Statistica Sinica **34** (2024), 481-503
doi:https://doi.org/10.5705/ss.202021.0392

COMMUNITY DETECTION IN CENSORED HYPERGRAPH

Mingao Yuan*[1], Bin Zhao[1] and Xiaofeng Zhao[2]

[1]*North Dakota State University and*
[2]*North China University of Water Resources and Electric Power*

Abstract: Community detection refers to the problem of clustering the nodes of a network (either a graph or a hypergrah) into groups. Various algorithms are available for community detection, all of which apply to uncensored networks. In practice, a network may have censored (or missing) values, which have been shown to have a non-negligible effect on the structural properties of a network. In this study, we examine community detection in a censored m-uniform hypergraph from an information-theoretic point of view. As such, we derive the information-theoretic threshold for the exact recovery of the community structure. Furthermore, we propose a polynomial-time algorithm to exactly recover the community structure up to the threshold. The proposed algorithm consists of a spectral algorithm plus a refinement step. It is also interesting to determine whether a single spectral algorithm without refinement achieves the threshold. To this end, we explore the semi-definite relaxation algorithm and analyze its performance.

Key words and phrases: Censored hypergraph, community detection, exact recovery, information-theoretic threshold.

1. Introduction

Many complex data sets can be modeled as a network of items (nodes). One of the most popular topics in network data mining is to understand which items are similar to each other. Community detection refers to the problem of clustering the nodes of a network into groups based on similarity. Community detection is widely used in analyses of, for example, social networks (Goldenberg et al. (2010); Zhao, Levina and Zhu (2011)), protein-to-protein interaction networks (Chen and Yuan (2006)), and image segmentation (Shi and Malik (1997)). Existing studies on community detection can be classified into two categories: (1) those that derive an information-theoretic threshold used to recover the community structure (Abbe, Bandeira and Hall (2016); Mossel, Neeman and Sly (2015, 2017); Chien, Lin and Wang (2018); Dhara et al. (2021); Hajek, Wu and Xu (2018); Yuan and Shang (2021a)); and (2) those that devise efficient algorithms to recover the community structure (Ghoshdastidar and Dukkipati (2014, 2017); Luo and Zhang (2020); Liu, Jan and Yan (2015); Ke, Shi and Xia (2020);

*Corresponding author.

Yuan and Qu (2021); Ahn, Lee and Suh (2018, 2019); Hajek, Wu and Xu (2016); Gao et al. (2018); Weng and Feng (2021); Zhen and Wang (2021); Lei and Rinaldo (2015); Jin (2015)); see Abbe (2018); Bi et al. (2021) for additional references. The aforementioned methods all apply to uncensored networks.

In practice, network data may have censored or missing values. For example, in a social network, the non-response of actors can cause missingness of ties (Huisman (2009); Gile and Handcock (2016)); and in an MRI network, missingness may be due to the high cost of PET scanning (Liu et al. (2018)). Missing values have non-negligible effects on the structural properties of a network (Huisman (2009); Smith, Moody and Morgan (2018)). Most existing algorithms for community detection apply to uncensored networks. Thus, a natural question is how to recover communities in a censored network. To the best of our knowledge, Abbe et al. (2014) was the first to examine community detection in a censored graph, obtaining an information-theoretic threshold for the exact recovery of communities. Recently, Dhara et al. (2021) showed that a spectral algorithm without a refinement step can exactly recover the community structure in censored graph, up to the information-theoretic threshold.

Many complex networks in the real world can be formulated as a hypergraph, where hyperedges are used to model higher-order interactions between individuals (Estrada and Rodriguez-velasquez (2005); Ouvrard, Goff and Marchand-Maillet. (2017); Ramasco, Dorogovtsev and Pastor-Satorras (2004); Newman (2001); Ghoshal et al. (2009); Ghoshdastidar and Dukkipati (2014)). For example, in a folksonomy network, a hyperedge may represent a triple (user, resource, annotation) structure (Ghoshal et al. (2009)), and in coauthorship networks, the coauthors of a paper form a hyperedge (Estrada and Rodriguez-velasquez (2005); Ouvrard, Goff and Marchand-Maillet. (2017); Ramasco, Dorogovtsev and Pastor-Satorras (2004); Newman (2001)). Hypergraph learning with missing values has recently attracted much attention (Hu and Shi (2015); Liu et al. (2017, 2018)). In this study, we are interested in detecting communities in censored hypergraphs. It is not immediately clear how the sharp threshold obtained by (Dhara et al. (2021)) changes in the case of a censored hypergraph, which motivated this research. Our contributions to the literature are summarized as follows. We derive an information-theoretic threshold for the exact recovery of a community structure in a censored hypergraph. Interestingly, the threshold is larger, in general, than that in the graph case. In this sense, community detection in a censored hypergraph is more difficult than it is in the case of a censored graph. In addition, we propose a polynomial-time algorithm that can exactly recover the community structure up to the information-theoretic threshold. The proposed algorithm consists of a spectral algorithm plus a refinement step. It is also interesting to study whether a single spectral algorithm without refinement can achieve the threshold as the censored graph case Dhara et al. (2021). To this end, we study the semi-definite relaxation algorithm, and provide a sufficient condition for the

algorithm to achieve exact recovery.

1.1. The censored hypergraph block model

For a positive integer n, let $\mathcal{V} = \{1, 2, \ldots, n\}$ denote a set of nodes and $\mathcal{E}$ be a set of subsets of $\mathcal{V}$. The pair $\mathcal{H}_m = (\mathcal{V}, \mathcal{E})$ is called an *undirected* m-uniform hypergraph if $|e| = m$, for every $e \in \mathcal{E}$. That is, each element $e \in \mathcal{E}$ (called a hyperedge) contains exactly m distinct nodes. The hypergraph $\mathcal{H}_m$ can be represented as an m-dimensional symmetric array $A = (A_{i_1,\ldots,i_m}) \in \{0,1\}^{\otimes n^m}$, where $A_{i_1 i_2 \ldots i_m} = 1$ if $\{i_1, i_2, \ldots, i_m\}$ is a hyperedge, and $A_{i_1 i_2 \ldots i_m} = 0$ otherwise. In addition, $A_{i_1 i_2 \ldots i_m} = A_{j_1 j_2 \ldots j_m}$ if $\{i_1, i_2, \ldots, i_m\} = \{j_1, j_2, \ldots, j_m\}$. In this study, a self-loop is not allowed, that is, $A_{i_1 i_2 \ldots i_m} = 0$ if $|\{i_1, i_2, \ldots, i_m\}| < m$. When $m = 2$, $\mathcal{H}_2$ is the usual graph that has been widely used in community detection problems (Abbe (2018)). A hypergraph is said to be random if elements of the adjacency tensor are random. Throughout this paper, we focus on the hypergraph generated from the censored m-uniform hypergraph stochastic block model (CHSBM) $\mathcal{H}_m(n, p, q, \alpha)$, defined below.

Definition 1 (Censored m-uniform Hypergraph Stochastic Block Model (CHSBM)). Each node $i \in \mathcal{V}$ is randomly and independently assigned a label σ_i, with

$$\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}.$$

Let $\sigma = (\sigma_1, \ldots, \sigma_n)^T$ be a column vector of labels, $\mathrm{I}_+(\sigma) = \{i | \sigma_i = +1\}$, and $\mathrm{I}_-(\sigma) = \{i | \sigma_i = -1\}$. The nodes in $\mathrm{I}_+(\sigma)$ and $\mathrm{I}_-(\sigma)$ constitute two communities. The distinct nodes $i_1, i_2, \ldots, i_m$ form a hyperedge with probability p if $\{i_1, i_2, \ldots, i_m\}$ is a subset of $\mathrm{I}_+(\sigma)$ or $\mathrm{I}_-(\sigma)$, and q otherwise. The status of each hyperedge is revealed independently with probability α. A hyperedge of the resulting hypergraph takes a value in $\{1, 0, *\}$, where $*$ means the hyperedge is censored or missing (the hyperedge status is not revealed). This model is denoted as $\mathcal{H}_m(n, p, q, \alpha)$.

The status of each hyperedge in $\mathcal{H}_m(n, p, q, \alpha)$ with $\alpha < 1$ can take one of three values: 1 (present), 0 (absent), or $*$ (censored or missing). When $\alpha = 1$, the hypergraph is uncensored and $\mathcal{H}_m(n, p, q, 1)$ is the usual hypergraph stochastic block model (Ghoshdastidar and Dukkipati (2014, 2017); Chien, Lin and Wang (2018); Kim, Bandeira and Goemans (2018); Ke, Shi and Xia (2020); Yuan and Shang (2021a)). The censored stochastic block model $CSBM(p, q, \alpha)$ studied in (Dhara et al. (2021)) corresponds to $\mathcal{H}_2(n, p, q, \alpha)$. Throughout this paper, we assume $p, q \in (0, 1)$ are fixed constants, $p > q$, and $\alpha = t \log n / n^{m-1}$, for some constant $t > 0$. We consider the $\log n / n^{m-1}$ order of α because: this is the smallest order for which exact recovery is possible; see Theorem 1 and Theorem 2.

Table 1. Regions for exact recovery.

Region	Exact Recovery
(a) $t < I_m(p,q)$	Exact recovery is impossible
(b) $t > I_m(p,q)$	Exact recovery is possible

1.2. Summary of our main result

Given a hypergraph A generated from $\mathcal{H}_m(n,p,q,\alpha)$, community detection refers to the problem of recovering the unknown true label vector σ, or equivalently, identifying the sets $I_+(\sigma)$ and $I_-(\sigma)$. We say an estimator $\hat{\sigma}$ is an exact recovery of σ, $\hat{\sigma}$ exactly recovers σ, or $\hat{\sigma}$ achieves an exact recovery if

$$\mathbb{P}(\exists s \in \{\pm 1\} : \hat{\sigma} = s\sigma) = 1 - o(1).$$

That is, the estimator $\hat{\sigma}$ is equal to σ or $-\sigma$ with probability $1 - o(1)$. If there exists an estimator $\hat{\sigma}$ that exactly recovers σ, we say an exact recovery is possible. Otherwise, we say that an exact recovery is not possible.

