Sparse Hypergraph Community Detection Thresholds in Stochastic Block Model

Erchuan Zhang
School of Science
Edith Cowan University
erchuan.zhang@ecu.edu.au

David Suter
School of Science
Edith Cowan University
d.suter@ecu.edu.au

Giang Truong
School of Science
Edith Cowan University
h.truong@ecu.edu.au

Syed Zulqarnain Gilani
School of Science
Edith Cowan University
s.gilani@ecu.edu.au

Abstract

Community detection in random graphs or hypergraphs is an interesting fundamental problem in statistics, machine learning and computer vision. When the hypergraphs are generated by a *stochastic block model*, the existence of a sharp threshold on the model parameters for community detection was conjectured by Angelini et al. 2015. In this paper, we confirm the positive part of the conjecture, the possibility of non-trivial reconstruction above the threshold, for the case of two blocks. We do so by comparing the hypergraph stochastic block model with its Erdös-Rényi counterpart. Furthermore, we show the negative part of the conjecture by relating the model with the so-called *multi-type Galton-Watson hypertrees* and considering the broadcasting problem on these hypertrees. The methods developed in this paper are generalised from the study of sparse random graphs by Mossel et al. 2015.

1 Introduction

Community detection, or clustering, aims to identify groups sharing similar properties from a global population. It is a fundamentally important problem in statistics, network analysis and computer vision [15, 36, 11, 31, 27]. Many clustering algorithms have been developed based on graphs in which each edge contains exactly two vertices representing pairwise relationships on data. However, real-world data are often much more complicated than the ordinary graph structures. For example, in the co-authorship network [9, 25, 28], the number of co-authors may vary from paper to paper so that one can hardly study edges consisting of two co-authors only.

To represent higher order relationships among data, *hypergraphs* have been proposed to model data and have been shown to have some advantages over graphs [38]. In the co-authorship example above, a generalised edge (called *hyperedge*) allows us to consider the connectivity of arbitrarily many co-authors. In recent years, hypergraphs have been widely used to model complex data relationships, for instance, in robust multi-structure clustering [27], bioinformatics [34], and social networks extraction [37]. However, from both theoretical and methodological sides, the complexity of hypergraphs creates much more challenges in the sense that many analysis tools developed in the graph case can not be easily generalised to the hypergraph case. In general, analysing the spectral properties of the adjacency tensor of a hypergraph is more difficult than studying the spectrum of the adjacency matrix of a graph.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

The *stochastic block model* (SBM), also known as the *planted partition model*, was initially proposed by Holland et al. [17] to model random graphs with community structures. The general idea behind this model is to label each vertex as a community member identically and independently, and then construct a random graph based on the membership and some given probabilities. It becomes the well-known Erdös-Rényi graph model if we ignore the membership restriction. The recovery or detection problem, based on the SBM, is to find the community membership when the number of vertices is sufficiently large. SBM often serves as a benchmark for clustering algorithms on graph data. Likewise, an analogous model called the *hypergraph stochastic block model* (HSBM) was proposed to model higher-order relationships for random hypergraphs [13]. In what follows, we will present a brief introduction to hypergraphs, HSBM and its associated detection problem.

1.1 The stochastic block model for hypergraphs

A *hypergraph* is a pair $H = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of vertices and $\mathcal{E}$ is a set of *hyperedges*, which is a collection of subsets of $\mathcal{V}$. The *degree* of a vertex $v \in \mathcal{V}$ is the number of hyperedges in H that contains v. The *degree* of a hyperedge $e \in \mathcal{E}$ is the number of vertices contained in e. The hypergraph H is called *d-uniform* if the degree of every hyperedge $e \in \mathcal{E}$ is d. A 2-uniform hypergraph is a graph. An *l-cycle* in H is a cyclic ordering of a subset of the vertex set, and of hyperedges, where any two adjacent hyperedges have exactly l common vertices. We call an l-cycle *loose* if $l = 1$. For instance, let $H_3 = (\mathcal{V}, \mathcal{E})$ be a 3-uniform hypergraph, where $\mathcal{V} = \{v_1, v_2, \cdots, v_7\}$, $\mathcal{E} = \{(v_i, v_j, v_k) | 1 \leq i < j < k \leq 7\}$. Then $\{(v_1, v_2, v_3), (v_3, v_4, v_5), (v_5, v_6, v_1)\}$ is a loose cycle.

There are many ways to represent a hypergraph. For instance, the *incidence matrix* of a hypergraph H is defined as $A \in \{0, 1\}^{|\mathcal{V}| \times |\mathcal{E}|}$, where $A(v, e) = 1$ if the vertex $v \in \mathcal{V}$ is contained in the hyperedge $e \in \mathcal{E}$, and 0 otherwise. The *adjacency tensor* of H is defined as T, where $T_{i_1, \cdots, i_d} = 1$ if $\{i_1, \cdots, i_d\}$ is a hyperedge of H, and 0 otherwise. The *adjacency matrix* of H is defined as $B = (B_{ij})_{|\mathcal{V}| \times |\mathcal{V}|}$, where $B_{ij} = \sum_{e: \{i,j\} \in e} T_e$ if $i \neq j$, and 0 otherwise, here e is a hyperedge and T_e is an element in its associated adjacency tensor. Unfortunately, most of the computations involving the adjacency tensor are NP-hard, for examples, determining the rank-1 approximation of a tensor, whether a tensor has a given spectral norm [16].

For simplicity, from now on, we specify our d-uniform hypergraph stochastic block model with two roughly balanced communities. Considering non-uniform hypergraphs or hypergraphs with multi-communities may require extra effort, which will be reserved for future work. Let $H = ([n], \mathcal{E})$ be a d-uniform hypergraph with vertex set $[n] := \{1, 2, \cdots, n\}$ and hyperedge set $\mathcal{E}$, $\sigma := (\sigma_1, \cdots, \sigma_n) = \{+1, -1\}^n$ the spins on $[n]$, which means each vertex $i \in [n]$ is assigned with a spin $\sigma_i \in \{+1, -1\}$. Let $\mathcal{S}_n$ be the set of all pairs (H, σ), we can generate a random pair (H, σ) from the finite set $\mathcal{S}_n$ as follows ([13, 14, 20, 26]):

- First generate i.i.d random variables $\sigma_i \in \{+1, -1\}$ uniformly for each $i \in [n]$.
- Then, for the obtained $\sigma = (\sigma_1, \cdots, \sigma_n)$, we generate a random d-uniform hypergraph H where an hyperedge $e = \{i_1, \cdots, i_d\}$ is included independently with probability p_n if $\sigma_{i_1} = \cdots = \sigma_{i_d}$, and with probability q_n otherwise, where $0 < q_n < p_n < 1$ (p_n, q_n possibly depending on n).

