The Annals of Statistics
2022, Vol. 50, No. 1, 147–169
https://doi.org/10.1214/21-AOS2099

TESTING COMMUNITY STRUCTURE FOR HYPERGRAPHS

BY MINGAO YUAN[1], RUIQI LIU[2], YANG FENG[3] AND ZUOFENG SHANG[4]

[1]*Department of Statistics, North Dakota State University, mingao.yuan@ndsu.edu*
[2]*Department of Mathematical Sciences, Texas Tech University, ruiqliu@ttu.edu*
[3]*Department of Biostatistics, New York University, yang.feng@nyu.edu*
[4]*Department of Mathematical Sciences, New Jersey Institute of Technology, zshang@njit.edu*

Many complex networks in the real world can be formulated as hypergraphs where community detection has been widely used. However, the fundamental question of whether communities exist or not in an observed hypergraph remains unclear. This work aims to tackle this important problem. Specifically, we systematically study when a hypergraph with community structure can be successfully distinguished from its Erdős–Rényi counterpart, and propose concrete test statistics when the models are distinguishable. The main contribution of this paper is threefold. First, we discover a phase transition in the hyperedge probability for distinguishability. Second, in the bounded-degree regime, we derive a sharp signal-to-noise ratio (SNR) threshold for distinguishability in the special two-community 3-uniform hypergraphs, and derive nearly tight SNR thresholds in the general two-community m-uniform hypergraphs. Third, in the dense regime, we propose a computationally feasible test based on sub-hypergraph counts, obtain its asymptotic distribution, and analyze its power. Our results are further extended to nonuniform hypergraphs in which a new test involving both edge and hyperedge information is proposed. The proofs rely on Janson's contiguity theory (*Combin. Probab. Comput.* **4** (1995) 369–405), a high-moments driven asymptotic normality result by Gao and Wormald (*Probab. Theory Related Fields* **130** (2004) 368–376), and a truncation technique for analyzing the likelihood ratio.

1. Introduction. Community detection is a fundamental problem in network data analysis. For instance, in social networks [24, 34, 58], protein to protein interactions [16], image segmentation [52], among others, many algorithms have been developed for identifying community structure. Theoretical studies on community detection have mostly been focusing on ordinary graph setting in which each possible edge contains exactly two vertices (see [4, 7, 13, 28, 50, 58, 59]). One common assumption made in these references is the existence of communities. Recently, a number of researchers have been devoted to testing this assumption, for example, [9, 10, 12, 26, 27, 40, 43, 55]. Besides, hypothesis testing has been used to test the number of communities in a network [12, 40].

Real-world networks are usually more complex than ordinary graphs. Unlike ordinary graphs where the data structure is typically unique, for example, edges only contain two vertices, *hypergraphs* demonstrate a number of possibly overlapping data structures. For instance, in coauthorship networks [22, 47–49], the number of coauthors varies across different papers so that one cannot consider edges consisting of two coauthors only. Instead, a new type of "edge," called *hyperedge*, must be considered which allows the connectivity of arbitrarily many coauthors. The complex structures of hypergraphs create new challenges in both theoretical and methodological studies. As far as we know, existing hypergraph literature mostly focuses on community detection in algorithmic aspects [4, 13, 17, 31, 39, 41,

Received June 2020; revised March 2021.
MSC2020 subject classifications. Primary 62G10; secondary 05C80.
Key words and phrases. Hypergraph, stochastic block model, hypothesis testing, contiguity, l-cycle.

50, 51]. Only recently, Ghoshdastidar and Dukkipati [31, 32] provided a statistical study in which a spectral algorithm based on adjacency tensor was proposed for identifying community structure and asymptotic results were developed. Nonetheless, the important problem of testing the existence of community structure in an observed hypergraph remains untreated.

In this paper, we aim to tackle the problem of testing community structure for hypergraphs. We first consider the relatively simpler but widely used uniform hypergraphs in which each hyperedge consists of an equal number of vertices. For instance, the (user, resource, annotation) structure in folksonomy may be represented as a uniform hypergraph where each hyperedge consists of three vertices [30]; the (user, remote host, login time, logout time) structure in the login-data can be modeled as a uniform hypergraph where each hyperedge contains four vertices [33]; the point-set matching problem is usually formulated as identifying a strongly connected component in a uniform hypergraph [17]. We provide various theoretical or methodological studies ranging from dense uniform hypergraphs to sparse ones and investigate the possibility of a successful test in each scenario. Our testing results in the dense case are then extended to the more general nonuniform hypergraph setting, in which a new test statistic involving both edge and hyperedge is proposed. One important finding is that our new test is more powerful than the classic one involving edge information only, showing the advantage of using hyperedge information to boost the testing performance. A more notable contribution is a nearly tight threshold for signal-to-noise ratio to examine the existence of community structure (Theorem 2.6).

1.1. *Review of hypergraph model and relevant literature.* In this section, we review some basic notions in hypergraphs and recent progress in the literature. Let us first review the notion of the uniform hypergraph. An *m-uniform hypergraph* $\mathcal{H}_m = (\mathcal{V}, \mathcal{E})$ consists of a vertex set $\mathcal{V}$ and a hyperedge set $\mathcal{E}$, where each hyperedge in $\mathcal{E}$ is a subset of $\mathcal{V}$ consisting of exactly m vertices. Two hyperedges are the same if they are equal as vertex sets. An *l-cycle* in $\mathcal{H}_m$ is a cyclic ordering $\{v_1, v_2, \ldots, v_r\}$ of a subset of the vertex set with hyperedges like $\{v_i, v_{i+1}, \ldots, v_{i+m-1}\}$ and any two adjacent hyperedges have exactly l common vertices. An l-cycle is *loose* if $l = 1$ and *tight* if $l = m - 1$. To better illustrate the notion, consider a 3-uniform hypergraph $\mathcal{H}_3 = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$, $\mathcal{E} = \{(v_i, v_j, v_t) | 1 \le i < j < t \le 7\}$. Then $(\{v_1, v_2, v_3, v_4, v_5, v_6\}, \{(v_1, v_2, v_3), (v_3, v_4, v_5), (v_5, v_6, v_1)\})$ is a *loose* cycle and $(\{v_1, v_2, v_3, v_4\}, \{(v_1, v_2, v_3), (v_2, v_3, v_4), (v_3, v_4, v_1), (v_4, v_1, v_2)\})$ is a *tight* cycle (see Figure 1).

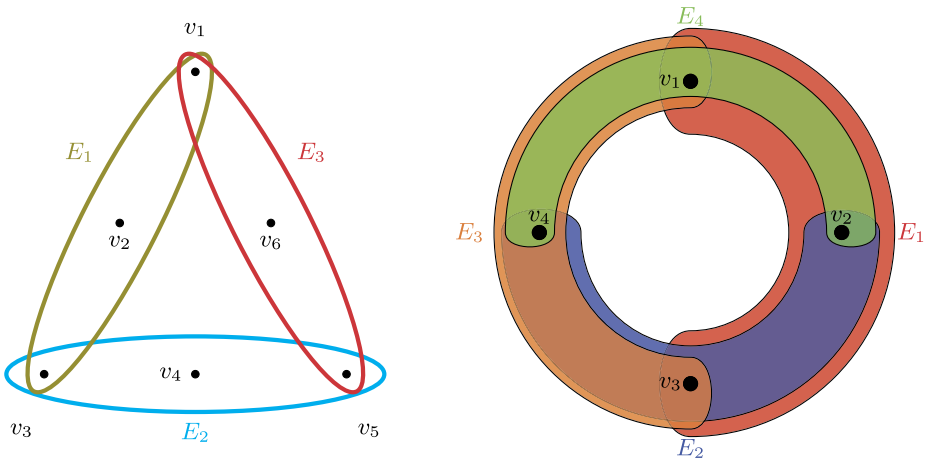

FIG. 1. *Left: a loose cycle of three edges E_1, E_2, E_3. Right: a tight cycle of four edges E_1, E_2, E_3, E_4. Both cycles are subgraphs of the 3-uniform hypergraph $\mathcal{H}_3(\mathcal{V}, \mathcal{E})$.*

Next, let us review uniform hypergraphs with a planted partitioning structure, also known as stochastic block model (SBM). For any positive integers n, m, k with $m, k \geq 2$, and positive sequences $0 < q_n < p_n < 1$ (possibly depending on n), let $\mathcal{H}_m^k(n, p_n, q_n)$ denote a m-uniform hypergraph of n vertices and k balanced communities, in which p_n (q_n) represents the hyperedge probability within (between) communities. More explicitly, any vertex $i \in [n] \equiv \{1, 2, \ldots, n\}$ is assigned, independently and uniformly at random, a label $\sigma_i \in [k] \equiv \{1, 2, \ldots, k\}$, and then each possible hyperedge $(i_1, i_2, \ldots, i_m)$ is included with probability p_n if $\sigma_{i_1} = \sigma_{i_2} = \cdots = \sigma_{i_m}$ and with probability q_n otherwise. In particular, $\mathcal{H}_2^2(n, p_n, q_n)$ (with $m = k = 2$) reduces to the ordinary bisection stochastic block models considered by [44, 55]. Let $A \in \{0, 1\}^{\overbrace{n \times n \times \cdots \times n}^{m}}$ denote the symmetric adjacency tensor of order m associated with $\mathcal{H}_m^k(n, p_n, q_n)$. By symmetry, we mean that $A_{i_1 i_2 \ldots i_m} = A_{\psi(i_1)\psi(i_2)\ldots\psi(i_m)}$ for any permutation ψ of $(i_1, i_2, \ldots, i_m)$. For convenience, assume $A_{i_1 i_2 \ldots i_m} = 0$ if $i_s = i_t$ for some distinct $s, t \in \{1, 2, \ldots, m\}$, that is, the hypergraph has no self-loops. Conditional on $\sigma_1, \ldots, \sigma_n$, the $A_{i_1 i_2 \ldots i_m}$'s, with $i_1, \ldots, i_m$ pairwise distinct, are assumed to be independent following the distribution below:

$$\mathbb{P}(A_{i_1 i_2 \ldots i_m} = 1 | \sigma) = p_{i_1 i_2 \ldots i_m}(\sigma), \qquad \mathbb{P}(A_{i_1 i_2 \ldots i_m} = 0 | \sigma) = q_{i_1 i_2 \ldots i_m}(\sigma), \tag{1}$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)$,

$$p_{i_1 i_2 \ldots i_m}(\sigma) = \begin{cases} p_n & \sigma_{i_1} = \cdots = \sigma_{i_m}, \\ q_n & \text{otherwise}, \end{cases} \qquad q_{i_1 i_2 \ldots i_m}(\sigma) = 1 - p_{i_1 i_2 \ldots i_m}(\sigma).$$

In other words, each possible hyperedge $(i_1, \ldots, i_m)$ is included with probability p_n if the vertices $i_1, \ldots, i_m$ belong to the same community, and with probability q_n otherwise. Let $\mathcal{H}_m(n, \frac{p_n + (k^{m-1}-1)q_n}{k^{m-1}})$ denote the m-uniform hypergraph without community structure, that is, an Erdős–Rényi model in which each possible hyperedge is included with common probability $\frac{p_n + (k^{m-1}-1)q_n}{k^{m-1}}$. We consider such a special choice of hyperedge probability in order to make the model have the same average degree as $\mathcal{H}_m^k(n, p_n, q_n)$. In particular, $\mathcal{H}_2(n, \frac{p_n + (k-1)q_n}{k})$ with $m = 2$ becomes the traditional Erdős–Rényi model that has been well studied in ordinary graph literature; see [14, 15, 21, 25, 54]. Nonuniform hypergraphs can be simply viewed as a superposition of uniform ones; see Section 3. Throughout this paper, we assume k and m are fixed constants independent of n.