For $m = 2$, Dhara et al. (2021) establishes the sharp information-theoretic threshold for exact recovery. The authors show that a spectral algorithm can exactly recover the true label without requiring a refinement step. It is not immediately clear how $m \geq 3$ changes the threshold for an exact recovery. More importantly, the spectral method in (Dhara et al. (2021)) cannot be extended straightforwardly to $m \geq 3$, because the spectral analysis of a tensor is still not well developed.

Here, we focus on $m \geq 3$ and derive the sharp information-theoretic threshold for an exact recovery. Define $I_m(p,q)$ as

$$I_m(p,q) = \frac{2^{m-1}(m-1)!}{(\sqrt{p}-\sqrt{q})^2 + (\sqrt{1-p}-\sqrt{1-q})^2}.$$

Theorem 1 shows that the maximum likelihood estimator (MLE) does not coincide with the true label with probability $1 - o(1)$ if $t < I_m(p,q)$. Theorem 2 states that the MLE succeeds with probability $1 - o(1)$ if $t > I_m(p,q)$. We also propose a spectral algorithm plus a refinement step that can achieve an exact recovery up to the information-theoretic threshold; see Theorem 3. Finally, we prove in Theorem 4 that the semi-definite relaxation algorithm can exactly recover the true label under mild conditions. Table 1 summarizes our main results. For $m = 2, 3$ and $q = 0.2$, Figure 1 displays the region in which an exact recovery is impossible; and the region in which an exact recovery is possible. Interestingly, with fixed q, the impossible region of $m = 3$ contains that of $m = 2$ as a proper subset. In this sense, an exact recovery becomes more difficult as m increases. For fixed q and m, $I_m(p,q)$ decreases as p goes to one; hence, an exact recovery becomes easier.

Throughout this paper, we adopt the Bachmann–Landau notation $o(1)$ and $O(1)$. For two positive sequences a_n and b_n, we write $a_n \sim b_n$ if $\lim_{n\to\infty} a_n/b_n = 1$. Denote $a_n \asymp b_n$ if $0 < c_1 \le a_n/b_n \le c_2 < \infty$, for constants c_1 and c_2. Denote $a_n \gg b_n$ or $b_n \ll a_n$ if $\lim_{n\to\infty} a_n/b_n = \infty$. For a square matrix M, $||M||$ denotes the operator norm of M, and $M \succeq 0$ means M is symmetric and positive semi-definite. Define $\langle M, N\rangle = \sum_{i,j} M_{ij}N_{ij}$.

2. Main Results

In this section, we present an information-theoretic threshold for the exact recovery of a CHSBM. First, we use the maximum likelihood method to show that an exact recovery is impossible if $t < I_m(p, q)$. Then, we prove that the MLE can exactly recover the true label if $t > I_m(p, q)$. Combining these two results yields the sharp information-theoretic threshold for exact recovery. This threshold provides a benchmark for developing practical recovery algorithms. Because the time complexity of an MLE is not polynomial in n, we propose a polynomial-time algorithm that achieves an exact recovery if $t > I_m(p, q)$.

2.1. Sharp threshold for exact recovery

In this subsection, we derive a sharp phase-transition threshold for exact recovery. The first result specifies a sufficient condition for the impossibility of exact recovery.

Theorem 1. *For each fixed integer $m \ge 2$, if $t < I_m(p, q)$, then $\mathbb{P}(\hat{\sigma} = \sigma) = o(1)$ for any estimator $\hat{\sigma}$. Here, $I_m(p, q)$ is defined as*

$$I_m(p, q) = \frac{2^{m-1}(m-1)!}{(\sqrt{p} - \sqrt{q})^2 + (\sqrt{1-p} - \sqrt{1-q})^2}. \tag{2.1}$$

Theorem 1 states that no estimator can exactly recover the true label if $t < I_m(p, q)$. For $m = 2$, $I_2(p, q)$ is just $t_c(p, q)$ in (Dhara et al. (2021)). Our result can be considered as a nontrivial extension of Theorem 2.1 in (Dhara et al. (2021)). Interestingly, with fixed p and q, the region $t < I_2(p, q)$ is smaller than $t < I_m(p, q)$ for $m \ge 3$. A similar phenomenon exists in the exact recovery of a community in an uncensored hypergraph stochastc block model (Kim, Bandeira and Goemans (2018)), although it differs significantly from that in hypothesis testing for communities. For example, Yuan and Shang (2021b) derived the sharp boundary for testing the presence of a dense subhypergraph. When the number of nodes in the dense subhypergraph is not too small, the region in which any test is asymptotically powerless for $m = 2$ is larger than $m \ge 3$.

The next result shows that the threshold $I_m(p, q)$ is actually sharp for exact recovery.

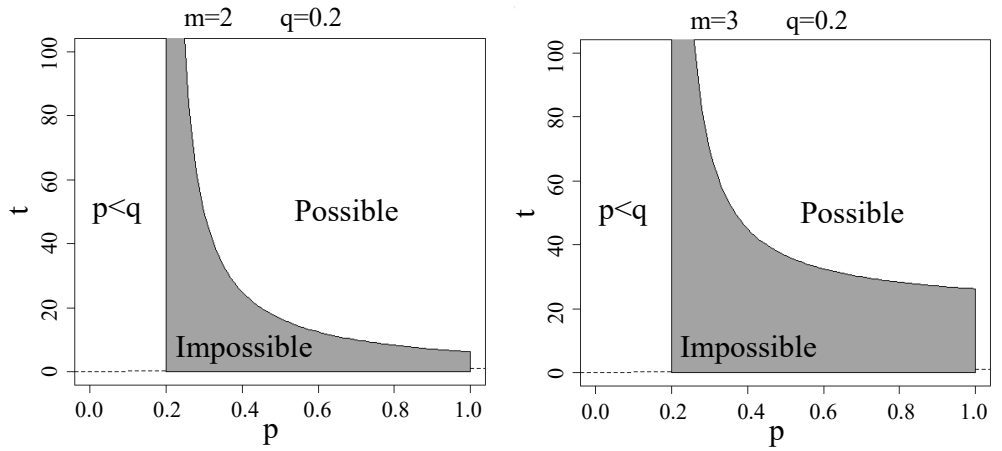

Figure 1. Regions for exact recovery with $m = 2, 3$ and $q = 0.2$. Impossible region: exact recovery is impossible. Possible region: exact recovery is possible.

Theorem 2. *For each fixed integer $m \geq 2$, if $t > I_m(p, q)$, with $I_m(p, q)$ defined in (2.1), then the MLE exactly recovers the true label with probability $1 - o(1)$.*

By Theorem 2, if $t > I_m(p, q)$, the true label can be exactly recovered by the MLE. Combining Theorem 1 and Theorem 2, we get the sharp boundary $t = I_m(p, q)$ for exact recovery, which is a surface in $\mathbb{R}^3$. For illustration, we visualize the regions $t > I_m(p, q)$ and $t < I_m(p, q)$ with $q = 0.2$ and $m = 2, 3$ in Figure 1. The red region represents $t < I_m(p, 0.2)$, where exact recovery is impossible. The green region corresponds to $t > I_m(p, 0.2)$, where exact recovery is possible. Clearly, the green region for $m = 3$ is smaller than that for $m = 2$. In this sense, an exact recovery becomes more difficult as m increases.

2.2. Efficient algorithm for exact recovery

Because the time complexity of an MLE is not polynomial in n, we propose an efficient algorithm for reconstructing two communities up to the information-theoretic threshold. The algorithm starts with a random splitting of the hypergraph A into two parts. Then, a spectral algorithm is applied to the first part, followed by a refinement based on the second part. We describe the algorithm in the following three steps.

In the first step, we randomly split the hypergraph A into two parts. Denote $M_m = \{(i_1, i_2, \ldots, i_m) \mid 1 \leq i_1 < \cdots < i_m \leq n\}$. Let S_1 be a random subset of M_m obtained by including each element of M_m in S_1 with probability $\log\log n / \log n$. Let S_2 be the complement of S_1 in M_m, that is, $S_2 = M_m - S_1$. Define a hypergraph $\tilde{A}$ as

$$\tilde{A}_{i_1 i_2 \ldots i_m} = \begin{cases} \mathbb{1}[A_{i_1 i_2 \ldots i_m} = 1], & \{i_1, i_2, \ldots, i_m\} \in S_1, \\ 0, & \text{otherwise.} \end{cases}$$

Here, $\mathbb{1}[E]$ is the indicator function of event E. Define hypergraph $\bar{A}$ as

$$\bar{A}_{i_1 i_2 \ldots i_m} = \begin{cases} A_{i_1 i_2 \ldots i_m}, & \{i_1, i_2, \ldots, i_m\} \in S_2, \\ *, & \text{otherwise.} \end{cases}$$

Then, hypergraph A is divided randomly into two independent hypergraphs, $\tilde{A}$ and $\bar{A}$.

In the second step, we apply the weak recovery algorithm HSC in (Ahn, Lee and Suh (2018)) to $\tilde{A}$. The HSC algorithm converts a hypergraph $\tilde{A}$ to an $n \times n$ similarity matrix B using $B_{ij} = \sum_{1 \le i_3 < i_4 < \cdots < i_m \le n} \tilde{A}_{iji_3 i_4 \ldots i_m}$, and then applies geometric two-clustering to the top two eigenvectors of B to output the communities $\tilde{I}_+(\sigma)$ and $\tilde{I}_-(\sigma)$. The sampling probability $\log\log n/\log n$ in the first step ensures that the hyperedge probability of $\tilde{A}$ has order $(\log\log n/\log n)\alpha = t\log\log n/n^{m-1}$ (Here, the $\log\log n$ factor can be replaced by any a_n with $a_n \to \infty$). According to Theorem 1 of Ahn, Lee and Suh (2018), $n - o(n)$ of the nodes are correctly labeled by the HSC algorithm with probability $1 - o(1)$.

