $\widehat{\Pi}^{(m)} = \Pi_0$, let $\mathcal{N}_1^{(m,0)}, \cdots, \mathcal{N}_m^{(m,0)}$ be the clusters defined by Π_0. Let $\theta^{(m,0)}$, $\Theta^{(m,0)}$ and $P^{(m,0)}$ be constructed similarly as in (2.1)-(2.2), except that $(A, \widehat{\Pi}^{(m)})$ and the vector $d = (d_1, d_2, \ldots, d_n)'$ are replaced by (Ω, Π_0) and $\Omega\mathbf{1}_n$, respectively. Let $\Omega^{(m,0)} = \Theta^{(m,0)}\Pi_0 P^{(m,0)}\Pi_0'\Theta^{(m,0)}$. Then, on the event $\widehat{\Pi}^{(m)} = \Pi_0$, $\Omega^{(m,0)}$ is a non-stochastic

[3] To shed light on why $\widehat{\Pi}^{(m)}$ has so many possible realizations, suppose we wish to group n *iid* samples from $N(0,1)$ into two clusters with the same size. We have $\exp(O(n))$ possible clustering results.

proxy of the refitted matrix $\widehat{\Omega}^{(m)}$. Recall that Ω is a non-stochastic proxy of the adjacency matrix A. We thus expect the RQ statistic in (2.4) to satisfy that

$$\begin{aligned} Q_n^{(m)} &\approx \sum_{i_1,i_2,i_3,i_4(dist)} (\Omega_{i_1i_2} - \Omega^{(m,0)}_{i_1i_2})(\Omega_{i_2i_3} - \Omega^{(m,0)}_{i_2i_3})(\Omega_{i_3i_4} - \Omega^{(m,0)}_{i_3i_4})(\Omega_{i_4i_1} - \Omega^{(m,0)}_{i_4i_1}) \\ &\approx \mathrm{tr}((\Omega - \Omega^{(m,0)})^4), \qquad \text{on the event of } \widehat{\Pi}^{(m)} = \Pi_0. \end{aligned} \tag{3.5}$$

Now, when $m = K$, it can be shown that $\widehat{\Pi}^{(m)} = \Pi$ except for a small probability. Note also that when $\Pi_0 = \Pi$, our re-fitting procedure guarantees that $\theta^{(m,0)} = \theta$, $P^{(m,0)} = P$, and so $\Omega^{(m,0)} = \Omega$. It follows that $\mathrm{tr}((\Omega - \Omega^{(m,0)})^4) = 0$. When $m < K$, $\Omega^{(m,0)}$ has a rank $m < K$ and Ω has a rank K. Recall that $\lambda_1, \ldots, \lambda_K$ are the nonzero eigenvalues of Ω (arranged in the descending order of magnitudes). By Weyl's theorem, the kth largest absolute eigenvalue of $\Omega - \Omega^{(m,0)}$ is always lower bounded by $|\lambda_{k+m}|$, for all $1 \le k \le K - m$. It follows that $\mathrm{tr}((\Omega - \Omega^{(m,0)})^4) = \sum_{k=1}^{K-m} |\lambda_k(\Omega - \Omega^{(m,0)})|^4 \ge \sum_{k=1}^{K-m} \lambda_{m+k}^4$. In summary,

$$\mathrm{tr}((\Omega - \Omega^{(m,0)})^4) = 0 \text{ if } m = K, \text{ and } \mathrm{tr}((\Omega - \Omega^{(m,0)})^4) \ge \textstyle\sum_{k=m+1}^{K} \lambda_k^4 \text{ if } m < K. \ [4] \tag{3.6}$$

Recall that $\psi_n^{(m)}$ is the standardized version of $Q_n^{(m)}$, and that except for a small probability, $\widehat{\Pi}$ has only one possible realization for $\widehat{\Pi}$ in the null case and has only finite realizations in the alternative case. Using the above and union bounds, we can show that

$$\begin{cases} \psi_n^{(m)} \to N(0,1), & \text{if } m = K, \\ \mathbb{E}[\psi_n^{(m)}] \asymp (\sum_{k=m+1}^{K} \lambda_k^4)/\lambda_1^2 \text{ and so } \psi_n^{(m)} \to \infty \text{ in prob.}, & \text{if } 1 \le m < K, \end{cases} \tag{3.7}$$

where $(\sum_{k=m+1}^{K} \lambda_k^4)/\lambda_1^2 \ge (\lambda_K/\sqrt{\lambda_1})^4$ when $m < K$. Therefore, with a proper threshold on $\psi_n^{(m)}$, StGoF stops at $m = K$ with an overwhelming probability and outputs a consistent estimate for K. The proofs for the NSP and (3.6)-(3.7) are technically demanding. See Section 4 and Section A of the supplement for detailed explanations and proofs.

[4] This explains why in StGoF we do not use the refitted triangle (RT) $T_n^{(m)} = \sum_{i_1,i_2,i_3(dist)} (A_{i_1i_2} - \widehat{\Omega}^{(m)}_{i_1i_2})(A_{i_2i_3} - \widehat{\Omega}^{(m)}_{i_2i_3})(A_{i_3i_1} - \widehat{\Omega}^{(m)}_{i_3i_1})$, which is comparably easier to analyze. While we may similarly derive $T_n^{(m)} \gtrsim \sum_{k=m+1}^{K} \lambda_k^3$ with large probability, $\lambda_{m+1}, \ldots, \lambda_K$ may have different signs and so may cancel with each other. We can not use $B_n^{(m)} = \sum_{i_1,i_2(dist)} (A_{i_1i_2} - \widehat{\Omega}^{(m)}_{i_1i_2})(A_{i_2i_1} - \widehat{\Omega}^{(m)}_{i_2i_1})$ either. The variance of $B_n^{(m)}$ is unappealingly large so the resultant procedure can not achieve the optimal phase transition; see Section 3.3. Also, see [23] for discussion on statistics similar to $T_n^{(m)}$ and $B_n^{(m)}$.

3.3 Information lower bound and phase transition

In Theorem 3.3 and Corollary 3.2, we require the SNR, $|\lambda_K|/\sqrt{\lambda_1}$, to tend to ∞ at a speed of at least $\sqrt{\log(n)}$. We show that such a condition cannot be significantly relaxed. There are relatively few studies on the lower bound for estimating K, and our results are new.

We say two DCBM models are asymptotically indistinguishable if for any test that tries to decide which model is true, the sum of Type I and Type II errors is no smaller than $1+o(1)$, as $n \to \infty$. Given a DCBM with K communities, our idea is to construct a DCBM with $(K+m)$ communities for any $m \geq 1$, and show that two DCBM are asymptotically indistinguishable, provided that the SNR of the latter is $o(1)$.

Fixing $K_0 \geq 1$, we consider a DCBM with K_0 communities that satisfies (1.1)-(1.3). Let $(\Theta, \widetilde{\Pi}, \widetilde{P})$ be the parameters of this DCBM, and let $\widetilde{\Omega} = \Theta\widetilde{\Pi}\widetilde{P}\widetilde{\Pi}'\Theta$. When $K_0 > 1$, let $(\beta', 1)'$ be the last column of $\widetilde{P}$, and let $S \in \mathbb{R}^{K_0-1,K_0-1}$ be the sub-matrix of $\widetilde{P}$ excluding the last row and the last column. Given $m \geq 1$ and $b_n \in (0,1)$, we construct a DCBM model with (K_0+m) communities as follows. We define a $(K_0+m) \times (K_0+m)$ matrix P:

$$P = \begin{bmatrix} S & \beta\mathbf{1}'_{m+1} \\ \mathbf{1}_{m+1}\beta' & \frac{m+1}{1+mb_n}M \end{bmatrix}, \qquad \text{where} \quad M = (1-b_n)I_{m+1} + b_n\mathbf{1}_{m+1}\mathbf{1}'_{m+1}. \tag{3.8}$$

When $K_0 = 1$, we simply let $P = \frac{m+1}{1+mb_n}M$. Let $\tilde{\ell}_i \in \{1, \ldots, K_0\}$ be the community label of node i defined by $\widetilde{\Pi}$. We generate labels $\ell_i \in \{1, \ldots, K_0+m\}$ by

$$\ell_i = \begin{cases} \tilde{\ell}_i, & \text{if } \tilde{\ell}_i \in \{1, \ldots, K_0-1\}, \\ \text{uniformly drawn from } \{K_0, K_0+1, \ldots K_0+m\}, & \text{if } \tilde{\ell}_i = K_0. \end{cases} \tag{3.9}$$

Let Π be the corresponding community label matrix. This gives rise to a DCBM model with (K_0+m) communities, where $\Omega = \Theta\Pi P\Pi'\Theta$. Though P does not have unit diagonals, we can re-parametrize so that it has unit diagonals: Let D be the $(K_0+m) \times (K_0+m)$ diagonal matrix with $D_{kk} = \sqrt{P_{kk}}$, $1 \leq k \leq K_0+m$. Now, if we let $P^* = D^{-1}PD^{-1}$, $\theta_i^* = \theta_i\|D\pi_i\|_1$, and $\Theta^* = \mathrm{diag}(\theta_1^*, \ldots, \theta_n^*)$, then P^* has unit-diagonals and $\Omega = \Theta^*\Pi P^*\Pi'\Theta^*$.