We denote by $\mathcal{H}_d(n, p_n, q_n)$ a d-uniform hypergraph generated by the process above. (With a little notation abuse, we sometimes denote the distribution that such hypergraphs follow by the same notation). In particular, $\mathcal{H}_2(n, p_n, q_n)$ is a random graph generated by the graph version of the SBM.

Suppose $\mathcal{C}_1 = \{i \in [n] | \sigma_i = +1\}$ and $\mathcal{C}_2 = \{i \in [n] | \sigma_i = -1\}$ are two communities in the hypergraph H. The goal of community detection is to estimate the unknown spin σ up to a sign flip by observing H only from a sample (H, σ) drawn from $\mathcal{H}_d(n, p_n, q_n)$, i.e., to find estimators $\hat{\sigma}_i \in \{+1, -1\}$ that are correlated with the true partition σ_i. By correlation, we mean there exists a constant $c \in (0, 1)$, for any $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}(\{|\mathcal{O}_n(\hat{\sigma}, \sigma) - c| > \epsilon\} \cap \{|\mathcal{O}_n(\hat{\sigma}, \sigma) + c| > \epsilon\}) = 0,$$

where $\mathcal{O}_n(\hat{\sigma}, \sigma) := \frac{1}{n} \sum_{i \in [n]} \sigma_i \hat{\sigma}_i$ is the empirical overlap between σ and $\hat{\sigma}$.

Recall the definitions of exact recovery and almost exact recovery of communities in the literature (for example, see [26]). By *exact recovery*, we mean finding a spin vector $\hat{\sigma} \in \{+1, -1\}^n$ such that

$$\lim_{n\to\infty} \mathbb{P}(\{\mathcal{O}_n(\hat{\sigma}, \sigma) = 1\} \cup \{\mathcal{O}_n(\hat{\sigma}, \sigma) = -1\}) = 1.$$

That is, the estimated spin vector $\hat{\sigma}$ is either exactly the true spin vector σ or exactly the negative true spin vector $-\sigma$. By *almost exact recovery*, we mean finding a spin vector $\hat{\sigma} \in \{+1, -1\}^n$ such that, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\{|\mathcal{O}_n(\hat{\sigma}, \sigma) - 1| > \epsilon\} \cap \{|\mathcal{O}_n(\hat{\sigma}, \sigma) + 1| > \epsilon\}) = 0.$$

That is, the empirical overlap $\mathcal{O}_n(\hat{\sigma}, \sigma)$ is in $[-1, -1+\epsilon) \cup (1-\epsilon, 1]$ asymptotically almost surely. This case almost exactly recovers the community, as indicated by the terminology, compared with the definition of exact recovery. Note that we only consider the detection (not (almost) exact recovery) of communities in this paper.

The study of *sparse* graphs/hypergraphs as well as those with constant average degree, is well motivated from the viewpoint of real networks. For instance, many large social networks like Twitter or LinkedIn, may have several millions of nodes but only ten times more edges [21]. A coauthorship network may have millions of nodes, but the average degree is no more than 20 in general [25]. Therefore, It is natural to consider the sparse HSBM with p_n and q_n of the order $\mathcal{O}\left(1/\binom{n}{d-1}\right)$.

Decelle et al. [7] conjectured the existence of a sharp threshold, called *Kesten-Stigun threshold*, of detection in the sparse graph SBM based on some ideas from statistical physics. More precisely, if $p_n = \frac{a}{n}$, $q_n = \frac{b}{n}$, where a, b are constants independent of n, then detection is possible if and only if $(a-b)^2 > 2(a+b)$. This conjecture was confirmed in [22, 23, 24, 6], for examples, by studying the spectra of the non-backtracking [6], self-avoiding [22] or graph powering [1] matrices. Moreover, some efficient algorithms were proposed to achieve detection at the threshold [6, 32, 1].

In the case of the hypergraph SBM, Angelini et al. [2] conjectured the existence of a sharp threshold on the model parameters, for community detection in the sparse hypergraph. If $p_n = \frac{a}{\binom{n}{d-1}}$, $q_n = \frac{b}{\binom{n}{d-1}}$, i.e., the hypergraph $H \sim \mathcal{H}_d(n, p_n, q_n)$ has constant expected vertex degree. Pal et al. [26] proved the positive part of the conjecture, the detection is possible if $\beta^2 > \alpha$, where $\alpha = (d-1)\frac{a+(2^{d-1}-1)b}{2^{d-1}}$, $\beta = (d-1)\frac{a-b}{2^{d-1}}$. Their method is based on the spectral analysis of the so-called *self-avoiding matrix* associated with the hypergraph, which is a generalisation to random hypergraphs of the method developed by Massoulié [22] for sparse random graphs. Moreover, the consistent estimators for model parameters a and b are still unknown.

In the present paper, we reconfirm the positive part of the conjecture by comparing the sparse HSBM with an appropriate Erdös-Rényi hypergraph model (see Theorem 1.1). Further, in the same theorem we give consistent estimators for a and b if the detection is possible. More importantly, we show that it is impossible (see Theorem 1.3) to detect communities if $\beta^2 < \alpha$ by coupling our hypergraph with the multi-type Galton-Watson hypertree and studying broadcasting problems on hypertree. The threshold of broadcasting on a (hyper)tree results in that of detection on the HSBM. Our methods are motivated by the study of the community detection problem on the graph SBM conducted by Mossel et al. [23]. The challenges in generalizing the results to hypergraphs and the way how we overcame them are summarized at the end of the supplementary material. The main results of the present paper are summarised as follows.

1.2 Main results

Let $\mathcal{H}_d\left(n, \frac{p_n+(2^{d-1}-1)q_n}{2^{d-1}}\right)$ be the Erdös-Rényi model in which each hyperedge is included with a common probability $\frac{p_n+(2^{d-1}-1)q_n}{2^{d-1}}$, which is chosen to make sure this model has the same average degree as $\mathcal{H}_d(n, p_n, q_n)$. In other words, $\mathcal{H}_d\left(n, \frac{p_n+(2^{d-1}-1)q_n}{2^{d-1}}\right)$ represents a d-uniform random hypergraph without any community structures. In particular, $\mathcal{H}_2\left(n, \frac{p_n+q_n}{2}\right)$ is the traditional Erdös-Rényi random graph that has been extensively studied in the graph literature [4, 12, 8].