Given an observed adjacency tensor A, *does A represent a hypergraph that exhibits community structure?* In the present setting, this problem can be formulated as testing the following hypothesis:

$$H_0 : A \sim \mathcal{H}_m\left(n, \frac{p_n + (k^{m-1}-1)q_n}{k^{m-1}}\right) \quad \textit{vs.} \quad H_1 : A \sim \mathcal{H}_m^k(n, p_n, q_n). \tag{2}$$

When $m = k = 2$, problem (2) has been well studied in the literature. Specifically, for extremely sparse scenario $p_n \asymp q_n \ll n^{-1}$, [44] show that H_0 and H_1 are always indistinguishable in the sense that all tests are asymptotically powerless; for bounded degree case $p_n \asymp q_n \asymp n^{-1}$, the two models are distinguishable if and only if the signal-to-noise ratio (SNR) is greater than 1 [44, 45, 55]; for dense scenario $p_n \asymp q_n \gg n^{-1}$, H_0 and H_1 are always distinguishable and a number of algorithms have been developed (see [3, 10, 12, 26, 27, 40]). When $m = 2$ and $k \geq 3$, the above statements remain true for extremely sparse and dense scenarios; but for bounded degree scenario, SNR > 1 is only a sufficient condition for successfully distinguishing H_0 from H_1 while a necessary condition remains an open problem (see [3, 11, 56]). Abbe [1] provides a comprehensive review of the recent development in this field. From the best of our knowledge, there is a lack of literature dealing with the testing problem (2) for general m. The literature on hypergraph analysis mainly focused on community detection (see [8, 17, 31, 32, 35, 39, 41, 42, 51]).

1.2. *Our contributions*. The aim of this paper is to provide a study on hypergraph testing under a spectrum of hyperedge probability scenarios. Our results consist of four major parts. Section 2.1 deals with the extremely sparse scenario $p_n \asymp q_n \ll n^{-m+1}$, in which we show that H_0 and H_1 are always indistinguishable in the sense of contiguity. Section 2.2 deals with bounded degree case $p_n \asymp q_n \asymp n^{-m+1}$, in which we show that H_0 and H_1 are distinguishable if the SNR of uniform hypergraph is greater than a certain threshold, but indistinguishable if the SNR is below another threshold. Interestingly, when $k = 2$, the two thresholds are nearly tight in that they are of the same order 2^{-m} (up to universal constants). We also construct a powerful test statistic when SNR is greater than one based on counting the "long loose cycles". Section 2.3 deals with dense scenario $p_n \asymp q_n \gg n^{-m+1}$. We propose a test based on counting the hyperedges, l-hypervees, and l-hypertriangles with l determined by the order of p_n (or q_n), and show that the power of the proposed test approaches one as the number of vertices goes to infinity. In Section 3, we extend some of the previous results to nonuniform hypergraph testing. We propose a new test involving both edge and hyperedge information and show that it is generally more powerful than the classic test using edge information only (see Remark 3.1). The results of the present paper can be viewed as nontrivial extensions of the ordinary graph testing results such as [26, 44, 45]. Section 4 provides numerical studies to support our theory. Possible extensions are discussed in Section 5 and proof of the main results are collected in [57].

Figure 2 displays a phase transition phenomenon in the special 3-uniform hypergraph, based on our results in Sections 2.1 and 2.3. We find that H_0 and H_1 are indistinguishable if the hyperedge probabilities satisfy $p_n, q_n = o(n^{-2})$ (see red zone), and are distinguishable if $p_n, q_n \gg n^{-2}$ (see green zone), which is consistent with [6] who showed that community detection with weak consistency is possible if and only if $p_n, q_n \gg n^{-2}$. Therefore, the seemingly different perspectives, that is, hypothesis testing and community detection, appear to coincide here. In contrast, the spectral algorithm proposed by [32] is able to detect communities with strong consistency if $p_n, q_n \gg n^{-2}(\log n)^2$ (later improved to $p_n, q_n \gg n^{-2}\log n$ by [5, 41]). For bounded degree case $p_n, q_n \asymp n^{-2}$, detection algorithms better than random guess were proposed by [6, 18, 23]. Overall, in the references [6, 18], the SNR conditions are not comparable to our $\kappa > 1$ since unknown constants are involved in their conditions. The SNR condition in [23] seems more restrictive than our condition $\kappa > 1$.

On the next page, Figure 3 demonstrates the distinguishable and indistinguishable regions for two-community graph (left) and two-community 3-uniform hypergraph (right) in the bounded degree regime, that is, $p_n = \frac{a}{n^{m-1}}$, $q_n = \frac{b}{n^{m-1}}$. The regions are characterized by (a, b) with $a > b > 0$. The left plot is based on [44] who derived the decision boundary $(a-b)^2 = 2(a+b)$. The right plot is based on our Theorem 2.6 with the decision boundary

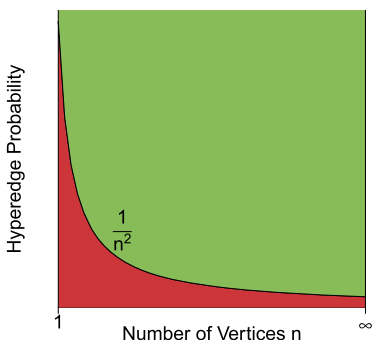

FIG. 2. *Phase transition for* 3*-uniform hypergraph. Red*: *indistinguishable*; *green*: *distinguishable*.

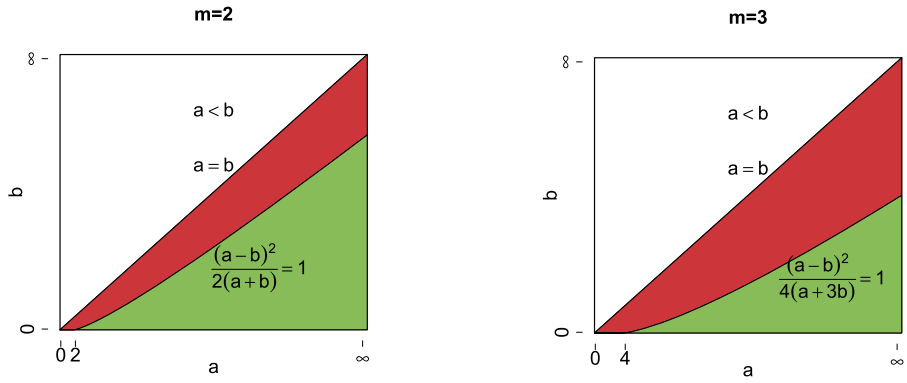

FIG. 3. *Phase transition in bounded degree case. Red*: *indistinguishable*; *green*: *distinguishable.*

$(a-b)^2 = 4(a+3b)$. It can be observed that $m = 3$ yields a larger indistinguishable region than $m = 2$, which reveals a substantial difference for hypothesis testing in the two models.

2. Main results. In this section, we present our main results in three parts, organized by the sparsity of the network. The contiguity theory for the extremely sparse case is presented in Section 2.1, followed by the contiguity and orthogonality result for the bounded degree case in Section 2.2. In Section 2.3, we construct a powerful test by counting the hyperedges, l-hypervees, and l-hypertriangles for the dense case. Throughout this paper, we assume k and m are fixed positive integers.

2.1. *A contiguity theory for extremely sparse case.* In this section, we consider the testing problem (2) with $p_n \asymp q_n \ll n^{-m+1}$, that is, the hyperedge probability of the hypergraph is extremely low. For technical convenience, we only consider $p_n = \frac{a}{n^\alpha}$ and $q_n = \frac{b}{n^\alpha}$ with constants $a > b > 0$ and $\alpha > m-1$. The results in this section may be extended to general orders of p_n and q_n with more cumbersome arguments. We will show that no test can successfully distinguish H_0 from H_1 in such a situation. The proof proceeds by showing that the probability measures associated with H_0 and H_1 are contiguous (see Theorem 2.1). We remark that contiguity has also been used to prove indistinguishability for ordinary graphs (see [44, 45]).

Let $\mathbb{P}_n$ and $\mathbb{Q}_n$ be sequences of probability measures on a common probability space $(\Omega_n, \mathcal{F}_n)$. We say that $\mathbb{P}_n$ and $\mathbb{Q}_n$ are mutually *contiguous* if for every sequence of measurable sets $A_n \subset \Omega_n$, $\mathbb{P}_n(A_n) \to 0$ if and only if $\mathbb{Q}_n(A_n) \to 0$ as $n \to \infty$. They are said to be *orthogonal* if there exists a sequence of measurable sets A_n such that $\mathbb{P}_n(A_n) \to 0$ and $\mathbb{Q}_n(A_n) \to 1$ as $n \to \infty$. According to [44], two probability models are indistinguishable if their associated probability measures are mutually contiguous, and two probability models are distinguishable if their associated probability measures are orthogonal. The following theorem shows that H_0 and H_1 are indistinguishable.

THEOREM 2.1. *If $\alpha > m-1$ and $a > b > 0$ are fixed constants, then the probability measures associated with H_0 and H_1 are mutually contiguous.*

The proof of Theorem 2.1 proceeds by showing that the ratio of the likelihood function of H_1 over H_0 converges in distribution to 1 under H_0, which implies the contiguity of H_1 and H_0 [36]. Theorem 2.1 says that the hypergraphs in H_0 and H_1 are indistinguishable, and hence, no test can successfully separate the two hypotheses. One intuitive explanation is that

when $\alpha > m-1$, the average degree of both hypergraph models converges to zero. To see this, the average degree is

$$\binom{n}{m-1}\frac{a+(k^{m-1}-1)b}{k^{m-1}n^{\alpha}}, \tag{3}$$

which goes to zero as $n \to \infty$ if $\alpha > m-1$. Therefore, the signals in both models are not strong enough to support a successful test. It is easy to see that the average degree becomes bounded when $\alpha = m-1$ which will be investigated in the next section.

2.2. *Bounded degree case.* In this section, we consider $p_n \asymp q_n \asymp n^{-m+1}$ which leads to bounded average degrees for the models in H_0 and H_1; see (3). For convenience, let us denote $p_n = \frac{a}{n^{m-1}}$ and $q_n = \frac{b}{n^{m-1}}$ for fixed $a > b > 0$. Define the signal to noise ratio (SNR) for H_0 and H_1 as

$$\kappa = \frac{(a-b)^2}{k^{m-1}(m-2)![a+(k^{m-1}-1)b]}. \tag{4}$$

When $m = k = 2$, it is easy to check that $\kappa = \frac{(a-b)^2}{2(a+b)}$ which becomes the classic SNR of ordinary stochastic block models considered by [2, 3, 44]. Hence, it is reasonable to view κ defined in (4) as a generalization of the classic SNR to the hypergraph model $\mathcal{H}_m^k(n, \frac{a}{n^{m-1}}, \frac{b}{n^{m-1}})$. Like the classic SNR, the value of κ characterizes the separability between communities. Intuitively, when κ is large which means that the communities are very different, the testing problem (2) becomes simpler. The following result shows that when $\kappa > 1$, successful testing becomes possible.

THEOREM 2.2. *Suppose that $a > b > 0$ are fixed constants, $m, k \geq 2$. If $\kappa > 1$, then the probability measures associated with H_0 and H_1 are orthogonal.*

We prove Theorem 2.2 by constructing a sequence of events dependent on the number of long loose cycles and showing that the probabilities of the events converge to 1 (or 0) under H_0 (or H_1), based on the high moments driven asymptotic normality theorem from Gao and Wormald [29]. Theorem 2.2 says that it is possible to distinguish the hypotheses H_0 and H_1 provided that $\kappa > 1$. Abbe and Sandon [3] obtained relevant results in the ordinary graph setting, that is, $m = 2$ and $k \geq 2$ in our case; see Corollary 2.8 therein which states that community detection in polynomial time becomes possible if SNR > 1. Whereas Theorem 2.2 holds for arbitrary $m, k \geq 2$. Hence, our result can be viewed as an extension of [3] to hypergraph setting.

Let us now propose a test statistic based on "long loose cycles" that can successfully distinguish H_0 and H_1 when $\kappa > 1$. Let ξ_n be a positive integer sequence diverging along with n. Let X_{ξ_n} be the number of loose cycles, each consisting of exactly ξ_n edges. Define

$$\mu_{n0} = \frac{\lambda_m^{\xi_n}}{2\xi_n}, \qquad \mu_{n1} = \mu_{n0} + \frac{k-1}{2\xi_n}\left[\frac{a-b}{k^{m-1}(m-2)!}\right]^{\xi_n},$$

where $\lambda_m = \frac{a+(k^{m-1}-1)b}{k^{m-1}(m-2)!}$ for any $m \geq 2$. Note that when $m = 2$, $\lambda_m = \frac{a+(k-1)b}{k}$ is the average degree [11]. Let $\mathbb{P}_{H_1}$ denote the probability measure induced by A under H_1. We have the following theorem about the asymptotic property of X_{ξ_n}.