The last step is to refine the communities $\tilde{I}_+(\sigma)$ and $\tilde{I}_-(\sigma)$ based on $\bar{A}$. For a set $S \subset [n]$, define $e(i, S)$ as

$$e(i,S) = \sum_{\substack{i_2,\ldots,i_m \in S\setminus\{i\} \\ i_2 < \cdots < i_m}} \left\{ \log\left(\frac{p}{q}\right) \mathbb{1}[\bar{A}_{ii_2\ldots i_m} = 1] + \log\left(\frac{1-p}{1-q}\right) \mathbb{1}[\bar{A}_{i_1 i_2 \ldots i_m} = 0] \right\}.$$

For each node $i \in \tilde{I}_+(\sigma)$, flip the label of i if

$$e\{i, \tilde{I}_+(\sigma)\} < e\{i, \tilde{I}_-(\sigma)\}.$$

For each node $j \in \tilde{I}_-(\sigma)$, flip the label of j if

$$e\{j, \tilde{I}_-(\sigma)\} < e\{j, \tilde{I}_-(\sigma)\}.$$

Let $\hat{I}_+(\sigma)$ and $\hat{I}_-(\sigma)$ be the resulting communities. If $|\hat{I}_+(\sigma)| \neq |\tilde{I}_+(\sigma)|$, output $\tilde{I}_+(\sigma)$ and $\tilde{I}_-(\sigma)$; otherwise, output $\hat{I}_+(\sigma)$ and $\hat{I}_-(\sigma)$.

The above algorithm is summarized in Algorithm 1.

Theorem 3. *For each fixed integer $m \ge 2$, if $t > I_m(p,q)$, with $I_m(p,q)$ defined in* (2.1)*, then Algorithm 1 exactly recovers the true label with probability* $1 - o(1)$.

Note that the time complexity of Algorithm 1 is at most $O(n^m)$. Specifically, the random splitting in Step 1 and the refinement in Step 3 have time complexity of at most $O(n^m)$. In Step 2, the weak recovery algorithm HSC of Ahn, Lee and Suh (2018) has time complexity $O(n^m)$ (see the comments below Remark 1 of Ahn, Lee and Suh (2018)). Hence, Theorem 3 states that the information-theoretic threshold can be attained by an algorithm with polynomial time complexity.

Algorithm 1: Spectral algorithm plus refinement for exact recovery.

Input: A censored m-uniform hypergraph A generated from $\mathcal{H}_m(n,p,q,\alpha)$.

Step 1: Random splitting

Randomly select elements in $M_m = \{(i_1,i_2,\ldots,i_m) \mid 1 \le i_1 < \cdots < i_m \le n\}$ with probability $\log\log n/\log n$ to form a subset $S_1 \subset M_m$, and let $S_2 = M_m - S_1$. Construct the hypergraph $\tilde{A}$ as $\tilde{A}_{i_1i_2\ldots i_m} = \mathbb{1}[A_{i_1i_2\ldots i_m} = 1]$ if $i_1,i_2,\ldots,i_m \in S_1$, and $\tilde{A}_{i_1i_2\ldots i_m} = 0$ otherwise. Construct the hypergraph $\bar{A}$ as $\bar{A}_{i_1i_2\ldots i_m} = A_{i_1i_2\ldots i_m}$ if $i_1,i_2,\ldots,i_m \in S_2$, and $\bar{A}_{i_1i_2\ldots i_m} = *$ otherwise.

Step 2: Spectral algorithm

Apply the weak recovery algorithm HSC in Ahn, Lee and Suh (2018) to $\tilde{A}$, and denote the community output as $\tilde{I}_+(\sigma)$ and $\tilde{I}_-(\sigma)$.

Step 3: Refinement

Flip the label of $i \in \tilde{I}_+(\sigma)$ if $e\{i,\tilde{I}_+(\sigma)\} < e\{i,\tilde{I}_-(\sigma)\}$.
Flip the label of $j \in \tilde{I}_-(\sigma)$ if $e\{j,\tilde{I}_-(\sigma)\} < e\{j,\tilde{I}_-(\sigma)\}$.
Let $\hat{I}_+(\sigma)$ and $\hat{I}_-(\sigma)$ be the resulting communities.

Output: If $|\hat{I}_+(\sigma)| \neq |\tilde{I}_+(\sigma)|$, output $\tilde{I}_+(\sigma)$ and $\tilde{I}_-(\sigma)$; otherwise, output $\hat{I}_+(\sigma)$ and $\hat{I}_-(\sigma)$.

2.3. Semi-definite relaxation algorithm

In subsection 2.2, we show that a spectral algorithm with a refinement step can achieve an exact recovery. It is also interesting to determine whether a single spectral algorithm without a refinement step can achieve the threshold. In the graph case ($m = 2$), the answer is yes and the semi-definite relaxation algorithm and the spectral algorithm are shown to succeed without a refinement step (Hajek, Wu and Xu (2016); Dhara et al. (2021)). In the hypergraph case ($m \ge 3$), either censored or uncensored, this remains an open problem. In this subsection, we study the semi-definite relaxation algorithm and analyze its performance. To this end, we define a new hypergraph based on the given hypergraph A, and transform it to a weighted graph. Then, we show that applying the semi-definite relaxation algorithm to the weighted graph can achieve an exact recovery.

Define the hypergraph $\widetilde{A}$ based on A as

$$\widetilde{A}_{i_1i_2\ldots i_m} = \mathbb{1}[A_{i_1i_2\ldots i_m} = 1],$$

and $\widetilde{A}_{i_1i_2\ldots i_m} = 0$ if $|\{i_1,i_2,\ldots,i_m\}| \le m-1$. Each hyperedge $\widetilde{A}_{i_1i_2\ldots i_m}$ takes a value in $\{1,0\}$. The hypergraph $\widetilde{A}$ shares the same community structure as that of A, because

$$\mathbb{E}(\widetilde{A}_{i_1i_2\ldots i_m}) = \begin{cases} p\alpha, & \{i_1,i_2,\ldots,i_m\} \subset I_+(\sigma) \text{ or } I_-(\sigma); \\ q\alpha, & \text{otherwise.} \end{cases}$$

Next, we construct a weighted graph $G = [G_{ij}]$ based on $\widetilde{A}$ by using

$$G_{ij} = \sum_{1 \le i_3 < \cdots < i_m \le n} \widetilde{A}_{iji_3 \ldots i_m}.$$

Define the semi-definite program problem (SDP) as

$$\begin{aligned} &\max_{Y} \langle G, Y \rangle \\ &\text{s.t. } Y \succeq 0 \\ &\langle Y, J \rangle = 0 \\ &Y_{ii} = 1, \ i \in [n], \end{aligned} \tag{2.2}$$

where J is an $n \times n$ all-one matrix. Suppose σ is the true label, and denote $Y = \sigma\sigma^T$. Let $\widehat{Y}$ be the solution to the SDP (2.2). The following result provides a sufficient condition under which $\widehat{Y}$ is an exact recovery of Y.

Theorem 4. *For each fixed integer $m \ge 2$, let*

$$J_m(p, q) = \frac{2^{m+2}(m-2)!\{mp - (m - 2^m)q\}}{(p-q)^2}.$$

If $t > J_m(p, q)$, then $\mathbb{P}(\widehat{Y} = Y) = 1 - o(1)$, where $Y = \sigma\sigma^T$, with true label σ.

Note that $J_m(p, q) > I_m(p, q)$, for each $m \ge 2$. When $m = 2$ and the graph is uncensored, $\widehat{Y}$ can exactly recover the true label up to the information-theoretic threshold (Hajek, Wu and Xu (2016)). However, for $m \ge 3$, it is unclear whether $\widehat{Y}$ succeeds in the range $I_m(p, q) < t < J_m(p, q)$. A similar gap exists in the uncensored hypergraph case (Kim, Bandeira and Goemans (2018)). The proof of Theorem 4 is provided in the Supplementary Material.

3. Proof of Theorem 1

In this section, we prove Theorem 1.

Let $l(\sigma)$ be the log-likelihood function of a label σ. Note that by Definition 1.1, the true label vector σ is uniformly and independently selected from $S = \{\pm 1\}^n$. By Proposition 4.1 in Dhara et al. (2021), if there are labels η_t $(1 \le t \le k_n)$ with $k_n \to \infty$ such that $l(\eta_1) = l(\eta_2) = \cdots = l(\eta_{k_n}) = l(\sigma)$, then the MLE fails to exactly recover the true label with probability $1 - o(1)$. Our proof proceeds by constructing labels η_t $(1 \le t \le k_n)$ with $k_n \to \infty$ under the condition $t < I_m(p, q)$.

First, we provide an explicit expression of the likelihood function. Note that for distinct nodes $i_1, i_2, \ldots, i_m$, we have

$$A_{i_1 i_2 \ldots i_m} = \begin{cases} 1\,, \\ 0\,, \\ *\,. \end{cases}$$

For convenience, let $\mathbb{1}[E]$ be the indicator function of event E and

$$\mathbb{1}_{i_1 i_2 \ldots i_m}(\sigma) = \mathbb{1}[\sigma_{i_1} = \sigma_{i_2} = \cdots = \sigma_{i_m}].$$

Then, the likelihood function for σ given an observation of hypergraph A from $\mathcal{H}_m(n, p, q, \alpha)$ is

$$\begin{aligned} L = & \prod_{1 \le i_1 < \cdots < i_m \le n} (p\alpha)^{\mathbb{1}[A_{i_1 i_2 \ldots i_m}=1]\mathbb{1}_{i_1 i_2 \ldots i_m}(\sigma)} \{\alpha(1-p)\}^{\mathbb{1}[A_{i_1 i_2 \ldots i_m}=0]\mathbb{1}_{i_1 i_2 \ldots i_m}(\sigma)} \\ & \times (q\alpha)^{\mathbb{1}\{A_{i_1 i_2 \ldots i_m}=1](1-\mathbb{1}_{i_1 i_2 \ldots i_m}(\sigma))} \{\alpha(1-q)\}^{\mathbb{1}[A_{i_1 i_2 \ldots i_m}=0](1-\mathbb{1}_{i_1 i_2 \ldots i_m}(\sigma))} \\ & \times (1-\alpha)^{\mathbb{1}[A_{i_1 i_2 \ldots i_m}=*]} \\ = & \prod_{1 \le i_1 < \cdots < i_m \le n} (1-\alpha)^{\mathbb{1}[A_{i_1 i_2 \ldots i_m}=*]} (q\alpha)^{\mathbb{1}[A_{i_1 i_2 \ldots i_m}=1]} \left(\frac{p}{q}\right)^{\mathbb{1}[A_{i_1 i_2 \ldots i_m}=1]\mathbb{1}_{i_1 i_2 \ldots i_m}(\sigma)} \\ & \times [\alpha(1-q)]^{\mathbb{1}[A_{i_1 i_2 \ldots i_m}=0]} \left(\frac{1-p}{1-q}\right)^{\mathbb{1}[A_{i_1 i_2 \ldots i_m}=0]\mathbb{1}_{i_1 i_2 \ldots i_m}(\sigma)} \\ = & \prod_{1 \le i_1 < \cdots < i_m \le n} (1-\alpha)^{\mathbb{1}[A_{i_1 i_2 \ldots i_m}=*]} (q\alpha)^{\mathbb{1}[A_{i_1 i_2 \ldots i_m}=1]} [\alpha(1-q)]^{\mathbb{1}[A_{i_1 i_2 \ldots i_m}=0]} \\ & \times \prod_{1 \le i_1 < \cdots < i_m \le n} \left(\frac{p}{q}\right)^{\mathbb{1}[A_{i_1 i_2 \ldots i_m}=1]\mathbb{1}_{i_1 i_2 \ldots i_m}(\sigma)} \left(\frac{1-p}{1-q}\right)^{\mathbb{1}[A_{i_1 i_2 \ldots i_m}=0]\mathbb{1}_{i_1 i_2 \ldots i_m}(\sigma)}. \end{aligned}$$