Let $\mathbb{P}_n$ and $\tilde{\mathbb{P}}_n$ denote the probability measures with respect to $\mathcal{H}_d(n, p_n, q_n)$ and $\mathcal{H}_d\left(n, \frac{p_n+(2^{d-1}-1)q_n}{2^{d-1}}\right)$, respectively. Then $\mathbb{P}_n$ and $\tilde{\mathbb{P}}_n$ are said to be *mutually contiguous* if for any

measurable sets A_n, $\mathbb{P}_n(A_n) \to 0$ if and only if $\tilde{\mathbb{P}}_n(A_n) \to 0$ as $n \to \infty$. $\mathbb{P}_n$ and $\tilde{\mathbb{P}}_n$ are said to be *asymptotically orthogonal* if there exists a sequence of measurable sets A_n such that $\mathbb{P}_n(A_n) \to 0$ and $\tilde{\mathbb{P}}_n(A_n) \to 1$ as $n \to \infty$. From now on, we fix the following notation,

$$p_n := \frac{a}{\binom{n}{d-1}}, \quad q_n := \frac{b}{\binom{n}{d-1}},$$

where $a > b > 0$ are some constants independent of n, $d \geq 2$ is an integer. Define

$$\alpha := (d-1)\frac{a + (2^{d-1}-1)b}{2^{d-1}}, \quad \beta := (d-1)\frac{a-b}{2^{d-1}},$$

where α measures the expected degree of any vertex and β measures the discrepancy between the number of neighbors with different signs of any vertex.

Theorem 1.1. *If $\beta^2 > \alpha$, then $\mathbb{P}_n$ and $\tilde{\mathbb{P}}_n$ are asymptotically orthogonal. Let X_{ζ_n} be the number of loose cycles of length ζ_n and define*

$$\hat{\alpha}_n := \frac{d|\mathcal{E}|}{\binom{n}{d-1}}, \quad \hat{\beta}_n := (2\zeta_n X_{\zeta_n} - \hat{\alpha}_n^{\zeta_n})^{\frac{1}{\zeta_n}},$$

where $\zeta_n = \lfloor \log^{1/4} n \rfloor$, $|\mathcal{E}|$ is the number of observed hyperedges, then $\hat{a}_n = \frac{1}{d-1}\left(\hat{\alpha}_n + (2^{d-1}-1)\hat{\beta}_n\right)$ and $\hat{b}_n = \frac{1}{d-1}\left(\hat{\alpha}_n - \hat{\beta}_n\right)$ are consistent estimators for a and b, respectively. Namely, $\hat{a}_n \to a$ and $\hat{b}_n \to b$ in probability as $n \to \infty$.

Theorem 1.2. *If $\beta^2 < \alpha$, then for any fixed vertices v_1 and v_2, $H \sim \mathcal{H}_d(n, p_n, q_n)$,*

$$\lim_{n\to\infty} \mathbb{P}_n(\sigma_{v_1} = +1 | H, \sigma_{v_2}) = \frac{1}{2}.$$

Theorem 1.3. *If $\beta^2 < \alpha$, then $\mathbb{P}_n$ and $\tilde{\mathbb{P}}_n$ are mutually contiguous. Further more, there is no consistent estimator for a and b if $\beta^2 < \alpha$.*

When $d = 2$, the condition $\beta^2 = \alpha$ becomes $(a-b)^2 = 2(a+b)$, which is the threshold for community detection in the (graph-based) SBM. Readers may refer to [22, 23, 24] for more detail. Intuitively speaking, Theorem 1.1 says that $\mathcal{H}_d(n, p_n, q_n)$ and $\mathcal{H}_d\left(n, \frac{p_n + (2^{d-1}-1)q_n}{2^{d-1}}\right)$ are asymptotically almost surely distinguishable above the threshold (i.e., can distinguish a community with structure from a random hypergraph with the same average degree). While Theorem 1.3 states that they are not distinguishable below the threshold. Figure 1 shows the phase transition of community detection in the sparse HSBM where $d = 3, 4, 5$. Consider any two random vertices of a hypergraph H, Theorem 1.2 claims that no algorithm can tell whether they have the same spin, i.e., the probability of identifying them with the same spin is asymptotically almost surely the same as the probability of finding them with different spins. Put loosely, it is not possible to do any better than random chance when assigning labels.

The remainder of the present paper is organised as follows: In section 2, we study the probability distribution of the number of ζ_n-loose cycle X_{ζ_n} and then prove Theorem 1.1. In section 3, based on the multi-type Galton-Watson hypertree proposed in [26], we investigate the connection between the detection problem on $\mathcal{H}_d(n, p_n, q_n)$ and the broadcasting problem on the hypertree, which provides some tools to prove Theorem 1.2. Mossel et al. in [23] applied Theorem 4.1 in [35] that gave sufficient conditions for contiguity to prove Theorem 1.3 where $d = 2$ (i.e., the graph SBM case). To prove Theorem 1.3 in the sparse HSBM case, we need to verify all conditions of Theorem 4.1 in [35]. Section 4 concludes the paper, including some further discussion.

2 The distribution of loose cycles

In this section, we mainly show that the number of ζ_n-loose cycles of $H \sim \mathcal{H}_d(n, p_n, q_n)$ or $H \sim \mathcal{H}_d\left(n, \frac{p_n + (2^{d-1}-1)q_n}{2^{d-1}}\right)$ is asymptotically approximately Poisson-distributed. In general inhomogeneous graphs, Bollobás et al. [5] demonstrated that the number of cycles asymptotically follows the Poisson distribution. We borrow their proof techniques to get a similar conclusion in the hypergraph case. Then, by comparing the first two moments of the number of ζ_n-loose cycles, we can find that $\mathcal{H}_d(n, p_n, q_n)$ and $\mathcal{H}_d\left(n, \frac{p_n + (2^{d-1}-1)q_n}{2^{d-1}}\right)$ are asymptotically orthogonal if $\beta^2 > \alpha$.