Here some rows of Π are random (so we may call the corresponding model the random-label DCBM), but this is conventional in the study of lower bounds. Let λ_k be the kth largest eigenvalue (in magnitude) of Ω. Since Ω is random, λ_k's are also random (but we can bound $|\lambda_K|/\sqrt{\lambda_1}$ conveniently). The following theorem is proved in the supplement.

Theorem 3.4. *Fix $K_0 \geq 1$ and consider a DCBM model with n nodes and K_0 communities, whose parameters $(\theta, \widetilde{\Pi}, \widetilde{P})$ satisfy (3.1)-(3.2). Let $(\beta', 1)'$ be the last column of $\widetilde{P}$, and let S be the sub-matrix of $\widetilde{P}$ excluding the last row and last column. We assume $|\beta' S^{-1}\beta - 1| \geq C$.*

- *Fix $m \geq 1$. Given any $b_n \in (0,1)$, we can construct a random-label DCBM model with $K = K_0 + m$ communities as in (3.8)-(3.9). Then, as $n \to \infty$, $|\lambda_K|/\sqrt{\lambda_1} \leq C\|\theta\|(1-b_n)$ with probability $1 - o(n^{-1})$. Moreover, if $(1-b_n)/|\lambda_{\min}(S)| = o(1)$, where $\lambda_{\min}(S)$ is the minimum eigenvalue (in magnitude) of S, then $|\lambda_K|/\sqrt{\lambda_1} \geq C^{-1}\|\theta\|(1-b_n)$ with probability $1 - o(n^{-1})$. Here $C > 1$ is a constant that does not depend on b_n.*
- *Fix $m_1, m_2 \geq 1$ with $m_1 \neq m_2$. As $n \to \infty$, if $\|\theta\|(1-b_n) \to 0$, then the two random-label DCBM models associated with m_1 and m_2 are asymptotically indistinguishable.*

Here, the condition $|\beta' S^{-1}\beta - 1| \geq C$ is used to bound the last diagonal entry of $\widetilde{P}^{-1}$, which is $1/(\beta' S^{-1}\beta - 1)$. By Theorem 3.4, starting from a (fixed-label) DCBM with K_0 communities, we can construct a collection of random-label DCBM, with $K_0+1, K_0+2, \ldots, K_0+m$ communities, respectively, where (a) for the model with (K_0+m) communities, $|\lambda_{K_0+m}|/\sqrt{\lambda_1} \asymp \|\theta\|(1-b_n)$, with an overwhelming probability, and (b) each pair of models are asymptotically indistinguishable if $\|\theta\|(1-b_n) = o(1)$. Therefore, for a broad class of DCBM with unknown K where SNR $= o(1)$ for some models, a consistent estimate for K does not exist.

Fixing $m_0 > 1$ and a sequence of numbers $a_n > 0$, let $\mathcal{M}_n(m_0, a_n)$ be the collection of DCBM for an n-node network with K communities, where $1 \leq K \leq m_0$, $|\lambda_K|/\sqrt{\lambda_1} \geq a_n$, and (3.1)-(3.2) hold. In Section 3.2, we show that if $a_n \geq C_0\sqrt{\log(n)}$ for a sufficiently large constant C_0, then for each DCBM in $\mathcal{M}_n(m_0, a_n)$, StGoF provides a consistent estimate for K. The following theorem says that, if we allow $a_n \to 0$, then $\mathcal{M}_n(m_0, a_n)$ is too broad, and a consistent estimate for K does not exist.

Theorem 3.5. *Fix $m_0 > 1$ and let $\mathcal{M}_n(m_0, a_n)$ be the class of DCBM as above. As $n \to \infty$, if $a_n \to 0$, then $\inf_{\hat{K}}\{\sup_{\mathcal{M}_n(m_0,a_n)} \mathbb{P}(\hat{K} \neq K)\} \geq (1/6 + o(1))$, where the probability is evaluated at any given model in $\mathcal{M}_n(m_0, a_n)$ and the supremum is over all such models.*

Combining Theorems 3.1, 3.5, and Corollary 3.2, we have a phase transition result (phase transition is a recent theoretical framework (e.g., [7, 24]). It is closely related to the classical minimax framework but can be more informative in many cases).

- *Impossibility.* If $a_n \to 0$, then $\mathcal{M}_n(m_0, a_n)$ defines a class of DCBM that is too broad where some pairs of models in the class are asymptotically indistinguishable. Therefore, no estimator can consistently estimate the number of communities for each model in the class (and we say "a consistent estimate for K does not exist" for short).
- *Possibility.* If $a_n \geq C_0\sqrt{\log(n)}$ for a sufficiently large C_0, then for every DCBM in $\mathcal{M}_n(m_0, a_n)$, StGoF provides a consistent estimate for the number of communities if the model only has moderate degree heterogeneity (i.e., $\theta_{max} \leq C\theta_{min}$). StGoF continues to be consistent in the presence of severe degree heterogeneity if the adjusted SNR satisfies that $s_n \geq C_0\sqrt{\log(n)}$ with a sufficiently large C_0.

The case of $C \leq a_n < C_0\sqrt{\log(n)}$ is more delicate. Sharp results are possible if we consider more specific models (e.g., for a scaling parameter $\alpha_n > 0$, (θ_i/α_n) are *iid* from a fixed distribution F, and the off-diagonals of P are the same). We leave this to the future.

Comparing with existing works, we have the following comments: (a) StGoF is the first method that is proved to achieve the optimal transition, (b) StGoF is the first method that is proved to have a (completely) explicit limiting null, (c) we prove the NSP of SCORE, and use it to derive sharp results that are not available before, (d) our settings are much broader and our regularity conditions are much weaker, and (e) we overcome the challenges of stepwise algorithms of this kind by using the sharp results we derive and by using a different stepwise scheme (so to avoid the analysis of the over-fitting case where the NSP does not hold). We now compare with [38, 34] with more details.

First, their approaches require a signal strength much stronger than ours, and so do not achieve the phase transition. When $\theta_{\max} \leq C\theta_{\min}$, our result requires $|\lambda_K|/\sqrt{\lambda_1} \geq C_0\sqrt{\log(n)}$, which matches the lower bound in Section 3.3. However, [38] needs $|\lambda_K|/\sqrt{\lambda_1} \gg n^{1/4}\sqrt{\log(n)}$ (see their Section 2.5), which is non-optimal. Also, [34] proves consistency under the condition of $\lambda_1 \geq C\log(n)$. Recall that they assume $P = \rho_n P_0$. In their setting, $|\lambda_1|, \ldots, |\lambda_K|$ are at the same order, and $\lambda_1 \geq C\log(n)$ indeed translates to $|\lambda_K|/\sqrt{\lambda_1} \geq C_0\sqrt{\log(n)}$. However, for general settings where $|\lambda_1|, \cdots, |\lambda_K|$ are at different orders, it is unclear whether their method is optimal (because the SNR is captured by $|\lambda_K|/\sqrt{\lambda_1}$, not $\sqrt{\lambda_1}$). In comparison, our result matches with the lower bound for all settings. Second, [38] only studies the SBM where θ_i's are all equal, and [34] assumes that $\theta_{\max} \leq C\theta_{\min}$ and $P = \rho_n P_0$, for a fixed matrix P_0; in this

setting, the degree heterogeneity is only moderate, and $|\lambda_1|, \cdots, |\lambda_K|$ are at the same order. This excludes many practical cases of interest. Last, besides the very mild condition of (3.2), we do not need any hard-to-check conditions on Π. In contrast, [38, 34] impose stringent conditions. For example, [34] defines a quantity $Q_K(k)$ by applying spectral clustering to Ω and then evaluating the change of residual sum of squares by further splitting one cluster. They impose conditions on $Q_K(k)$ for every $1 \leq k \leq K-1$ (see their Assumption 3). These conditions are hard to check in practice. Moreover, when $P = \rho_n P_0$ does not hold for a fixed P_0, the conditions on $\{Q_K(k)\}_{1 \leq k \leq K-1}$ are easy to violate (e.g., in our Example 1).

The advantage of our theory partially comes from the way our algorithm is designed. StGoF only assesses one candidate of K in each step, instead of comparing two adjacent values of K. It helps avoid the analysis of the over-fitting case, and it also avoids imposing stringent conditions on Π. Another advantage comes from our new proof ideas. We do not need $\widehat{\Pi}^{(m)}$ to converge to a non-stochastic matrix, because our proof is *not* based on Taylor expansion. For example, a key component of our analysis is the NSP of SCORE. We develop the NSP under very weak conditions where $\widehat{\Pi}^{(m)}$ can be non-tractable, non-unique, and depending on a data-driven rotation matrix (see Section 4).