The MLE is obtained by maximizing L with respect to σ. The first product factor of L does not involve σ. Hence, we need only maximize the second product factor of L to obtain the MLE. Denote

$$\begin{aligned} l(\sigma) = \sum_{1 \le i_1 < \cdots < i_m \le n} \Bigg\{ & \log\left(\frac{p}{q}\right) \mathbb{1}[A_{i_1 i_2 \ldots i_m} = 1] \mathbb{1}_{i_1 i_2 \ldots i_m}(\sigma) \\ & + \log\left(\frac{1-p}{1-q}\right) \mathbb{1}[A_{i_1 i_2 \ldots i_m} = 0] \mathbb{1}_{i_1 i_2 \ldots i_m}(\sigma) \Bigg\}. \end{aligned}$$

The log-likelihood function is equal to

$$\log L = R_n + l(\sigma), \tag{3.1}$$

where R_n is independent of σ.

Below, we construct labels η_t $(1 \le t \le k_n)$ with $k_n \to \infty$ under the condition $t < I_m(p, q)$. Because R_n is independent of σ, we need only focus on $l(\sigma)$.

Note that

$$
\begin{aligned}
&l(\sigma) = \\
&\left\{\log\left(\frac{p}{q}\right)\mathbb{1}[A_{i_1\ldots i_m}=1] + \log\left(\frac{1-p}{1-q}\right)\mathbb{1}[A_{i_1\ldots i_m}=0]\right\}\mathbb{1}[\sigma_{i_1}=\cdots=\sigma_{i_m}=+1] \\
&+\left\{\log\left(\frac{p}{q}\right)\mathbb{1}[A_{i_1\ldots i_m}=1] + \log\left(\frac{1-p}{1-q}\right)\mathbb{1}[A_{i_1\ldots i_m}=0]\right\}\mathbb{1}[\sigma_{i_1}=\cdots=\sigma_{i_m}=-1].
\end{aligned}
$$

Suppose $i_0 \in I_+(\sigma)$ has exactly m_1 present hyperedges and m_2 absent hyperedges in $I_+(\sigma)$, and has exactly m_1 present hyperedges and m_2 absent hyperedges in $I_-(\sigma)$. Furthermore, suppose $j_0 \in I_-(\sigma)$ has exactly m_1 present hyperedges and m_2 absent hyperedges in $I_+(\sigma)$, and has exactly m_1 present hyperedges and m_2 absent hyperedges in $I_-(\sigma)$. Then, $l(\sigma)$ remains the same if we flip the labels of i_0 and j_0. Let $\tilde{\sigma}$ be the labels obtained from σ by flipping the labels of i_0 and j_0. We verify that $l(\sigma) = l(\tilde{\sigma})$. To prove this, let $T_1 = \log(p/q)$ and $T_2 = \log((1-p)/(1-q))$; then,

$$
\begin{aligned}
l(\sigma) = &\left(T_1 \sum_{i_1 i_2 \ldots i_m} \mathbb{1}[A_{i_1 i_2 \ldots i_m}=1] + T_2 \sum_{i_1 i_2 \ldots i_m} \mathbb{1}[A_{i_1 i_2 \ldots i_m}=0]\right) \\
&\mathbb{1}[\sigma_{i_1}=\cdots=\sigma_{i_m}=+1] \\
&+\left(T_1 \sum_{i_1 i_2 \ldots i_m} \mathbb{1}[A_{i_1 i_2 \ldots i_m}=1] + T_2 \sum_{i_1 i_2 \ldots i_m} \mathbb{1}[A_{i_1 i_2 \ldots i_m}=0]\right) \\
&\mathbb{1}[\sigma_{i_1}=\cdots=\sigma_{i_m}=-1].
\end{aligned}
$$

Further, $l(\sigma)$ can be written as

$$
\begin{aligned}
l(\sigma) = &T_1 \sum_{\substack{i_1 i_2 \ldots i_m \in I_+(\sigma) \\ i_1 i_2 \ldots i_m \neq i_0}} \mathbb{1}[A_{i_1 i_2 \ldots i_m}=1] + T_1 \sum_{\substack{i_2 \ldots i_m \in I_+(\sigma) \\ i_2 \ldots i_m \neq i_0}} \mathbb{1}[A_{i_0 i_2 \ldots i_m}=1] \\
&+T_2 \sum_{\substack{i_1 i_2 \ldots i_m \in I_+(\sigma) \\ i_1 i_2 \ldots i_m \neq i_0}} \mathbb{1}[A_{i_1 i_2 \ldots i_m}=0] + T_2 \sum_{\substack{i_2 \ldots i_m \in I_+(\sigma) \\ i_2 \ldots i_m \neq i_0}} \mathbb{1}[A_{i_0 i_2 \ldots i_m}=0] \\
&+T_1 \sum_{\substack{i_1 i_2 \ldots i_m \in I_-(\sigma) \\ i_1 i_2 \ldots i_m \neq j_0}} \mathbb{1}[A_{i_1 i_2 \ldots i_m}=1] + T_1 \sum_{\substack{i_2 \ldots i_m \in I_-(\sigma) \\ i_2 \ldots i_m \neq j_0}} \mathbb{1}[A_{j_0 i_2 \ldots i_m}=1] \\
&+T_2 \sum_{\substack{i_1 i_2 \ldots i_m \in I_-(\sigma) \\ i_1 i_2 \ldots i_m \neq j_0}} \mathbb{1}[A_{i_1 i_2 \ldots i_m}=0] + T_2 \sum_{\substack{i_2 \ldots i_m \in I_-(\sigma) \\ i_2 \ldots i_m \neq j_0}} \mathbb{1}[A_{j_0 i_2 \ldots i_m}=0],
\end{aligned}
$$

and

$$
l(\tilde{\sigma}) = T_1 \sum_{\substack{i_1 i_2 \ldots i_m \in I_+(\sigma) \\ i_1 i_2 \ldots i_m \neq j_0}} \mathbb{1}[A_{i_1 i_2 \ldots i_m}=1] + T_1 \sum_{\substack{i_2 \ldots i_m \in I_+(\sigma) \\ i_2 \ldots i_m \neq j_0}} \mathbb{1}[A_{j_0 i_2 \ldots i_m}=1]
$$

$$
\begin{aligned}
&+ T_2 \sum_{\substack{i_1 i_2 \dots i_m \in I_+(\sigma) \\ i_1 i_2 \dots i_m \neq j_0}} \mathbb{1}[A_{i_1 i_2 \dots i_m} = 0] + T_2 \sum_{\substack{i_2 \dots i_m \in I_+(\sigma) \\ i_2 \dots i_m \neq j_0}} \mathbb{1}[A_{j_0 i_2 \dots i_m} = 0] \\
&+ T_1 \sum_{\substack{i_1 i_2 \dots i_m \in I_-(\sigma) \\ i_1 i_2 \dots i_m \neq i_0}} \mathbb{1}[A_{i_1 i_2 \dots i_m} = 1] + T_1 \sum_{\substack{i_2 \dots i_m \in I_-(\sigma) \\ i_2 \dots i_m \neq i_0}} \mathbb{1}[A_{i_0 i_2 \dots i_m} = 1] \\
&+ T_2 \sum_{\substack{i_1 i_2 \dots i_m \in I_-(\sigma) \\ i_1 i_2 \dots i_m \neq i_0}} \mathbb{1}[A_{i_1 i_2 \dots i_m} = 0] + T_2 \sum_{\substack{i_2 \dots i_m \in I_-(\sigma) \\ i_2 \dots i_m \neq i_0}} \mathbb{1}[A_{i_0 i_2 \dots i_m} = 0].
\end{aligned}
$$

Then, $l(\sigma) = l(\tilde{\sigma})$ by the assumption of i_0 and j_0.

Next, we show there are k_n ($k_n \to \infty$) such pairs. More specifically, we show that there exist $i_1, i_2, \dots, i_k \in I_+(\sigma)$ and $j_1, j_2, \dots, j_k \in I_-(\sigma)$ with $k \gg 1$ such that the likelihood function remains unchanged if we flip the label of a pair (i_t, j_t), for $t = 1, 2, \dots, k$. Let η_t be the label obtained by flipping the label of i_t, j_t in σ. Then, $l(\eta_t) = l(\sigma)$, for $1 \le t \le k \to \infty$.

Let $n_1 = |I_+(\sigma)|$ and $n_2 = |I_-(\sigma)|$. Then, $n_1, n_2 = (n/2)\{1 + O(n^{-1/3})\}$ with probability $1 - o(1)$. Hence, we take $n_1 = n_2 = n/2$ below. Let $S_+ \subset I_+(\sigma)$ be a random subset with $|S_+| = n/\log^2 n$, and $S_- \subset I_-(\sigma)$ be a random subset with $|S_-| = n/\log^2 n$. Denote $S = S_+ \cup S_-$. Define

$$
S_0 = \left\{ i \in S | any\ i_2, \dots, i_t \in S, i_{t+1}, \dots, i_m \in S^c,\ s.t.\ A_{i i_2 \dots i_t i_{t+1} \dots i_m} = *,\ t \ge 2 \right\}.
$$

For each node $i \in S_0$, the hyperedge $A_{i i_2 \dots i_m}$ is possibly revealed if and only if $\{i_2, \dots, i_m\} \subset I_+(\sigma) - S$ or $\{i_2, \dots, i_m\} \subset I_-(\sigma) - S$.