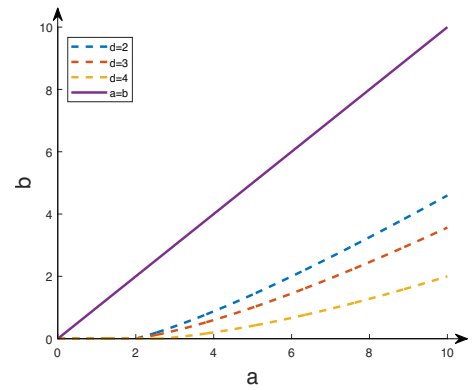

Figure 1: Phase transition of community detection in the sparse HSBM: the detection is possible below the dashed curves, while impossible above the dashed curves (and below the solid line). The larger d is, the larger the area of impossible detection is, which can be viewed from the threshold $\beta^2 = \alpha$, i.e., $b = a + \frac{2^{2d-3}-2^{d-2}}{d-1} - \sqrt{\frac{2^{2d-2}}{d-1}a + \left(\frac{2^{2d-3}-2^{d-2}}{d-1}\right)^2}$.

Theorem 2.1. *Let X_{ζ_n} be the number of ζ_n-loose cycle of a hypergraph H. Suppose $\zeta_n = \mathcal{O}\left(\log^{1/4}(n)\right)$.*

(1) If $H \sim \mathcal{H}_d(n, p_n, q_n)$, then $X_{\zeta_n} \xrightarrow{d}$ Pois $\left(\frac{\alpha^{\zeta_n}+\beta^{\zeta_n}}{2\zeta_n}\right)$;

(2) If $H \sim \mathcal{H}_d\left(n, \frac{p_n+(2^{d-1}-1)q_n}{2^{d-1}}\right)$, then $X_{\zeta_n} \xrightarrow{d}$ Pois $\left(\frac{\alpha^{\zeta_n}}{2\zeta_n}\right)$,

where $\xrightarrow{d}$ denotes convergent in distribution.

For any non-negative integer s, let $[X]_s$ be the product $X(X-1)\cdots(X-s+1)$. By Lemma 2.8 in [35] (Lemma A.1 in the Appendix), if we can prove the expectation $\mathbb{E}[X_{\zeta_n}]_r \to \mu^r$, then $X_{\zeta_n} \xrightarrow{d} \text{Pois}(\mu)$. Refer to the Appendix for the detailed proof.

With the preparation provided by establishing Theorem 2.1, now we can prove our first main result, Theorem 1.1. We use $\mathbb{P}_n(\cdot)$, $\mathbb{E}(\cdot)$, $\text{Var}(\cdot)$ to denote probability, mean and variation with respect to $\mathcal{H}_d(n, p_n, q_n)$ while notations with tilde above denote their counterparts with respect to $\mathcal{H}_d\left(n, \frac{p_n+(2^{d-1}-1)q_n}{2^{d-1}}\right)$.

Proof of Theorem 1.1. Since $\beta^2 > \alpha$, there exists a constant ρ satisfying $\beta > \rho > \sqrt{\alpha}$. Then $\mathbb{E}X_{\zeta_n} = \text{Var}X_{\zeta_n} = o(\rho^{2\zeta_n})$ and $\tilde{\mathbb{E}}X_{\zeta_n} = \tilde{\text{Var}}X_{\zeta_n} = o(\rho^{2\zeta_n})$ as $n \to \infty$.

Let $A_n := \{X_{\zeta_n} \le \tilde{\mathbb{E}}X_{\zeta_n} + \rho^{\zeta_n}\}$, by Chebyshev's inequality,

$$\tilde{\mathbb{P}}_n(A_n) = \tilde{\mathbb{P}}_n\left(\frac{X_{\zeta_n} - \tilde{\mathbb{E}}X_{\zeta_n}}{\sqrt{\tilde{\text{Var}}X_{\zeta_n}}} \le \frac{\rho^{\zeta_n}}{\sqrt{\tilde{\text{Var}}X_{\zeta_n}}}\right) \ge 1 - \frac{\tilde{\text{Var}}X_{\zeta_n}}{\rho^{2\zeta_n}} \to 1, \quad \text{as } n \to \infty. \tag{1}$$

Since $\mathbb{E}X_{\zeta_n} - \tilde{\mathbb{E}}X_{\zeta_n} \sim \frac{\beta^{\zeta_n}}{2\zeta_n} = \omega(\rho^{\zeta_n})$, then for larger enough ζ_n, $\mathbb{E}X_{\zeta_n} - \rho^{\zeta_n} \ge \tilde{\mathbb{E}}X_{\zeta_n} + \rho^{\zeta_n}$, which implies $\mathbb{P}_n(A_n) \le \mathbb{P}_n(X_{\zeta_n} \le \mathbb{E}X_{\zeta_n} - \rho^{\zeta_n})$. By Chebyshev's inequality,

$$\mathbb{P}_n(A_n) \le \mathbb{P}_n\left(\frac{X_{\zeta_n} - \mathbb{E}X_{\zeta_n}}{\sqrt{\text{Var}X_{\zeta_n}}} \le \frac{-\rho^{\zeta_n}}{\sqrt{\text{Var}X_{\zeta_n}}}\right) \le \frac{\text{Var}X_{\zeta_n}}{\rho^{2\zeta_n}} \to 0, \quad \text{as } n \to \infty. \tag{2}$$

By (1) and (2), we have shown that $\mathbb{P}_n$ and $\tilde{\mathbb{P}}_n$ are asymptotically orthogonal.

Now we show that Theorem 2.1 can provide consistent estimators for a and b. First of all, we have a consistent estimator $\hat{\alpha}_n$ for α by simply counting the number of hyperedges. If we can prove that β

can be consistently estimated by $\hat{\beta}_n = \left(2\zeta_n X_{\zeta_n} - \hat{\alpha}_n^{\zeta_n}\right)^{\frac{1}{\zeta_n}}$, then $\hat{a}_n = \frac{1}{d-1}\left(\hat{\alpha}_n + (2^{d-1}-1)\hat{\beta}_n\right)$ and $\hat{b}_n = \frac{1}{d-1}\left(\hat{\alpha}_n - \hat{\beta}_n\right)$ are consistent estimators for a and b, respectively.