4 The non-splitting property (NSP) of SCORE

To prove the NSP of SCORE, we face technical challenges. In the SCORE step of StGoF, for each $2 \leq m \leq K$, we cluster n rows of the matrix $\widehat{R}^{(m)}$ into m clusters. We find that for any two rows of $\widehat{R}^{(m)}$, the distance critically depends on a non-tractable data-dependent rotation matrix $\widehat{\Gamma}$ dictated by the David-Kahn $\sin(\theta)$ theorem [6], and it may vary significantly as $\widehat{\Gamma}$ changes from one realization to another. This poses an unconventional setting for clustering. To overcome the challenge, we first discover a new distance-based quantity that is *semi-invariant* with respect to $\widehat{\Gamma}$: the quantity remains at the same order of $O(1)$ as $\widehat{\Gamma}$ varies from one realization to another. We then develop a new k-means theorem (Theorem 4.1) and use it to prove the NSP. The proof of Theorem 4.1 is non-trivial: our setting is an unconventional clustering setting and we do not want to impose unrealistic and strong conditions. Note that the literature on SCORE has been focused on the null case of $m = K$, but our primary interest is in the under-fitting case of $m < K$.

4.1 Row-wise large-deviation bounds and Ideal polytope

Recall that for $2 \le m \le K$, $\widehat{R}^{(m)}$ is an $n \times (m-1)$ matrix constructed from the eigenvectors $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_m$ by taking entry-wise ratios between $\hat{\xi}_2, \ldots, \hat{\xi}_m$ and $\hat{\xi}_1$; see Section 2. Let λ_k be the k-th largest (in magnitude) eigenvalue of Ω and let ξ_k be the corresponding eigenvector. Under our assumptions (e.g., see condition (3.4)), there exists a $(K-1) \times (K-1)$ orthogonal matrix $\widehat{\Gamma}$ such that $[\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K] \approx [\xi_1, \xi_2, \ldots, \xi_K] \cdot \mathrm{diag}(1, \widehat{\Gamma})$. The rotation matrix $\widehat{\Gamma}$ is dictated by the Davis-Kahan $\sin(\theta)$ theorem in spectral analysis. It is well-known that the matrix is data dependent and hard to track. Even if $\lambda_1, \ldots, \lambda_K$ are distinct (so each vector in $\{\xi_1, \ldots, \xi_K\}$ is unique up to a ± 1 factor), $\widehat{\Gamma}$ can still take arbitrary values on the Stiefel manifold and does not concentrate on any non-stochastic rotation matrix on the manifold. [5]

Therefore, we must consider all possible realizations $\widehat{\Gamma} = \Gamma$. This not only poses analytical challenges (see Section 4.2) but also makes notations more complicate. Fix a (non-stochastic) orthogonal matrix Γ. For $2 \le k \le K$, let $\xi_k(\Gamma)$ be the kth column of $[\xi_1, \xi_2, \ldots, \xi_K] \cdot \mathrm{diag}(1, \Gamma)$,[6] and let $\xi_k(j, \Gamma)$ be the jth entry of $\xi_k(\Gamma)$, $1 \le j \le n$. Define $R^{(m)}(\Gamma) \in \mathbb{R}^{n, m-1}$ by

$$R^{(m)}(i, \ell; \Gamma) = \xi_{\ell+1}(i; \Gamma)/\xi_1(i), \qquad 1 \le i \le n,\ 1 \le \ell \le m-1. \tag{4.1}$$

Comparing (4.1) with the definition of $\widehat{R}^{(m)}$ in Section 2, it is seen that $R^{(m)}(\Gamma)$ is the population counterpart of $\widehat{R}^{(m)}$ on the event of $\widehat{\Gamma} = \Gamma$. Lemma 4.1 provides a sharp row-wise large-deviation bound for $\widehat{R}^{(m)} - R^{(m)}(\Gamma)$ and is proved in the supplemental material.

Lemma 4.1 (Row-wise bounds). *Consider a DCBM model where (3.1)-(3.4) hold. Let $s_n = a_0(\theta)(|\lambda_K|/\sqrt{\lambda_1})$, where $a_0(\theta)$ is as in Section 3. For each $1 < i \le n$, let $(r_i^{(m)}(\Gamma))'$ and $(\hat{r}_i^{(m)})'$ denote the i-th row of $R^{(m)}(\Gamma)$ and $\widehat{R}^{(m)}$, respectively. As $n \to \infty$, with probability $1 - O(n^{-3})$, for all $1 \le m \le K$ and $1 \le i \le n$ and all $(K-1) \times (K-1)$ orthogonal matrix Γ, $\|\hat{r}_i^{(m)} - r_i^{(m)}(\Gamma)\| \le \|\hat{r}_i^{(K)} - r_i^{(K)}(\Gamma)\| \le C s_n^{-1}\sqrt{\log(n)}$ over the event $\widehat{\Gamma} = \Gamma$.*

Under our assumptions, $s_n^{-1}\sqrt{\log(n)}$ is upper bounded by a sufficiently small constant. It implies that each $\hat{r}_i^{(m)}$ is sufficiently close to $r_i^{(m)}(\Gamma)$ on the event $\widehat{\Gamma} = \Gamma$.

It remains to study the geometry underlying $\{r_i^{(m)}(\Gamma)\}_{1 \le i \le n}$ for *an arbitrary* rotation matrix $\Gamma \in \mathbb{R}^{K-1, K-1}$. Recall that $H \in \mathbb{R}^{K,K}$ is the diagonal matrix with $H_{kk} = \|\theta^{(k)}\|/\|\theta\|$,

[5] $\widehat{\Gamma}$ is tractable only if we impose a strong eigen-gap condition. However, this excludes many practical settings of interest, especially when the signals are weak and $|\lambda_1|, \ldots, |\lambda_K|$ are at different orders.

[6] By Perron's theorem, ξ_1 is uniquely defined and a strictly positive vector. Vectors $\xi_2, \ldots, \xi_K$ are not necessarily unique, but we can select an arbitrary candidate of $\xi_2, \ldots, \xi_K$ to define $\xi_2(\Gamma), \ldots, \xi_K(\Gamma)$.

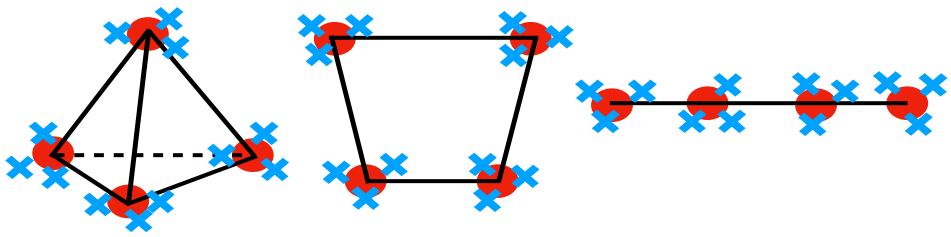

Figure 2: An example ($K = 4$). From left to right: $m = 4, 3, 2$. Red dots: the 4 distinct rows of $R^{(m)}$, which are $v_1^{(m)}, v_2^{(m)}, v_3^{(m)}, v_4^{(m)}$. Blue crosses: the rows of $\widehat{R}^{(m)}$. The red dots are the vertices of a tetrahedron when $m = 4$, vertices of a quadrilateral when $m = 3$, and scalars when $m = 2$. For each m, the n rows of $\widehat{R}^{(m)}$ form K clusters, each corresponding to a true community. The figure is only for illustration, and we should not have the wrong impression that the K clusters are always well-separated.

$1 \le k \le K$. For each $1 \le k \le K$, let μ_k be the k-th largest (in magnitude) eigenvalue of HPH and let $\eta_k \in \mathbb{R}^K$ be the associated (unit-norm) eigenvectors, respectively. By Lemma B.1 of the supplement, η_1 is unique and all entries are strictly positive. Also, while $(\eta_2, \ldots, \eta_K)$ may be non-unique, there is a one-to-one correspondence between the choice of $(\eta_2, \ldots, \eta_K)$ and the choice of $(\xi_2, \ldots, \xi_K)$; see the paragraph above (4.1). Fix Γ. For each $2 \le k \le K$, let $\eta_k(\Gamma)$ be the $(k-1)$-th column of $[\eta_2, \eta_3, \ldots, \eta_K]\Gamma$, and let $\eta_k(i, \Gamma)$ denote the i-th entry of $\eta_k(\Gamma)$, $1 \le i \le K$. Define a $K \times (m-1)$ matrix $V^{(m)}(\Gamma)$ by

$$V^{(m)}(k, \ell; \Gamma) = \eta_{\ell+1}(k; \Gamma)/\eta_1(k), \qquad 1 \le k \le K,\ 1 \le \ell \le m-1. \tag{4.2}$$

Let $(v_k^{(m)}(\Gamma))'$ be the k-th row of $V^{(m)}(\Gamma)$. Lemma 4.2 is proved in the supplement.