We show that $|S_0| = 2n\{1 + o(1)\}/(\log^2 n)$ with probability $1 - o(1)$. Let

$$
T = \sum_{t=2}^{m} \sum_{\substack{i_1, \dots, i_t \in S \\ i_{t+1}, \dots, i_m \in S^c}} \mathbb{1}[A_{i_1 i_2 \dots i_t i_{t+1} \dots i_m} \neq *].
$$

The expectation of T is

$$
\begin{aligned}
\mathbb{E}T &= \sum_{t=2}^{m} \binom{\frac{2n}{\log^2 n}}{t} \binom{n - \frac{2n}{\log^2 n}}{m - t} \alpha \\
&= \sum_{t=2}^{m} \binom{\frac{2n}{\log^2 n}}{t} \binom{n - \frac{2n}{\log^2 n}}{m - t} \frac{t \log n}{n^{m-1}} \\
&= \frac{c \cdot n^m}{\log^4 n} \frac{t \log n}{n^{m-1}} \\
&\asymp \frac{n}{\log^3 n}.
\end{aligned}
$$

Hence, by the Markov inequality, we have

$$\mathbb{P}\left(T \geq \frac{n}{\log^2 n\sqrt{\log n}}\right) \leq \frac{1}{n/(\log^2 n\sqrt{\log n})}\frac{c\cdot n}{\log^3 n} = \frac{\sqrt{\log n}}{\log n} = o(1).$$

Then, $T < n/(\log^2 n\sqrt{\log n})$ with probablity $1-o(1)$. Hence, $|S_0| = (2n/\log^2 n)$ $(1+o(1))$ with probability $1-o(1)$.

Let $m_1 = \sqrt{pq}t\log n/\{2^{m-1}(m-1)!\}$ and $m_2 = \sqrt{(1-p)(1-q)}t\log n/\{2^{m-1}(m-1)!\}$. For some $k \gg 1$, we show that there exists $i_t \in S_0 \cap S_+$, for $(1 \leq t \leq k)$, such that i_t has m_1 present hyperedges and m_2 absent hyperedges in $I_+(\sigma)$ and $I_-(\sigma)$. Denote

$$\widetilde{n}_1 = \binom{n_1 - \frac{2n}{\log^2 n}}{m-1} \sim \frac{n^{m-1}}{2^{m-1}(m-1)!}.$$

Let $i_0 \in S_0 \cap S_+$. The probability that i_0 has m_1 present hyperedges and m_2 absent hyperedges in $I_+(\sigma)$ and $I_-(\sigma)$ is

$$\begin{aligned}
p_0 &= \frac{\widetilde{n}_1!}{m_1!m_2!(\widetilde{n}_1-m_1-m_2)!}\cdot(\alpha p)^{m_1}\{\alpha(1-p)\}^{m_2}(1-\alpha)^{(\widetilde{n}_1-m_1-m_2)}\\
&\quad\times\frac{\widetilde{n}_1!}{m_1!m_2!(\widetilde{n}_1-m_1-m_2)!}\cdot(\alpha q)^{m_1}\{\alpha(1-q)\}^{m_2}(1-\alpha)^{(\widetilde{n}_1-m_1-m_2)}\\
&\sim \frac{1}{m_1!^2m_2!^2}\left\{\frac{\widetilde{n}_1^{\widetilde{n}_1+1/2}e^{-\widetilde{n}_1}}{(\widetilde{n}_1-m_1-m_2)^{\widetilde{n}_1-m_1-m_2+1/2}e^{-\widetilde{n}_1+m_1+m_2}}\right\}^2(\alpha^2pq)^{m_1}\\
&\quad\times\{\alpha^2(1-p)(1-q)\}^{m_2}(1-\alpha)^{2(\widetilde{n}_1-m_1-m_2)}\\
&= \frac{1}{m_1!^2m_2!^2}\left[\frac{(\widetilde{n}_1-m_1-m_2)^{m_1+m_2}}{e^{m_1+m_2}\{1-(m_1+m_2)/\widetilde{n}_1\}^{\widetilde{n}_1+1/2}}\right]^2(\alpha^2pq)^{m_1}\\
&\quad\times\{\alpha^2(1-p)(1-q)\}^{m_2}(1-\alpha)^{2(\widetilde{n}_1-m_1-m_2)}\\
&= \frac{1}{m_1!^2m_2!^2}\left\{\frac{\widetilde{n}_1^{m_1+m_2}}{e^{m_1+m_2}e^{-(m_1+m_2)}}\right\}^2(\alpha^2pq)^{m_1}\\
&\quad\times\left\{\alpha^2(1-p)(1-q)\right\}^{m_2}e^{-t\log n/\{2^{m-2}(m-1)!\}}\\
&= \frac{\tilde{n}_1^{2(m_1+m_2)}}{m_1!^2m_2!^2}e^{-t\log n/\{2^{m-2}(m-1)!\}}(\alpha^2pq)^{m_1}\{\alpha^2(1-p)(1-q)\}^{m_2}\\
&= \frac{n^{-t/\{2^{m-2}(m-1)!\}}}{m_1!^2m_2!^2}(\alpha^2\tilde{n}_1^2pq)^{m_1}\{\alpha^2\tilde{n}_1^2(1-p)(1-q)\}^{m_2}\\
&= n^{-t/\{2^{m-2}(m-1)!\}}\frac{e^{2(m_1+m_2)}}{4\pi^2m_1m_2}\left(\frac{\alpha^2\tilde{n}_1^2pq}{m_1^2}\right)^{m_1}\left\{\frac{\alpha^2\tilde{n}_1^2(1-p)(1-q)}{m_2^2}\right\}^{m_2}\\
&= \frac{1}{4\pi^2m_1m_2}n^{-t/\{2^{m-2}(m-1)!\}}e^{[\{\sqrt{pq}+\sqrt{(1-p)(1-q)}\}/\{2^{m-2}(m-1)!\}t\log n}\\
&= \frac{1}{4\pi^2m_1m_2}n^{-[t/\{2^{m-2}(m-1)!\}]\{1-\sqrt{pq}-\sqrt{(1-p)(1-q)}\}}
\end{aligned}$$

$$= \frac{1}{4\pi^2 m_1 m_2} n^{-t\cdot[\{(\sqrt{p}-\sqrt{q})^2+(\sqrt{1-p}-\sqrt{1-q})^2\}/\{2^{m-1}(m-1)!\}]}.$$

If $t < 2^{m-1}(m-1)!/\{(\sqrt{p}-\sqrt{q})^2 + (\sqrt{1-p}-\sqrt{1-q})^2\}$, then $p_0 \gg n^{1-\epsilon}/n$, for some $\epsilon \in (0,1)$. Similarly, the probability that $j_0 \in S_0 \cap S_-$ has m_1 present hyperedges and m_2 absent hyperedges in $I_+(\sigma)$ and $I_-(\sigma)$ is equal to p_0.

For $i \in S_0$, let $\mathbb{1}_i$ denote the event that i has m_1 present hyperedges and m_2 absent hyperedges in $I_+(\sigma)$ and $I_-(\sigma)$. Define two random variables,

$$X = \sum_{i \in S_0 \cap S_+} \mathbb{1}_i \ \ and \ \ Y = \sum_{i \in S_0 \cap S_-} \mathbb{1}_i.$$

If $\mathbb{1}_i = \mathbb{1}_j = 1$, for $i \in S_0 \cap S_+$ and $j \in S_0 \cap S_-$, then the likelihood function remains unchanged if we flip the labels of i and j. By Chebyshev's inequality, given $|S_0 \cap S_+|$, we have

$$\begin{aligned}
&\mathbb{P}\left(X \le (1-\epsilon)\frac{2n}{\log^2 n}p_0\right) \\
&= \mathbb{P}\left(X \le (1-\epsilon)\frac{2n}{\log^2 n}p_0 \middle| |S_0 \cap S_+| \ge \frac{2n}{\log^2 n}\{1-o(1)\}\right) \\
&\quad \cdot \mathbb{P}\left(|S_0 \cap S_+| \ge \frac{2n}{\log^2 n}\{1-o(1)\}\right) \\
&\quad + \mathbb{P}\left(X \le (1-\epsilon)\frac{2n}{\log^2 n}p_0 \middle| |S_0 \cap S_+| < \frac{2n}{\log^2 n}\{1-o(1)\}\right) \\
&\quad \cdot \mathbb{P}\left(|S_0 \cap S_+| < \frac{2n}{\log^2 n}\right) \\
&\le \mathbb{P}\left(X \le (1-\epsilon)|S_0 \cap S_+|p_0 \middle| |S_0 \cap S_+| \ge \frac{2n}{\log^2 n}\{1-o(1)\}\right) + o(1) \\
&\le \frac{1}{\epsilon^2 |S_0 \cap S_+| p_0} + o(1).
\end{aligned}$$

Note that $p_0 \gg n^{1-\epsilon}/n$ for some $\epsilon > 0$ and $|S_0 \cap S_+| \ge (2n/\log^2 n)\{1-o(1)\}$. Then, $X \ge |S_0 \cap S_+|p_0 \to +\infty$ with probability $1-o(1)$. Similarly, $Y \ge |S_0 \cap S_+|p_0 \to +\infty$ with probability $1-o(1)$. As a result, we have pairs (i_t, j_t) $(1 \le t \le k \to \infty)$. For each t, the likelihood remains constant after flipping the labels of i_t and j_t. The proof is complete by Proposition 4.1 in Dhara et al. (2021).

4. Proof of Theorem 2

Let σ be the MLE. Recall the log-likelihood function in (3.1). The MLE fails to exactly recover the true label if there exists a label η such that $l(\eta) \ge l(\sigma)$ with probability δ, for some constant $\delta > 0$. Our proof proceeds by showing that the probability that the MLE fails is $o(1)$.