THEOREM 2.3. *Suppose $\kappa > 1$ and $1 \ll \xi_n \leq \delta_0 \log_{\lambda_m} \log_\gamma n$, where $\gamma > 1$ and $0 < \delta_0 < 2$ are constants. Then, under H_l for $l = 0, 1$, $\frac{X_{\xi_n}-\mu_{nl}}{\sqrt{\mu_{nl}}} \xrightarrow{d} N(0,1)$ as $n \to \infty$. Furthermore, for any constant $C > 0$, $\mathbb{P}_{H_1}(|\frac{X_{\xi_n}-\mu_{n0}}{\sqrt{\mu_{n0}}}| > C) \to 1$ as $n \to \infty$.*

TABLE 1
Minimal n to achieve a desirable value of ξ_n

Desirable ξ_n	3	4	5	6
Minimal n	2	3	25	29,786

The proof is based on the asymptotic normality theory developed by [29]. According to Theorem 2.3, we propose the following test statistic:

$$T_{\xi_n} = \frac{X_{\xi_n} - \mu_{n0}}{\sqrt{\mu_{n0}}}.$$

We remark that computation of T_{ξ_n} is typically in super-polynomial time since it requires to find X_{ξ_n} which has complexity $n^{O(\xi_n)}$. By Theorem 2.3, $T_{\xi_n} \xrightarrow{d} N(0,1)$ under H_0. Hence, we construct the following testing rule at significance level $\alpha \in (0,1)$:

$$\text{reject } H_0 \quad \text{if and only if} \quad |T_{\xi_n}| > z_{\alpha/2},$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of $N(0,1)$. It follows by Theorem 2.3 that $\mathbb{P}_{H_1}(|T_{\xi_n}| > z_{\alpha/2}) \to 1$, that is, the power of T_{ξ_n} approaches one when $\kappa > 1$.

Theorem 2.3 requires $\xi_n \to \infty$ and to grow slower than an iterative logarithmic order. This is due to the use of [29] which requires ξ_n to diverge with $\xi_n \lambda_m^{\xi_n} = o(\log n)$. In practice, we suggest choosing $\xi_n = \lfloor \delta_0 \log_{\lambda_m} \log_\gamma n \rfloor$ with γ close to 1 and δ_0 close to 2. Such γ and δ_0 will make ξ_n suitably large so that the test statistic T_{ξ_n} becomes valid. For instance, Table 1 demonstrates the values of ξ_n along with n with $\delta_0 = 1.99$, $\gamma = 1.01$, $\lambda_m = 10$. We can see that, for a moderate range of n, the values of ξ_n are sufficiently large to make the test valid. When $\xi_n = l$ is fixed and the exact α-level test is needed, we should use Poisson distribution as the null limiting distribution. In this case, the number of l-loose cycle X_l converges in distribution to Poisson distribution with mean $\mu_0 = \frac{\lambda_m^l}{2l}$ under H_0 (It's implied by the proof of Theorem 2.5). It should be mentioned that the calculation of T_{ξ_n} requires known values of a and b. When a and b are unknown, motivated by the ordinary graph [44], they can be estimated as follows. Define

$$\widehat{\lambda}_m = \frac{n^{m-1}|\mathcal{E}|}{(m-2)!\binom{n}{m}}, \qquad \widehat{f} = (2\xi_n X_{\xi_n} - \widehat{\lambda}_m^{\xi_n})^{\frac{1}{\xi_n}},$$

where $|\mathcal{E}|$ is the number of observed hyperedges and X_{ξ_n} is the number of loose cycles of length ξ_n. Let $\widehat{a}_n = (m-2)![\widehat{\lambda}_m + (k^{m-1}-1)(k-1)^{-\frac{1}{\xi_n}}\widehat{f}]$ and $\widehat{b}_n = (m-2)![\widehat{\lambda}_m - (k-1)^{-\frac{1}{\xi_n}}\widehat{f}]$. The following theorem says that $\widehat{a}_n$ and $\widehat{b}_n$ are consistent estimators of a and b, respectively.

THEOREM 2.4. *Suppose $\kappa > 1$ and ξ_n satisfies the condition in Theorem 2.3. Then $\widehat{a}_n \to a$ and $\widehat{b}_n \to b$ in probability.*

Another interesting question is to investigate for what values of κ a successful test becomes impossible. When $m = k = 2$, [44] showed that no test can successfully distinguish H_0 from H_1 provided $\kappa < 1$; and a successful test becomes possible provided $\kappa > 1$. It is substantially challenging to obtain such a sharp result when k becomes larger. For instance, in the ordinary graph setting, [46] obtained a (nonsharp) condition in terms of SNR when $k \geq 3$ under which successful test becomes impossible. In Theorem 2.5 below, we address a similar question in the hypergraph setting. For any integers $m \geq 3$, $k \geq 2$, define

$\tau_1(m,k) = \binom{m}{2}^{-1} \sum_{i=1}^{\lceil \frac{m}{2}-1 \rceil} \frac{1}{k^{2i-1}} \binom{m}{i+2}$ and $\tau_2(m,k) = 1 + \binom{m}{2}^{-1} \sum_{i=1}^{m-2} \frac{1}{k^{2i}} \binom{m}{i+2}$. The quantities $\tau_1(m,k)$ and $\tau_2(m,k)$ will jointly characterize a spectrum of (m,k,κ) such that successful test does not exist.

THEOREM 2.5. *Suppose that $m \geq 3$, $k \geq 2$ are integers satisfying $\tau_1(m,k) \leq 1$, $a > b > 0$ are fixed constants and $\alpha = m - 1$. If*

$$0 < \kappa < \frac{1}{\tau_2(m,k)(k^2-1)}, \tag{5}$$

then the probability measures associated with H_0 and H_1 are mutually contiguous.

The proof of Theorem 2.5 relies on Janson's contiguity theory [36]. Theorem 2.5 says that when $\tau_1(m,k) \leq 1$ and κ falls in the range (5), there is no test that can successfully distinguish the hypotheses H_0 and H_1. It should be emphasized that the condition $\tau_1(m,k) \leq 1$ holds for a broad range of pairs (m,k). For instance, such condition holds for any $k \geq 2$ and $3 \leq m \leq 6$. To see this, for any $k \geq 2$, $\tau_1(3,k) = \frac{1}{3k} < 1$, $\tau_1(4,k) = \frac{2}{3k} < 1$, $\tau_1(5,k) = \frac{1}{k} + \frac{1}{2k^3} < 1$ and $\tau_1(6,k) = \frac{4}{3k} + \frac{1}{k^3} < 1$. Note that $m \leq 6$ covers most of the practical cases (see [32]).

Combining Theorems 2.5 and 2.2, it is still unknown whether H_0 and H_1 are distinguishable when $\frac{1}{\tau_2(m,k)(k^2-1)} \leq \kappa \leq 1$. Such result can be further improved for the special case $k = 2$, and we close the gap if in addition $m = 3$, as presented in the following theorem.

THEOREM 2.6. *For $k = 2$, the following results hold.*

1. *For any $m \geq 2$, if $0 < \kappa < 2^{2-m}$, then H_0 and H_1 are indistinguishable. Moreover, for any given constant κ_0 such that $\kappa_0 > \frac{m(m-1)\log 2}{2^{m-1}-1}$, there exist $a > b > 0$ such that the SNR κ for the hypotheses H_0 and H_1 is equal to κ_0, and H_0 and H_1 are distinguishable by the likelihood ratio test.*
2. *For any $m \geq 2$, if $0 < \kappa < \frac{m(m-1)}{2N_m}$, where $N_m = [3^m + (-1)^m]/4 - 2^{m-1} + 1/2$, then H_0 and H_1 are indistinguishable.*

Specifically, Part 1 indicates that, when SNR is below 2^{2-m}, H_0 and H_1 are indistinguishable; while they are possible to be distinguishable when SNR is greater than $\frac{m(m-1)\log 2}{2^{m-1}-1}$. Essentially, Part 1 implies that the derived SNR upper and lower bounds satisfy the following relationship:

$$0 < \sup_{m \geq 1} \frac{1}{m^2} \cdot \frac{\text{SNR upper bound}}{\text{SNR lower bound}} < \infty,$$

and

$$0 < \min_{m \geq 1} \frac{\text{SNR lower bound}}{2^{-m}} \leq \max_{m \geq 1} \frac{\text{SNR lower bound}}{2^{-m}} < \infty.$$

Part 2 in Theorem 2.6 provides an SNR interval for H_1 and H_0 to be indistinguishable. When $m = 3$, $N_3 = 3$ leads to an SNR interval $0 < \kappa < 1$ which is sharp since $\kappa > 1$ implies that H_0 and H_1 are distinguishable thanks to Theorem 2.2. For $k = 2$ and general m, the upper bound $\frac{m(m-1)}{2N_m}$ may be less sharp as m grows. In particular, when m is large, the upper bound is of order $m^2 3^{-3}$, which can actually be improved as shown in Part 1 of Theorem 2.6. Specifically, Part 1 indicates that, when SNR is below 2^{2-m}, H_0 and H_1 are indistinguishable; while they are possible to be distinguishable when SNR is greater than $m^2 2^{-m}$ up to a constant in the special case $k = 2$. Note that for $3 \leq m \leq 8$, $\frac{m(m-1)}{2N_m} > 2^{2-m}$ and for $m \geq 9$, $\frac{m(m-1)}{2N_m} < 2^{2-m}$.

The proof of Theorem 2.6 relies on a truncation technique to show the stochastic boundedness of the likelihood ratio and a delicate derivation of a lower bound for the truncated likelihood ratio. An interesting consequence of Part 1 is that the likelihood ratio test is possible to distinguish H_0 and H_1 even when κ is below 1 (but greater than $\frac{m(m-1)\log 2}{2^{m-1}-1}$). However, the computation of the likelihood ratio is NP-hard. When $\kappa > 1$, the l-cycle based test can distinguish H_0 and H_1 as well (see Theorem 2.3), and is computationally less expensive.

REMARK 2.1. We provide more details about why truncation technique is needed in our setting. The proof of Theorem 2.6 relies on the first moment technique which requires the analysis of $\mathbb{E}_1 Y_n$ where $\mathbb{E}_1$ is the expectation taken under H_1 and $Y_n = \frac{dP_1}{dP_0}$ is the likelihood ratio of H_1 to H_0. We find that the expression of $\mathbb{E}_1 Y_n$ includes terms like

$$\mathbb{E}_\sigma \exp\Bigg(\sum_{i_1<\cdots<i_m} \text{poly}(\sigma_{i_1}, \ldots, \sigma_{i_m}) \Bigg), \tag{6}$$

where $\text{poly}(\sigma_{i_1}, \ldots, \sigma_{i_m})$ is an mth-order polynomial of $\sigma_{i_1}, \ldots, \sigma_{i_m} \in \{\pm 1\}$. When $m = 2$, (6) becomes a second-order polynomial which is asymptotically χ^2 by CLT. And so, (6) is heuristically $\mathbb{E}\exp(\text{const} \times \chi^2)$ which is finite. This is why no truncation technique is needed here.

However, when $m = 3$, the above polynomial is third-order which is asymptotically Z^3 where $Z \sim N(0, 1)$. And as a result, (6) is heuristically $\mathbb{E}\exp(\text{const} \times Z^3)$ which is infinite. This is why we used the truncation technique, that is, to truncate the likelihood ratio on an even with high probability so that the higher-order polynomials are well controlled, and the truncated likelihood ratio has a finite expectation.

2.3. *A powerful test for dense uniform hypergraph.* In this section, we consider the problem of testing community structure in dense m-uniform hypergraphs with $p_n \asymp q_n \gg n^{-m+1}$. Our approach is based on counting the hyperedges, l-hypervees, and l-hypertriangles in the observed hypergraph. To ensure the success of our test, l needs to be properly selected according to the hyperedge probability of the model. Under such correct selection, we derive asymptotic normality for the test and analyze its power. We also discuss the effect of misspecified l in Remark 2.2. Our method can be viewed as a generalization of [26, 27] from ordinary graph testing. The different features of the hypergraph cycles make our generalization nontrivial.