Take $\rho \in (\sqrt{\alpha}, \beta)$, by Chebyshev's inequality,

$$\begin{aligned}\mathbb{P}_n(2\zeta_n X_{\zeta_n} - \alpha^{\zeta_n} \geq \beta^{\zeta_n} - \rho^{\zeta_n}) &= \mathbb{P}_n\left(\frac{X_{\zeta_n} - \frac{\alpha^{\zeta_n}+\beta^{\zeta_n}}{2\zeta_n}}{\sqrt{\mathrm{Var}X_{\zeta_n}}} \geq \frac{-\frac{\rho^{\zeta_n}}{2\zeta_n}}{\sqrt{\mathrm{Var}X_{\zeta_n}}}\right) \\ &\geq 1 - \left(\frac{2\zeta_n\sqrt{\mathrm{Var}X_{\zeta_n}}}{\rho^{\zeta_n}}\right)^2 \to 1, \quad \text{as } \zeta_n \to \infty.\end{aligned} \tag{3}$$

$$\begin{aligned}\mathbb{P}_n(2\zeta_n X_{\zeta_n} - \alpha^{\zeta_n} \leq \beta^{\zeta_n} + \rho^{\zeta_n}) &= \mathbb{P}_n\left(\frac{X_{\zeta_n} - \frac{\alpha^{\zeta_n}+\beta^{\zeta_n}}{2\zeta_n}}{\sqrt{\mathrm{Var}X_{\zeta_n}}} \leq \frac{\frac{\rho^{\zeta_n}}{2\zeta_n}}{\sqrt{\mathrm{Var}X_{\zeta_n}}}\right) \\ &\geq 1 - \left(\frac{2\zeta_n\sqrt{\mathrm{Var}X_{\zeta_n}}}{\rho^{\zeta_n}}\right)^2 \to 1, \quad \text{as } \zeta_n \to \infty.\end{aligned} \tag{4}$$

Thus, $2\zeta_n X_{\zeta_n} - \alpha^{\zeta_n} \in [\beta^{\zeta_n} - \rho^{\zeta_n}, \beta^{\zeta_n} + \rho^{\zeta_n}]$ a.a.s. Since $\rho^{\zeta_n} = o(\beta^{\zeta_n})$ as $\zeta_n \to \infty$, we have $2\zeta_n X_{\zeta_n} - \alpha^{\zeta_n} = (1+o(1))\beta^{\zeta_n}$ a.a.s. Since $\hat{\alpha}_n^{\zeta_n} \to \alpha^{\zeta_n}$ as $\zeta_n \to \infty$, we have $2\zeta_n X_{\zeta_n} - \hat{\alpha}_n^{\zeta_n} = (1+o(1))\beta^{\zeta_n}$ a.a.s. Therefore, $\hat{\beta}_n = (2\zeta_n X_{\zeta_n} - \hat{\alpha}_n^{\zeta_n})^{\frac{1}{\zeta_n}} \to \beta$ as $\zeta_n \to \infty$. □

3 Non-reconstruction of the community structures

To prove Theorem 1.2, we use a multi-type Galton-Watson hypertree (proposed in [26]) that is coupled with $\mathcal{H}_d(n, p_n, q_n)$. Intuitively speaking, a neighborhood of a vertex in $\mathcal{H}_d(n, p_n, q_n)$ looks like the hypertree with high probability. We take advantage of the threshold of broadcasting on the hypertree to show the threshold for community detection on the HSBM.

Let T be an infinite rooted hypertree with root ρ constructed as follows: Generate a root ρ with spin $\tau_\rho \in \{+1, -1\}$ and then generate $\mathrm{Pois}(\alpha)$ many hyperedges that only intersects at ρ. For $0 \leq s \leq d-1$, a hyperedge is called *type* s if there are s many children (i.e., the vertices in the hyperedge) with spin τ_ρ and $d-1-s$ many children with spin $-\tau_\rho$ in the hyperedge. Hyperedges of type $d-1$ are generated with probability $\frac{(d-1)a}{\alpha 2^{d-1}}$ and of type s with probability $\frac{(d-1)b\binom{d-1}{s}}{\alpha 2^{d-1}}$ for $0 \leq s \leq d-2$. It is easy to check that

$$\frac{(d-1)a}{\alpha 2^{d-1}} + \sum_{s=0}^{d-2} \frac{(d-1)b\binom{d-1}{s}}{\alpha 2^{d-1}} = 1.$$

We generate hyperedges of different types i.i.d. by Poisson distribution. For each hyperedge of type s, in the first generation, we uniformly and randomly pick s vertices with spin τ_ρ and the rest of the $d-1-s$ vertices with spin $-\tau_\rho$. The subsequent generations are generated by induction. The generated random hypertree, denoted by (T, ρ, τ), is called a *multi-type Galton-Watson hypertree*. Figure 2 shows an example of a multi-type Galton-Watson hypertree where $d = 3$.

The broadcasting problem on (T, ρ, τ) is whether the spin τ_ρ could be deduced from the spins at generation l of the hypertree, where l is sufficiently large. Such a problem on general infinite trees was initially proposed by Kesten and Stigum in 1966 [19], although they used somewhat different terminologies. This problem was solved by Blecher et al. [3] and Evans et al. [10].

Let T_l be the vertices of generations no more than l in T, ∂T_l the vertices of generation l in T, then we have (refer to the Appendix for the proof):

Theorem 3.1. *Suppose (T, ρ, τ) is a multi-type Galton-Watson hypertree where the offspring distribution has mean $\alpha > 1$, if $\beta^2 < \alpha$, then*

$$\lim_{l\to\infty} \mathbb{P}(\tau_\rho = +1 | \tau_{\partial T_l}) = \frac{1}{2} \quad \textit{a.s.} \tag{5}$$

where $\tau_{\partial T_l} = \{\tau_v | v \in \partial T_l\}$.

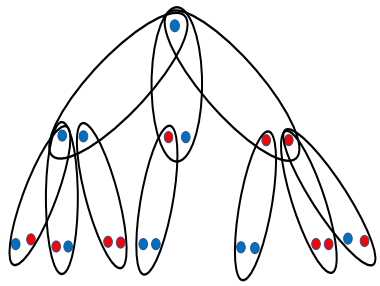

Figure 2: A multi-type Galton-Watson hypertree with $d = 3$, where vertices with spin $+1$ are in blue and with spin -1 are in red.

Let (H, ρ) be a rooted hypergraph, i.e., a hypergraph H with a distinguished vertex $\rho \in \mathcal{V}(H)$. Two rooted hypergraphs (H_1, ρ_1) and (H_2, ρ_2) are said to be *isomorphic* if and only if there is a bijection $\varphi : \mathcal{V}(H_1) \to \mathcal{V}(H_2)$ such that $\varphi(\rho_1) = \rho_2$ and $e \in \mathcal{E}(H_1)$ if and only if $\varphi(e) := \{\varphi(i) | i \in e\} \in \mathcal{E}(H_2)$.