The MLE is obtained by maximizing $\log L$ in (3.1) with respect to σ. The first term of $\log L$ does not involve σ. Hence, we need only maximize the second term of $\log L$ to obtain the MLE. Let σ be the MLE. Recall that the MLE fails if there exists a label η such that $l(\eta) \geq l(\sigma)$ with probability δ, for some constant $\delta > 0$. Below, we show the probability that the MLE fails is $o(1)$.

Let k be an even number and $1 \leq k \leq n/2$. Define the Hamming distance between two labels σ and η as

$$d(\sigma, \eta) = \min \left\{ \sum_{i=1}^{n} \mathbb{1}[\sigma_i \neq \eta_i], \sum_{i=1}^{n} \mathbb{1}[\sigma_i \neq -\eta_i] \right\}.$$

Let η be a label such that $d(\sigma, \eta) = k$, and denote

$$C_{i_1 i_2 \ldots i_m}(A) = \log\left(\frac{p}{q}\right) \mathbb{1}[A_{i_1 i_2 \ldots i_m} = 1] + \log\left(\frac{1-p}{1-q}\right) \mathbb{1}[A_{i_1 i_2 \ldots i_m} = 0].$$

Then, the log-likelihood difference at η and σ is

$$l(\eta) - l(\sigma) = \sum_{1 \leq i_1 < \cdots < i_m \leq n} C_{i_1 i_2 \ldots i_m}(A) \{ \mathbb{1}_{i_1 \ldots i_m}(\eta) - \mathbb{1}_{i_1 \ldots i_m}(\sigma) \}.$$

We show that

$$\mathbb{P}(\exists k \textit{ and } d(\sigma, \eta) = k, \textit{ s.t. } l(\eta) - l(\sigma) \geq 0) = o(1).$$

Recall $\mathrm{I}_+(\sigma)$ and $\mathrm{I}_-(\sigma)$. Denote $\mathbb{1}_{i_1 \ldots i_m}(\eta) = I[\eta_{i_1} = \eta_{i_2} = \cdots = \eta_{i_m}]$. Note that

$$\begin{aligned} &\mathbb{1}_{i_1 \ldots i_m}(\eta) - \mathbb{1}_{i_1 \ldots i_m}(\sigma) \\ &= \begin{cases} 1, & i_1 \ldots i_m \subset \mathrm{I}_+(\eta) \text{ or } \mathrm{I}_-(\eta),\ i_1 \ldots i_m \not\subset \mathrm{I}_+(\sigma)\ , \mathrm{I}_-(\sigma); \\ -1, & i_1 \ldots i_m \subset \mathrm{I}_+(\sigma) \text{ or } \mathrm{I}_-(\sigma),\ i_1 \ldots i_m \not\subset \mathrm{I}_+(\eta), \mathrm{I}_-(\eta); \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Hence, $l(\eta) - l(\sigma)$ is written as

$$l(\eta) - l(\sigma) = \sum_{\substack{i_1 \ldots i_m \\ i_1 \ldots i_m \subset I_+(\eta) \text{ or } \mathrm{I}_-(\eta) \\ i_1 \ldots i_m \not\subset I_+(\sigma), I_-(\sigma)}} C_{i_1 \ldots i_m}(A) - \sum_{\substack{i_1 \ldots i_m \\ i_1 \ldots i_m \subset I_+(\sigma) \text{ or } \mathrm{I}_-(\sigma) \\ i_1 \ldots i_m \not\subset I_+(\eta), I_-(\eta)}} C_{i_1 \ldots i_m}(A).$$

It is easy to verify that there are $n_k = 2\left\{\binom{n/2}{m} - \binom{k/2}{m} - \binom{(n-k)/2}{m}\right\}$ hyperedges $\{i_1, \ldots, i_m\}$ such that $\{i_1 \ldots i_m\} \subset \mathbb{1}_+(\eta)$ or $\mathbb{1}_-(\eta)$ and $\{i_1 \ldots i_m\} \not\subset \mathbb{1}_+(\sigma), \mathbb{1}_-(\sigma)$. For convenience, define random variables X and Y as

$$\begin{aligned} &\mathbb{P}(X = 1) = \alpha p, \quad \mathbb{P}(X = 0) = \alpha(1-p), \quad \mathbb{P}(X = -1) = 1 - \alpha. \\ &\mathbb{P}(Y = 1) = \alpha q, \quad \mathbb{P}(Y = 0) = \alpha(1-q), \quad \mathbb{P}(Y = -1) = 1 - \alpha. \end{aligned}$$

Let X_i, Y_i be independent and identically distributed (i.i.d.) copies of X, Y, respectively, and

$$W_i = \log\left(\frac{p}{q}\right)\mathbb{1}[X_i = 1] + \log\left(\frac{1-p}{1-q}\right)\mathbb{1}[X_i = 0]$$
$$V_i = \log\left(\frac{p}{q}\right)\mathbb{1}[Y_i = 1] + \log\left(\frac{1-p}{1-q}\right)\mathbb{1}[Y_i = 0].$$

For any $r > 0$, by the Markov inequality, we have

$$\begin{aligned}
\mathbb{P}\{l(\eta) - l(\sigma) \geq 0\} &= \mathbb{P}\left\{\sum_{i=1}^{n_k}(V_i - W_i) \geq 0\right\} \\
&= \mathbb{P}\left\{\sum_{i=1}^{n_k}(W_i - V_i) \leq 0\right\} \\
&= \mathbb{P}\left\{e^{\sum_{i=1}^{n_k}(-r)(W_i - V_i)} \geq 1\right\} \\
&\leq \left\{\mathbb{E}\left(e^{-rW_1}\right)\mathbb{E}\left(e^{rV_1}\right)\right\}^{n_k}.
\end{aligned}$$

Next, we find explicit expressions for the expectations $\mathbb{E}\left(e^{-rW_1}\right)$ and $\mathbb{E}\left(e^{rV_1}\right)$:

$$\begin{aligned}
\mathbb{E}(e^{-rW_1}) &= \mathbb{E}e^{-r[\log(p/q)\mathbb{1}[X_i=1]+\log\{(1-p)/(1-q)\}\mathbb{1}[X_i=0]]} \\
&= e^{-r\log(p/q)}\alpha p + e^{-r\log\{(1-p)/(1-q)\}}\alpha(1-p) + (1-\alpha) \\
&= \left(\frac{q}{p}\right)^r \alpha p + \left(\frac{1-q}{1-p}\right)^r \alpha(1-p) + (1-\alpha) \\
\mathbb{E}(e^{rV_1}) &= \mathbb{E}e^{r[\log(p/q)\mathbb{1}[Y_i=1]+\log\{(1-p)/(1-q)\}\mathbb{1}[Y_i=0]]} \\
&= e^{r\log(p/q)}\alpha q + e^{r\log\{(1-p)/(1-q)\}}\alpha(1-q) + (1-\alpha) \\
&= \left(\frac{p}{q}\right)^r \alpha q + \left(\frac{1-p}{1-q}\right)^r \alpha(1-q) + (1-\alpha).
\end{aligned}$$

Taking $r = 1/2$ yields

$$\begin{aligned}
\mathbb{E}(e^{-rW_1}) &= \alpha\sqrt{pq} + \alpha\sqrt{(1-p)(1-q)} + (1-\alpha) \\
&= 1 + \alpha\{\sqrt{pq} + \sqrt{(1-p)(1-q)} - 1\}, \\
\mathbb{E}(e^{rV_1}) &= \alpha\sqrt{pq} + \alpha\sqrt{(1-p)(1-q)} + (1-\alpha) \\
&= 1 + \alpha\{\sqrt{pq} + \sqrt{(1-p)(1-q)} - 1\}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\log\mathbb{P}\{l(\eta) - l(\sigma) \geq 0\} &\leq n_k \log\mathbb{E}(e^{-rW_1}) + n_k \log\mathbb{E}(e^{rV_1}) \\
&\leq n_k\{2\alpha(\sqrt{pq} + \sqrt{(1-p)(1-q)} - 1)\}
\end{aligned}$$

$$= n_k\alpha\left[(-1)\left\{(\sqrt{p}-\sqrt{q})^2+(\sqrt{1-p}-\sqrt{1-q})^2\right\}\right]$$
$$= -n_k\alpha\left[(\sqrt{p}-\sqrt{q})^2+(\sqrt{1-p}-\sqrt{1-q})^2\right]. \quad (4.1)$$

For $k \geq n/\log\log n$, it is easy to check that $n_k \geq (1/2^{m-1})(n/\log\log n)\binom{n-1}{m-1}$. Hence, by (4.1), we obtain

$$\begin{aligned}
&\mathbb{P}\{l(\eta)-l(\sigma)\geq 0\}\\
&\leq e^{-\{(\sqrt{p}-\sqrt{q})^2+(\sqrt{1-p}-\sqrt{1-q})^2\}(t\log n/n^{m-1})(1/2^{m-1})(n/\log\log n)\{n^{m-1}/(m-1)!\}}\\
&= e^{-\{(\sqrt{p}-\sqrt{q})^2+(\sqrt{1-p}-\sqrt{1-q})^2\}\{t/2^{m-1}(m-1)!\}(n\log n/\log\log n)}\\
&= e^{-c(n\log n/\log\log n)},
\end{aligned}$$

for some positive constant c. Clearly, there are $\binom{n/2}{k/2}^2$ choices for η, with $d(\sigma,\eta)=k$. Note that $\binom{n/2}{k/2}^2 \leq 2^n$. Then, the probability that there exists η with $d(\sigma,\eta)=k$ for $k \geq n/\log\log n$ is upper bounded by

$$\frac{n}{2}\cdot 2^n\cdot e^{-c(n\log n/\log\log n)} = e^{n\log 2+\log(n/2)-cn(\log n/\log\log n)} = o(1).$$