Lemma 4.2 (The ideal polytope)**.** *Consider a DCBM model where (3.4) holds. Fix $1 < m \le K$. We have that $r_i^{(m)}(\Gamma) = v_k^{(m)}(\Gamma)$ for all $i \in \mathcal{N}_k$ and $1 \le k \le K$.*

Combining Lemmas 4.1-4.2 gives the following claim. Viewing $\{\hat{r}_i^{(m)}\}_{1 \le i \le n}$ as a point cloud in $\mathbb{R}^{m-1}$, we have that with overwhelming probability, for any realization of $\widehat{\Gamma} = \Gamma$ and each $1 < m \le K$, there are K clusters in the point cloud, corresponding to K true communities, where $v_1^{(m)}(\Gamma), \ldots, v_K^{(m)}(\Gamma)$ are the cluster centers (see Figure 2).

4.2 Challenges in proving NSP and our approach

Given the results in previous section, one may think that NSP is easy to prove. Unfortunately, this is not the case: even with the results in the previous section, how to prove NSP remains a non-trivial problem, especially when $m < K$. We now provide a detailed explanation.

Recall that $\widehat{\Gamma}$ is data dependent and hard to track, so we have to consider all realizations of $\widehat{\Gamma} = \Gamma$. Therefore, to prove the claim, we need to show that the NSP holds *uniformly for all* Γ *in the Stiefel manifold* $\mathcal{O}_{K-1}$. Given a realization of $\widehat{\Gamma} = \Gamma$, let $\widehat{B}^{(m)}$ and $B^{(m)}$ be the sub-matrices of $\widehat{\Gamma}$ and Γ, consisting of the first $(m-1)$ columns. Introduce a matrix $V_0 \in \mathbb{R}^{K,K-1}$ by $V_0(i,k) = \eta_{k+1}(i)/\eta_1(i)$, $1 \le i \le K, 1 \le k \le K-1$, where η_k's are as in the previous section. For each $1 < m \le K$, by our notations, the K cluster centers in Lemma 4.2 are the K rows of the matrix $V^{(m)}(\Gamma) \in \mathbb{R}^{K,m-1}$, and $V^{(m)}(\Gamma)$ is related to V_0 by $V^{(m)}(\Gamma) = V_0 B^{(m)}$. For any $K \times (m-1)$ matrix M, let $d_K(M)$ be the minimum pairwise Euclidean distance of the K rows of M. In [19], it was shown that $d_K(V_0) \ge \sqrt{2}$. We now discuss the null case and the under-fitting case separately.

In the null case, $m = K$, and $B^{(m)} = \Gamma$ is a rotation matrix. Since the Euclidean distances remain unchanged for rotation, the relative position of the K cluster centers is *invariant* with respect to Γ, and especially, $d_K(V^{(m)}(\Gamma)) = d_K(V_0) \ge \sqrt{2}$. Combining this with Lemmas 4.1-4.2, we have: (a) With high probability, the n rows of $\widehat{R}^{(m)}$ split into K clusters; for each row, the distance to the closest cluster center is $\le O(s_n\sqrt{\log(n)}) = o(1)$. (b) The K cluster centers are well-separated by a distance of $\sqrt{2}$. (c) $m = K$, so the number of clusters assumed in k-means matches the number of true clusters. In this case, the cluster labels estimated by k-means match with the true cluster labels (up to a permutation) so the NSP follows.

The under-fitting case is unfortunately much harder to prove. For $m < K$, $B^{(m)}$ is not a square matrix (it is not a rotation matrix even in the simplest case where Γ is the identity matrix). Compared to the null case, we have some major differences. First, even in the case where Γ is the identity matrix, we may have $d_K(V^{(m)}) = 0$ so the K cluster centers are not well-separated. Second, for any two rows of $V^{(m)}(\Gamma)$ (each is one of the K cluster centers), the Euclidean distance critically depends on Γ. As Γ varies continuously in the Stiefel manifold, the distance may vary from $O(1)$ to 0. Therefore, the relative positions of the K cluster centers critically depend on Γ, and may vary significantly from one case to another (we may have $d_K(V^{(m)}(\Gamma)) \ge C$ for one Γ and $d_K(V^{(m)}(\Gamma)) = 0$ for another Γ). Note also that since $m < K$, the number of clusters fed into the k-means algorithm is smaller than the number of true clusters. Seemingly, this is an unconventional clustering setting, especially as our goal is to show that when we apply k-means, the NPS holds uniformly for all Γ. To overcome the challenges, (1) we propose a new metric for the relative positions of the K cluster centers,

and (2) we develop a new k-means theorem specifically for the setting we have. Here, (1) is motivated by the observation that, the main reason NSP is easier to prove in the null case is that $d_K(V^{(m)}(\Gamma))$ (minimum pairwise distance of the K cluster centers) is invariant to Γ and so the K clusters are always well-separated, uniformly for all Γ. In the under-fitting case, $d_K(V^{(m)}(\Gamma))$ is not invariant to Γ, but there may exist a different measure that is invariant to Γ. This motivates us to define $d_m(V^{(m)}(\Gamma))$ as a new measure for the relative positions of the K cluster centers, which is *semi-invariant* to Γ (i.e., there are constants $c_2 > c_1 > 0$ such that $c_1 \leq d_m(V^{(m)}(\Gamma)) \leq c_2$ for all Γ in the Stiefel manifold). In detail, for any $1 < m \leq K$ and any given K points in $\mathbb{R}^{m-1}$, we have the following definition, which is an extension of the *minimum pairwise distance*.

Definition 4.1 (Distance-based metrics defined by bottom up pruning)**.** *Fixing $K > 1$ and $1 < m \leq K$, consider a $K \times (m-1)$ matrix $U = [u_1, u_2, \ldots, u_K]'$. First, let $d_K(U)$ be the minimum pairwise distance of all K rows. Second, let u_k and u_ℓ $(k < \ell)$ be the pair that satisfies $\|u_k - u_\ell\| = d_K(U)$ (if this holds for multiple pairs, pick the first pair in the lexicographical order). Remove row ℓ from the matrix U and let $d_{K-1}(U)$ be the minimum pairwise distance for the remaining $(K-1)$ rows. Repeat this step and define $d_{K-2}(U), d_{K-3}(U), \ldots, d_2(U)$ recursively. Note that $d_K(U) \leq d_{K-1}(U) \leq \ldots \leq d_2(U)$.*

For each fixed Γ, $d_K(V^{(m)}(\Gamma))$ is the minimum pairwise distance between the K cluster centers $v_1^{(m)}(\Gamma), \ldots, v_K^{(m)}(\Gamma)$, and $d_m(V^{(m)}(\Gamma))$ is the minimum pairwise distance of the m remaining cluster centers after we prune $(m-K)$ cluster centers in the bottom-up fashion as above. When Γ ranges continuously in $\mathcal{O}_{K-1}$, $d_K(V^{(m)}(\Gamma))$ may range continuously from $O(1)$ to 0, but fortunately $d_m(V^{(m)}(\Gamma))$ remains at the same order of $O(1)$, and so is *semi-invariant*. This is the following lemma, which is proved in the supplement.

Lemma 4.3. *Consider a DCBM model where (3.2) and (3.4) hold. Fix $1 \leq m \leq K$. There is a constant $C > 0$ (which may depend on m), such that $\min_{\Gamma \in \mathcal{O}_{K-1}} \{d_m(V^{(m)}(\Gamma))\} \geq C$.*

We now discuss (2). To prove that NSP holds uniformly for all Γ, it remains to develop a new k-means theorem. We can have two versions of the k-means theorem: a "weaker" version where we assume $d_K(V^{(m)}(\Gamma)) \geq C$, for a constant $C > 0$, and a "stronger" version where we only require $c_1 \leq d_m(V^{(m)}(\Gamma) \leq c_2$, and $d_K(V^{(m)}(\Gamma))$ may be as large as $O(1)$ or as small as 0. As $d_K(V^{(m)}(\Gamma)) = 0$ for many Γ, the "weaker" version is inadequate for our setting. Theorem

4.1 is a "stronger" version of the k-means theorem, and is proved in the supplement. The "weaker" version is implied by Theorem 4.1 and so the proof is skipped.

Theorem 4.1 (The "stronger" version of the k-means theorem)**.** *Fix $1 < m \leq K$ and let n be sufficiently large. Consider the non-stochastic vectors $x_1, \ldots, x_n$ that take only K values in $u_1, \ldots, u_K$. Write $U = [u_1, \ldots, u_K]'$. Let $F_k = \{1 \leq i \leq n : x_i = u_k\}$, $1 \leq k \leq K$. Suppose for some constants $0 < \alpha_0 < 1$ and $C_0 > 0$, $\min_{1\leq k\leq K} |F_k| \geq \alpha_0 n$ and $\max_{1\leq k\leq K} \|u_k\| \leq C_0 \cdot d_m(U)$. We apply the k-means clustering to a set of n points $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$ assuming $\leq m$ clusters, and denote by $\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_m$ the obtained clusters (if the solution is not unique, pick any of them). There exists a constant $c > 0$, which only depends on (α_0, C_0, m), such that, if $\max_{1\leq i\leq n} \|\hat{x}_i - x_i\| \leq c \cdot d_m(U)$, then $\#\{1 \leq j \leq m : \hat{S}_j \cap F_k \neq \emptyset\} = 1$, for each $1 \leq k \leq K$.*

To prove the NSP of SCORE, we apply Theorem 4.1 with $U = V^{(m)}(\Gamma)$, $x_i = r_i^{(m)}(\Gamma)$, and $\hat{x}_i = \hat{r}_i^{(m)}$, and the main condition we need is $c_1 \leq d_m(V^{(m)}(\Gamma)) \leq c_2$ uniformly for all Γ. But by Lemma 4.3, this is implied, so we do not need extra conditions to show the NSP. If however we use a "weaker" version of the k-means theorem, then we need conditions such as $d_K(V^{(m)}(\Gamma)) \geq C$ for all Γ (as explained above, the condition can be violated easily). The formal proof of the NSP (i.e., Theorem 3.2) is given in Section B.1 of the supplement, where we combine Lemmas 4.1-4.3, Theorem 4.1, and some elementary probability.