For convenience, let us denote $p_n = \frac{a_n}{n^{m-1}}$ and $q_n = \frac{b_n}{n^{m-1}}$ with diverging a_n, b_n. Therefore, (2) becomes the following hypothesis testing problem:

$$H_0' : A \sim \mathcal{H}_m\Big(n, \frac{a_n + (k^{m-1}-1)b_n}{k^{m-1}n^{m-1}}\Big) \quad \textit{vs.} \quad H_1' : A \sim \mathcal{H}_m^k\Big(n, \frac{a_n}{n^{m-1}}, \frac{b_n}{n^{m-1}}\Big). \tag{7}$$

We temporarily assume that there exists an integer $1 \le l \le \frac{m}{2}$ such that $n^{l-1} \ll a_n \asymp b_n \ll n^{l-\frac{2}{3}}$. Such a requirement will be relaxed by invoking a sparsification technique. Note that model (7) allows $1 \ll a_n \asymp b_n \ll n^{1/3}$ (with $l = 1$), compared with spectral algorithm [32] which requires $a_n \gg (\log n)^2$ or $a_n \gg \log n$ in [41].

We consider the following degree-corrected SBM in [19, 20, 26] which is more general than (1) and generalizes its counterpart in ordinary graphs. Let $\{W_i, i = 1, \ldots, n\}$ be nonnegative i.i.d. random variables with $\mathbb{E}(W_1^2) = 1$ and $\{\sigma_i, i = 1, \ldots, n\}$ be i.i.d. random variables from multinomial distribution $\text{Mult}(k, 1, 1/k)$. Assume that W_i's and σ_i's are independent.

Given W_i's and σ_i's, the $A_{i_1 i_2 \ldots i_m}$'s, with pairwise distinct $i_1, \ldots, i_m$, are conditional independent satisfying

$$\begin{aligned} \mathbb{P}(A_{i_1 i_2 \ldots i_m} = 1 | W, \sigma) &= W_{i_1} \ldots W_{i_m} p_{i_1 i_2 \ldots i_m}(\sigma), \\ \mathbb{P}(A_{i_1 i_2 \ldots i_m} = 0 | W, \sigma) &= 1 - W_{i_1} \ldots W_{i_m} p_{i_1 i_2 \ldots i_m}(\sigma), \end{aligned} \tag{8}$$

where $W = (W_1, \ldots, W_n)$,

$$p_{i_1 i_2 \ldots i_m}(\sigma) = \begin{cases} \dfrac{a_n}{n^{m-1}} & \sigma_{i_1} = \cdots = \sigma_{i_m}, \\ \dfrac{b_n}{n^{m-1}} & \text{otherwise.} \end{cases}$$

We call (8) the degree-corrected SBM in hypergraph setting. The degree-correction weights W_i's can capture the degree of inhomogeneity exhibited in many social networks. When $m = 2$, (8) reduces to the classical degree-corrected SBM for ordinary graphs (see [19, 20, 26]). For ordinary graphs, [26] proposed a test through counting small subgraphs to distinguish the degree-corrected SBM from an Erdős–Rényi model. In what follows, we generalize their results to hypergraphs through counting small sub-hypergraphs, including hyperedges, l-hypervee, and l-hypertriangles, with definitions given below.

DEFINITION 2.1. An *l-hypervee* consists of two hyperedges with l common vertices. An *l-hypertriangle* is an l-cycle consisting of three hyperedges.

For example, in Figure 4, the hyperedge set $\{(v_1, v_2, v_3, v_4), (v_3, v_4, v_5, v_6)\}$ is a 2-hypervee, and $\{(v_1, v_2, v_3, v_4), (v_3, v_4, v_5, v_6), (v_5, v_6, v_1, v_2)\}$ is a 2-hypertriangle.

Consider the following probabilities of hyperedge, hypervee and hypertriangle in $\mathcal{H}_m^k(n, \frac{a_n}{n^{m-1}}, \frac{a_n}{n^{m-1}})$:

$$\begin{aligned} E &= \mathbb{P}(A_{i_1 i_2 \ldots i_m} = 1), \\ V &= \mathbb{P}(A_{i_1 i_2 \ldots i_m} A_{i_{m-l+1} \ldots i_{2m-l}} = 1), \\ T &= \mathbb{P}(A_{i_1 i_2 \ldots i_m} A_{i_{m-l+1} \ldots i_{2m-l}} A_{i_{2m-2l+1} \ldots i_{3(m-l)} i_1 \ldots i_l} = 1). \end{aligned}$$

It follows from direct calculations that

$$E = (\mathbb{E}W_1)^m \frac{a_n + (k^{m-1} - 1) b_n}{n^{m-1} k^{m-1}},$$

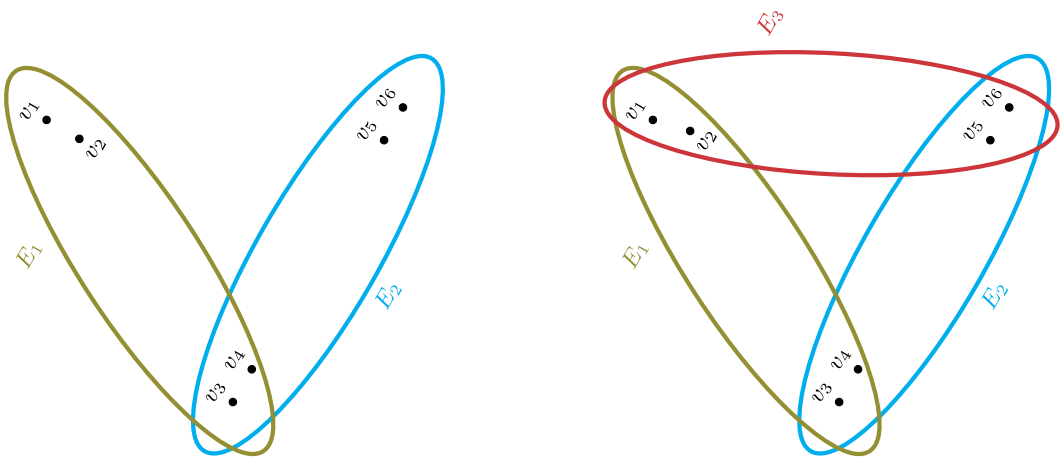

FIG. 4. *Examples of hypervee (left) and hypertriangle (right) with two common vertices between consecutive hyperedges.*

$$V = (\mathbb{E}W_1)^{2(m-l)}\left(\frac{(a_n-b_n)^2}{n^{2(m-1)}k^{2m-l-1}} + \frac{2(a_n-b_n)b_n}{n^{2(m-1)}k^{m-1}} + \frac{b_n^2}{n^{2(m-1)}}\right),$$

$$T = (\mathbb{E}W_1)^{3(m-2l)}\left(\frac{(a_n-b_n)^3}{n^{3(m-1)}k^{3(m-l)-1}} + \frac{3(a_n-b_n)^2b_n}{n^{3(m-1)}k^{2m-l-1}} + \frac{3(a_n-b_n)b_n^2}{n^{3(m-1)}k^{m-1}} + \frac{b_n^3}{n^{3(m-1)}}\right).$$

Define $\mathcal{T} = T - (\frac{V}{E})^3$. The following result demonstrates a strong relationship between $\mathcal{T}$ and H_0', H_1'.

PROPOSITION 2.7. *Under H_0', $\mathcal{T} = 0$ and under H_1', $\mathcal{T} \neq 0$.*

Proposition 2.7 says that H_0' holds if and only if $\mathcal{T} = 0$. Hence, it is reasonable to use an empirical version of $\mathcal{T}$, namely, $\widehat{\mathcal{T}}$, as a test statistic for (7).

Prior to constructing $\widehat{\mathcal{T}}$, let us introduce some notation. For convenience, we use $i_1 : i_m$ to represent the ordering $i_1 i_2 \dots i_m$. Also define $C_{2m-l}(A)$ and $C_{3(m-l)}(A)$ for any adjacency tensor A as follows.

$$\begin{aligned} C_{2m-l}(A) &= A_{i_1:i_m}A_{i_{m-l+1}:i_{2m-l}} + A_{i_2:i_{m+1}}A_{i_{m-l+2}:i_{2m-l}i_1} + \cdots + A_{i_{2m-l}i_1:i_{m-1}}A_{i_{m-l}:i_{2m-l-1}}, \\ C_{3(m-l)}(A) &= A_{i_1:i_m}A_{i_{m-l+1}:i_{2m-l}}A_{i_{2m-2l+1}:i_{3(m-l)}i_1:i_l} \\ &\quad + A_{i_2:i_{m+1}}A_{i_{m-l+2}:i_{2m-l+1}}A_{i_{2m-2l+2}:i_{3(m-l)}i_1:i_{l+1}} \\ &\quad + \cdots + A_{i_{m-l}:i_{2m-l-1}}A_{i_{2(m-l)}:i_{3(m-l)}i_1:i_{l-1}}A_{i_{3(m-l)}i_1:i_{m-1}}. \end{aligned}$$

Note that $C_{2m-l}(A)$ is the number of hypervees in the given vertex ordering $i_1 i_2 \dots i_{2m-l}$, while $C_{3(m-l)}(A)$ counts the number of hypertriangles in the given vertex ordering $i_1 i_2 \dots i_{3(m-l)}$. Define $\widehat{E}$, $\widehat{V}$, $\widehat{T}$ as the empirical versions of E, V, T:

$$\begin{aligned} \widehat{E} &= \frac{1}{\binom{n}{m}} \sum_{i \in c(m,n)} A_{i_1:i_m}, \\ \widehat{V} &= \frac{1}{\binom{n}{2m-l}} \sum_{i \in c(2m-l,n)} \frac{C_{2m-l}(A)}{2m-l}, \\ \widehat{T} &= \frac{1}{\binom{n}{3(m-l)}} \sum_{i \in c(3(m-l),n)} \frac{C_{3(m-l)}(A)}{m-1}, \end{aligned} \tag{9}$$

where, for any positive integers s, t, $c(s,t) = \{(i_1, \dots, i_s) : 1 \le i_1 < \cdots < i_s \le t\}$. We have the following asymptotic normality result.

THEOREM 2.8. *Suppose $\mathbb{E}W_1^4 = O(1)$ and $n^{l-1} \ll a_n \asymp b_n \ll n^{l-\frac{2}{3}}$ for some integer $1 \le l \le \frac{m}{2}$. Moreover, let*

$$\delta := \frac{\sqrt{\binom{n}{3(m-l)}(m-l)}}{\sqrt{T}}\left[T - \left(\frac{V}{E}\right)^3\right] \in [0, \infty). \tag{10}$$

Then we have, as $n \to \infty$,

$$\frac{\sqrt{\binom{n}{3(m-l)}(m-l)}[\widehat{T} - (\frac{\widehat{V}}{\widehat{E}})^3]}{\sqrt{\widehat{\mathcal{T}}}} - \delta \xrightarrow{d} N(0,1), \tag{11}$$

$$2\sqrt{\binom{n}{3(m-l)}(m-l)}\left[\sqrt{\widehat{\mathcal{T}}} - \left(\frac{\widehat{V}}{\widehat{E}}\right)^{\frac{3}{2}}\right] - \delta \xrightarrow{d} N(0,1). \tag{12}$$

When $l = 1$ and $m = 2$, Theorem 2.8 becomes Theorem 2.2 of [26].

Following (11) in Theorem 2.8, we can construct a test statistic for (7) as

$$\widehat{\mathcal{T}}_m = \frac{\sqrt{\binom{n}{3(m-l)}(m-l)}[\widehat{T} - (\frac{\widehat{V}}{\widehat{E}})^3]}{\sqrt{\widehat{T}}}. \tag{13}$$

In practice, $\widehat{T}$ might be close to zero which may cause computational instability, an alternative test can be constructed based on (12) as

$$\widehat{\mathcal{T}}'_m = 2\sqrt{\binom{n}{3(m-l)}(m-l)}\left[\sqrt{\widehat{T}} - \left(\frac{\widehat{V}}{\widehat{E}}\right)^{\frac{3}{2}}\right]. \tag{14}$$

We remark that computation of $\widehat{\mathcal{T}}_m$ and $\widehat{\mathcal{T}}'_m$ is in polynomial time since the computations of $\widehat{T}$, $\widehat{V}$, and $\widehat{E}$ all have complexity $O(n^{3(m-l)})$. Theorem 2.8 proves asymptotic normality for $\widehat{\mathcal{T}}_m$ and $\widehat{\mathcal{T}}'_m$ under both H'_0 and H'_1. Under H'_0, that is, $\delta = 0$, both $\widehat{\mathcal{T}}_m$ and $\widehat{\mathcal{T}}'_m$ are asymptotically standard normal. Under H'_1, both $\widehat{\mathcal{T}}_m$ and $\widehat{\mathcal{T}}'_m$ are asymptotically normal with mean $\delta > 0$ and unit variance. When $\widehat{T}$ has a large magnitude, both test statistics can be used to construct valid rejection regions.