For $k < n/\log\log n$, we have $n_k = (k/2^{m-1})\binom{n-1}{m-1}$. Then,

$$\begin{aligned}
\mathbb{P}\{l(\eta)-l(\sigma)\geq 0\} &\leq e^{-\{(\sqrt{p}-\sqrt{q})^2+(\sqrt{1-p}-\sqrt{1-q})^2\}(t\log n/n^{m-1})(k/2^{m-1})\{n^{m-1}/(m-1)!\}}\\
&= e^{-[\{(\sqrt{p}-\sqrt{q})^2+(\sqrt{1-p}-\sqrt{1-q})^2\}/\{2^{m-1}(m-1)!\}]tk\log n}\\
&= n^{-[\{(\sqrt{p}-\sqrt{q})^2+(\sqrt{1-p}-\sqrt{1-q})^2\}/\{2^{m-1}(m-1)!\}]/tk}.
\end{aligned}$$

There are $\binom{n/2}{k/2}^2 \leq n^k$ choices for η with $d(\sigma,\eta)=k$. Then, the probability that there exists η with $d(\sigma,\eta)=k$ for $k < n/\log\log n$ is upper bounded by

$$\begin{aligned}
k\cdot\binom{\frac{n}{2}}{\frac{k}{2}}^2\mathbb{P}\{l(\eta)-l(\sigma)\geq 0\} &\leq kn^k\cdot n^{-[\{(\sqrt{p}-\sqrt{q})^2+(\sqrt{1-p}-\sqrt{1-q})^2\}/\{2^{m-1}(m-1)!\}]tk}\\
&\leq kn^k n^{-(1+\epsilon)k}\\
&= \frac{k}{n^{\epsilon k}} = o(1),
\end{aligned}$$

where ϵ is a constant such that $[\{(\sqrt{p}-\sqrt{q})^2+(\sqrt{1-p}-\sqrt{1-q})^2\}/\{2^{m-1}(m-1)!\}]t = 1+\epsilon$. This is possible by the condition $t > 2^{m-1}(m-1)!/\{(\sqrt{p}-\sqrt{q})^2+(\sqrt{1-p}-\sqrt{1-q})^2\}$. The proof is complete.

5. Proof of Theorem 3

This proof proceeds by showing that the probability that there exists a mislabeled node goes to zero. By the definition of the hypergraph $\tilde{A}$, we have

$$\mathbb{P}(\tilde{A}_{i_1 i_2 \ldots i_m} = 1) = \begin{cases} \frac{\log\log n}{\log n} \cdot \alpha p, & \{i_1, i_2, \ldots, i_m\} \subset I_+(\sigma) \text{ or } I_-(\sigma), \\ \frac{\log\log n}{\log n} \cdot \alpha q, & \text{otherwise.} \end{cases}$$
$$= \begin{cases} \frac{tp\log\log n}{n^{m-1}}, & \{i_1, i_2, \ldots, i_m\} \subset I_+(\sigma) \text{ or } I_-(\sigma), \\ \frac{tq\log\log n}{n^{m-1}}, & \text{otherwise.} \end{cases}$$

Then, $\tilde{A}$ has the same community structure as the original hypergraph A, and in $\tilde{A}$, the order of a hyperedge probability is $\log\log n/n^{m-1}$. With probability $1 - o(1)$, the weak recovery algorithm in Ahn, Lee and Suh (2018) recovers the true labels of $(1-\delta)n$ nodes of $\tilde{A}$, with $\delta = o(1)$. Denote the communities as $\tilde{I}_+(\sigma)$ and $\tilde{I}_-(\sigma)$. Hence, with probability $1 - o(1)$, $(\delta/2)n$ nodes in $\tilde{I}_+(\sigma)$ and $\tilde{I}_-(\sigma)$ are mislabeled. In the refinement step, a node i among the correctly labeled $\{(1-\delta)/2\}n$ nodes in $\tilde{I}_+(\sigma)$ is mislabeled if

$$e\{i, \tilde{I}_+(\sigma)\} < e\{i, \tilde{I}_-(\sigma)\}.$$

A node among the mislabeled $(\delta/2)n$ nodes in $\tilde{I}_+(\sigma)$ remains mislabeled if

$$e\{i, \tilde{I}_+(\sigma)\} \geq e\{i, \tilde{I}_-(\sigma)\}.$$

A similar result holds for nodes in $\tilde{I}_-(\sigma)$. Let X_i, Y_i, W_i and V_i be defined as in the proof of Theorem 2, and let W_i' and V_i' be *i.i.d.* copies of W_i and V_i, respectively. Then, a node i being mislabeled is equivalent to

$$\sum_{i=1}^{\binom{(\delta/2)n}{m-1}} W_i + \sum_{i=1}^{\binom{(n/2)}{m-1}-\binom{(\delta/2)n}{m-1}} V_i \geq \sum_{i=1}^{\binom{(1-\delta)(n/2)}{m-1}} W_i' + \sum_{i=1}^{\binom{(n/2)}{m-1}-\binom{(1-\delta)(n/2)}{m-1}} V_i'.$$

We bound the probability that node i is mislabeled and then apply the union bound. Let $r = 1/\{\delta\sqrt{\log(1/\delta)}\}$. Then, we have

$$\begin{aligned} p_i &= \mathbb{P}(\text{node } i \text{ is mislabeled}) \\ &= \mathbb{P}\left\{ \sum_{i=1}^{\binom{(\delta/2)n}{m-1}} W_i + \sum_{i=1}^{\binom{(n/2)}{m-1}-\binom{(\delta/2)n}{m-1}} V_i \geq \sum_{i=1}^{\binom{(1-\delta)(n/2)}{m-1}} W_i' + \sum_{i=1}^{\binom{(n/2)}{m-1}-\binom{(1-\delta)(n/2)}{m-1}} V_i' \right\} \\ &= \mathbb{P}\left\{ \sum_{i=1}^{\binom{(n/2)}{m-1}-\binom{(\delta/2)n}{m-1}} (V_i - W_i') + \sum_{i=1}^{\binom{(\delta/2)n}{m-1}} W_i \geq \sum_{i=1}^{\binom{(n/2)}{m-1}-\binom{(1-\delta)(n/2)}{m-1}} V_i' \right. \\ &\qquad \left. - \sum_{i=1}^{\binom{(n/2)}{m-1}-\binom{(\delta n/2)}{m-1}-\binom{(1-\delta)(n/2)}{m-1}} W_i' \right\} \end{aligned}$$

$$\leq \mathbb{P}\left\{\sum_{i=1}^{\binom{(n/2)}{m-1}-\binom{(\delta/2)n}{m-1}} (V_i - W_i') \geq -r\delta\log n\right\} +$$
$$\mathbb{P}\left\{\sum_{i=1}^{\binom{(\delta/2)n}{m-1}} W_i + \sum_{i=1}^{\binom{(n/2)}{m-1}-\binom{(\delta n/2)}{m-1}-\binom{(1-\delta)(n/2)}{m-1}} W_i' - \sum_{i=1}^{\binom{(n/2)}{m-1}-\binom{(1-\delta)(n/2)}{m-1}} V_i' \geq r\delta\log n\right\}$$
$$= (I) + (II).$$

Next, we show $(II) = O\left(n^{-2}\right)$ and $(I) = O\left(n^{-t/I_m(p,q)}\right)$. It is easy to verify that

$$(II) \leq \mathbb{P}\left(\sum_{i=1}^{\binom{(\delta/2)n}{m-1}} W_i \geq \frac{r\delta}{3}\log n\right) + \mathbb{P}\left(\sum_{i=1}^{\binom{(n/2)}{m-1}-\binom{(\delta n/2)}{m-1}-\binom{(1-\delta)(n/2)}{m-1}} W_i' \geq \frac{r\delta}{3}\log n\right)$$
$$+\mathbb{P}\left(\sum_{i=1}^{\binom{(n/2)}{m-1}-\binom{(1-\delta)(n/2)}{m-1}} -V_i' \geq \frac{r\delta}{3}\log n\right).$$

Because $p > q > 0$, it follows that $1-q > 1-p$, and then

$$\begin{aligned} W_i &= \log\left(\frac{p}{q}\right)\mathbb{1}[X_i = 1] + \log\left(\frac{1-p}{1-q}\right)\mathbb{1}[X_i = 0] \\ &\leq \log\left(\frac{p}{q}\right)\mathbb{1}[X_i = 1]. \end{aligned}$$

Then, by the multiplicative Chernoff bound, we have

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^{\binom{(\delta/2)n}{m-1}} W_i \geq \frac{r\delta}{3}\log n\right) &\leq \mathbb{P}\left\{\sum_{i=1}^{\binom{(\delta/2)n}{m-1}} \mathbb{1}[X_i = 1] \geq \frac{r\delta\log n}{3\log(p/q)}\right\} \\ &\leq \left\{\frac{(r/\delta^{m-2})2^{m-1}(m-1)!}{e\cdot 3pt\log(p/q)}\right\}^{-r\delta\log n/\{3\log(p/q)\}} \\ &= e^{-[\log n/\{3\log(p/q)\sqrt{\log(1/\delta)}\}][\log(1/\delta)+(m-2)\log(1/\delta)\{1+o(1)\}]} \\ &= e^{-[(m-1)\sqrt{\log(1/\delta)}/\{3\log(p/q)\}]\log n\ \{1+o(1)\}} \\ &= O\left(n^{-2}\right). \end{aligned}$$

Similarly, we have

$$\mathbb{P}\left(\sum_{i=1}^{\binom{(n/2)}{m-1}-\binom{(\delta n/2)}{m-1}-\binom{(1-\delta)(n/2)}{m-1}} W_i' \geq \frac{r\delta}{3}\log n\right) = O\left(n^{-2}\right).$$

Note that

$$\begin{aligned}-V_i' &= \log\left(\frac{1-p}{1-q}\right)\mathbb{1}[A_i=0] - \log\left(\frac{p}{q}\right)\mathbb{1}[A_i=1]\\ &\le \log\left(\frac{1-p}{1-q}\right)\mathbb{1}[A_i=0].\end{aligned}$$

Hence, by the multiplicative Chernoff bound, it follows that

$$\begin{aligned}&\mathbb{P}\left(\sum_{i=1}^{\binom{(n/2)}{m-1}-\binom{(1-\delta)(n/2)}{m-1}}(-V_i') \ge \frac{r\delta}{3}\log n\right)\\ &\le \mathbb{P}\left(\sum_{i=1}^{\binom{(n/2)}{m-1}-\binom{(1-\delta)(n/2)}{m-1}}\mathbb{1}[A_i=0] \ge \frac{r\delta\log n}{3\log\{(1-q)/(1-p)\}}\right)\\ &\le \left[\frac{(r/\delta^{m-2})2^{m-1}(m-1)!}{e\cdot 3(1-p)t\log\{(1-q)/(1-p)\}}\right]^{-r\delta\log n/[3\log\{(1-q)/(1-p)\}]}\\ &= e^{-\{(1-\delta\log n)/[3\log\{(1-q)/(1-p)\}]\}[(m-1)\log(1/\delta)\{1+o(1)\}]}\\ &= e^{-\{(m-1)\sqrt{\log(1/\delta)}\log n/[3\log\{(1-q)/(1-p)\}]\}\{1+o(1)\}}\\ &= O\left(n^{-2}\right).\end{aligned}$$

Thus, we conclude that $(II) = O(n^{-2})$.

