Statistica Sinica Preprint No: SS-2021-0392	
Title	Community Detection in Censored Hypergraph
Manuscript ID	SS-2021-0392
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0392
Complete List of Authors	Mingao Yuan, Bin Zhao and Xiaofeng Zhao
Corresponding Authors	Mingao Yuan
E-mails	mingao.yuan@ndsu.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

COMMUNITY DETECTION IN CENSORED HYPERGRAPH

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

[1]*North Dakota State University,USA*

[2] *North China University of Water Resources and Electric Power,China*

Abstract: Community detection refers to the problem of clustering the nodes of a network (either graph or hypergrah) into groups. Various algorithms are available for community detection and all these methods apply to uncensored networks. In practice, a network may has censored (or missing) values and it is shown that censored values have non-negligible effect on the structural properties of a network. In this paper, we study community detection in censored m-uniform hypergraph from information-theoretic point of view. We derive the information-theoretic threshold for exact recovery of the community structure. Besides, 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 study whether a single spectral algorithm without refinement achieves the threshold. To this end, we also explore the semi-definite relaxation algorithm and analyze its performance.

Key words and phrases: Community detection, Information-theoretic threshold, Censored hypergraph, Exact recovery.

1. Introduction

Many complex data sets can be modelled as a network of items (nodes). One of the most popular topic 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 network into groups based on similarity. Community detection is widely used in analysis of social networks (Goldenberg et al. (15); Zhao et al. (41)), protein-to-protein interaction networks ((8)), image segmentation ((36)) and so on. Existing literature in community detection can be roughly classified into two categories: (1) derive information-theoretic threshold for recovering the community structure ((2; 30; 31; 7; 9; 18; 39)); (2) devise efficient algorithms to recover the community structure ((11; 12; 28; 29; 22; 38; 4; 5; 17; 14; 37; 42; 25; 20)). See (1; 6) for more references. All these methods apply to uncensored networks.

In practice, network data may have censored or missing values. For example, in social network, non-response of actors can cause missingness of ties ((21; 13)); in MRI network, missingness may be due to the high cost involved with PET scanning ((27)). Missing values have non-negligible effects on the structural properties of a network ((21; 35)). Most existing algorithms for community detection apply to uncensored networks. A natural question is how to recover communities in a censored network. As far as we know, (3) is the first to deal with community detection in censored graph and obtains the information-theoretic threshold for exact recovery of communities. Recently, (9) shows spectral algorithm without 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 hypergraphs, where hyperedge is used to model higher-order interaction among individuals((10; 33; 34; 32; 16; 11)). For example, in folksonomy network, an hyperedge may represent a triple (user, resource, annotation) structure ((16)); in coauthorship networks, the coauthors of a paper form a hyperedge((10; 33; 34; 32)). Hypergraph learning with missing values has recently attracted much attention in literature ((19; 26; 27)). In this paper, we are interested in detecting communities in censored hypergraphs. It is not immediately clear how the sharp threshold obtained by ((9)) changes in the censored hypergraph case. This motivates us to study this problem. Our contributions are summarized as follows. We derive the information-theoretic threshold for exact recovery of community struture in censored hypergraph. Interestingly, the threshold is generally larger than that in the graph case. In this sense, community detection in censored hypergraph is harder than in censored graph. Besides, 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 in the censored graph case (9). 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 hyperedge) contains exactly m distinct nodes. The hypergraph $\mathcal{H}_m$ can be represented as a 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. Besides, $A_{i_1 i_2 \ldots i_m} = A_{j_1 j_2 \ldots j_m}$ if $\{i_1, i_2, \ldots, i_m\} = \{j_1, j_2, \ldots, j_m\}$. In this paper, 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 just the usual graph that has been widely used in community detection problems ((1)). A hypergraph is said to be random if elements of the adjacency tensor are random. Throughout this paper, we focus on hypergraph genereated from the Censored m-uniform Hypergraph Stochastic Block Model (CHSBM) $\mathcal{H}_m(n, p, q, \alpha)$ defined below.

Definition 1.1 (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. Each hyperedge status is revealed independently with probability α. The hyperedge of the resulting hypergraph takes value in $\{1, 0, *\}$, where $*$ means a hyperedge is censored or missing (the hyperedge status is not revealed). This model is denoted as $\mathcal{H}_m(n, p, q, \alpha)$.

