Network Global Testing by Counting Graphlets

Jiashun Jin [1] **Zheng Tracy Ke** [2] **Shengming Luo** [1]

Abstract

Consider a large social network with possibly severe degree heterogeneity and mixed-memberships. We are interested in testing whether the network has only one community or there are more than one communities. The problem is known to be non-trivial, partially due to the presence of severe degree heterogeneity. We construct a class of test statistics using the numbers of short paths and short cycles, and the key to our approach is a general framework for canceling the effects of degree heterogeneity. The tests compare favorably with existing methods. We support our methods with careful analysis and numerical study with simulated data and a real data example.

1. Introduction

Given a large symmetrical network, we are interested in the global testing problem where we use the adjacency matrix of the network to test whether the network consists of only one community or that it consists of multiple communities, where some nodes may have mixed memberships.

Real networks frequently have *severe degree heterogeneity*. The Stochastic Block Model (SBM) is well-known, but does not accommodate severe degree heterogeneity. To tackle the problem, Karrer and Newman (2011) proposed the Degree-Corrected Block Model (DCBM). DCBM strikes a better balance between theory and practice than SBM, and has become increasingly more popular recently.

We adopt a *Degree-Corrected Mixed-Membership (DCMM)* model (Jin et al., 2017). DCMM can be viewed as an extension of DCBM, but allows for mixed memberships. Suppose the network has n nodes and K perceivable communities

$$\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K.$$

For each node, we assign a Probability Mass Function (PMF) $\pi_i = (\pi_i(1), \pi_i(2), \cdots, \pi_i(K))'$, where $\pi_i(k)$ is the "weight" that node i puts on community $\mathcal{C}_k$, $1 \leq k \leq K$. We call node i "pure" if π_i is degenerate and "mixed" otherwise. Let $A \in \mathbb{R}^{n,n}$ be the adjacency matrix, where $A_{ij} = 1$ if nodes i and j have an edge, and $A_{ij} = 0$ otherwise (all diagonal entries of A are 0 as we don't treat a node as connecting to itself). In DCMM, we think the upper triangle of A contains independent Bernoulli random variables. Moreover, for n degree heterogeneity parameters $\theta_1, \theta_2, \ldots, \theta_n$ and a non-singular, irreducible matrix $P \in \mathbb{R}^{K,K}$,

$$\mathbb{P}(A_{ij} = 1) = \theta_i \theta_j \sum_{k=1}^{K} \sum_{\ell=1}^{K} \pi_i(k) \pi_j(\ell) P_{k\ell}. \qquad (1)$$

Also, we assume (a) all diagonals of P are 1, and (b) each of the K communities has at least one pure node. With such constraints, DCMM is identifiable (Jin et al., 2017).

Let Θ be the $n \times n$ diagonal matrix $\Theta = \mathrm{diag}(\theta_1, \ldots, \theta_n)$ and let Π be the $n \times K$ matrix $\Pi = [\pi_1, \pi_2, \ldots, \pi_n]'$. Let $\Omega = \Theta \Pi P \Pi' \Theta$ (recall that $P \in \mathbb{R}^{K,K}$):

$$\Omega = \begin{bmatrix} \theta_1 & & \\ & \ddots & \\ & & \theta_n \end{bmatrix} \begin{bmatrix} \pi_1' \\ \vdots \\ \pi_n' \end{bmatrix} P \, [\pi_1, \ldots, \pi_n] \begin{bmatrix} \theta_1 & & \\ & \ddots & \\ & & \theta_n \end{bmatrix}.$$

Let $\mathrm{diag}(\Omega)$ be the diagonal matrix where the i-th diagonal entry is Ω_{ii}, and let $W = A - [\Omega - \mathrm{diag}(\Omega)]$. We have

$$A = [\Omega - \mathrm{diag}(\Omega)] + W = \text{“}signal\text{”} + \text{“}noise\text{”}.$$

Remark 1. Many recent works use the "Random Degree Parameter (RDP)" model (which is narrower than ours): fixing a scaling parameter $\alpha_n > 0$ and a density function f over $(0, \infty)$ where the first a few moments of f are finite, and especially the second moment is 1, we assume $(\theta_i/\alpha_n) \stackrel{iid}{\sim} f$, $i = 1, 2, \ldots, n$. We call the resultant DCMM model the DCMM-RDP model. In applications where we don't know how θ_i's are correlated, it is preferable to treat θ as non-random as before, and it is safer to use the original DCMM than DCMM-RDP.

The testing problem can be cast as testing a null hypothesis $H_0^{(n)}$ against a specific hypothesis $H_1^{(n)}$ in its complement:

$$H_0^{(n)} : K = 1 \qquad vs. \qquad H_1^{(n)} : K > 1, \qquad (2)$$

*Equal contribution [1]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, USA [2]Department of Statistics, University of Chicago, Chicago, USA. Correspondence to: Jiashun Jin <jiashun@stat.cmu.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

where under $H_1^{(n)}$, DCMM holds for some eligible $(P, \theta_1, ..., \theta_n, \pi_1, ..., \pi_n)$. Note that under $H_0^{(n)}$, $P = 1$, $\pi_1, ..., \pi_n$ are all degenerate, and $\mathbb{P}(A(i,j) = 1) = \theta_i \theta_j$.

Most existing literatures for the testing problem (2) have been focused on the special case where

- *No degree heterogeneity*. $\theta_1 = \theta_2 = \ldots = \theta_n$.
- *No mixed-membership*. All π_i are degenerate PMFs.

Many tests were proposed for this special case, including but are not limited to the likelihood ratio test approach (Wang & Bickel, 2017) and the spectral approach (Bickel & Sarkar, 2016; Lei, 2016; Banerjee & Ma, 2017).

However, in our setting, $\theta_1, \theta_2, \ldots, \theta_n$ vary significantly from one to another, and it is unclear how to extend the methods above to the current setting. The likelihood ratio test is not applicable, for there are many unknown parameters $(\theta_1, \pi_1), (\theta_2, \pi_2), ..., (\theta_n, \pi_n)$. The spectral approach also faces challenges, because the distributions of such test statistics depend on unknown parameters in a complicated way, and it is nontrivial to figure out the rejection region.

Also, it may be tempting to adapt the recent approaches on estimating K to our testing problem (Saldana et al., 2017; Chen & Lei, 2017; Le & Levina, 2015), but similarly, due to severe degree heterogeneity, the null distributions of such statistics are not tractable, so they cannot be used directly.

Recently, Gao and Lafferty (2017) (see also Bubeck et al. (2016) which is related but on different settings) proposed a new test called the Erdős-Zuckerberg (EZ) test for DCMM, with the following constraints.

- (GL1) No mixed-membership: all π_i are degenerate.
- (GL2) The community labels are uniformly drawn so the K communities have roughly equal sizes.
- (GL3) The DCMM-RDP holds (i.e., θ_i are iid samples), and the $K \times K$ matrix P has the special form of

$$P = \begin{bmatrix} a & b \cdots & b \\ \vdots & \ddots & \vdots \\ b & b \cdots & a \end{bmatrix}.$$

Note that the last bullet point is particularly restrictive.

Gao and Lafferty (2017) made an interesting observation that the effect of the degree heterogeneity parameters $\theta_1, \theta_2, \ldots, \theta_n$ is largely canceled out in the EZ test, and the test statistic approximately equals to 0 under the null, regardless of what $\theta_1, \theta_2, \ldots, \theta_n$ are. This allows us to find a convenient way to map out the rejection region.