Theorem 4.1 is quite general and may be useful for many other unsupervised learning settings (e.g., [11]). The proof of the theorem is non-trivial and we now briefly explain the reason. As the objective function of the k-means is nonlinear and we do not have an explicit formula for the k-means solution, we prove by contradiction. Let $\hat{\ell}$ be the estimated cluster label vector by k-means and $RSS(\hat{\ell})$ be the associated objective function, we aim to show that, when NSP does not hold for $\hat{\ell}$, we can always find a cluster label vector ℓ such that $RSS(\ell) < RSS(\hat{\ell})$ (a contradiction). The key is finding such an ℓ and evaluating $RSS(\ell)$. However, except for a lower bound on $d_m(U)$, we have little information about the K true cluster centers. Since $d_K(U)$ can take any value in $[0, d_m(U)]$, a pair of true cluster centers may be well-separated, moderately close, sufficiently close, or exactly overlapping (correspondingly, their distance is much larger than, comparable with, or much smaller than $\max_{1\leq i\leq n} \|\hat{x}_i - x_i\|$, or exactly zero). With the *infinitely many* configurations of true cluster centers, the main challenge in the proof is pinning down a strategy of constructing ℓ that guarantees a decrease

of RSS for *every* possible configuration. One might think that the oracle k-means solution ℓ^* (k-means applied to $x_1, x_2, \ldots, x_n$) can help guide the construction of ℓ, but unfortunately this does not work: first, we do not have an explicit form of ℓ^*; second, in some of our settings, $\hat{\ell}$ can be significantly different from ℓ^*. The way we construct ℓ and evaluate $RSS(\ell)$ subtly utilizes the definition of $d_m(U)$ and properties of k-means objective, which is highly non-trivial (see the supplemental material). Note that while [38, 34] proved special cases of the "weaker" version of the k-means theorem, they used assumptions (i) true cluster centers are mutually well separated, (ii) the oracle solution ℓ^* is mathematically tractable, and (iii) $\hat{\ell}$ is exactly the same as ℓ^*. As none of (i)-(iii) holds in our setting, it is unclear how to generalize their proofs. We deal with a much harder setting (the "stronger" version), and our proof is different.

We conjecture that Theorem 4.1 (and so the NSP of SCORE) continues to hold if we replace the k-means step in SCORE by (say) the ϵ-approximation k-means (e.g., [26]). Let $\hat{\ell}$ be the ϵ-approximate k-means solution. We have $RSS(\hat{\ell}) \leq (1+\epsilon)\min_\ell RSS(\ell)$. For an appropriately small ϵ, if the NSP does not hold, then by a similar proof as that of Theorem 4.1, we can first construct an $\tilde{\ell}$ such that $RSS(\tilde{\ell}) < RSS(\hat{\ell}) - O(d_m(U))$, and then use it to deduce a contradiction. For reasons of space, we leave this to future.

5 Simulations

In Experiments 1-3, we compare StGoF with the BIC approach [38],[7] the ECV approach [30], and the NCV approach [4]. We use the R package "randnet" to implement these other methods. In Experiment 4, we compare StGoF with the RPLR approach [34]. In Experiment 5, we consider settings with comparably larger values of K. In all simulations, we fix $\alpha = 0.05$ in StGoF. Given (n, K), a scalar $\beta_n > 0$ that controls the sparsity, a symmetric non-negative matrix $P \in \mathbb{R}^{K\times K}$, a distribution $f(\theta)$ on $(0,\infty)$, and a distribution $g(\pi)$ on the standard simplex of $\mathbb{R}^K$, we generate the adjacency matrix $A \in \mathbb{R}^{n,n}$ as follows: First, generate $\tilde{\theta}_1, ..., \tilde{\theta}_n$ *iid* from $f(\theta)$. Let $\theta_i = \beta_n\tilde{\theta}_i/\|\tilde{\theta}\|$ and $\Theta = \mathrm{diag}(\theta_1, ..., \theta_n)$. Next, generate $\pi_1, ..., \pi_n$ iid from $g(\pi)$, and let $\Pi = [\pi_1, ..., \pi_n]'$. Last, let $\Omega = \Theta\Pi P\Pi'\Theta$ and generate A from Ω, for 100 times independently. For each algorithm, we measure the performance by the fraction of times it

[7] [38] primarily focused on the SBM model. Their algorithm has an ad-hoc extension to DCBM, which has no theoretical guarantee. We use this extension, instead of the original BIC approach.

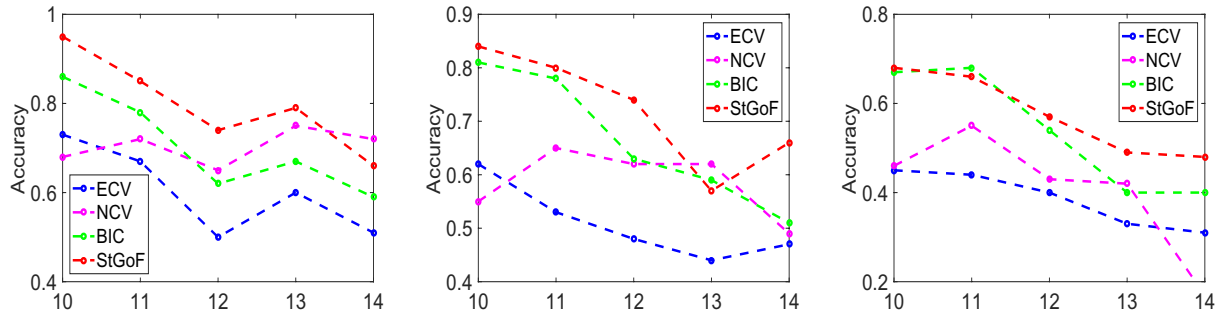

Figure 3: Experiment 1a (left), 1b (middel) and 1c (right), where from 1a to 1c, the degree heterogeneity is increasingly more severe. For all three panels, x-axis is $\|\theta\|$ (sparsity level), and y-axis is the estimation accuracy over 100 repetitions ($(n, K) = (600, 4)$).

correctly estimates K (i.e., accuracy). Note that $\|\theta\| = \beta_n$, and SNR $\asymp \|\theta\|(1 - b_n)$. For the experiments, we let β_n range so to cover many different sparsity levels, but keep $\|\theta\|(1 - b_n)$ fixed (so the problem of estimating K is not too difficult or too easy; see details below).

Experiment 1. We study how degree heterogeneity affect the results and comparisons. Fixing $(n, K) = (600, 4)$, we let $P \in \mathbb{R}^{4,4}$ be a Toeplitz matrix with $P(k, \ell) = 1 - [(1 - b_n)(|k - \ell| + 1)]/K$ in the off-diagonal and 1 in the diagonal. Let $g(\pi)$ be the uniform distribution over e_1, e_2, e_3, e_4 (standard basis vectors). We consider three sub-experiments, Exp 1a-1c. In these sub-experiments, we keep $(1 - b_n)\|\theta\|$ fixed at 9.5 so the SNR's are roughly at the same level. We let β_n range from 10 to 14 so to cover both the more sparse and the more dense cases. Moreover, for the three sub-experiments, we take $f(\theta)$ to be Unif$(2, 3)$, Pareto$(8, .375)$ (8 is the shape parameter and .375 is the scale parameter), and two point mixture $0.95\delta_1 + 0.05\delta_2$ (δ_a is a point mass at a), respectively (from Exp 1a to Exp 1c, the degree heterogeneity gets increasingly more severe). See Figure 3. StGoF consistently outperforms other approaches.