Each hyperedge in $\mathcal{H}_m(n, p, q, \alpha)$ with $\alpha < 1$ has three status: 1 (present), 0 (absent) or $*$ (censored or missing). When $\alpha = 1$, the hypergraph is uncensored and $\mathcal{H}_m(n, p, q, 1)$ is just the usual hypergraph stochastic block model ((11; 12; 7; 24; 22; 39)). The Censored Stochastic Block Model $CSBM(p,q,\alpha)$ studied in ((9)) 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 $\alpha = \frac{t\log n}{n^{m-1}}$ for some constant $t>0$. The reason for considering the $\frac{\log n}{n^{m-1}}$ order of α is: this is the smallest order that exact recovery is possible. Please see Theorem 2.1 and Theorem 2.2.

1.2 Summary of 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 or $\hat{\sigma}$ exactly recovers σ or $\hat{\sigma}$ achieves 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 exact recovery is possible. Otherwise, we say exact recovery is impossible.

For $m=2$, (9) establishes the sharp information-theoretic threshold for exact recovery. The authors show that spectral algorithm can exactly recover the true label without refinement step. It is not immediately clear how $m \geq 3$ changes the threshold for exact recovery. More importantly, the spectral method in ((9)) can not be straightforwardly extended to $m \geq 3$, since the spectral analysis of tensor is still not well developed.

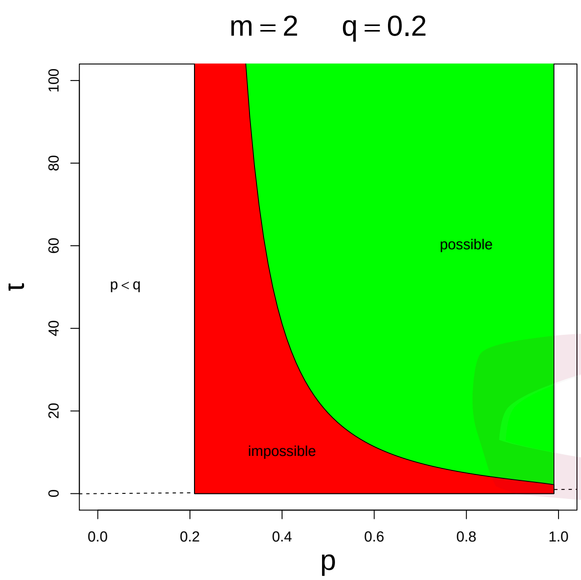
In this paper, we focus on $m \geq 3$ and derive the sharp information-theoretic threshold for exact recovery. Define $I_m(p,q)$ as

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

Theorem 2.1 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 2.2 says that MLE succeeds with probability $1-o(1)$ if $t > I_m(p,q)$. For efficient algorithms, we propose a spectral algorithm plus refinement step that can achieve exact recovery up to the information-theoretic threshold, see Theorem 2.3. Finally, we prove in Theorem 2.4 that the semidefinite relaxation algorithm can exactly recover the true label under mild conditions. The following Table 1 summarizes our main results. For $m = 2, 3$ and $q = 0.2$, Figure 1 displays the region in which exact recovery is impossible (red region) and the region where exact recovery is possible (green region). Interestingly, with fixed q, the red region of $m = 3$ contains that of $m = 2$ as a proper subset. In this sense, exact recovery gets harder as m increases. For fixed q, m, $I_m(p,q)$ decreases as p goes to one, hence exact recovery becomes easier.

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



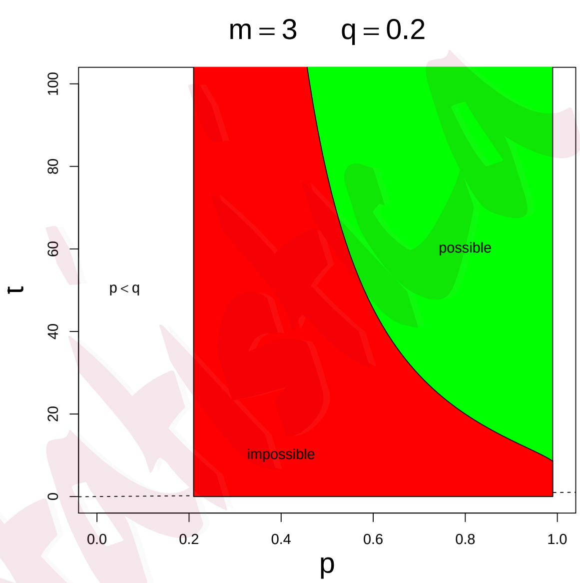


Figure 1: *Regions for exact recovery with* $m = 2, 3$ *and* $q = 0.2$. *Red: exact recovery is impossible. Green: exact recovery is possible.*

Throughout this paper, we adopt the Bachmann-Landau notation $o(1), O(1)$.