The following Theorem 2.9 says that the power of our test tends to one if δ goes to infinity.

THEOREM 2.9. *Suppose* $\mathbb{E}W_1^4 = O(1)$ *and* $n^{l-1} \ll a_n \asymp b_n \ll n^{l-\frac{2}{3}}$ *for some integer* $1 \le l \le \frac{m}{2}$. *Under* H'_1, *as* $n, \delta \to \infty$, $\mathbb{P}(|\widehat{\mathcal{T}}_m| > z_{\alpha/2}) \to 1$. *The same result holds for* $\widehat{\mathcal{T}}'_m$.

REMARK 2.2. When there are multiple possible choices for l, Theorem 2.8 and Theorem 2.9 may fail if l is misspecified. For example, if $m = 4$ and the "correct" value is $l_0 = 2$ (corresponding to the true hyperedge probability), but we count 1-cycle. Then under H_0, the test statistic in (11) or (12) is of order $O_p(n^{\frac{3}{2}})$, that is, the limiting distribution does not exist. Whereas, if the correct value is $l_0 = 1$ but we count 2-cycle, then the test statistic in (11) or (12) have the same limiting distribution (if it exists) under H_0 and H_1, that is, the power of the test does not approach one. In practice, we recommend using the hyperedge proportion to get a rough estimate for l.

Theorem 2.8 and Theorem 2.9 work for relatively sparse hypergraphs. For denser hypergraphs, we propose a sparsification procedure so that Theorem 2.8 and Theorem 2.9 are valid. For any index $i_1 < i_2 < \cdots < i_m$, generate $\epsilon_{i_1 i_2 \ldots i_m} \overset{\text{iid}}{\sim} \text{Bernoulli}(r_n)$. Consider a new hypergraph with adjacency tensor $\tilde{A}$ defined by $\tilde{A}_{i_1 i_2 \ldots i_m} = \epsilon_{i_1 i_2 \ldots i_m} A_{i_1 i_2 \ldots i_m}$, where $A_{i_1 i_2 \ldots i_m}$ are the elements of the original observed adjacency tensor. Under H'_0, we have

$$\mathbb{E}[\tilde{A}_{i_1 i_2 \ldots i_m}] = (\mathbb{E}W_1)^m \frac{(r_n a_n) + (k^{m-1} - 1)(r_n b_n)}{k^{m-1} n^{m-1}}.$$

Set $\tilde{a}_n = r_n a_n$ and $\tilde{b}_n = r_n b_n$. For dense hypergraphs, we could replace A, a_n and b_n in (7) by $\tilde{A}$, $\tilde{a}_n$ and $\tilde{b}_n$ respectively. Note that the hypergraphs $\tilde{A}$ and A have the same global community structure. A properly selected r_n will make Theorem 2.8 and Theorem 2.9 valid.

COROLLARY 2.10. *Suppose* $\mathbb{E}W_1^4 = O(1)$ *and* $1 \ll a_n \asymp b_n \le n^{m-1}$. *If* $r_n = o(1)$ *and* $n^{l-1} \ll r_n a_n \asymp r_n b_n \ll n^{l-\frac{2}{3}}$ *for some integer* $1 \le l \le \frac{m}{2}$. *Then the results of Theorems* 2.8 *and* 2.9 *based on* l*-cycle continue to hold based on the sparsified hypergraph* $\tilde{A}$.

Note that Corollary 2.10 is valid for a broad range of hyperedge probabilities $\frac{1}{n^{m-1}} \ll p_n \asymp q_n \le 1$. Since H_0 and H_1 are indistinguishable when $p_n \asymp q_n \ll \frac{1}{n^{m-1}}$ (see Section 2.1), it covers all density regimes of interest. One just needs to select the sparsification factor r_n to ensure that $r_n a_n$ and $r_n b_n$ fall into the range $n^{l-1} \ll r_n a_n \asymp r_n b_n \ll n^{l-\frac{2}{3}}$, provided that one wants to use l-cycles to construct the test. The selection of l has been discussed in Remark 2.2.

REMARK 2.3. In some literature, the degree correction variable W_i in (8) are assumed to be deterministic [28, 37, 38]. In this case, Theorem 2.8 still holds under mild conditions and the proof goes through with slight modifications. To illustrate this, we consider $m = 3$. Let $W = (W_1, \dots, W_n)$ be a given and deterministic degree correction vector and denote $\|W\|_t^t = \sum_{i=1}^n W_i^t$ for positive integer t. Let $\widehat{T}_1$, $\widehat{E}_1$, $\widehat{V}_1$ be defined as

$$\widehat{T}_1 = \frac{\sum_{i_1,\dots,i_6:\text{distinct}} A_{i_1 i_2 i_3} A_{i_3 i_4 i_5} A_{i_5 i_6 i_1}}{n^6},$$

$$\widehat{V}_1 = \frac{\sum_{i_1,\dots,i_5:\text{distinct}} A_{i_1 i_2 i_3} A_{i_3 i_4 i_5}}{n^5},$$

$$\widehat{E}_1 = \frac{\sum_{i_1,i_2,i_3:\text{distinct}} A_{i_1 i_2 i_3}}{n^3}.$$

Then we have the following result.

PROPOSITION 2.11. *Suppose* $1 \ll \|W\|_t^t = O(\|W\|_1)$ *for* $2 \le t \le 12$, $\|W\|_1 \asymp \|W\|_2^2 = O(n)$, $p_0\|W\|_1^2 \gg 1$ *and* $p_0^2\|W\|_1^3 = o(1)$. *Then under* H_0' *we have*

$$\widehat{T}_3 = \sqrt{\frac{n^6}{\widehat{T}_1}}\left[\widehat{T}_1 - \left(\frac{\widehat{V}_1}{\widehat{E}_1}\right)^3\right] \xrightarrow{d} N(0,1). \tag{15}$$

Further, if $1 \ll a_n \asymp b_n \ll n^{\frac{1}{3}}$, *then the power of the test* $\widehat{T}_3$ *goes to* 1 *as* $\delta_1 \to \infty$, *where*

$$\delta_1 := \sqrt{\frac{n^6}{T_1}}\left[T_1 - \left(\frac{V_1}{E_1}\right)^3\right],$$

$$E_1 = \frac{a_n + (k^2-1)b_n}{n^2k^2}\frac{\|W\|_1^3}{n^3},$$

$$V_1 = \left(\frac{(a_n-b_n)^2}{n^4k^4} + \frac{2(a_n-b_n)b_n}{n^4k^2} + \frac{b_n^2}{n^4}\right)\frac{\|W\|_2^2\|W\|_1^4}{n^5},$$

$$T_1 = \left(\frac{(a_n-b_n)^3}{n^6k^5} + \frac{3(a_n-b_n)^2b_n}{n^6k^4} + \frac{3(a_n-b_n)b_n^2}{n^6k^2} + \frac{b_n^3}{n^6}\right)\frac{\|W\|_2^6\|W\|_1^3}{n^6}.$$

The proof of Proposition 2.11 is given in the supplement [57]. In Proposition 2.11, the conditions $p_0\|W\|_1^2 \gg 1$ and $p_0^2\|W\|_1^3 = o(1)$ require the hypergraph to be moderately sparse. At first glance, the conditions $1 \ll \|W\|_t^t = O(\|W\|_1) = O(n)$ for $2 \le t \le 12$ and $\|W\|_1 \asymp \|W\|_2^2$ seem very restrictive. However, these conditions are easy to satisfy and can accommodate severe degree heterogeneity. For example, when $W_i = \frac{i}{n}$ for $i = 1, 2, \dots, n$, we have $\|W\|_t^t = \frac{n}{t+1}(1+o(1))$ for any positive integer t. In this case, the average degrees d_1 and d_n for vertices 1 and n are

$$d_1 = \sum_{1<j<k} W_1 W_j W_k p_0 = p_0 W_1\left(0.5\left(\sum_{j=2}^n W_j\right)^2 - \sum_{j=2}^n W_j^2\right) = \frac{np_0}{8}(1+o(1)),$$

$$d_n = \sum_{j<k<n} W_n W_j W_k p_0$$
$$= p_0 W_n \left(0.5 \left(\sum_{j=1}^{n-1} W_j \right)^2 - \sum_{j=1}^{n-1} W_j^2 \right) = \frac{n^2 p_0}{8}(1 + o(1)).$$

Clearly, $d_n \asymp n d_1$ and hence the hypergraph is highly heterogeneous. Another example is to take $W_i \in [c_1, c_2]$ with positive constants $c_1 < c_2$, which yields a hypergraph with less heterogeneous degrees.

3. Extensions to nonuniform hypergraph. Nonuniform hypergraph can be viewed as a superposition of a collection of uniform hypergraphs, introduced by [32] in which the authors proposed a spectral algorithm for community detection. In this section, we study the problem of testing community structure over a nonuniform hypergraph.

Let $\mathcal{H}^k(n, M)$ be a nonuniform hypergraph over n vertices, with the vertices uniformly and independently partitioned into k communities, and $M \geq 2$ is an integer representing the maximum length of the hyperedges. Following [32], we can write $\mathcal{H}^k(n, M) = \bigcup_{m=2}^{M} \mathcal{H}_m^k(n, \frac{a_{mn}}{n^{m-1}}, \frac{b_{mn}}{n^{m-1}})$, where $\mathcal{H}_m^k(n, \frac{a_{mn}}{n^{m-1}}, \frac{b_{mn}}{n^{m-1}})$ are independent uniform hypergraphs with degree-corrected vertices introduced in Section 2.3. Correspondingly, define $\mathcal{H}(n, M) = \bigcup_{m=2}^{M} \mathcal{H}_m(n, \frac{a_{mn}+(k^{m-1}-1)b_{mn}}{k^{m-1}n^{m-1}})$ as a superposition of Erdős–Rényi models. Clearly, each Erdős–Rényi model in $\mathcal{H}(n, M)$ has the same average degree as its counterpart in $\mathcal{H}^k(n, M)$, and $\mathcal{H}(n, M)$ has no community structure. Let A_m denote the adjacency tensor for m-uniform sub-hypergraph and $A = \{A_m, m = 2, \ldots, M\}$ is a collection of A_m's. We are interested in the following hypotheses:

$$H_0'' : A \sim \mathcal{H}(n, M) \quad \textit{vs.} \quad H_1'' : A \sim \mathcal{H}^k(n, M). \tag{16}$$

3.1. *Nonuniform homogeneous hypergraphs with bounded degree.* To enhance readability, we assume $M = 3$, that is, $\mathcal{H} = \mathcal{H}_2 \cup \mathcal{H}_3$, and the hypergraphs are homogeneous without degree correction. The results are extendable to arbitrary M with more tedious arguments. The following Corollary 3.1, extending Theorem 2.1, shows that it is impossible to distinguish H_0'' and H_1'' in extremely sparse regime. The proof is essentially the same as Theorem 2.1 which also relies on the conditional independence of $\mathcal{H}_2$ and $\mathcal{H}_3$.

COROLLARY 3.1. *If $a_{mn} \asymp b_{mn} = o(1)$, then H_0'' and H_1'' are mutually contiguous.*

The following Corollary 3.2 extends the bounded degree results from Section 2.2. Let $a_{mn} = a_m$, $b_{mn} = b_m$ be positive constants, and $\kappa_m = \frac{(a_m - b_m)^2}{k^{m-1}[a_m + (k^{m-1}-1)b_m]}$.

COROLLARY 3.2. *If $\kappa_2 > 1$ or $\kappa_3 > 1$, then H_0'' and H_1'' are asymptotically orthogonal. If*

$$\left[\kappa_2 + \frac{\kappa_3}{3} \left(1 + \frac{1}{3k^2} \right) \right] (k^2 - 1) < 1,$$

then H_0'' and H_1'' are mutually contiguous. Furthermore, the results of Theorems 2.3 and 2.4 still hold with the corresponding quantities therein replaced by those in $\mathcal{H}_m$.

3.2. *Nonuniform hypergraph with growing degree.* Assume that, for $2 \le m \le M$, a_{mn}, b_{mn} are proxies of the hyperedge densities satisfying $n^{l_m-1} \ll a_{mn} \asymp b_{mn} \ll n^{l_m-\frac{2}{3}}$, for some integer $1 \le l_m \le \frac{m}{2}$.