Experiment 2. We study how the relative sizes of different communities affect the results and comparisons. Given $b_n > 0$, we set $(n, K) = (1200, 3)$, $f(\theta)$ as Pareto$(10, 0.375)$, and let P be such that $P(k, \ell) = 1 - \frac{|k-\ell|(1-b_n)}{2}$, for $1 \le k, \ell \le 3$. We let β_n range in $\{12, 13, ..., 17\}$ and keep $(1 - b_n)\|\theta\|$ fixed at 10 so the SNR's are roughly at the same level. We take $g(\pi)$ as the distribution with weights a, b, and $(1 - a - b)$ on vectors e_1, e_2, e_3, respectively. Consider three sub-experiments, Exp 2a-2c, where we take $(a, b) = (.30, .35), (.25, .375)$, and $(.20, .40)$, respectively, so three communities are slightly unbalanced, moderately unbalanced, and very unbalanced, respectively. See Figure 4. First, StGoF consistently outperforms NCV, ECV and

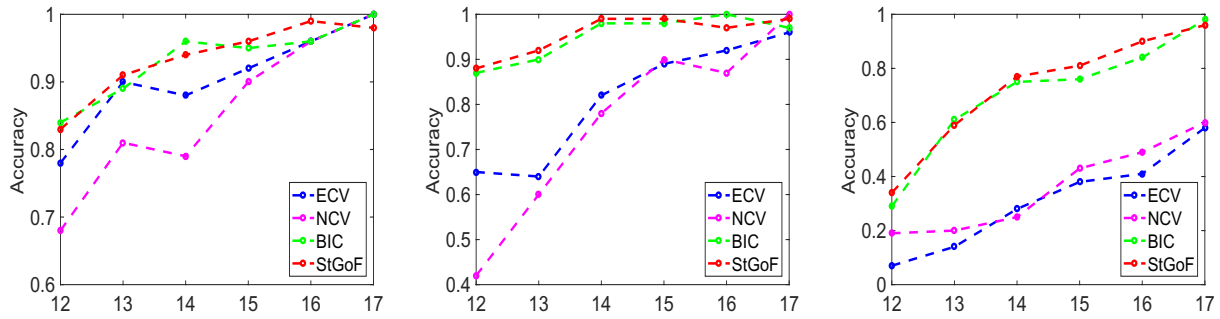

Figure 4: Experiment 2a (left), 2b (middle), and 2c (right), where from 2a to 2c, the communities sizes are more and more unbalanced. For all three panels, x-axis is $\|\theta\|$ (sparsity level), and y-axis is the estimation accuracy over 100 repetitions ($(n, K) = (1200, 3)$).

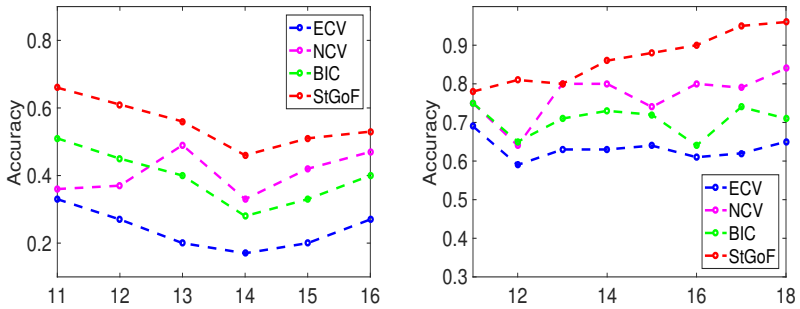

Figure 5: Experiment 3a (left) and 3b (right), where 3a allows for mixed memberships and 3b allows for outlier nodes. For both panels, x-axis is $\|\theta\|$ (sparsity level), and y-axis is the estimation accuracy over 100 repetitions ($(n, K) = (600, 4)$).

BIC. Second, when three communities get increasingly unbalanced, all methods become less accurate, suggesting that estimating K gets increasingly harder. Last, the performances of ECV and NCV are close to that of StGoF when communities are relatively balanced (e.g., Exp 2a), but is more unsatisfactorily when communities are more unbalanced (e.g., Exp 2b-2c).

Experiment 3. We study robustness of the algorithms under model misspecification. Fix $(n, K) = (600, 4)$. Let P have unit diagonals and $P(k, \ell) = 1 - \frac{(1-b_n)(|k-\ell|+1)}{K}$ as off-diagonals. Let $f(\theta)$ be Unif$(2, 3)$. We consider two sub-experiments, Exp 3a-3b. For sparsity, we let β_n range from 11 to 16 in Exp 3a and from 11 to 18 in Exp 3b, while fixing $(1-b_n)\|\theta\| = 10.5$. In Exp 3a, we allow mixed-memberships. Let $g(\pi)$ to be the mixing distribution with probability .2 on each of e_1, e_2, e_3, e_4 and probability .2 on Dirichlet($\mathbf{1}_4$). Once we have θ_i, π_i, and P, let $\Omega_{ij} = \theta_i\theta_j\pi_i' P\pi_j$, similar to that in DCBM. In Exp 3b, we allow outliers. Let $g(\pi)$ be the mixing distribution with a point mass .25 on each of e_1, e_2, e_3, e_4, and obtain Ω as in

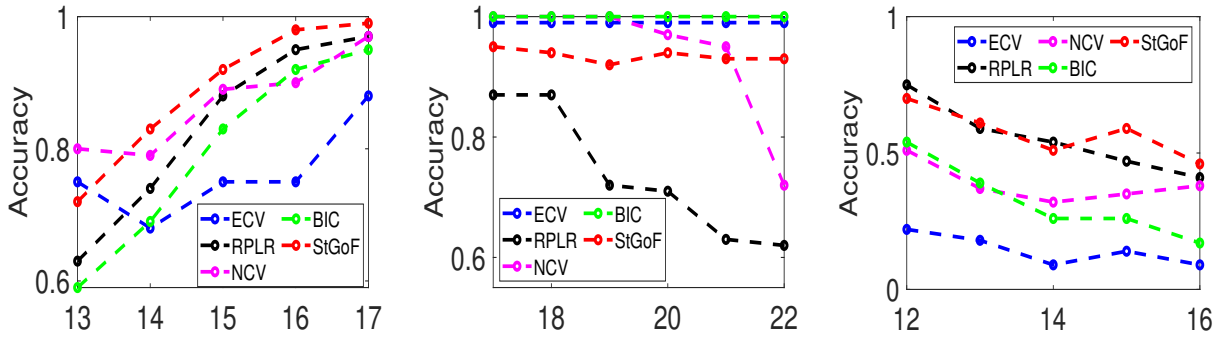

Figure 6: Experiment 4a (left), 4b (middle), and 4c (right), where $(n, K) = (600, 3), (1200, 3), (1200, 4)$ respectively. For all three panels, x-axis is $\|\theta\|$ (sparsity level), and y-axis is the estimation accuracy over 100 repetitions.

DCBM. Let $\rho_n = \frac{1}{n}\sum_{1\le i,j\le n}\Omega_{ij}$. We then randomly select 10% of nodes and re-set $\Omega_{ij} = \rho_n$ if either of (i, j) is selected. ECV and NCV are not model based so should be less sensitive to model misspecification; we use their results as benchmarks to evaluate StGoF and BIC. Figure 5 shows that StGoF is not sensitive to model misspecification, and that BIC behaves less satisfactory here than in Experiment 1-2, and so is more sensitive to model misspecification.

Experiment 4. We compare StGoF with RPLR [34] (and also BIC, ECV, and NCV). RPLR has tuning parameters $(K_{\max}, c_\eta, h_n)$. Following [34], we set $(K_{\max}, c_\eta, h_n) = (10, 1, \bar{d}^{-1/2})$, where $\bar{d}$ is the average node degree. Consider three sub-experiments, Exp 4a-4b, covering different combinations of (n, K, Θ, P). In Exp 4a, $(n, K) = (600, 3)$, $P(k, \ell) = 1 - \frac{(1-b_n)(|k-\ell|+1)}{K}$ if $k \neq \ell$ and 1 otherwise. We let $\|\theta\|$ vary and select b_n such that $(1-b_n)\|\theta\| = 9$, and let $f(\theta)$ be Unif$(2, 3)$. In Exp 4b, $(n, K) = (1200, 3)$, $P(k, \ell) = 1$ if $k = \ell$ and b_n otherwise. We let $\|\theta\|$ vary while keeping $(1-b_n)\|\theta\| = 4.75$, and let $f(\theta)$ be Unif$(3, 4)$. In Exp 4c, $(n, K) = (1200, 4)$, and P is the same as in Exp 4a. We let $\|\theta\|$ vary while keeping $(1 - b_n)\|\theta\| = 10.5$, and let $f(\theta)$ be Pareto$(10, .375)$. We take $g(\pi)$ to be the mixing distribution which puts probability .2 on each of e_1, e_2, e_3, e_4 and .2 on Dirichlet$(\mathbf{1}_4)$ (the model does not satisfy DCBM so we have a model misspecification). See Figure 6. RPLR underperforms StGoF, especially in Exp 4b (where the first two eigenvalues of Ω have a relatively large gap). This is because RPLR tends to estimate K as the index that has the largest eigen-gap. If the largest eigen-gap happens at an index smaller than K, RPLR tends to underestimate (see Section 6 for more discussion).