Unfortunately, the authors did not make it clear whether the cancellation is "coincidental" and is due to the symmetry they imposed on the model (see GL1-GL3), or is "inherent" and holds for much broader settings.

In this paper, we introduce a class of test statistics by counting the number of graphlets in the network. Fixing a small $m \geq 1$, we count two kinds of graphlets: length-m paths and m-cycles. Our main contributions are as follows:

- *Ideation*. For a K-vector η and a $K \times K$ matrix $G^{1/2} P G^{1/2}$ to be introduced, we derive succinct proxies for the number of length-m paths and m-cycles, using η and the eigenvalues and eigenvectors of $G^{1/2} P G^{1/2}$. The proxies motivate a systematic way of constructing tests where the degree heterogeneity is largely removed so the distributions are more tractable.

  To the best of our knowledge, such proxies have not been discovered in the literature.
- *Methods and theory*. We propose a class of graphlet count (GC) test statistics, and derive their asymptotic distributions, under the null and alternative hypotheses.

  We try to be as general as possible, and our methods and theory are for DCMM with minimal constraints.

The way we construct our statistics is to use the proxies aforementioned, and thus is different from that in Gao and Lafferty (2017), which uses calculations that heavily depend on the imposed constraints GL1-GL3. See Section 3.2 for more comparison.

Our findings support the philosophy of Jin (2015) which introduced the community detection algorithm SCORE. Jin (2015) pointed out that $\theta_1, \theta_2, \ldots, \theta_n$ are required to model severe degree heterogeneity, but they turn out to be nuisance parameters, the effects of which can be largely removed with a *proper construction* of statistics.

While our tests are designed for global testing, the idea is also useful for tackling other problems. For example, we can combine our idea with those in community detection (e.g. Jin (2015), Chen et al. (2018), Qin and Rohe (2013)) to estimate the number of communities K.

2. A Class of Graphlet Count (GC) Statistics

The testing problem (2) is hard for there are so many unknown parameters: $P, \theta_1, \ldots, \theta_n, \pi_1, \ldots, \pi_n$. The parameters $\theta_1, \theta_2, \ldots, \theta_n$, which are required to model the severe degree heterogeneity of real networks, are especially hard to deal with for they vary significantly from one to another. What we need is therefore a smart test statistic that

- does not vary significantly as $\theta = (\theta_1, \theta_2, \ldots, \theta_n)'$ varies, and has a tractable limiting distribution (so it is easy to map out the rejection region),
- is powerful in differentiating the null and alternative.

Our idea is to use the graphlet-count statistics. In a network, we say a path is a "self-avoiding path" if it doesn't intersect with itself, and a "cycle" if it is a closed path that does not intersect with itself.

Definition 2.1 *For $0 \le m \le n$, let $B_{n,m} = \prod_{s=0}^{m-1}(n-s)$.*

Definition 2.2 *For $m \ge 1$, we define the "density of length-m self-avoiding paths" by*

$$\widehat{L}_m = \frac{1}{B_{n,m+1}} \sum_{\substack{1\le i_1,\ldots,i_{m+1}\le n \\ i_1,\cdots,i_{m+1} \text{ are distinct}}} A_{i_1 i_2} A_{i_2 i_3} \cdots A_{i_m i_{m+1}}.$$

and for $m \ge 3$, we define the "density of m-cycles" by

$$\widehat{C}_m = \frac{1}{B_{n,m}} \sum_{\substack{1\le i_1,\ldots,i_{m}\le n \\ i_1,\cdots,i_{m} \text{ are distinct}}} A_{i_1 i_2} A_{i_2 i_3} \cdots A_{i_m i_1}.$$

We propose the family of test statistics, called the **Graphlet Count (GC)** test statistics:

$$\widehat{\chi}^{(m)}_{gc} = \widehat{C}_m - (\widehat{L}_{m-1}/\widehat{L}_{m-2})^m, \qquad m = 3, 4, \ldots. \quad (3)$$

Remark 2. Using the adjacency matrix A, $\widehat{C}_m$ and $\widehat{L}_m$ can be conveniently computed (e.g., $\widehat{C}_4 = \frac{1}{24\binom{n}{4}}\big[\mathrm{tr}(A^4) - 2(1_n' A^2 1_n) + 1_n' A 1_n\big]$). See supplemental material.

2.1. The Key Idea: Why the GC Test Statistics Work

Recall that $\Omega = \Theta\Pi P \Pi'\Theta$. Let 1_n be the n-dimensional vector of 1's and let $\theta = (\theta_1, \theta_2, \ldots, \theta_n)'$. We use $(\cdot,\cdot)$ to denote the inner product of two vectors. The $K \times K$ matrix $G \equiv \Pi'\Theta^2\Pi$ and the vector $\eta \in \mathbb{R}^K$ play a key role.

Definition 2.3 *Denote the vector $G^{-1/2}\Pi'\Theta 1_n$ by η.*

Definition 2.4 *For $1 \le k \le K$, let λ_k be the k-th largest (in absolute value) eigenvalue of $G^{1/2}PG^{1/2}$, and let ξ_k be the corresponding eigenvector.*

It turns out that $\lambda_1, \ldots, \lambda_K$ (eigenvalues of $G^{1/2}PG^{1/2}$) are the nonzero eigenvalues of Ω. The following results, which will be made precise in Theorems 3.1-3.2, play the key role:

$$n^m \cdot \widehat{C}_m \approx \mathrm{tr}(\Omega^m) = \sum_{k=1}^K \lambda_k^m, \quad (4)$$

$$n^{m+1} \cdot \widehat{L}_m \approx 1_n'\Omega^m 1_n = \sum_{k=1}^K (\eta,\xi_k)^2 \lambda_k^m. \quad (5)$$

We explain why these equations motivate the test statistic $\widehat{\chi}^{(m)}_{gc}$. Recall that we hope to have a statistic that does not vary too much as θ varies, so first, it is desirable to remove the terms $(\eta,\xi_k)^2$, which not only depend on θ but are also not very tractable. Under the alternative hypothesis, it is unclear how to cancel these terms, but under the null hypothesis, $K = 1$, and the right hand side of (5) reduces to $(\eta,\xi_1)^2\lambda_1^m$, and there are many ways to do the cancellation. One such way is to use the following ratio:

$$\frac{n^m \widehat{L}_{m-1}}{n^{m-1}\widehat{L}_{m-2}} \approx \frac{\sum_{k=1}^K (\eta,\xi_k)^2\lambda_k^{m-1}}{\sum_{k=1}^K (\eta,\xi_k)^2\lambda_k^{m-2}} \begin{cases} = \lambda_1, \text{under } H_0^{(n)}, \\ \le \lambda_1, \text{under } H_1^{(n)}. \end{cases} \quad (6)$$

Therefore, at least under $H_0^{(n)}$, we have managed to cancel the terms $(\eta,\xi_k)^2$.