For any $2 \le m \le M$, let $\widehat{\mathcal{T}}_m$ and δ_m be defined as in (13) and (10), respectively, based on the m-uniform sub-hypergraph. We define a test statistic for (16) as

$$\widehat{\mathcal{T}} = \sum_{m=2}^{M} c_m \widehat{\mathcal{T}}_m, \tag{17}$$

where c_m are constants with normalization $\sum_{m=2}^{M} c_m^2 = 1$. As a simple consequence of Theorems 2.8 and 2.9, we get the asymptotic distribution of $\widehat{\mathcal{T}}$ as follows.

COROLLARY 3.3. *Suppose that the degree-correction weights satisfy the same conditions as in Theorem 2.8, and for any* $2 \le m \le M$, $n^{l_m-1} \ll a_{mn} \asymp b_{mn} \ll n^{l_m-\frac{2}{3}}$, *for some integer* $1 \le l_m \le \frac{m}{2}$. *Then, as* $n \to \infty$, $\widehat{\mathcal{T}} - \sum_{m=2}^{M} c_m \delta_m \xrightarrow{d} N(0,1)$. *Furthermore, for any constant* $C > 0$, *under* H_1'', $\mathbb{P}(|\widehat{\mathcal{T}}| > C) \to 1$, *provided that* $\sum_{m=2}^{M} c_m \delta_m \to \infty$ *as* $n \to \infty$.

Under H_0'', that is, each m-uniform subhypergraph has no community structure, we have $\delta_m = 0$ by Proposition 2.7. Corollary 3.3 says that $\widehat{\mathcal{T}}$ is asymptotically standard normal. Hence, an asymptotic testing rule at significance α would be

$$\text{reject } H_0'' \quad \text{if and only if} \quad |\widehat{\mathcal{T}}| > z_{\alpha/2}.$$

The quantity $\sum_{m=2}^{M} c_m \delta_m$ may represent the degree of separation between H_0'' and H_1''. By Corollary 3.3, under H_1'', the test will achieve high power when $\sum_{m=2}^{M} c_m \delta_m$ is large.

REMARK 3.1. According to Corollary 3.3, to make $\widehat{\mathcal{T}}$ having the largest power, we need to maximize the value of $\sum_{m=2}^{M} c_m \delta_m$ subject to $\sum_{m=2}^{M} c_m^2 = 1$. The maximizer is $c_m^* = \frac{\delta_m}{\sqrt{\sum_{m=2}^{M} \delta_m^2}}$, $m = 2, 3, \ldots, M$. The corresponding test $\widehat{\mathcal{T}}^* = \sum_{m=2}^{M} c_m^* \widehat{\mathcal{T}}_m$ becomes asymptotically the most powerful among (17). In particular, $\widehat{\mathcal{T}}^*$ is more powerful than $\widehat{\mathcal{T}}_m$ for a single m. This can be explained by the additional hyperedge information involved in the test. This intuition is further confirmed by numerical studies in Section 4. Note that $\widehat{\mathcal{T}}_2$ ($m = 2$) is the classic test proposed by [26] in ordinary graph settings.

4. Numerical studies. In this section, we provide a simulation study in Section 4.1 and real data analysis in Section 4.2 to assess the finite sample performance of our tests.

4.1. *Simulation.* We generated a nonuniform hypergraph $\mathcal{H}^2(n,3) = \mathcal{H}_2^2(n, a_2, b_2) \cup \mathcal{H}_3^2(n, a_3, b_3)$, with $n = 100$ under various choices of $\{(a_m, b_m), m = 2, 3\}$. In each scenario, we calculated $Z_2 := \widehat{\mathcal{T}}_2'$ and $Z_3 := \widehat{\mathcal{T}}_3'$ by (14). Note that $Z_2 = \widehat{\mathcal{T}}_2'$ is the test for ordinary graph considered in [26]. For testing the community structure on the nonuniform hypergraph, we calculated the statistic $Z := \widehat{\mathcal{T}} = (\widehat{\mathcal{T}}_2' + \widehat{\mathcal{T}}_3')/\sqrt{2}$. In addition, we considered a strategy similar to [6] by first reducing the hypergraph to a weighted graph and applying a test designed for weighted graphs in [53]. Specifically, given an m-uniform hypergraph with hyperedges $e_1, e_2, \ldots, e_M$, we first transformed it to a weighted graph with an adjacency matrix $\widetilde{A} = [\widetilde{A}_{ij}]_{1 \le i,j \le n}$ in which $\widetilde{A}_{ij} = \sum_{k=1}^{M} I(\{i, j\} \subset e_k)$ for $i \ne j$ and $\widetilde{A}_{ij} = 0$ for $i = j$. In other words, $\widetilde{A}_{ij}$ is the total number of hyperedges containing vertices i and j. Next, we

TABLE 2
Choices of r_2, r_3, b_3 for δ to range from 1 to 10

b_3	δ	0	1	2	3	4	5	6	7	8	9	10
0.01	r_3	1	2.26	2.65	2.93	3.17	3.38	3.58	3.75	3.91	4.06	4.21
	r_2	1	2.07	2.43	2.71	2.95	3.16	3.35	3.53	3.71	3.87	4.02
0.005	r_3	1	2.89	3.51	3.98	4.39	4.75	5.08	5.38	5.67	5.94	6.20
	r_2	1	2.66	3.29	3.79	4.22	4.61	4.97	5.31	5.64	5.94	6.24
0.001	r_3	1	6.50	8.83	10.73	12.41	13.95	15.39	16.76	18.03	19.28	20.48
	r_2	1	6.57	9.31	11.59	13.64	15.51	17.26	18.92	20.51	22.00	23.46

generated a new weighted graph with an adjacency matrix $A = [A_{ij}]_{1\le i,j\le n}$ by zeroing out row s and column s of $\widetilde{A}$ if $\sum_{j=1}^{n} \widetilde{A}_{sj} > c_{thr} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \widetilde{A}_{ij}$. Here $c_{thr} > 0$ is a prespecified threshold constant.[1] We then applied the test method proposed by [53] to the weighted graph A, where the test statistic is denoted by Z_T.

We examined the size and power of each test by calculating the rejection proportions based on 500 independent replications at 5% significance level. Let δ_m denote the quantity defined in (10) which is the main factor that affects power.

Our study consists of two parts. In the first part, we investigated the power change of the four testing procedures when $\delta_2 = \delta_3 = \delta$ increases from 0 to 10. Specifically, we set $b_2 = 10b_3$, where $b_3 = 0.01, 0.005, 0.001$ represents the dense, moderately dense and sparse network, respectively; $a_m = r_m b_m$ for $m = 2, 3$ with the values of r_m summarized in Table 2. It can be checked that such choice of (a_m, b_m) indeed makes δ range from 0 to 10. We also considered both balanced and imbalanced networks with the probability (ς) of the smaller community takes the value of 0.5 and 0.3, respectively.

The rejection proportions under various settings are summarized in Figures 5, 6, and 7. Several interesting findings should be discussed. First, the rejection proportions of all test statistics except the Z_T (based on the graph transformation) at $\delta = 0$ are close to the nominal level 0.05 under different choices of ς and b_3, which demonstrates that these three test statistics are valid. We observe that the size (corresponding to $\delta = 0$) and power (corresponding to

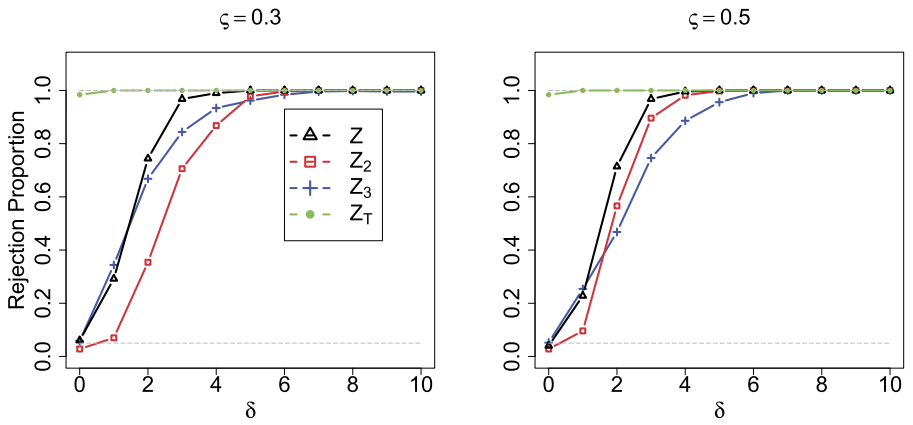

FIG. 5. *Rejection proportions in dense case with $b_3 = 0.1 \times b_2 = 0.01$.*

[1] According to the proof of Lemmas 5 and 7 in [6], c_{thr} is a large enough constant such that $\log(1+m^2x) - m^2 \ge \frac{1}{2}\log(x)$ for all $x \ge c_{thr}$. In the simulation studies, we chose c_{thr} to be the largest root of $\log(1+m^2x) - m^2 = \frac{1}{2}\log(x)$

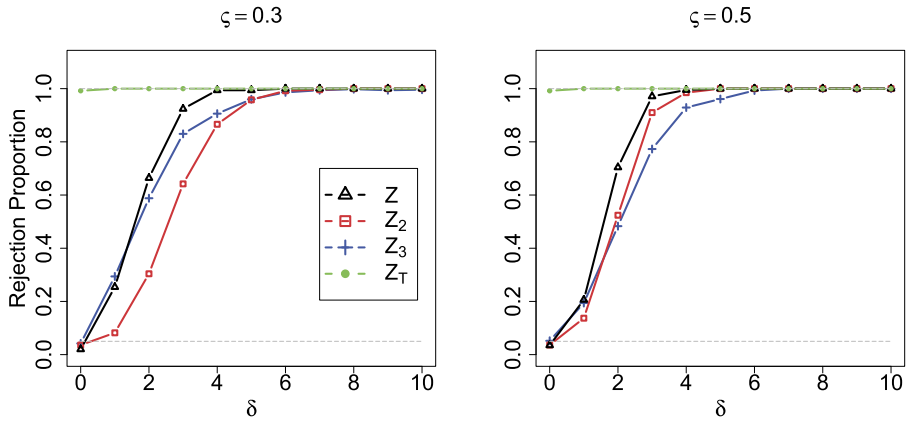

FIG. 6. *Rejection proportions in moderately dense case with* $b_3 = 0.1 \times b_2 = 0.005$.

$\delta > 0$) of the graph-transformation test are almost 100% regardless of the choice of b_3, which implies that the testing procedure Z_T is asymptotically invalid. Second, as expected, the rejection proportions of all tests increase with δ, regardless of the choices of b_3 and ς. Third, in most cases, the testing procedure based on nonuniform hypergraph (Z) has larger power than the one based only on the 3-uniform hypergraph (Z_3) or the ordinary graph (Z_2). This agrees with our theoretical finding since more information has been used in the combined test; see Remark 3.1 for a detailed explanation.

REMARK 4.1. The failure of the graph-transformation-based testing procedure Z_T is possibly due to the dependence between the edges of the transformed graph. Given the number of communities k, many existing community detection algorithms do not require the independence assumption about the edges. However, this assumption is important to derive the limit distributions of the corresponding statistics in the hypothesis testing problems about k (e.g., see [12, 26, 40, 53]). The graph-transformation-based method might still be promising for testing hypergraphs, but new asymptotic theory based on dependent edges seems necessary.

In the second part, we investigated how the powers of the tests change along with the hyperedge probability. For convenience, we report the results based on the log-scale of b_3 which ranges from -8 to -6. We chose $\delta = 1$ and 3, $\varsigma = 0.3$ and 0.5, $b_2 = 10b_3$. Similar to

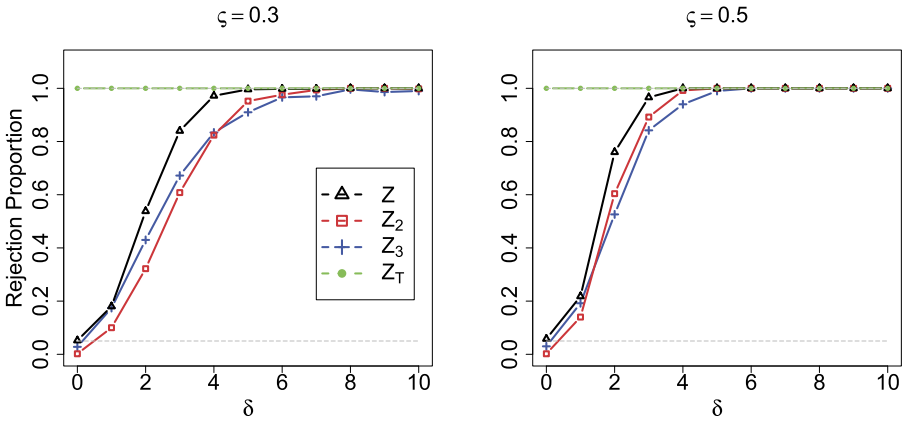

FIG. 7. *Rejection proportions in sparse case with* $b_3 = 0.1 \times b_2 = 0.001$.