Experiment 5. We study two sub-experiments, Exp 5a-5b, for settings with a larger K. In Exp 5a, $(n, K) = (600, 6)$. We let P have 1 in the diagonal and $P(k, \ell) = 1 - \frac{(1-b_n)(|k-\ell|+K-1}{2K}$

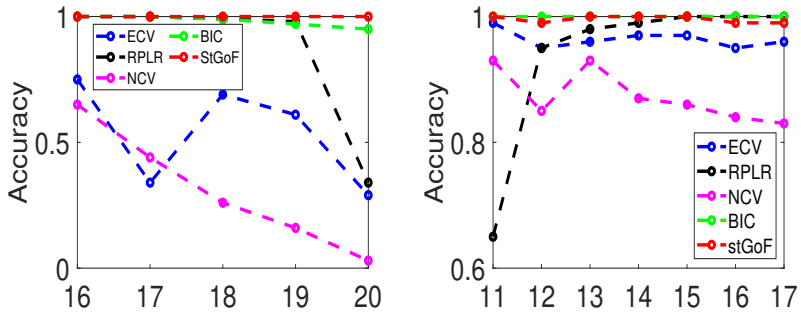

Figure 7: Experiment 5a (left) and 5b (right), where $(n, K) = (600, 6)$ for 5a and $(n, K) = (600, 8)$ for 5b. For both panels, x-axis is $\|\theta\|$ (sparsity level), and y-axis is the estimation accuracy over 100 repetitions.

in the off-diagonal, and take $f(\theta)$ as the two-point mixture $0.95\delta_1 + 0.05\delta_2$. We vary $\|\theta\|$ and select b_n such that $(1 - b_n)\|\theta\| = 15.5$. In Exp 5b, $(n, K) = (600, 8)$, P has unit diagonals and b_n in the off-diagonal, and $f(\theta)$ is the same as in Exp 5a. As $\|\theta\|$ vary, we select b_n such that $(1 - b_n)\|\theta\| = 10.5$. See Figure 7. Note that ECV and NCV are cross-validation approaches, which may be less satisfactory for larger K.

Remark 3. StGoF may estimate K incorrectly if some regularity conditions are violated. If this happens, we may either underestimate or overestimate K, depending on the data set. E.g., if the network has very weak signals (i.e., $|\lambda_K|/\sqrt{\lambda_1}$ is small), StGoF may underestimate K, and if the model is misspecified (say, due to many outliers), StGoF may overestimate K.

6 Real data analysis

In theory, a good approximation for the null distribution of $\psi_n^{(m)}$ is $N(0, 1)$ (see Theorem 3.1), but such a result requires some model assumptions, which may be violated in real applications (e.g., outliers, artifacts). We thus propose a modification of StGoF using the idea of empirical null [8]. Under model misspecification, a good approximation for the null distribution of $\psi_n^{(m)}$ is no longer $N(0, 1)$ (i.e., theoretical null), but $N(u, \sigma^2)$ (i.e., empirical null) for some $(u, \sigma) \neq (0, 1)$. Efron [8] argued that due to artifacts or model misspecification, the *empirical null* frequently works better for real data than the *theoretical null*. The problem is then how to estimate the parameters (u, σ^2) of the empirical null.

We propose a bootstrap approach to estimating (u, σ^2). Recall that $\hat{\lambda}_k$ is the k-th largest eigenvalue of A and $\hat{\xi}_k$ is the corresponding eigenvector. Fixing $N > 1$ and $m > 1$, letting

$\widehat{M}^{(m)} = \sum_{k=1}^{m} \hat{\lambda}_k \hat{\xi}_k \hat{\xi}_k'$ and let $\widehat{S}^{(m)} = A - \widehat{M}^{(m)}$. For $b = 1, 2, \ldots, N$, we simultaneously permute the rows and columns of $\widehat{S}^{(m)}$ and denote the resultant matrix by $\widehat{S}^{(m,b)}$. Truncating all entries of $(\widehat{M}^{(m)} + \widehat{S}^{(m,b)})$ at 1 at the top and 0 at the bottom, and denote the resultant matrix by $\widehat{\Omega}^{(b)}$. Generate an adjacency matrix $A^{(b)}$ such that for all $1 \le i < j \le n$, $A_{ij}^{(b)}$ are independent Bernoulli samples with parameters $\widehat{\Omega}_{ij}^{(b)}$ (we may need to repeat this step until the network is connected). Apply StGoF to $A^{(b)}$ and denote the resultant statistic by $Q_n^{(b)}$. We estimate u and σ by the empirical mean and standard deviation of $\{Q_n^{(b)}\}_{b=1}^N$, respectively. Denote the estimates by $\hat{u}^{(m)}$ and $\hat{\sigma}^{(m)}$, respectively. The bootstrap StGoF statistic is then $\psi_n^{(m,*)} = [Q_n^{(m)} - \hat{u}^{(m)}]/\hat{\sigma}^{(m)}$, $m = 1, 2, \ldots$, where $Q_n^{(m)}$ is the same as in (2.8). Similarly, we estimate K as the smallest integer m such that $\psi_n^{(m,*)} \le z_\alpha$, for the same z_α in StGoF. We recommend $N = 25$, as it usually gives stable estimates for $\hat{u}^{(m)}$ and $\hat{\sigma}^{(m)}$. We call this method the bootstrap StGoF (StGoF*).

We consider 6 data sets as in Table 6, which can be downloaded from http://www-personal.umich.edu/~mejn/netdata/. We now discuss the true K. For the dolphin network, it was argued in [32] that both $K = 2$ or $K = 4$ are reasonable. For UKfaculty network, we symmetrize the network by ignoring the directions of the edges. There are 4 school affiliations for the faculty members so we take $K = 4$. For the Football network, we take $K = 11$. The network was manually labelled as 12 groups, but the 12^{th} group only consists of the 5 "independent" teams that do not belong to any conference and do not form a conference themselves. For the Polbooks network, Le and Levina [28] suggest that $K = 3$, but it was argued by [22] that a more appropriate model for the network is a degree corrected mixed-membership (DCMM) model with two communities, so $K = 2$ is also appropriate.

We compare StGoF and StGoF* with the BIC [38], BH [28], ECV [30], NCV [4] and RPLR [34]. The first 4 methods are implemented via the R package "randnet". Among them, ECV and NCV are cross validation (CV) approaches and the results vary from one repetition to the other. Therefore, we run each method for 25 times and report the mean and SD. The StGoF* uses bootstrapping mean and standard deviation and is also random, but the SDs are 0 for five data sets. Most methods require a feasible range of K a priori (say, $\{1, 2, \ldots, k_{max}\}$, where k_{max} is a prescribed upper bound for K. For the 6 data sets considered here, the largest (true) K is 11, so we take $k_{max} = 15$.

In Section 5, we mention that RPLR tends to underestimate K if the largest eigen-gap

Name	n	K	BIC	BH	ECV	NCV	RPLR	StGoF	StGoF*
Dolphins	62	2, 4	2	2	3.08(.91) [2,5]	2.20(2.71) [1,15]	2	2	3
Football	115	11	10	10	11.28(.61) [11,13]	12.36(1.15) [11,15]	2	10	10
Karate	34	2	2	2	2.60(1.0) [1,6]	2.56(.58) [2,4]	*	2	2
UKfaculty	81	4	4	3	5.56(1.61) [3,11]	2.40(.28) [2,3]	4	4	4
Polblogs	1222	2	6	8	4.88(1.13) [4, 8]	2(0) [2, 2]	2	2*	2
Polbooks	105	2, 3	3	4	7.56(2.66) [2,15]	2.08(.71) [2,5]	3	5	2.4(.25) [2,3]

Table 1: Comparison of estimated K. Take ECV for Dolphins for example: for 25 independent repetitions, the mean and SD of estimated K are 3.08 and 0.91, ranging from 2 to 5 (the SDs of StGoF* are 0 for the first 5 data sets). For Karate, RPLR (with $k_{max} = 15$) reports an error message without an estimated K; the error messages disappear if we take $k_{max} = 5$, where the estimated K is 2. See the text for more discussion.

of Ω happens at an index smaller than K. This seems to be the case for Football, where RPLR significantly underestimates. RPLR also has a (seemingly fixable) coding issue: the code (generously shared by the authors) may report an error message and does not output an estimate for K (e.g., if we apply it to the Dolphins, Karate, and UKfaculty with $k_{max} = 15$ for 500 times, then in 63%, 100%, and 96% of the times respectively, the code reports an error and does not output an estimate for K). If we take $k_{max} = 10$ for Dolphins and UKfaculty and take $k_{max} = 5$ for Karate, then the error messages disappear and the estimated K are $2, 4, 2$ for Dophis, UKfaculty, and Karate, respectively.

The Polblogs network is suspected to have outliers, so most methods do not work well. For this particular network, the mean of StGoF is much larger than expected, so we choose to estimate K by the m that minimizes $\psi_n^{(m)}$ for $1 \leq m \leq 15$ (for this reason, we put a $*$ next to 2 in the table). Note that StGoF* correctly estimates K as 2. The Polbooks network is suspected to have a signifiant faction of mixed nodes [22], which explains why StGoF overestimates K. Fortunately, for both data sets, StGoF* estimates K correctly, suggesting that the bootstrapping means and standard deviations help standardize $Q_n^{(m)}$.