Let (H, ρ, σ) be a rooted hypergraph with a spin vector σ, i.e., each vertex $i \in \mathcal{V}(H)$ is given a spin $\sigma(i) \in \{+1, -1\}$. Two rooted hypergarphs (H_1, ρ_1, σ_1) and (H_2, ρ_2, σ_2) are said to be *spin-preserving isomorphic*, denoted by $(H_1, \rho_1, \sigma_1) \equiv (H_2, \rho_2, \sigma_2)$, if and only if there is an isomorphism $\varphi : (H_1, \rho_1) \to (H_2, \rho_2)$ such that $\sigma_1(j) = \sigma_2(\varphi(j))$ for any $j \in \mathcal{V}(H_1)$.

Let $(H, \rho, \sigma)_l$ be the rooted hypergraph (H, ρ, σ) truncated at distance l from ρ and $(T, \rho, \tau)_l$ the rooted hypertree (T, ρ, τ) truncated at generation (distance) l from ρ. If $l = c\log(n)$ with $c\log(\alpha) < 1/4$ and c is a constant, then for sufficiently large n (see Theorem 5.2 in [26]),

(1) if $\sigma_\rho = +1$, there exists a coupling between (H, ρ, σ) and (T, ρ, τ) such that

$$\mathbb{P}(\{(H, \rho, \sigma)_l \equiv (T, \rho, \tau)_l\}) \geq 1 - n^{-1/5}.$$

(2) if $\sigma_\rho = -1$, there exists a coupling between (H, ρ, σ) and $(T, \rho, -\tau)$ such that

$$\mathbb{P}(\{(H, \rho, \sigma)_l \equiv (T, \rho, -\tau)_l\}) \geq 1 - n^{-1/5}.$$

Namely, we have

$$\lim_{n\to\infty} \mathbb{P}(\{(H, \rho, \sigma)_l \equiv (T, \rho, \sigma_\rho \cdot \tau)_l\}) = 1,$$

which means a neighborhood in H looks like a broadcasting process on a multi-type Galton-Watson hypertree T.

The statement of Theorem 1.2 is equivalent to $\mathbb{E}(\sigma_\rho | H, \sigma_v) \to 0$, and also equivalent to $\mathrm{Var}(\sigma_\rho | H, \sigma_v) \to 1$. In order to prove Theorem 1.2, we can consider the variance $\mathrm{Var}(\sigma_\rho | H, \sigma_{\partial H_l})$, which asymptotically almost surely is equal to $\mathrm{Var}(\tau_\rho | T, \tau_{\partial T_l})$. By the property of monotonicity of variance, we can add conditions on the spins σ_v and $\sigma_{\partial H_l}$, which can make the reconstruction easier. Then we wish to remove the condition σ_v and such action will not lose much information, i.e., $\sigma_{\partial H_l}$ gives much more information in the process. To fill the gap that once we have conditioned on $\sigma_{\partial H_l}$, we can remove the condition on σ_v, the following lemma is required.

Lemma 3.2. *Let V_1, V_2, V_3 be a random partition of the vertex set $\mathcal{V}(H)$ such that V_2 separates V_1 and V_3 (i.e., any path connecting vertices in V_1 and V_3 must pass through vertices in V_2). If $|V_1 \cup V_2| = o(\sqrt{n})$ for a.a.e H, then*

$$\mathbb{P}(\sigma_{V_1} | H, \sigma_{V_2}) = (1 + o(1))\mathbb{P}(\sigma_{V_1} | H, \sigma_{V_2 \cup V_3}) \tag{6}$$

for a.a.e. H and σ.

Now we can establish the second main result, Theorem 1.2.

Proof of Theorem 1.2. By the property of the monotonicity of conditional variance, for $v \notin H_l$, we have

$$\mathrm{Var}(\sigma_\rho | H, \sigma_v, \sigma_{\partial H_l}) \leq \mathrm{Var}(\sigma_\rho | H, \sigma_v). \tag{7}$$

By Lemma 3.2, σ_ρ and σ_v are a.a.s. conditionally independent given $\sigma_{\partial H_l}$ and H, which means $\mathrm{Var}(\sigma_\rho|H,\sigma_v,\sigma_{\partial H_l}) \to \mathrm{Var}(\sigma_\rho|H,\sigma_{\partial H_l})$. By Theorem 5.2 in [26], $\mathrm{Var}(\sigma_\rho|H,\sigma_{\partial H_l}) \to \mathrm{Var}(\tau_\rho|T,\tau_{\partial T_l})$. By Theorem 3.1, $\mathrm{Var}(\tau_\rho|T,\tau_{\partial T_l}) \to 1$ a.a.s. Therefore, $\mathrm{Var}(\sigma_\rho|H,\sigma_v) \to 1$ a.a.s. □

In the rest of this section, we will show our last main result, Theorem 1.3. The second half of Theorem 1.3 can be derived from its first half as follows: If $\beta^2 < \alpha$, suppose $\hat{a}$ and $\hat{b}$ are estimators for a and b which converge in probability, we claim their limits depend on $a + (2^{d-1}-1)b$ only. To verify this, let A, B be another choice of the model parameters with $(d-1)(A-B)^2 < 2^{d-1}(A+(2^{d-1}-1)B)$ and $A + (2^{d-1}-1)B = a + (2^{d-1}-1)b$. Let $\mathbb{P}'_n$ be the probability with respect to the model $\mathcal{H}_d\left(n, \frac{A}{\binom{n}{d-1}}, \frac{B}{\binom{n}{d-1}}\right)$, and suppose

$$\hat{a} \to a^* \quad \mathbb{P}_n - a.a.s, \quad \hat{a} \to A^* \quad \mathbb{P}'_n - a.a.s.$$

For any $\epsilon > 0$, let A_n be the event that $|\hat{a} - a^*| > \epsilon$, then $\mathbb{P}_n(A_n) \to 0$. By the first half of Theorem 1.3, $\tilde{\mathbb{P}}_n(A_n) \to 0$. Since $A + (2^{d-1}-1)B = a + (2^{d-1}-1)b$, using the first half of Theorem 1.3 again, we have $\mathbb{P}'_n(A_n) \to 0$, which implies $A^* = a^*$. Therefore, $\hat{a}$ converges to the same limit under $\mathbb{P}_n$ and $\mathbb{P}'_n$, i.e., $\hat{a}$ is not a consistent estimator for a (same argument for b).

The first half of Theorem 1.3 is a bit difficult to prove. Here, we give some outline of how to establish it, and the lengthy proof is put in the Appendix. For our purpose, we will use the small subgraph conditioning method, which was initially proposed by Robinson et al. [29, 30] and then applied by Janson [18] to prove contiguity. The follow theorem from [35] is a crucial tool to show the contiguity between $\mathcal{H}_d(n, p_n, q_n)$ and $\mathcal{H}_d\left(n, \frac{p_n + (2^{d-1}-1)q_n}{2^{d-1}}\right)$ when $\beta^2 < \alpha$.