Next, we bound (I). Note that $\binom{n/2}{m-1} - \binom{(\delta/2)n}{m-1} = [n^{m-1}/\{2^{m-1}(m-1)!\}]\{1+o(1)\}$. By Markov's inequality, we have

$$\begin{aligned}(I) &= \mathbb{P}\left[e^{1/2\sum_{i=1}^{\binom{(n/2)}{m-1}-\binom{(\delta/2)n}{m-1}}(V_i-W_i')} \ge e^{-r\delta\log n/2}\right]\\ &\le e^{r\delta(\log n/2)}[\mathbb{E}\{e^{(1/2)V_1}e^{-(1/2)W_1'}\}]^{n^{m-1}/\{2^{m-1}(m-1)!\}}\\ &= e^{r\delta(\log n/2)}\{e^{-(1/2)\log(p/q)}\alpha p\\ &\quad + e^{-(1/2)\log\{(1-p)/(1-q)\}}\alpha(1-p)+(1-\alpha)\}^{n^{m-1}/\{2^{m-1}(m-1)!\}}\\ &\quad \times[e^{(1/2)\log(p/q)}\alpha q + e^{(1/2)\log\{(1-p)/(1-q)\}}\alpha(1-q)+(1-\alpha)]^{n^{m-1}/\{2^{m-1}(m-1)!\}}.\end{aligned}$$

Taking the logarithm of both sides yields

$$\begin{aligned}\log(I) &\le \frac{1}{2}r\delta\log n + \frac{n^{m-1}\alpha}{2^{m-1}(m-1)!}\{2\sqrt{pq}+2\sqrt{(1-p)(1-q)}-2\}\\ &= \frac{1}{2}\frac{\log n}{\sqrt{\log(1/\delta)}} - \frac{t\log n}{2^{m-1}(m-1)!}\{(\sqrt{p}-\sqrt{q})^2+(\sqrt{1-p}-\sqrt{1-q})^2\}.\end{aligned}$$

Hence,

$$(I) \leq n^{-t[\{(\sqrt{p}-\sqrt{q})^2+(\sqrt{1-p}-\sqrt{1-q})^2\}/\{2^{m-1}(m-1)!\}]\{1+o(1)\}} = n^{-\{t/I_m(p,q)\}\{1+o(1)\}}.$$

Because $t > I_m(p,q)$, by assumption, we have $(I) \leq n^{-(1+\epsilon)}$, for some small constant $\epsilon > 0$, and hence

$$p_i \leq (I) + (II) \leq n^{-(1+\epsilon)}.$$

By the union bound, the probability that a mislabeled node exists is bounded by $n^{-\epsilon} = o(1)$. The proof is complete.

Supplementary Material

The supplement provides more detail on the proof of theorem 4, and demonstrates the results of the simulation study for the SDP algorithm and the refined spectral algorithm proposed in this paper.

Acknowledgments

The authors are grateful to the co-editor, associate editor, and reviewers for their valuable comments and suggestions.

References

Abbe, E. (2018). Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research* **18**, 1–86.

Abbe, E., Bandeira, A. S., Bracher, A. and Singer, A. (2014). Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery. *IEEE Transactions on Network Science and Engineering* **1**, 10–22.

Abbe, E, Bandeira, A. S. and Hall, G. (2016). Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory* **62**, 471–487.

Ahn, K., Lee, K. and Suh, C. (2018). Hypergraph spectral clustering in the weighted stochastic block model. *IEEE Journal of Selected Topics in Signal Processing* **12**, 959–974.

Ahn, K., Lee, K. and Suh, C. (2019). Community recovery in hypergraphs. *IEEE Transactions on Information Theory* **12**, 6561–6578.

Bi, X., Tang, X., Yuan, Y., Zhang, Y. and Qu, A. (2021). Tensors in statistics. *Annual Review of Statistics and Its Application* **8**, 345–368.

Chen, J. and Yuan, B. (2006). Detecting functional modules in the yeast proteinprotein interaction network. *Bioinformatics* **22**, 2283–2290.

Chien, I., Lin, C. and Wang, I.(2018). Community detection in hypergraphs: Optimal statistical limit and efficient algorithms. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* **84**, 871–879.

Dhara, S., Gaudio, J., Mossel, E. and Sandon, C. (2021). Spectral recovery of binary cencored block models. Web: `https://arxiv.org/pdf/2107.06338.pdf`.

Estrada, E. and Rodriguez-velasquez, J. (2005). Complex networks as hypergraphs. Web: `https://arxiv.org/ftp/physics/papers/0505/0505137.pdf`.

Gao, C., Ma, Z., Zhang, A. Y. and Zhou, H. H. (2018). Community detection in degree-corrected block models. *The Annals of Statistics* **46**, 2153–2185

Ghoshal, G., Zlatic, V., Caldarelli, G. and Newman, M. E. J. (2009). Random hypergraphs and their applications. *Physical Review E* **79**.

Ghoshdastidar, D. and Dukkipati, A. (2014). Consistency of spectral partitioning of uniform hypergraphs under planted partition model. *Advances in Neural Information Processing Systems (NIPS)* **2014**, 397–405.

Ghoshdastidar, D. and Dukkipati A. (2017). Consistency of spectral hypergraph partitioning under planted partition model. *The Annals of Statistics* **45**, 289–315.

Gile, K and Handcock M. (2016). Analysis of networks with missing data with application to the National Longitudinal Study of Adolescent Health, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **66**, 501–519.

Goldenberg, A., Zheng, A. X. S., Fienberg, E. and Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning* **2**, 129–233.

Hajek, B., Wu, Y. and Xu, J. (2016). Achieving exact cluster recovery threhold via semidefinite programming. *IEEE Transaction on Information Theory* **62**, 2788–2796.

Hajek, B., Wu, Y. and Xu, J. (2018). Recovering a hidden community beyond the Kesten Stigum threshold in $O(|E| \log |V|)$ time. *Journal of Applied Probability* **55**, 325–352.

Hu, F. and Shi, J.(2015). Neighborhood hypergraph based classification algorithm for incomplete information system. *Mathematical Problems in Engineering*, Article ID 735014.

Huisman, M. (2009). Imputation of missing network data: Some simple procedures. *Journal of Social Structure* **10**, 1–29.

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

Ke, Z., Shi, F. and Xia, D. (2020). Community detection for hypergraph networks via regularized tensor power iteration. Web: `https://arxiv.org/pdf/1909.06503.pdf`.

Kim, C., Bandeira, A. and Goemans, M. (2018). Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach.

Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics* **43**, 215–237.

Liu, H., Jan, L. and Yan, S.(2015). Dense subgraph partition of positive hypergraphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**, 541–554.

Liu, M., Gao, Y., Yap, P. and Shen, D.(2018). Multi-Hypergraph learning for incomplete multimodality data. *IEEE Journal of Biomedical and Health Informatics* **22**, 1197–1208.

Liu, M., Zhang, J., Yap, P. and Shen, D. (2017). View-aligned hypergraph learning for Alzheimer's disease diagnosis with incomplete multi-modality data. *Medical Image Analysis* **36**, 123–134.

Luo, Y. and Zhang, A. (2020). Open problem: Average-case hardness of hypergraphic planted clique detection. In *Proceedings of 33rd Conference on Learning Theory* **PMLR 125**, 3852–3856.

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

Mossel, E., Neeman, J. and Sly, A. (2017). A proof of the block model threshold conjecture. *Combinatorica* **38**, 665–708.

Newman, M. (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E* **64**, 016–131.

Ouvrard, X., Goff, J. and Marchand-Maillet, S. (2017). Networks of collaborations: Hypergraphmodeling and visualisation. Web: `https://arxiv.org/pdf/1707.00115.pdf`.

Ramasco, J., Dorogovtsev, S. N. and Pastor-Satorras, R. (2004). Self-organization of

collaboration networks, *Physical Review E* **70**, 036–106.

Shi, J. and Malik, J. (1997). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 888–905.

Smith, J., Moody, J. and Morgan, J. (2018). Network sampling coverage II: The effect of non-random missing data on network measurement. *Social Networks* **48**, 78–99.

Weng, H. and Feng, Y.(2021). Community detection with nodal information: Likelihood and its variational approximation. *Stat*, e428.

Yuan, M. and Shang, Z. (2021a). Informatin limits for detection a subhypergraph. *Stat*, e407.

Yuan, M. and Shang, Z. (2021b). Sharp detection boundaries on testing dense subhypergraph. *Bernoulli* **28**, 2459–2491.

Yuan,Y. and Qu, A.(2021). Community detection with dependent connectivity. *The Annals of Statistics* **49**, 2378–2428.

Zhao, Y., Levina, E. and Zhu, J.(2011). Community extraction for social networks. *Proceedings of the National Academy of Sciences* **108**, 7321–7326.

Zhen, Y. and Wang, J.(2021). Community detection in general hypergraph via graph embedding. Web: `https://arxiv.org/pdf/2103.15035.pdf`.

Mingao Yuan

Department of Statistics, North Dakota State University, Fargo, ND 58108-6050, USA.

E-mail: mingao.yuan@ndsu.edu

Bin Zhao

Department of Statistics, North Dakota State University, Fargo, ND 58108-6050, USA.

E-mail: bin.zhao@ndsu.edu

Xiaofeng Zhao

School of Mathematics and Statistics, North China University of Water Resources and Electric Power, Zhengzhou, Henan, China.

E-mail: zxfstat@ncwu.edu.cn

(Received November 2021; accepted July 2022)