Next, it is also desirable to cancel the term λ_1, at least under the null hypothesis. Comparing (4) and (6), there are many ways to do this, and one such way is to use the $\widehat{\chi}^{(m)}_{gc}$ statistic aforementioned:

$$n^m \widehat{\chi}^{(m)}_{gc} = n^m \widehat{C}_m - \big[(n^m \widehat{L}_{m-1})/(n^{m-1}\widehat{L}_{m-2})\big]^m.$$

In fact, by (4) and (6), we have that up to a negligible term (i.e., of a smaller order than that of the standard deviation of the statistic under $H_0^{(n)}$),

$$\widehat{\chi}^{(m)}_{gc} \begin{cases} = 0, & \text{under } H_0^{(n)}, \\ \ge \frac{1}{n^m}\sum_{k=2}^K \lambda_k^m, & \text{under } H_1^{(n)}. \end{cases}$$

On the one hand, the effects of degree heterogeneity are largely canceled in the statistic, so it does not vary significantly as the vector θ varies (this is particularly important for we wish to have a rejection region that is relatively insensitive to θ). On the other hand, the statistic $\widehat{\chi}^{(m)}_{gc}$ is able to differentiate the null hypothesis and the alternative hypothesis, through the term $\sum_{k=2}^K \lambda_k^m$. This suggests that $\widehat{\chi}^{(m)}_{gc}$ is a reasonable test statistic.

Note that $\widehat{\chi}^{(m)}_{gc}$ is only one of many test statistics with the desired properties above, but seemingly one of the simplest.

Remark 3. Recall that $\lambda_1, \ldots, \lambda_K$ are the eigenvalues of $G^{1/2}PG^{1/2}$, and they are also the eigenvalues of PG (nonnegative, irreducible). By Perron's theorem (Horn & Johnson, 1985), λ_1 is positive and $\lambda_1 > |\lambda_k|$ for all $2 \le k \le K$.

Remark 4. Our tests include the EZ test as a special case (i.e., $\widehat{\chi}^{(m)}_{gc}$ with $m = 3$), but our idea is by no means a straightforward extension of that in Gao and Lafferty (2017). The EZ test was derived by calculations that depend heavily on the constraints GL1-GL3 imposed on P, θ, etc., and it was unclear whether the core idea of the EZ test is only valid when these constraints hold, or in more general settings. The GC tests are derived by (4)-(6), where the relationship between the test statistics and θ, η, and the eigenvalues and eigenvectors of $G^{1/2}PG^{1/2}$ has not been discovered before, even in cases where the constraints GL1-GL3 hold.

Remark 5. Can we simply use $\sum_{k=2}^K \widehat{\lambda}_k^m$ as the test statistic, where $\widehat{\lambda}_k$ is the k-th eigenvalue of A? We can not, as K is unknown. Can we simply use $\widehat{\lambda}_2$ as the test statistic? We can, but the asymptotic distribution of $\widehat{\lambda}_2$ is much harder to derive than that of $\widehat{\chi}^{(m)}_{gc}$ (which is Gaussian; see below), so it is challenging to determine the rejection region.

3. Main Results

In theory, we use n as the driving asymptotic parameter, let the matrices (Θ, Π, P) change with n, and consider a

sequence of problems where we test $H_0^{(n)} : K = 1$ vs. $H_1^{(n)} : K > 1$ (K is unknown but does not change with n). Recall node i is pure if π_i is degenerate. A pure node can be in any of the K communities. For $1 \le k \le K$, we let

$$\mathcal{N}_k = \{1 \le i \le n : \pi_i \text{ is degenerate and } \pi_i(k) = 1\}$$

be the set of all pure nodes in the community k. Assume

$$\|\theta\| \to \infty, \qquad \|\theta\|_3 \to 0. \tag{7}$$

Since $\theta_{\max} \le \|\theta\|_3$, this implies $\theta_{\max} \to 0$. Suppose there is a constant $c_1 > 0$ so that for any $1 \le k \le K$,

$$\frac{\sum_{i\in\mathcal{N}_k}\theta_i^2}{\sum_{i=1}^n \theta_i^2} \ge c_1. \tag{8}$$

(this roughly says each community has sufficiently many pure nodes). Also, assume for some constant $c_2 \in (0,1)$,

$$\text{all singular values of } P \text{ fall between } c_2 \text{ and } c_2^{-1}. \tag{9}$$

Denote $C_m = \mathbb{E}[\widehat{C}_m]$ and $L_m = \mathbb{E}[\widehat{L}_m]$ and introduce a non-stochastic counterpart of $\widehat{\chi}_{gc}^{(m)}$ by

$$\chi_{gc}^{(m)} = C_m - (L_{m-1}/L_{m-2})^m. \tag{10}$$

Theorem 3.1 *Consider the DCMM model* (1) *where* (7)-(9) *hold. As* $n \to \infty$,

$$\chi_{gc}^{(m)} = \frac{1}{n^m}\left\{\sum_{k=1}^K \lambda_k^m - \left[\frac{\sum_{k=1}^K(\eta,\xi_k)^2\lambda_k^{m-1}}{\sum_{k=1}^K(\eta,\xi_k)^2\lambda_k^{m-2}}\right]^m\right\} + O(n^{-m}\|\theta\|_4^4\|\theta\|^{2m-4}).$$

The last term is of a smaller order of the standard deviation of $\widehat{\chi}_{gc}^{(m)}$ and is thus negligible. Theorem 3.1 solidifies what we mentioned in Section 2.1 and is proved in Section 6.

Theorem 3.2 *Consider the DCMM model* (1) *where* (7)-(9) *hold. As* $n \to \infty$, *for* $m = 3, 4$, *under either* $H_0^{(n)}$ *or* $H_1^{(n)}$,

$$\sqrt{\frac{B_{n,m}}{2m}}\cdot\widehat{C}_m^{-1/2}\left[\widehat{\chi}_{gc}^{(m)} - \chi_{gc}^{(m)}\right] \xrightarrow{d} N(0,1).$$

The proof for other fixed m is similar, but significantly more tedious so we leave it as future work. Theorem 3.2 is proved in the supplemental material. Compared to existing literature, our theorems are for a much broader setting where existing works have very little theory and understanding.

Remark 6. Conditions (8)-(9) are only for $H_1^{(n)}$ since they naturally hold under $H_0^{(n)}$; the conditions are only mild. Condition (7) is also only mild. Take the DCMM-RDP model for example (see Remark 1): $\|\theta\| \asymp \sqrt{n}\alpha_n$ and $\|\theta\|_3 \asymp n^{1/3}\alpha_n$, so Condition (7) requires $n^{-1/2} \ll \alpha_n \ll n^{-1/3}$. The case $\alpha_n \gg n^{-1/3}$ corresponds to the "strong signal" case, the analysis of which is different and we leave it to the forthcoming manuscript.

3.1. Testing Power

By Theorems 3.1-3.2, we expect to see that in distribution,

$$\sqrt{\frac{B_{n,m}}{2m}}\cdot\widehat{C}_m^{-1/2}\widehat{\chi}_{gc}^{(m)} \approx N\Big(\sqrt{\frac{B_{n,m}}{2m}}C_m^{-1/2}\chi_{gc}^{(m)},\ \ 1\Big).$$

Motivated by Theorem 3.1 and equation (4), we introduce a proxy of $\sqrt{\frac{B_{n,m}}{2m}}\cdot C_m^{-1/2}\chi_{gc}^{(m)}$, defined as

$$\delta_{gc}^{(m)} = \frac{(2m)^{-1/2}}{\sqrt{\sum_{k=1}^K\lambda_k^m}}\left[\sum_{k=1}^K\lambda_k^m - \left(\frac{\sum_{k=1}^K(\eta,\xi_k)^2\lambda_k^{m-1}}{\sum_{k=1}^K(\eta,\xi_k)^2\lambda_k^{m-2}}\right)^m\right], \tag{11}$$

and we expect to see that in distribution,

$$\sqrt{\frac{B_{n,m}}{2m}}\cdot\widehat{C}_m^{-1/2}\widehat{\chi}_{gc}^{(m)} \approx N(\delta_{gc}^{(m)}, 1). \tag{12}$$

It is noteworthy that under $H_0^{(n)}$, $\delta_{gc}^{(m)} = 0$.