TABLE 3
Choices of r_2, r_3, and δ for $\log(b_3)$ to range from -8 to -6

δ	$\log(b_3)$	-8	-7	-6
1	r_3	14.18	6.88	3.93
	r_2	15.78	7.03	3.72
3	r_3	26.37	11.51	5.82
	r_2	30.68	12.54	5.83

the first part, we set $a_m = r_m b_m$ with $m = 2$ and 3 to guarantee that $\log b_3$ indeed ranges from -8 to -6. The values of r_m were summarized in Table 3. Figures 8 and 9 report the rejection proportions for $\delta = 1$ and 3 under various hyperedge densities. We note that the rejection proportion of Z_T is always 100% under all settings. Moreover, Z is more powerful than Z_2 and Z_3 in the cases $\varsigma = 0.3, 0.5$ and $\delta = 3$. For the remaining scenarios, all procedures have satisfactory performance.

4.2. *Analysis of coauthorship data.* In this section, we applied our testing procedure to study the community structure of a coauthorship network dataset, available at https://static.aminer.org/lab-datasets/soinf/. The dataset contains a 2-author ordinary graph and a 3-author hypergraph. After removing vertices with degrees less than ten or larger than 20, we obtained a hypergraph (hereinafter referred to as global network) with 58 nodes, 110 edges, and 40 hyperedges. The vertex-removal process aims to obtain a suitably sparse network so that our testing procedure is applicable. We examined our procedures based on the global network and subnetworks. To do this, we first performed the spectral algorithm proposed by [32] to partition the global network into four subnetworks which consist of 7, 13, 14, 24 vertices, respectively (see Figure 10). In Figure 11, we plotted the incidence matrices of the 2- and 3-uniform hypergraphs, denoted 2-UH and 3-UH, respectively, as well as their superposition (Non-UH). The black dots represent vertices within the same communities. The red crosses represent vertices between different communities. An edge or hyperedge is drawn between the black dots or red crosses that are vertically aligned. It is observed that the between-community (hyper)edges are sparser than the within-community ones, indicating the validity of the partitioning.

We conducted testing procedures based on Z_2, Z_3, and Z at significance level 0.05 (similar to Section 4.1) to both global network and subnetworks. The values of the test statistics are

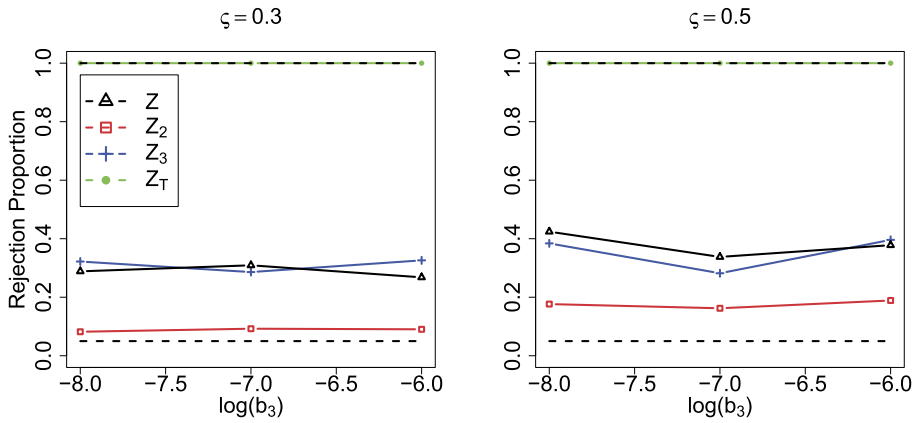

FIG. 8. *Rejection proportions when $\delta = 1$ and $b_2 = 10b_3$.*

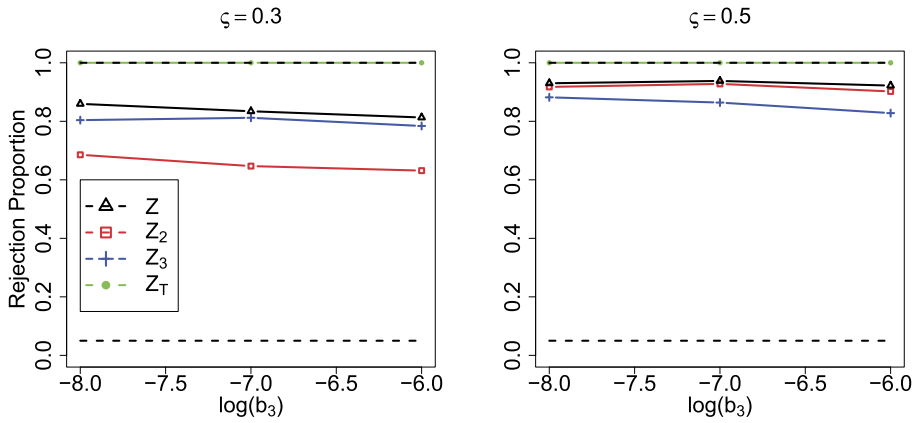

FIG. 9. *Rejection proportions when* $\delta = 3$ *and* $b_2 = 10b_3$.

summarized in Table 4. Observe that Z_2 and Z yield very large test values for the global network indicating strong rejection of the null hypothesis. For subnetwork testing, Z_2 rejects the null hypothesis for subnetwork 3; while Z_3 and Z do not reject the null hypotheses for any subnetworks. This demonstrates that the community detection results are reasonable in general, and the subnetworks may no longer have finer community structures.

5. Discussion. In the context of community testing for hypergraphs, we systematically considered various scenarios in terms of hyperedge densities and investigated distinguishability or indistinguishability of the hypotheses in each scenario. Extensions of our results are possible.

The first line is to extend the test statistic in Section 2.3 to tackle the model selection problem for SBM in hypergraphs. In particular, one possibility is to study the hypothesis testing problem of $H_0 : k = k_0$ vs. $H_1 : k > k_0$ for $k_0 = 1, 2, \ldots$ sequentially and stop when observing a rejection. The second line is to extend the current results to the increasingly

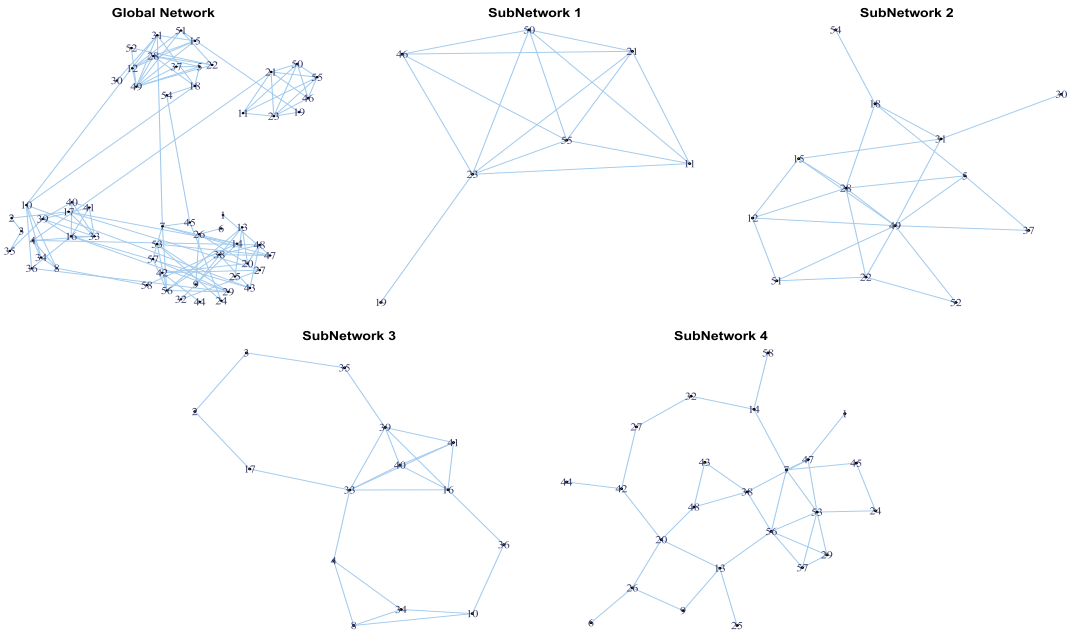

FIG. 10. *Global network and four subnetworks based on coauthorship data.*

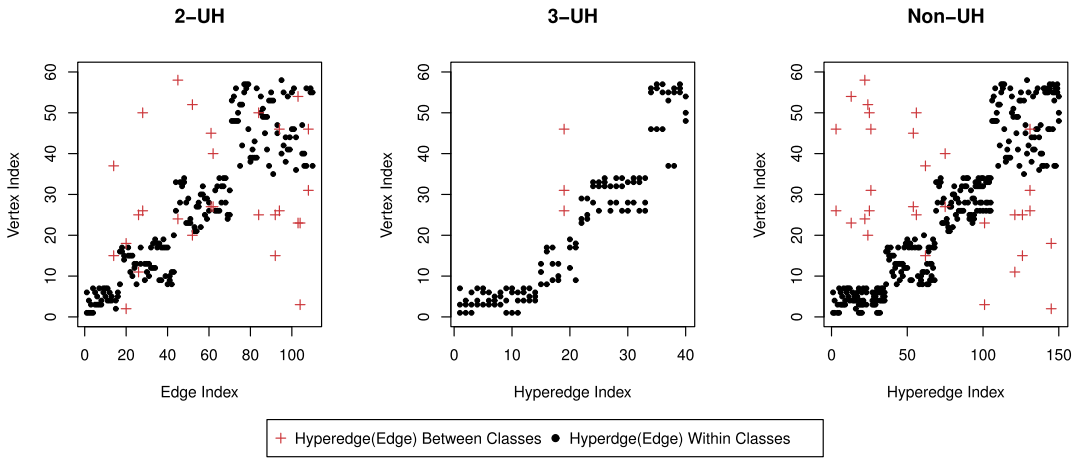

FIG. 11. *Incidence matrices based on coauthorship data. Left*: 2-*uniform hypergraph*; *Middle*: 3-*uniform hypergraph*; *Right*: *nonuniform hypergraph.*

popular degree-corrected stochastic block models. However, based on the current second-moment technique, the bounded degree results are not easy to establish. The main reason is that the moments of the likelihood ratio do not have an explicit expression in terms of a, b, κ. Even in the ordinary graph setting with $m = 2$, this is already very difficult. To see this, when $\sigma_i = \sigma_j$ and $\eta_i = \eta_j$ for all $i < j$, it can be shown that

$$
\begin{aligned}
&E_W E_{W^*} \prod_{i<j} \left(\frac{p_{ij}(\sigma, W) p_{ij}(\eta, W^*)}{p_0} + \frac{q_{ij}(\sigma, W) q_{ij}(\eta, W^*)}{q_0} \right) \\
&\approx E_W E_{W^*} \exp\left(\beta \sum_{i<j} (W_i W_j a - d)(W_i^* W_j^* a - d) \right),
\end{aligned} \tag{18}
$$

where $\beta = \frac{1}{dn^\alpha} + \frac{1}{n^{2\alpha}}$. The expected value (18) seems difficult to analyze under general random weights W, W^*, even for the above special choice of σ, η. Hence, a precise contiguity region in terms of a, b, κ is not available using the current second-moment method.

The third line is to test more general and complicated hypotheses. The current paper only deals with the relatively simple Erdős–Rényi null hypotheses, whereas the proposed methods may be extended to more general settings. For instance, in light of Theorem 2.3, the test statistics based on long loose cycles may also test the null hypothesis that the hypergraph is an SBM with k communities in which $k > 1$ is given; in light of Theorems 2.8 and 2.9, the test statistics based on sub-hypergraph counts may be extended to the null hypothesis that the model is a degree-corrected SBM with k communities. It is a worthwhile project to investigate the validity of these methods, especially in real-world situations.