Theorem 3.3 (Theorem 4.1 in [35]). *Let $\mathbb{P}_n$ and $\tilde{\mathbb{P}}_n$ be two fixed sequences of probability measures on a common measurable space, $Y_n = \frac{\mathbb{P}_n}{\tilde{\mathbb{P}}_n}$ the density of $\mathbb{P}_n$ with respect to $\tilde{\mathbb{P}}_n$. For $i \geq 1$, let $\lambda_i > 0$, $\delta_i \geq -1$, for each n, suppose random variables X_{in} satisfy*

(C1) $X_{in} \xrightarrow{d} W_i$ as $n \to \infty$ jointly for all i under $\tilde{\mathbb{P}}_n$, where $W_i \sim Pois(\lambda_i)$ are independent Poisson variables;

(C2) For every non-negative integers $s_1, \cdots, s_k$,

$$\tilde{\mathbb{E}}(Y_n[X_{1n}]_{s_1} \cdots [X_{kn}]_{s_k})/\tilde{\mathbb{E}}Y_n \to \prod_{i=1}^{k} (\lambda_i(1+\delta_i))^{s_i}.$$

(C3) $\sum_{i\geq 1} \lambda_i \delta_i^2 < \infty$;

(C4) $\tilde{\mathbb{E}}Y_n^2/(\tilde{\mathbb{E}}Y_n)^2 \to \exp(\sum_{i\geq 1} \lambda_i \delta_i^2)$.

Then, $\tilde{\mathbb{P}}_n$ and $\mathbb{P}_n$ are contiguous.

In order to apply the theorem above, we will take Y_n to be the likelihood ratio of the adjacency tensor A of H, and X_{in} to be the number of i-loose cycles in H. To shorten notation, we adopt $i_1 * i_d$ to represent the ordering $i_1 i_2 \cdots i_d$. $A_{i_1 * i_d} := A_{i_1 i_2 \cdots i_d}$. $I[\sigma_{i_1} * \sigma_{i_d}] := I[\sigma_{i_1} = \cdots = \sigma_{i_d}]$, which is 1 if $\sigma_{i_1} = \cdots = \sigma_{i_d}$, and 0 otherwise. $\varrho(s,t) = \{(i_1, \cdots, i_s) | 1 \leq i_1 \leq \cdots \leq i_s \leq t\}$. $\hat{d} := \frac{a + (2^{d-1}-1)b}{2^{d-1}}$. $p_0 := \frac{\hat{d}}{\binom{n}{d-1}}$, $q_0 := 1 - p_0$. Then, for $H \sim \mathcal{H}_d(n, p_n, q_n)$, its hyperedge probabilities $p_{i_1 * i_d}(\sigma)$ and $q_{i_1 * i_d}(\sigma)$ can be given by

$$p_{i_1 * i_d}(\sigma) = \mathbb{P}(A_{i_1 * i_d} = 1|\sigma) = p_n^{I[\sigma_{i_1} * \sigma_{i_d}]} q_n^{1 - I[\sigma_{i_1} * \sigma_{i_d}]},$$
$$q_{i_1 * i_d}(\sigma) = \mathbb{P}(A_{i_1 * i_d} = 0|\sigma) = 1 - p_{i_1 * i_d}(\sigma).$$

For $H \sim \mathcal{H}_d(n, p_0)$, its hyperedge probabilities can be given by $\mathbb{P}(\{i_1 * i_d\}) = p_0^{A_{i_1 * i_d}} q_0^{1 - A_{i_1 * i_d}}$.

Therefore, the likelihood ratio $Y_n = \frac{\mathbb{P}_n(A)}{\tilde{\mathbb{P}}_n(A)}$ of the adjacency tensor A can be written as

$$Y_n = 2^{-n} \sum_{\sigma \in \{\pm 1\}^n} \prod_{i \in \varrho(d,n)} \left(\frac{p_{i_1 * i_d}(\sigma)}{p_0}\right)^{A_{i_1 * i_d}} \left(\frac{q_{i_1 * i_d}(\sigma)}{q_0}\right)^{1 - A_{i_1 * i_d}}. \tag{8}$$

Obviously, we have $\tilde{\mathbb{E}}Y_n = 1$. The condition (C1) is easy to verify as we can see from Theorem 2.1, which also implies that λ_i can be chosen as $\frac{\alpha^i}{2i}$. Then, we need to estimate the asymptotics of $\mathbb{E}(Y_n[X_{1n}]_{s_1} \cdots [X_{kn}]_{s_k})$ and $\tilde{\mathbb{E}}Y_n^2$, where the calculation of first moment is somewhat similar to calculations in the proof of Theorem 2.1. The following lemma gives an estimation for the second moment of Y_n.

Lemma 3.4.

$$\tilde{\mathbb{E}}Y_n^2 = (1+o(1))\exp\left\{\frac{1}{\hat{d}\binom{n}{d-1}}\left(\binom{n}{d}(b-\hat{d})^2 + (a-b)^2\sum_{i\in\varrho(d,n)} I[\sigma_{i_1} * \sigma_{i_d}]I[\tau_{i_1} * \tau_{i_d}]\right.\right.$$
$$\left.\left. +(a-b)(b-\hat{d})\sum_{i\in\varrho(d,n)}(I[\sigma_{i_1} * \sigma_{i_d}] + I[\tau_{i_1} * \tau_{i_d}])\right)\right\}. \tag{9}$$

After verifying all the conditions of Theorem 3.3 (see the Appendix) the first half of Theorem 1.3 follows directly.

4 Conclusion

Community detection is of great interest in many fields such as network analysis, statistics and computer vision. Hypergraphs have been widely used to model higher-order relationships among data. The SBM (graph-based) model is popular in community detection but only directly models pair-wise interactions. Its hypergraph version, HSBM, likewise serves as a benchmark for clustering algorithms on higher order interaction (hypergraph) data.

This paper provides theoretical analyses on the community detection thresholds in the sparse HSBM. Specifically, we address a previous conjecture that there exists a sharp threshold on the model parameters for community detection to be possible. We proved that above the threshold, it is possible to asymptotically almost surely construct a bi-partition of the hypergraph positively correlated with the true bi-partition. Moreover, below the threshold, we show that community detection is information-theoretically impossible. Importantly, we also provide consistent estimators for the model parameters, when the detection is possible.