Fixing $0 < \alpha < 1$, let z_α be the $(1-\alpha)$-quantile of $N(0,1)$. Consider the Graphlet Count (GC) test where we

$$\text{reject } H_0^{(n)} \iff \sqrt{\frac{B_{n,m}}{2m}}\cdot\widehat{C}_m^{-1/2}\widehat{\chi}_{gc}^{(m)} > z_\alpha. \tag{13}$$

The theorem below is proved in the supplemental material.

Theorem 3.3 *Consider the DCMM model* (1) *where* (7)-(9) *hold. As* $n \to \infty$, *for* $m = 3, 4$, *the level and the power of the Graphlet Count test are respectively* $\alpha + o(1)$ *and* $\Phi\big(\delta_{gc}^{(m)} - z_\alpha\big) + o(1)$. *Moreover, if* $\delta_{gc}^{(m)} \to \infty$ *as* $n \to \infty$, *then the power* $\to 1$.

3.2. Comparison of $\widehat{\chi}_{gc}^{(3)}$ and $\widehat{\chi}_{gc}^{(4)}$

One of the key messages is that, $\widehat{\chi}_{gc}^{(3)}$ may be powerless for some non-null cases, even when the "signals" are strong and the test is relatively easy. In comparison, $\widehat{\chi}_{gc}^{(4)}$ has successfully avoided such a pitfall. Note that translated to our terms, the EZ test by Gao and Lafferty (2017) is $\widehat{\chi}_{gc}^{(3)}$.

In detail, by Remark 3, λ_1 is positive and has the largest absolute value among all λ_k's, so

$$[\sum_{k=1}^K(\eta,\xi_k)^2\lambda_k^{m-1}]/[\sum_{k=1}^K(\eta,\xi_k)^2\lambda_k^{m-2}] \le \lambda_1. \tag{14}$$

It follows that under $H_1^{(n)}$

$$\delta_{gc}^{(m)} \ge (\sum_{k=2}^K\lambda_k^m)/[2m\sum_{k=1}^K\lambda_k^m]^{1/2}, \tag{15}$$

where "=" is achievable; see Section 3.3 for examples.

It can be shown that $\sum_{k=1}^{K}\lambda_k^m > 0$, no matter m is odd or even. The numerator $\sum_{k=2}^{K}\lambda_k^m$, however, is more tricky.

- When m is even, the numerator of (15) is positive.
- When m is odd, the numerator of (15) can be negative or 0. In fact, $\delta_{gc}^{(m)}$ may be 0 under some $H_1^{(n)}$, even when signals are strong; see Section 3.3 for an example. If additionally P (and so $G^{1/2}PG^{1/2}$) is positive definite, then $\delta_{gc}^{(m)}$ is positive, either m is odd or even.

Corollary 3.1 *Consider the DCMM model* (1) *where* (7)-(9) *hold and the Graphlet Count test* (13). *As* $n \to \infty$.

- *When* $m = 3$, *for some configurations of the non-null case,* $\delta_{gc}^{(3)} = 0$ *and the power of the test is* $\alpha + o(1)$. *If additionally the* $K \times K$ *matrix* P *is positive definite, then there is a constant* $c_3 > 0$ *such that* $\delta_{gc}^{(3)} \geq c_3\|\theta\|^3$ *and so the power of the test* $\gtrsim \Phi\big(c_3\|\theta\|^3 - z_\alpha\big)$, *which tends to* 1 *since* $\|\theta\| \to \infty$ *in our setting.*
- *If* $m = 4$, *then there is a constant* $c_4 > 0$ *such that* $\delta_{gc}^{(4)} \geq c_4\|\theta\|^4$ *and the power of the test* $\gtrsim \Phi\big(c_4\|\theta\|^4 - z_\alpha\big)$, *which tends to* 1 *as* $\|\theta\| \to \infty$ *in our setting.*

See the supplement for the proof. In comparison, $\widehat{\chi}_{gc}^{(3)}$ has a two-fold disadvantage: it may lose power in some non-null configurations even when $\|\theta\|$ is large, and as $\|\theta\| \to \infty$, its power grows to 1 in a speed slower than that of $\widehat{\chi}_{gc}^{(4)}$.

Remark 7. In the most subtle case where $\|\theta\| \asymp 1$, for any $m \geq 3$, $\widehat{\chi}_{gc}^{(m)}$ has a non-trivial power since the signal to noise ratio $\delta_{gc}^{(m)} \asymp \|\theta\|^m$. The form is reminiscent of the results on likelihood ratio (Mossel et al., 2015) which is unfortunately only for SBM (much narrower than DCMM).

Remark 8. The computational complexity of the GC test is $O(nd^2)$ for $m = 3$ and $O(nd^3)$ for $m = 4$ (Schank & Wagner, 2005), where d is the maximum degree. Many large networks are reasonably sparse where $d \ll n$, and the complexity is reasonably modest in such cases.

Remark 9. Our work is connected to Maugis et al. (2017), which characterizes the expected number of closed walks. However, their work does not study the expected number non-closed walks, and the standard deviations of close and non-closed walks, so how to apply their results to our setting is unclear. Our work is also closed to the notion of clustering coefficient (Holland & Leinhardt, 1971; Watts & Strogatz, 1998), which in our notation equals to $3\widehat{C}_3/\widehat{L}_2$. To use this as a test, the challenge is how to normalize the statistic properly so the limiting distribution is more tractable. Also, the test may lose power in some settings. In Section 3.3, we provide an example where $C_3/L_2 = \lambda_1 + (\sum_{k=2}^{K}\lambda_k^3)/\lambda_1^2$, so when $\sum_{k=2}^{K}\lambda_k^3 = 0$, asymptotically, the test is powerless while the GC test may still have good power.

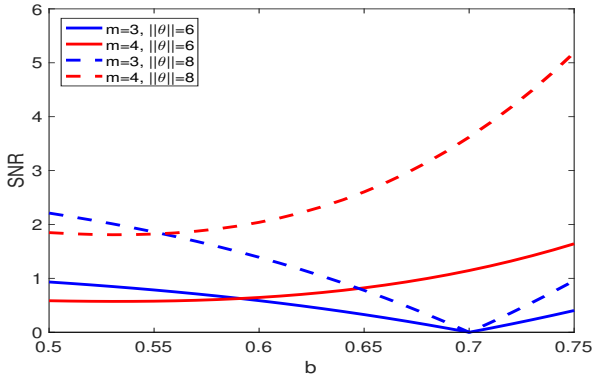

Figure 1. Plot of $\delta_{gc}^{(m)}$ in the setting of Section 3.3 ($(K, a) = (10, .25)$). When $b = .7$, $\delta_{gc}^{(3)} = 0$ even when $\|\theta\|$ is very large, and the EZ test is powerless. In comparison, $\widehat{\chi}_{gc}^{(4)}$ has overcome such a pitfall. Also, $\widehat{\chi}_{gc}^{(4)}$ is more powerful when $\|\theta\|$ is large.

3.3. An Example

It is instructive to consider an example where $\delta_{gc}^{(m)}$ can be further simplified. Consider a setting where

- the DCMM-RDP model holds (see Remark 1), i.e., $(\theta_i/\alpha_n) \overset{i.i.d.}{\sim} f$, where the $2nd$ moment of f is 1,
- all π_i's are degenerate, dividing to K equal-size communities,
- the rows of P have an equal sum.