TABLE 4

*Values of test statistics based on global network and four subnetworks. Symbols ** and * indicate the strength of rejection, that is, p-value* < 0.001 *and p-value* < 0.05 *respectively*

	Global Network	SubNetwork 1	SubNetwork 2	SubNetwork 3	SubNetwork 4
n	58	7	13	14	24
Z_2	8.360**	0.161	−0.030	2.667*	1.661
Z_3	1.451	−0.100	−0.211	−0.289	−0.052
Z	6.938**	0.043	−0.171	1.682	1.137

Acknowledgement. We are grateful to the co-Editor Professor Richard Samworth, the Associate Editor, and referees for their insightful comments which greatly improved the quality and scope of the paper. We thank Sumit Mukherjee for suggesting the truncation technique which motivates Theorem 2.6.

Funding. The third author was supported by NSF CAREER Grant DMS-2013789. The fourth author was supported by NSF Grants DMS-1764280 and DMS-1821157.

SUPPLEMENTARY MATERIAL

Supplement to "Testing community structure for hypergraphs" (DOI: 10.1214/21-AOS2099SUPP; .pdf). The supplementary material [57] contains all proofs to the theorems and lemmas in this paper.

REFERENCES

[1] ABBE, E. (2017). Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18** 177. MR3827065

[2] ABBE, E. and SANDON, C. (2016). Detection in the stochastic block model with multiple clusters: Proof of the achievability conjectures, acyclic BP, and the information-computation gap. https://arxiv.org/pdf/1512.09080.pdf.

[3] ABBE, E. and SANDON, C. (2018). Proof of the achievability conjectures for the general stochastic block model. *Comm. Pure Appl. Math.* **71** 1334–1406. MR3812075 https://doi.org/10.1002/cpa.21719

[4] AGARWAL, S., BRANSON, K. and BELONGIE, S. (2006). Higher order learning with graphs. In *Proceedings of the International Conference on Machine Learning* 17–24.

[5] AHN, K., LEE, K. and SUH, C. (2016). Community recovery in hypergraphs. In 54*th Annual Allerton Conference on Communication*, *Control*, *and Computing* (*Allerton*). https://doi.org/10.1109/ALLERTON.2016.7852294

[6] AHN, K., LEE, K. and SUH, C. (2018). Hypergraph spectral clustering in the weighted stochastic block model. *IEEE J. Sel. Top. Signal Process.* **12**.

[7] AMINI, A. A., CHEN, A., BICKEL, P. J. and LEVINA, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41** 2097–2122. MR3127859 https://doi.org/10.1214/13-AOS1138

[8] ANGELINI, M., CALTAGIRONE, F., KRZAKALA, F. and ZDEBOROVA, L. (2015). Spectral detection on sparse hypergraphs. In *Allerton Conference on Communication*, *Control*, *and Computing* 66–73.

[9] BANERJEE, D. (2018). Contiguity and non-reconstruction results for planted partition models: The dense case. *Electron. J. Probab.* **23** 18. MR3771755 https://doi.org/10.1214/17-EJP128

[10] BANERJEE, D. and MA, Z. (2017). Optimal hypothesis testing for stochastic block models with growing degrees. Available at https://arxiv.org/pdf/1705.05305.pdf.

[11] BANKDS, J., MOORE, C., NEEMAN, J. and NETRAPALLI, P. (2016). Information-theoretic thresholds for community detection in sparse networks. *J. Mach. Learn. Res. Workshop Conf. Proc.* **49** 1–34.

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

[13] BOLLA, M. (1993). Spectra, Euclidean representations and clusterings of hypergraphs. *Discrete Math.* **117** 19–39. MR1226129 https://doi.org/10.1016/0012-365X(93)90322-K

[14] BOLLOBÁS, B. (2001). *Random Graphs*, 2nd ed. *Cambridge Studies in Advanced Mathematics* **73**. Cambridge Univ. Press, Cambridge. MR1864966 https://doi.org/10.1017/CBO9780511814068

[15] BOLLOBÁS, B. and ERDŐS, P. (1976). Cliques in random graphs. *Math. Proc. Cambridge Philos. Soc.* **80** 419–427. MR0498256 https://doi.org/10.1017/S0305004100053056

[16] CHEN, J. and YUAN, B. (2006). Detecting functional modules in the yeast proteinprotein interaction network. *Bioinformatics* **22** 2283–2290.

[17] CHERTOK, M. and KELLER, Y. (2010). Efficient high order matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **32** 2205–2215.

[18] CHIEN, I., LIN, C. and WANG, I. (2018). Community detection in hypergraphs: Optimal statistical limit and efficient algorithms. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* **84** 871–879.

[19] DALL'AMICO, L. and COUILLET, R. (2019). Community detection in sparse reallistic graphs: Improving the Bethe Hessian. In *ICASSP* 2019 18778248.

[20] DALL'AMICO, L., COUILLET, R. and TREMBLAY, N. (2019). Revisiting the Bethe-Hessian: Improved community detection in sparse heterogeneous graphs. In *NIPS* 2019.
[21] ERDŐS, P. and RÉNYI, A. (1960). On the evolution of random graphs. *Magy. Tud. Akad. Mat. Kut. Intéz. Közl.* **5** 17–61. MR0125031
[22] ESTRADA, E. and RODRIGUEZ-VELASQUEZ, J. (2005). Complex networks as hypergraphs. Available at https://arxiv.org/ftp/physics/papers/0505/0505137.pdf.
[23] FLORESCU, L. and PERKINS, W. (2016). Spectral thresholds in the bipartite stochastic block model. In 29*th Annual Conference on Learning Theory* **49** 943–959.
[24] FORTUNATO, S. (2010). Community detection in graphs. *Phys. Rep.* **486** 75–174. MR2580414 https://doi.org/10.1016/j.physrep.2009.11.002
[25] FRIEZE, A. and KARONSKI, M. (2015). *Introduction to Random Graphs*. Cambridge Univ. Press, Cambridge.
[26] GAO, C. and LAFFERTY, J. (2017). Testing for global network structure using small subgraph statistics. Available at https://arxiv.org/pdf/1710.00862.pdf.
[27] GAO, C. and LAFFERTY, J. (2017). Testing network structure using relations between small subgraph probabilities. Available at https://arxiv.org/pdf/1704.06742.pdf.
[28] GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2018). Community detection in degree-corrected block models. *Ann. Statist.* **46** 2153–2185. MR3845014 https://doi.org/10.1214/17-AOS1615
[29] GAO, Z. and WORMALD, N. C. (2004). Asymptotic normality determined by high moments, and submap counts of random maps. *Probab. Theory Related Fields* **130** 368–376. MR2095934 https://doi.org/10.1007/s00440-004-0356-9
[30] GHOSHAL, G., ZLATIĆ, V., CALDARELLI, G. and NEWMAN, M. E. J. (2009). Random hypergraphs and their applications. *Phys. Rev. E* **79** 066118. MR2551286 https://doi.org/10.1103/PhysRevE.79.066118
[31] GHOSHDASTIDAR, D. and DUKKIPATI, A. (2014). Consistency of spectral partitioning of uniform hypergraphs under planted partition model. *Adv. Neural Inf. Process. Syst.* 397–405.
[32] GHOSHDASTIDAR, D. and DUKKIPATI, A. (2017). Consistency of spectral hypergraph partitioning under planted partition model. *Ann. Statist.* **45** 289–315. MR3611493 https://doi.org/10.1214/16-AOS1453
[33] GIBSON, D., KLEINBERG, J. and RAGHAVAN, P. (2000). Clustering categorical data: An approach based on dynamical systems. *VLDB J.* **8** 222–236.
[34] GOLDENBERG, A., ZHENG, A. X. S., FIENBERG, E. and AIROLDI, E. M. (2010). A survey of statistical network models. *Found. Trends Mach. Learn.* **2** 129–233.
[35] GOVINDU, V. M. (2005). A tensor decomposition for geometric grouping and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) 1150–1157.
[36] JANSON, S. (1995). Random regular graphs: Asymptotic distributions and contiguity. *Combin. Probab. Comput.* **4** 369–405. MR1377557 https://doi.org/10.1017/S0963548300001735
[37] KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83** 016107. MR2788206 https://doi.org/10.1103/PhysRevE.83.016107
[38] KE, Z., SHI, F. and XIA, D. (2020). Community Detection for Hypergraph Networks via Regularized Tensor Power Iteration. Available at https://arxiv.org/pdf/1909.06503.pdf.
[39] KIM, C., BANDEIRA, A. and GOEMANS, M. (2017). Community detection in hypergraphs, spiked tensor models, and sum-of-squares. In 2017 *International Conference on Sampling Theory and Applications* (*SampTA*) 124–128.
[40] LEI, J. (2016). A goodness-of-fit test for stochastic block models. *Ann. Statist.* **44** 401–424. MR3449773 https://doi.org/10.1214/15-AOS1370
[41] LIN, C., CHIEN, I. and WANG, I. (2017). On the fundamental statistical limit of community detection in random hypergraphs. In *Information Theory* (*ISIT*), 2017 *IEEE International Symposium* 2178–2182.
[42] MICHOEL, T. and NACHTERGAELE, B. (2012). Alignment and integration of complex networks by hypergraph-based spectral clustering. *Phys. Rev. E* **86**.
[43] MONTANARI, A. and SEN, S. (2016). Semidefinite programs on sparse random graphs and their application to community detection. In *STOC'*16*—Proceedings of the* 48*th Annual ACM SIGACT Symposium on Theory of Computing* 814–827. ACM, New York. MR3536616 https://doi.org/10.1145/2897518.2897548
[44] MOSSEL, E., NEEMAN, J. and SLY, A. (2015). Reconstruction and estimation in the planted partition model. *Probab. Theory Related Fields* **162** 431–461. MR3383334 https://doi.org/10.1007/s00440-014-0576-6
[45] MOSSEL, E., NEEMAN, J. and SLY, A. (2018). A proof of the block model threshold conjecture. *Combinatorica* **38** 665–708. MR3876880 https://doi.org/10.1007/s00493-016-3238-8
[46] NEEMAN, J. and NETRAPALLI, P. (2014). Non-reconstructability in the stochastic block model. Available at https://arxiv.org/abs/1404.6304.

[47] NEWMAN, M. (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E* **64** 016–131.

[48] OUVRARD, X., GOFF, J. and MARCHAND-MAILLET, S. (2017). Networks of Collaborations: Hypergraph modeling and visualisation. Available at https://arxiv.org/pdf/1707.00115.pdf.

[49] RAMASCO, J., DOROGOVTSEV, S. N. and PASTOR-SATORRAS, R. (2004). Self-organization of collaboration networks. *Phys. Rev. E* **70** 036–106.

[50] RODRÍGUEZ, J. A. (2009). Laplacian eigenvalues and partition problems in hypergraphs. *Appl. Math. Lett.* **22** 916–921. MR2523606 https://doi.org/10.1016/j.aml.2008.07.020

[51] ROTA BULO, S. and PELILLO, M. (2013). A game-theoretic approach to hypergraph clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 1312–1327.

[52] SHI, J. and MALIK, J. (1997). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 888–905.

[53] TOKUDA, T. (2018). Statistical test for detecting community structure in real-valued edge-weighted graphs. *PLoS ONE* **13** e0194079. https://doi.org/10.1371/journal.pone.0194079

[54] WORMALD, N. C. (1999). Models of random regular graphs. In *Surveys in Combinatorics*, 1999 (*Canterbury*). *London Mathematical Society Lecture Note Series* **267** 239–298. Cambridge Univ. Press, Cambridge. MR1725006

[55] YUAN, M., FENG, Y. and SHANG, Z. (2018). A likelihood-ratio type test for stochastic block models with bounded degrees. Available at https://arxiv.org/pdf/1807.04426.pdf.

[56] YUAN, M., FENG, Y. and SHANG, Z. (2018). Inference on multi-community stochastic block models with bounded degree. Manuscript.

[57] YUAN, M., LIU, R., FENG, Y. and SHANG, Z. (2022). Supplement to "Testing community structure for hypergraphs." https://doi.org/10.1214/21-AOS2099SUPP

[58] ZHAO, Y., LEVINA, E. and ZHU, J. (2011). Community extraction for social networks. *Proc. Natl. Acad. Sci. USA* **108** 7321–7326.

[59] ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40** 2266–2292. MR3059083 https://doi.org/10.1214/12-AOS1036