7 Discussions

How to estimate K is a fundamental problem in network analysis. We propose StGoF as a new stepwise algorithm for estimating K, which (a) has $N(0, 1)$ as its limiting null, (b) is uniformly consistent in a setting much broader than those considered in the literature, and (c) achieves the optimal phase transition. The results, especially (a) and (c), do not exist before. Analysis of stepwise algorithms of this kind is known to face challenges. We overcome

them by using a different stepwise scheme and by deriving sharp results, where the key is to prove the NSP of SCORE; we prove the NSP with new ideas and techniques.

We discuss some open questions. First, in this paper, we are primarily interested in DCBM, but the idea can be extended to the broader DCMM, where mixed-memberships exist. To this end, we need to replace SCORE by Mixed-SCORE [22] (an adapted version of SCORE for networks with mixed memberships), and modify the refitting step accordingly. In this case, whether NSP continues to hold is unclear, but we may have a revised version of NSP that holds for all pure nodes (i.e., nodes without mixed-memberships) and we can then use it to study the mixed nodes. The analysis of the resultant procedure is much more challenging so we leave it to the future. Second, in this paper, we assume K is fixed. For diverging K, the main idea of our paper continues to be valid, but we need to revise several things (e.g., definition of consistency and SNR, some regularity conditions, phase transition) to reflect the role of K. The proof for the case of diverging K can be much more tedious, but aside from that, we do not see a major technical hurdle. Especially, the NSP of SCORE continues to hold for a diverging K. Then, with some mild conditions, we can show that $\widehat{\Pi}^{(m)}$ has very few realizations, so the analysis of StGoF is readily extendable. That we assume K as fixed is not only for simplicity but also for practical relevance. For example, real networks may have hierarchical tree structure, and in each layer, the number of leaves (i.e., clusters) is small (e.g., [17]). Therefore, we have small K in each layer when we perform hierarchical network analysis. Also, the goal of real applications is to have interpretable results. For example, for community detection, results with a large K is hard to interpret, so we may prefer a DCBM with a small K to an SBM with a large K. In this sense, a small K is practically more relevant. Last, while the NSP of SCORE largely facilitates the analysis, it does not mean that StGoF ceases to work well once NSP does not hold; it is just harder to analyze in such cases. Our study suggests that StGoF continues to behave well even when NSP does not hold exactly. How to analyze StGoF in such cases is an interesting problem for the future.

Acknowledgments. The authors thank the Associate Editor and referees for very helpful comments. They also thank Shujie Ma for sharing the code of RPLR. The research of J. Jin is supported in part by NSF Grant DMS-2015469, and the research of Z. T. Ke is supported in part by NSF Grant DMS-1943902.

Supplemental Material: Supplemental material contains the proofs of main theorems

and lemmas.

References

[1] Bollobas, B. (1998). *Morden graph theory*, Volume 184. Springer Science & Business Media.

[2] Cammarata, L. and Z. T. Ke (2021). Power enhancement and phase transitions for global testing of mixed membership stochastic block models. *Manuscript*.

[3] Chen, E. Y., J. Fan, and X. Zhu (2020). Community network auto-regression for high-dimensional time series. *arXiv:2007.05521*.

[4] Chen, K. and J. Lei (2018). Network cross-validation for determining the number of communities in network data. *J. Amer. Statist. Assoc. 113*(521), 241–251.

[5] Daudin, J.-J., F. Picard, and S. Robin (2008). A mixture model for random graphs. *Stat. Comput. 18*(2), 173–183.

[6] Davis, C. and W. M. Kahan (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM J. Numer. Anal. 7*(1), 1–46.

[7] Donoho, D. and J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist. 32*, 962–994.

[8] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc. 99*(465), 96–104.

[9] Fan, J., Y. Fan, X. Han, and J. Lv (2019). SIMPLE: Statistical inference on membership profiles in large networks. *J. R. Stat. Soc. Ser. B. (to appear)*.

[10] Gao, C. and J. Lafferty (2017). Testing for global network structure using small subgraph statistics. *arXiv:1710.00862*.

[11] Han, X., X. Tong, and Y. Fan (2021). Eigen selection in spectral clustering: a theory guided practice. *J. Amer. Statist. Assoc. (to appear)*.

[12] Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning* (2nd ed.). Springer.

[13] Horn, R. and C. Johnson (1985). *Matrix Analysis*. Cambridge University Press.

[14] Hu, J., J. Zhang, H. Qin, T. Yan, and J. Zhu (2021). Using maximum entry-wise deviation to test the goodness-of-fit for stochastic block models. *J. Amer. Statist. Assoc. 116*(535), 1373–1382.

[15] Huang, D., X. Zhu, R. Li, and H. Wang (2021). Feature screening for network autoregression model. *Statist. Sinica 31*, 1239–1259.

[16] Huang, S., H. Weng, and Y. Feng (2020). Spectral clustering via adaptive layer aggregation for multi-layer networks. *arXiv:2012.04646*.

[17] Ji, P. and J. Jin (2016). Coauthorship and citation networks for statisticians (with discussions). *Ann. Appl. Statist. 10*, 1779–1812.

[18] Jiang, B., J. Li, and Q. Yao (2020). Autoregressive networks. *arXiv:2010.04492*.

[19] Jin, J. (2015). Fast community detection by SCORE. *Ann. Statist. 43*(1), 57–89.

[20] Jin, J., T. Ke, and J. Liang (2021). Sharp impossibility results for hyper-graph testing. *Advances in Neural Information Processing Systems 34*.

[21] Jin, J., Z. T. Ke, and S. Luo (2018). Network global testing by counting graphlets. In *International Conference on Machine Learning*, pp. 2333–2341. PMLR.

[22] Jin, J., Z. T. Ke, and S. Luo (2021a). Estimating network memberships by simplex vertex hunting. *arXiv:1708.07852*.

[23] Jin, J., Z. T. Ke, and S. Luo (2021b). Optimal adaptivity of signed-polygon statistics for network testing. *Ann. Statist. 49*(6), 3408–3433.

[24] Jin, J., Z. T. Ke, and W. Wang (2017). Phase transitions for high dimensional clustering and related problems. *Ann. Statist. 45*(5), 2151–2189.

[25] Karrer, B. and M. Newman (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E 83*(1), 016107.

[26] Kumar, A., Y. Sabharwal, and S. Sen (2004). A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pp. 454–462. IEEE.

[27] Latouche, P., E. Birmele, and C. Ambroise (2012). Variational bayesian inference and complexity control for stochastic block models. *Stat. Model. 12*(1), 93–115.

[28] Le, C. M. and E. Levina (2015). Estimating the number of communities in networks by spectral methods. *arXiv:1507.00827*.

[29] Lei, J. (2016). A goodness-of-fit test for stochastic block models. *Ann. Statist. 44*(1), 401–424.

[30] Li, T., E. Levina, and J. Zhu (2020). Network cross-validation by edge sampling. *Biometrika 107*(2), 257–276.

[31] Liu, F., D. Choi, L. Xie, and K. Roeder (2017). Global spectral clustering in dynamic networks. *Proc. Natl. Acad. Sci. 115*(5), 927–932.

[32] Liu, W., X. Jiang, M. Pellegrini, and X. Wang (2016, 03). Discovering communities in complex networks by edge label propagation. *Scientific Reports 6*, 22470.

[33] Liu, Y., Z. Hou, Z. Yao, Z. Bai, J. Hu, and S. Zheng (2019). Community detection based on the l_∞ convergence of eigenvectors in DCBM. *arXiv:1906.06713*.

[34] Ma, S., L. Su, and Y. Zhang (2021). Determining the number of communities in degree-corrected stochastic block models. *J. Mach. Learn. Res. 22*(69), 1–63.

[35] Rohe, K., S. Chatterjee, and B. Yu (2011, 08). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist. 39*(4), 1878–1915.

[36] Saldana, D. F., Y. Yu, and Y. Feng (2017). How many communities are there? *J. Comput. Graph Stat. 26*(1), 171–181.

[37] Tang, M., J. Cape, and C. E. Priebe (2021). Asymptotically efficient estimators for stochastic blockmodels: The naive MLE, the rank-constrained MLE, and the spectral. *Bernoulli Journal (to appear)*.

[38] Wang, Y. R. and P. J. Bickel (2017). Likelihood-based model selection for stochastic block models. *Ann. Statist. 45*(2), 500–528.

[39] Yuan, Y. and A. Qu (2021). Community detection with dependent connectivity. *Ann. Statist. 49*(4), 2378–2428.

[40] Zhang, J. and Y. Chen (2020). Modularity based community detection in heterogeneous networks. *Statist. Sinica 30*(2), 601–629.

[41] Zhang, J., W. W. Sun, and L. Li (2020). Mixed-effect time-varying network model and application in brain connectivity analysis. *J. Amer. Statist. Assoc. 115*(532), 2022–2036.

[42] Zhao, Y., E. Levina, and J. Zhu (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist. 40*(4), 2266–2292.

[43] Zhu, X., X. Chang, R. Li, and H. Wang (2019). Portal nodes screening for large scale social networks. *J. Econometrics 209*(2), 145–157.