It follows that approximately: (a). $\|\theta\| = \sqrt{n}\alpha_n$, (b). $G \propto$ the $K \times K$ identity matrix, so ξ_1 (the first eigenvector of $G^{1/2}PG^{1/2}$) is approximately proportional to the vector of ones 1_K. (c). $\eta = G^{-1/2}\Pi'\Theta 1_n \propto 1_K$, due to the random model for θ_i. Therefore, $(\eta, \xi_k) = (\xi_1, \xi_k)$, which equals to 1 if $k = 1$ and 0 otherwise, and so by basic linear algebra,

$$\delta_{gc}^{(m)} = \frac{(2m)^{-1/2}}{\sqrt{\sum_{k=1}^{K}\lambda_k^m}}\Big[\sum_{k=1}^{K}\lambda_k^m - \lambda_1^m\Big] = \frac{\sum_{k=2}^{K}\lambda_k^m}{\sqrt{2m\sum_{k=1}^{K}\lambda_k^m}}.$$

We now consider a setting where we can spell out λ_k more explicitly. Suppose K is even and the $K \times K$ matrix P calibrating the community structure is a 2×2 block-wise matrix having the form of $P = \begin{bmatrix} D & C \\ C & D \end{bmatrix}$, where $C, D \in \mathbb{R}^{K/2, K/2}$ and

$$D = \begin{bmatrix} 1 & a\cdots & a \\ \vdots & \ddots & \vdots \\ a & a\cdots & 1 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} b & b\cdots & b \\ \vdots & \ddots & \vdots \\ b & b\cdots & b \end{bmatrix}.$$

In this case, $\lambda_1 = \frac{n\alpha_n^2}{K}\big\{(1-a) + \frac{K}{2}(a+b)\big\}$, and the other $(K-1)$ eigenvalues are (which one is λ_2 depends on (a, b))

$$\frac{n\alpha_n^2}{K}[(1-a) + \frac{K}{2}(a-b)], \frac{n\alpha_n^2}{K}(1-a), \ldots, \frac{n\alpha_n^2}{K}(1-a).$$

If we let $A_K(a,b) = (1-a) + \frac{K}{2}(a-b)$ and $B_K(a,b) = (1-a) + \frac{K}{2}(a+b)$, recalling $\|\theta\| = \sqrt{n}\alpha_n$, then

$$\delta_{gc}^{(3)} = \frac{(K^{-\frac{3}{2}}\|\theta\|^3)\cdot[A_K^3(a,b) + (K-2)(1-a)^3]}{\sqrt{[A_K^3(a,b) + B_K^3(a,b) + (K-2)(1-a)^3}},$$

$$\delta_{gc}^{(4)} = \frac{(K^{-2}\|\theta\|^4)\cdot[A_K^4(a,b) + (K-2)(1-a)^4]}{\sqrt{A_K^4(a,b) + B_K^4(a,b) + (K-2)(1-a)^4}}.$$

Note that $\delta_{gc}^{(4)}$ is always positive, but $\delta_{gc}^{(3)}$ can be either positive or negative, and in particular,

$\delta_{gc}^{(3)} = 0$ when $b = w + (1-w)a$, where $w = \frac{1+(K-2)^{1/3}}{K/2}$.

Figure 1 compares $\delta_{gc}^{(3)}$ and $\delta_{gc}^{(4)}$ for $(K,a) = (10,.25)$ and various $(b,\|\theta\|)$. Regardless of $\|\theta\|$, $\delta_{gc}^{(3)}$ gets small when b is close to 0.7. However, $\delta_{gc}^{(4)}$ has no such issue.

3.4. The Lower Bound

The lower bound is not discussed here but is studied in the extended version (Jin et al., 2018), where we show that under DCMM, if $\|\theta\| \le c$ for a sufficiently small constant, then the risk (sum of type-I and type-II errors) of any test converges to 1 as $n \to \infty$. See also Massoulié (2014), Mossel (2015), Abbe & Sandon (2016), Gao & Lafferty (2017). Note by Corollary 3.1, if the level $\alpha \to 0$, then the risk of the GC test $\to 0$ as $\|\theta\| \to \infty$. A closely related work is Jin and Ke (2017). Under the DCMM and the assumption of $\theta_{\max} \le C\theta_{\min}$, they showed that when $\|\theta\| \not\to \infty$ it is impossible to successfully estimate the mixed memberships.

4. Simulations

We investigate $\widehat{\chi}_{gc}^{(m)}$ for $m = 3, 4$. Recall that when $m = 3$, it coincides with the EZ test (Gao & Lafferty, 2017). The methods (Bickel & Sarkar, 2016; Lei, 2016; Banerjee & Ma, 2017; Wang & Bickel, 2017) are for SBM, which do not apply to our settings and so we skip them for study.

Experiment 1 (checking for asymptotic normality). Fixing $n = 200$, we consider a null setting where θ_i's are from $\sqrt{10}\theta_i \overset{iid}{\sim}$ Pareto$(4, 0.375)$;[1] (note: severe degree heterogeneity!) We also consider an alternative case where θ_i's are from $\sqrt{2}\theta_i \overset{iid}{\sim}$ Pareto$(4, 0.375)$, $K = 3$, and P is the matrix with unit diagonals and all off-diagonals are $1/3$. Among the 200 nodes, 180 are pure with 60 in each community, and the remaining 20 nodes have a mixed membership $(1/3, 1/3, 1/3)$. The results of 500 repetitions are in Figure 2, which suggest that the claimed asymptotic normality is valid even for relatively small n.

Experiment 2 (power comparison). Fix $(n,K) = (300, 10)$. All nodes are pure with 30 in each community. For (a,b) and $h > 0$, let the matrix P be the same as in Section 3.3. Set $\theta_i = (h/\|\tilde{\theta}\|)\tilde{\theta}_i$, where $\tilde{\theta}_i \overset{iid}{\sim} Pareto(4, 0.375)$; we note that $\|\theta\| = h$.

[1] In Pareto(α, x_m), α is for shape, x_m is for scale.

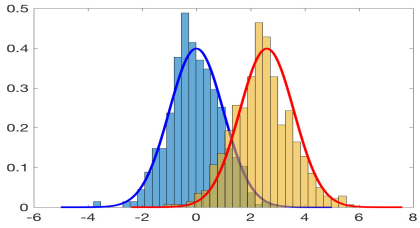

Figure 2. Histograms of $\sqrt{(B_{n,4}/8)}\cdot\widehat{C}_4^{-1/2}\widehat{\chi}_{gc}^{(4)}$ for the null hypothesis (light blue) and the alternative hypothesis (yellow). The blue and red curves are densities of $N(0,1)$ and $N(\delta_{gc}^{(4)},1)$, respectively, which support our results on asymptotical normality.

For $\|\theta\|$ ranging in $\{5, 6, \ldots, 10\}$, we consider three different settings for (a,b), (i)-(iii). See Table 1 for the results. In (i), $(a,b) = (.15,.52)$ and the second eigenvalue λ_2 of $G^{1/2}PG^{1/2}$ is moderately large, and $\widehat{\chi}_{gc}^{(4)}$ uniformly outperforms $\widehat{\chi}_{gc}^{(3)}$ (the EZ test). In (ii), $(a,b) = (.30,.54)$, λ_2 is relatively small and the testing problem is more challenging than (i). In this case, $\widehat{\chi}_{gc}^{(4)}$ performs better again. In (iii), we investigate a case where $\delta_{gc}^{(3)} = 0$ and see how $\widehat{\chi}_{gc}^{(3)}$ deals with this most challenging case (the case poses no challenge to $\widehat{\chi}_{gc}^{(4)}$ as $\delta_{gc}^{(4)}$ is always > 0; see Section 3.3). In this case, $\widehat{\chi}_{gc}^{(3)}$ loses power, and significantly underperforms $\widehat{\chi}_{gc}^{(4)}$.