Our estimators require the counting of the number of loose cycles in the hypergraph. To our best knowledge, there is no efficient algorithm to count the number of loose cycles in hypergraphs in general. Therefore, from the algorithmic side, it might be more practical to count the number of non-backtracking/self-avoiding walks of the same length [26, 33], which can be computed by matrix multiplication efficiently. Nevertheless, from the theoretical side, confirming the positive side of the threshold conjecture by analysing the number of loose cycles has not been done before. Moreover, providing more practical estimators (or examining whether one can estimate the number of loose cycles sufficiently accurate to still use the estimators of model parameters we provide), is a worthy topic of future research.

Note that the non-uniform hypergraphs can be viewed as a 'superposition' of a collection of uniform hypergraphs. Thus it may be possible to generalise our proofs to cover non-uniform hypergraphs in future work. Considering other variants of the hypergraph community detection problem, for instances, multi-community structures, extremely sparse scenario ($p_n, q_n \ll 1/n^{d-1}$), moderate sparse scenario ($p_n, q_n = \mathcal{O}(\log(n)/n^{d-1})$), dense scenario ($p_n, q_n \gg 1/n^{d-1}$), are future research directions that would complete the work presented here.

Acknowledgments and Disclosure of Funding

The authors would like to express their gratitude to three anonymous referees for their helpful suggestions and comments. This work was partly supported by the Australian Research Council grants DP200103448 (D. Suter & E. Zhang).

References

[1] Emmanuel Abbe, Enric Boix-Adsera, Peter Ralli, and Colin Sandon. Graph powering and spectral robustness. *SIAM Journal on Mathematics of Data Science*, 2(1):132–157, 2020.

[2] Maria Chiara Angelini, Francesco Caltagirone, Florent Krzakala, and Lenka Zdeborová. Spectral detection on sparse hypergraphs. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 66–73. IEEE, 2015.

[3] Pavel M Bleher, Jean Ruiz, and Valentin A Zagrebnov. On the purity of the limiting Gibbs state for the Ising model on the Bethe lattice. *Journal of Statistical Physics*, 79(1):473–482, 1995.

[4] Béla Bollobás. Random graphs. In *Modern graph theory*, pages 215–252. Springer, 1998.

[5] Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

[6] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1347–1357. IEEE, 2015.

[7] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

[8] Paul Erdös, Alfréd Rényi, et al. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[9] Ernesto Estrada and Juan A Rodriguez-Velazquez. Complex networks as hypergraphs. *arXiv preprint physics/0505137*, 2005.

[10] William Evans, Claire Kenyon, Yuval Peres, and Leonard J Schulman. Broadcasting on trees and the Ising model. *Annals of Applied Probability*, pages 410–433, 2000.

[11] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[12] Alan Frieze and Michał Karoński. *Introduction to random graphs*. Cambridge University Press, 2016.

[13] Debarghya Ghoshdastidar and Ambedkar Dukkipati. Consistency of spectral partitioning of uniform hypergraphs under planted partition model. *Advances in Neural Information Processing Systems*, 27, 2014.

[14] Debarghya Ghoshdastidar and Ambedkar Dukkipati. Consistency of spectral hypergraph partitioning under planted partition model. *The Annals of Statistics*, 45(1):289–315, 2017.

[15] Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, Edoardo M Airoldi, et al. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

[16] Christopher J Hillar and Lek-Heng Lim. Most tensor problems are NP-hard. *Journal of the ACM (JACM)*, 60(6):1–39, 2013.

[17] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[18] Svante Janson. Random regular graphs: asymptotic distributions and contiguity. *Combinatorics, Probability and Computing*, 4(4):369–405, 1995.

[19] Harry Kesten and Bernt P Stigum. A limit theorem for multidimensional Galton-Watson processes. *The Annals of Mathematical Statistics*, 37(5):1211–1223, 1966.

[20] Chiheon Kim, Afonso S Bandeira, and Michel X Goemans. Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. *arXiv preprint arXiv:1807.02884*, 2018.

[21] Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th international conference on World Wide Web*, pages 695–704, 2008.

[22] Laurent Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703, 2014.

[23] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3):431–461, 2015.

[24] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018.

[25] Xavier Ouvrard, Jean-Marie Le Goff, and Stéphane Marchand-Maillet. Networks of collaborations: Hypergraph modeling and visualisation. *arXiv preprint arXiv:1707.00115*, 2017.

[26] Soumik Pal and Yizhe Zhu. Community detection in the sparse hypergraph stochastic block model. *Random Structures & Algorithms*, 59(3):407–463, 2021.

[27] Pulak Purkait, Tat-Jun Chin, Alireza Sadri, and David Suter. Clustering with hypergraphs: the case for large hyperedges. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1697–1711, 2016.

[28] José J Ramasco, Sergey N Dorogovtsev, and Romualdo Pastor-Satorras. Self-organization of collaboration networks. *Physical review E*, 70(3):036106, 2004.

[29] Robert W. Robinson and Nicholas C. Wormald. Almost all cubic graphs are Hamiltonian. *Random Structures & Algorithms*, 3(2):117–125, 1992.

[30] Robert W. Robinson and Nicholas C. Wormald. Almost all regular graphs are Hamiltonian. *Random Structures & Algorithms*, 5(2):363–374, 1994.

[31] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[32] Ludovic Stephan and Laurent Massoulié. Robustness of spectral methods for community detection. In *Conference on Learning Theory*, pages 2831–2860. PMLR, 2019.

[33] Ludovic Stephan and Yizhe Zhu. Sparse random hypergraphs: Non-backtracking spectra and community detection. *arXiv preprint arXiv:2203.07346*, 2022.

[34] Ze Tian, TaeHyun Hwang, and Rui Kuang. A hypergraph-based learning algorithm for classifying gene expression and arrayCGH data with prior knowledge. *Bioinformatics*, 25(21):2831–2838, 2009.

[35] Nicholas C Wormald et al. Models of random regular graphs. *London Mathematical Society Lecture Note Series*, pages 239–298, 1999.

[36] Mingao Yuan, Ruiqi Liu, Yang Feng, and Zuofeng Shang. Testing community structure for hypergraphs. *The Annals of Statistics*, 50(1):147–169, 2022.

[37] Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108(18):7321–7326, 2011.

[38] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. *Advances in neural information processing systems*, 19, 2006.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section 2.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See the abstract and introduction.
   (b) Did you describe the limitations of your work? [Yes] See the last two paragraphs of the conclusion section.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We have read the ethics review guidelines.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See the main results subsection.
   (b) Did you include complete proofs of all theoretical results? [Yes] Some proofs are added in the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]