These results are consistent with our theoretical results, especially those of Section 3.3.

Table 1. Powers of $\widehat{\chi}_{gc}^{(4)}$ (GC) and $\widehat{\chi}_{gc}^{(3)}$ (EZ).

(a,b)	$\|\theta\|$	5	6	7	8	9	10
$(.15,.52)$	GC	.13	.46	.77	.99	1.0	1.0
	EZ	.07	.17	.28	.64	.93	.99
$(.30,.54)$	GC	.11	.12	.27	.55	.85	.99
	EZ	.07	.10	.22	.25	.51	.78
$(.15,.66)$	GC	.09	.27	.72	.94	1.0	1.0
	EZ	.08	.05	.07	.06	.16	.43

5. Application to a Football Network

In the college football network (Girvan & Newman, 2002), each node is a Division I-A college team and two nodes have an edge if and only if they played ≥ 1 games during the Fall 2000 season. There are a total of 115 teams. Except for 5 "independent" teams, all teams are manually divided into 11 conferences for administration purposes; we treat these "manually labeled communities" as the ground truth.

First, we consider a relatively easy setting where we test whether the whole network has only 1 community or has multiple communities. As expected, for both $m = 3$ and 4, our test $\widehat{\chi}_{gc}^{(m)}$ rejects the null with extremely small p-values.

Next, we consider a more subtle problem, where for each of the 11 *manually labeled communities* aforementioned, we test whether it can't be further divided into multiple communities (null) or it can (alternative). The results of all 11 testing settings are in Table 2.

For the first 10 test settings, despite some differences in the p-values, two tests, $\widehat{\chi}^{(3)}_{gc}$ and $\widehat{\chi}^{(4)}_{gc}$, agree with each other and both accept the null. For the last setting (corresponding to the Western Athletic Conference (WAC)), however, $\widehat{\chi}^{(4)}_{gc}$ votes for rejection and $\widehat{\chi}^{(3)}_{gc}$ votes for acceptance. It turns out that one team ("BioseState") in the WAC is an outlier, which did not play any game in the data range. After removing the outlier, both tests vote for acceptance. These results are consistent with the ground truth, suggesting (a) both tests yield reasonable testing results even for small-size networks, and (b) $\widehat{\chi}^{(4)}_{gc}$ is more effective in detecting outliers.

Table 2. The 11 testing results (each corresponds to a conference). Column 3-4: test scores $\sqrt{B_{n,m}/(2m)}\widehat{C}_m^{-1/2}\widehat{\chi}^{(m)}_{gc}$ and corresponding p-values with $m = 3$. Column 5-6: same but for $m = 4$.

Conference	size	score	p-value	score	p-value
Atlantic Coast	9	0.00	1.00	0.00	0.50
Big East	8	0.00	1.00	0.00	0.50
Big Ten	11	-0.07	0.94	-0.31	0.62
Big Twelve	12	-0.02	0.98	-0.48	0.68
Conference USA	10	0.26	0.80	1.23	0.11
Mid-American	13	0.65	0.51	0.24	0.41
Mountain West	8	0.00	1.00	0.00	0.50
Pacific Ten	10	-0.04	0.97	-0.19	0.58
Southeastern	12	-0.06	0.95	-0.40	0.65
Sun Belt	7	1.48	0.14	1.06	0.15
Western Athletic	10	0.51	0.61	**2.48**	**0.01**

6. Proof of Theorem 3.1

First, we show that for any $m \geq 3$,

$$B_{n,m}C_m = \sum_{k=1}^{K} \lambda_k^m + O\big(\|\theta\|_4^4\|\theta\|^{2m-4}\big), \tag{16}$$

and for any $m \geq 1$, $B_{n,m+1}L_m$ equals to

$$\sum_{k=1}^{K} \lambda_k^m(\eta'\xi_k)^2 + O\big(\|\theta\|_1^2\|\theta\|_4^4\|\theta\|^{2m-6}\big). \tag{17}$$

Consider (16). By definition, $B_{n,m}C_m = B_{n,m}\mathbb{E}[\widehat{C}_m] = \sum_{i_1,\ldots,i_m}\mathbb{E}[A_{i_1i_2}\cdots A_{i_mi_1}] = \sum_{i_1,\ldots,i_m}\Omega_{i_1i_2}\cdots\Omega_{i_mi_1}$, where the sum is over distinct indices $i_1, ..., i_m$. As a result,

$$B_{n,m}C_m = \mathrm{tr}(\Omega^m) - \sum_{\substack{\text{non-distinct}\\ i_1,\ldots,i_m}} \Omega_{i_1i_2}\Omega_{i_2i_3}\cdots\Omega_{i_mi_1}.$$

We calculate the term $\mathrm{tr}(\Omega^m)$. From the DCMM model, $\Omega = \Theta\Pi P\Pi'\Theta$. It follows that

$$\begin{aligned}\Omega^m &= \Theta\Pi P(\Pi'\Theta^2\Pi)P(\Pi'\Theta^2\Pi)\cdots P(\Pi'\Theta^2\Pi)P\Pi'\Theta\\ &= \Theta\Pi(PGPG\cdots PGP)\Pi'\Theta\\ &= (\Theta\Pi G^{-1/2})(G^{1/2}PG^{1/2})^m(G^{-1/2}\Pi'\Theta).\end{aligned}$$

For any matrices A and B, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. As a result,

$$\begin{aligned}\mathrm{tr}(\Omega^m) &= \mathrm{tr}\big[(P^{1/2}GP^{1/2})^m(G^{-1/2}\Pi'\Theta)(\Theta\Pi G^{-1/2})\big]\\ &= \mathrm{tr}\big[(P^{1/2}GP^{1/2})^m\big] = \sum_k \lambda_k^m.\end{aligned} \tag{18}$$

We then bound the remainder term. Note that $\Omega_{ij} = \theta_i\theta_j(\pi_i'P\pi_j) \leq C\theta_i\theta_j$, where the last inequality is from Condition (9). Hence,

$$\begin{aligned}&\sum_{\substack{\text{non-distinct}\\ i_1,\ldots,i_m}} \Omega_{i_1i_2}\Omega_{i_2i_3}\cdots\Omega_{i_mi_1} \leq \sum_{\substack{\text{non-distinct}\\ i_1,\ldots,i_m}} C\theta_{i_1}^2\theta_{i_2}^2\cdots\theta_{i_m}^2\\ &\leq \sum_{i_1,\ldots,i_{m-1}} C\theta_{i_1}^4\theta_{i_2}^2\cdots\theta_{i_{m-1}}^2 \leq C\|\theta\|_4^4\|\theta\|^{2(m-2)}.\end{aligned}$$

Combining the above gives (16).

Consider (17). Similarly, we have

$$B_{n,m+1}L_m = 1_n'\Omega^m 1_n - \sum_{\substack{\text{non-distinct}\\ i_1,\ldots,i_{m+1}}} \Omega_{i_1i_2}\Omega_{i_2i_3}\cdots\Omega_{i_mi_{m+1}}.$$

Since $\Omega = \Theta\Pi P\Pi'\Theta$, it follows that

$$\begin{aligned}1_n'\Omega^m 1_n &= 1_n'\Theta\Pi P(\Pi'\Theta^2\Pi)P(\Pi'\Theta^2\Pi)\cdots P\Pi'\Theta 1_n\\ &= 1_n'\Theta\Pi(PGPG\cdots P)\Pi'\Theta 1_n\\ &= 1_n'\Theta\Pi G^{-1/2}(G^{1/2}PG^{1/2})^m G^{-1/2}\Pi'\Theta 1_n\\ &= \eta'(G^{1/2}PG^{1/2})^m\eta.\end{aligned}$$

The eigen-decomposition $G^{1/2}PG^{1/2} = \sum_{k=1}^K \lambda_k\xi_k\xi_k'$ implies that $(G^{1/2}PG^{1/2})^m = \sum_{k=1}\lambda_k^m\xi_k\xi_k'$. As a result,

$$1_n'\Omega^m 1_n = \eta'\big[\sum_{k=1}^K \lambda_k^m\xi_k\xi_k'\big]\eta = \sum_{k=1}^K \lambda_k^m(\eta'\xi_k)^2. \tag{19}$$

We then bound the remainder term. Since $\Omega_{ij} \leq C\theta_i\theta_j$, $\Omega_{i_1i_2}\cdots\Omega_{i_mi_{m+1}} \leq C\theta_{i_1}\theta_{i_{m+1}}\theta_{i_2}^2\cdots\theta_{i_m}^2$. It follows that

$$\begin{aligned}&\sum_{\substack{\text{non-distinct}\\ i_1,\ldots,i_{m+1}}} \Omega_{i_1i_2}\Omega_{i_2i_3}\cdots\Omega_{i_mi_{m+1}}\\ &\leq\Big(\sum_{i_1=i_{m+1}} + \sum_{i_2=i_{m+1}} + \sum_{\substack{\text{non-distinct}\\ i_2,\ldots,i_m}}\Big)C\theta_{i_1}\theta_{i_{m+1}}\theta_{i_2}^2\cdots\theta_{i_m}^2\\ &\leq \sum_{i_1,\ldots,i_m} C\theta_{i_1}^2\cdots\theta_{i_m}^2 + \sum_{i_1,\ldots,i_m} C\theta_{i_1}\theta_{i_2}^3\theta_{i_3}^2\cdots\theta_{i_m}^2\\ &\quad + \sum_{i_1,\ldots,i_{m-1},i_{m+1}} \theta_{i_1}\theta_{i_{m+1}}\theta_{i_2}^4\theta_{i_3}^2\cdots\theta_{i_{m-1}}^2\\ &\leq C\Big[\|\theta\|^{2m} + \|\theta\|_1\|\theta\|_3^3\|\theta\|^{2(m-2)} + \|\theta\|_1^2\|\theta\|_4^4\|\theta\|^{2(m-3)}\Big]\end{aligned}$$

$\leq C\|\theta\|^{2(m-3)}\big(\|\theta\|^6 + \|\theta\|_1\|\theta\|_3^3\|\theta\|^2 + \|\theta\|_1^2\|\theta\|_4^4\big)$.

We need to compare the three terms in the brackets. First, applying Holder's inequality with $p = 3$ and $q = 3/2$, we have $\sum_i \theta_i^2 = \sum_i \theta_i^{\frac{4}{3}}\theta_i^{\frac{2}{3}} \leq (\sum_i \theta_i^{\frac{4p}{3}})^{\frac{1}{p}}(\sum_i \theta_i^{\frac{2q}{3}})^{\frac{1}{q}}$. It implies $\|\theta\|^2 \leq \|\theta\|_4^{4/3}\|\theta\|_1^{2/3}$. As a result,

$$\|\theta\|^6 \leq \|\theta\|_4^4\|\theta\|_1^2.$$

This means the first term above is dominated by the last term. Second, by Cauchy-Schwartz inequality, $\sum_i \theta_i^3 \leq (\sum_i \theta_i^2)^{1/2}(\sum_i \theta_i^4)^{1/2}$, which means $\|\theta\|_3^3 \leq \|\theta\|\|\theta\|_4^2$. As a result, $\|\theta\|_1\|\theta\|_3^3\|\theta\|^2 \leq \|\theta\|_1\|\theta\|_4^2\|\theta\|^3$. Furthermore, we have proved $\|\theta\|^3 \leq \|\theta\|_4^2\|\theta\|_1$. Combining the above gives $\|\theta\|_1\|\theta\|_3^3\|\theta\|^2 \leq \|\theta\|_1^2\|\theta\|_4^4$. Hence, the second term is dominated by the last term. In summary, the remainder term is $O(\|\theta\|_1^2\|\theta\|_4^4\|\theta\|^{2(m-3)})$. This proves (17).

Next, we use (16)-(17) to show the claim. Write for short

$$\chi_{gc,0}^{(m)} = \frac{1}{B_{n,m}}\left\{\sum_k \lambda_k^m - \left[\frac{\sum_k(\eta,\xi_k)^2\lambda_k^{m-1}}{\sum_k(\eta,\xi_k)^2\lambda_k^{m-2}}\right]^m\right\}.$$

Write $\widetilde{C}_m = B_{n,m}C_m$, $\widetilde{C}_m^0 = \mathrm{tr}(\Omega^m)$, $\widetilde{L}_m = B_{n,m+1}L_m$, and $\widetilde{L}_m^0 = 1_n'\Omega^m 1_n$, for all m. By definition and (18)-(19),

$$\begin{aligned}
B_{n,m}\chi_{gc,0}^{(m)} &= \widetilde{C}_m^0 - (\widetilde{L}_{m-1}^0/\widetilde{L}_{m-2}^0)^m,\\
B_{n,m}\chi_{gc}^{(m)} &= B_{n,m}[C_m - (L_{m-1}/L_{m-2})^m]\\
&= \widetilde{C}_m - \tfrac{(B_{n,m-1})^m}{(B_{n,m})^{m-1}}\cdot(\widetilde{L}_{m-1}/\widetilde{L}_{m-2})^m.
\end{aligned}$$

As a result,

$$\begin{aligned}
&B_{n,m}|\chi_{gc}^{(m)} - \chi_{gc,0}^{(m)}|\\
&\leq |\widetilde{C}_m - \widetilde{C}_m^0| + \Big|\tfrac{(\widetilde{L}_{m-1})^m}{(\widetilde{L}_{m-2})^m} - \tfrac{(\widetilde{L}_{m-1}^0)^m}{(\widetilde{L}_{m-2}^0)^m}\Big|\\
&\quad + \tfrac{(\widetilde{L}_{m-1})^m}{(\widetilde{L}_{m-2})^m}\Big|\tfrac{(B_{n,m-1})^m}{(B_{n,m})^{m-1}} - 1\Big|\\
&\equiv I_1 + I_2 + I_3. \qquad (20)
\end{aligned}$$

We now bound these three terms. By (16)-(17),

$$\begin{aligned}
|\widetilde{C}_m - \widetilde{C}_m^0| &= O(\|\theta\|_4^4\|\theta\|^{2m-4}),\\
|\widetilde{L}_m - \widetilde{L}_m^0| &= O(\|\theta\|_1^2\|\theta\|_4^4\|\theta\|^{2m-6}). \qquad (21)
\end{aligned}$$

Hence,

$$I_1 = O(\|\theta\|_4^4\|\theta\|^{2m-4}).$$

To bound I_2, we need the following lemma, which is proved in the supplemental material.

Lemma 6.1 *Under conditions of Theorem 3.3,* $|\lambda_k| \asymp \|\theta\|^2$ *for* $1 \leq k \leq K$, *and* $\max_{1\leq k\leq K}|\eta'\xi_k| \asymp \|\theta\|^{-1}\|\theta\|_1$.

By (19), $\widetilde{L}_m^0 = \sum_k(\eta'\xi_k)^2\lambda_k^m$. For m even, it then follows from Lemma 6.1 that

$$\widetilde{L}_m^0 \geq c\|\theta\|^{2m-2}\|\theta\|_1^2 \qquad (22)$$

for a constant $c > 0$. For m odd, this is still true; see the proof of Lemma B.1 in the supplemental material. Additionally, by Cauchy-Schwarz inequality, $\|\theta\|^4 \leq n\|\theta\|_4^4$; hence, for $m \leq 3$, $|\widetilde{L}_m - \widetilde{L}_m^0|$ is negligible compared to the order of $\widetilde{L}_m^0$, so we also have $\widetilde{L}_m \geq c\|\theta\|^{2m-2}\|\theta\|_1^2$. Now,

$$\begin{aligned}
&\Big|\tfrac{\widetilde{L}_{m-1}^0}{\widetilde{L}_{m-2}^0} - \tfrac{\widetilde{L}_{m-1}}{\widetilde{L}_{m-2}}\Big| \leq \tfrac{|\widetilde{L}_{m-1}-\widetilde{L}_{m-1}^0|}{\widetilde{L}_{m-2}^0} + \tfrac{\widetilde{L}_{m-1}}{\widetilde{L}_{m-2}}\tfrac{|\widetilde{L}_{m-2}-\widetilde{L}_{m-2}^0|}{\widetilde{L}_{m-2}^0}\\
&= O\Big(\tfrac{\|\theta\|_1^2\|\theta\|_4^4\|\theta\|^{2m-8}}{\|\theta\|^{2m-6}\|\theta\|_1^2}\Big) + O\Big(\|\theta\|^2\tfrac{\|\theta\|_1^2\|\theta\|_4^4\|\theta\|^{2m-10}}{\|\theta\|^{2m-6}\|\theta\|_1^2}\Big)\\
&= O(\|\theta\|_4^4\|\theta\|^{-2}).
\end{aligned}$$

Since $|x^m - y^m| \leq C|x-y|(|x|+|y|)^{m-1}$,

$$\begin{aligned}
I_2 &\leq C\Big|\tfrac{\widetilde{L}_{m-1}^0}{\widetilde{L}_{m-2}^0} - \tfrac{\widetilde{L}_{m-1}}{\widetilde{L}_{m-2}}\Big|\cdot\Big|\tfrac{\widetilde{L}_{m-1}^0}{\widetilde{L}_{m-2}^0}\Big|^{m-1}\\
&= O(\|\theta\|_4^4\|\theta\|^{-2}\cdot\|\theta\|^{2m-2}) = O(\|\theta\|_4^4\|\theta\|^{2m-4}).
\end{aligned}$$

Consider I_3. Note that $\frac{(B_{n,m-1})^m}{(B_{n,m})^{m-1}} = \frac{n(n-1)\cdots(n-m+2)}{(n-m+1)^{m-1}} = \prod_{j=1}^{m-1}(1+\frac{j}{n-m+1}) = 1 + O(\frac{1}{n})$. So,

$$I_3 = O(\|\theta\|^{2m}\cdot n^{-1}) = O(\|\theta\|_4^4\|\theta\|^{2m-4}),$$

where we have used the universal inequality $\|\theta\|^4 \leq n\|\theta\|_4^4$.

We plug the above results into (20) and note that $B_{n,m} \sim n^m$. The claim follows immediately. □

7. Conclusion

We consider a hard testing problem in the rather general DCMM model where the challenge is severe degree heterogeneity. We discover a systematic way to cancel the effects of degree heterogeneity, and propose a family of tests, with careful analysis and numerical support. Compared to literature, our tests have competitive powers and are applicable in much broader settings. Our theory is also for very broad settings where existing works have very limited understanding. We point out an unappealing feature of the EZ test (Gao & Lafferty, 2017), and shows a new test in our family has successfully overcome the problem that the EZ test faces.

In our theorems, we require K to be fixed, but the results continue to hold if $K \to \infty$ reasonably slowly. We also require the singular values of P are in the same order, but this is mostly for simplicity in presentation and can be replaced by weaker conditions. We also assume $\|\theta\|_3 \to 0$. The case $\|\theta\|_3 \to \infty$ is related to the "dense network" case, the analysis of which can be done but is different and we leave it as future work.

References

Abbe, E. and Sandon, C. Achieving the KS threshold in the general stochastic block model with linearized acyclic

belief propagation. In *Advances in Neural Information Processing Systems*, pp. 1334–1342, 2016.

Airoldi, E., Blei, D., Fienberg, S., and Xing, E. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

Banerjee, D. and Ma, Z. Optimal hypothesis testing for stochastic block models with growing degrees. *arXiv:1705.05305*, 2017.

Bickel, P. J. and Sarkar, P. Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78 (1):253–273, 2016.

Bubeck, S., Ding, J., Eldan, R., and Rácz, M. Z. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, 49(3):503–532, 2016.

Chen, K. and Lei, J. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, pp. 1–11, 2017.

Chen, Y., Li, X., and Xu, J. Convexified modularity maximization for degree-corrected stochastic block models. *The Annals of Statistics, to appear*, 2018.

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

Girvan, M. and Newman, M. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

Holland, P. W. and Leinhardt, S. Transitivity in structural models of small groups. *Comparative Group Studies*, 2 (2):107–124, 1971.

Horn, R. and Johnson, C. *Matrix Analysis*. Cambridge University Press, 1985.

Jin, J. Fast community detection by SCORE. *The Annals of Statistics*, 43(1):57–89, 2015.

Jin, J. and Ke, Z. T. A sharp lower bound for mixed-membership estimation. *arXiv:1709.05603*, 2017.

Jin, J., Ke, Z. T., and Luo, S. Estimating network memberships by simplex vertices hunting. *arXiv:1708.07852*, 2017.

Jin, J., Ke, Z. T., and Luo, S. Network global testing by counting graphlets (extended version). *Manuscript*, 2018.

Karrer, B. and Newman, M. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83 (1):016107, 2011.

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

Lei, J. A goodness-of-fit test for stochastic block models. *The Annals of Statistics*, 44(1):401–424, 2016.

Massoulié, L. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 694–703. ACM, 2014.

Maugis, P.-A. G., Olhede, S. C., and Wolfe, P. J. Topology reveals universal features for network comparison. *1705.05677*, 2017.

Mossel, E., Neeman, J., and Sly, A. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015.

Qin, T. and Rohe, K. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pp. 3120–3128, 2013.

Saldana, D., Yu, Y., and Feng, Y. How many communities are there? *Journal of Computational and Graphical Statistics*, 26(1):171–181, 2017.

Schank, T. and Wagner, D. Approximating clustering coefficient and transitivity. *Journal of Graph Algorithms and Applications*, 9(2):265–275, 2005.

Wang, Y. R. and Bickel, P. J. Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, 45(2):500–528, 2017.

Watts, D. J. and Strogatz, S. H. Collective dynamics of small-world networks. *Nature*, 393, 1998.