For two positive sequences a_n, b_n, we write $a_n \sim b_n$ if $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$. Denote $a_n \asymp b_n$ if $0 < c_1 \leq \frac{a_n}{b_n} \leq c_2 < \infty$ for constants c_1, c_2. Denote $a_n \gg b_n$ or $b_n \ll a_n$ if $\lim_{n\to\infty} \frac{a_n}{b_n} = \infty$. For a square matrix M, $||M||$ denotes the operator norm of M and $M \succeq 0$ means M is symmetric and positive semidefinite. Define $\langle M, N\rangle = \sum_{i,j} M_{ij}N_{ij}$.

2. Main Results

In this section, we present information theoretic threshold for exact recovery on the censored hypergraph stochastic block model. Firstly, we use the maximum likelihood method to show that exact recovery is impossible if $t < I_m(p, q)$. Then we prove that 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. Since the time complexity of MLE is not polynomial in n, we propose a polynomial-time algorithm that achieves 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 impossibility of exact recovery.

Theorem 2.1. For each fixed integer $m \geq 2$, 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 2.1 states that any estimator can not 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 ((9)). Our result can be considered as a nontrivial extension of Theorem 2.1 in ((9)). Interestingly, with fixed p, q, the region $t < I_2(p,q)$ is smaller than $t < I_m(p,q)$ for $m \geq 3$. Similar phenomenon exists in exact recovery of community in uncensored hypergraph stochastc block model ((24)). However, this phenomenon significantly differs from that in hypothesis testing for communities. For example, (40) 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 where any test is asymptotically powerless for $m = 2$ is larger than $m \geq 3$.

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

Theorem 2.2. For each fixed integer $m \geq 2$, if $t > I_m(p,q)$ with $I_m(p,q)$ defined in (2.1), then the maximum likelihood estimator exactly recovers the true label with probability $1 - o(1)$.

By Theorem 2.2, if $t > I_m(p, q)$, the true label can be exactly recovered by the maximum likelihood estimator estimator. Combining Theorem 2.1 and Theorem 2.2, 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 $m = 2$. In this sense, exact recovery gets harder as m increases.

2.2 Efficient algorithm for exact recovery

Since the time complexity of MLE is not polynomial in n, we propose an efficient algorithm to reconstruct the 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 $\frac{\log\log n}{\log n}$. Let S_2 be the compliment 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 randomly divided into two independent hypergraphs $\tilde{A}$ and $\bar{A}$.

In the second step, we apply the weak recovery algorithm HSC in ((4)) to $\tilde{A}$. The HSC algorithm proceeds by converting hypergraph $\tilde{A}$ to a $n \times n$ similarity matrix B by $B_{ij} = \sum_{1 \leq i_3 < i_4 < \cdots < i_m \leq n} \tilde{A}_{ij i_3 i_4 \ldots i_m}$ and then applying geometric 2-clustering to the top 2 eigenvectors of B to output the communities $\tilde{I}_+(\sigma)$ and $\tilde{I}_-(\sigma)$. The sampling probability $\frac{\log\log n}{\log n}$ in the first step makes sure the hyeredge probability of $\tilde{A}$ has order $\frac{\log\log n}{\log n}\alpha = \frac{t \log\log n}{n^{m-1}}$ (Here the $\log\log n$ factor can be replaced by any a_n with $a_n \to \infty$). According to Theorem 1 of (4), $n - o(n)$ of the nodes are correctly labelled by HSC 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

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

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 2.3. For each fixed integer $m \geq 2$, 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 refinement in Step 3 have time complexity at most $O(n^m)$. In Step 2, the weak recovery algorithm HSC in Ahn et all. (2018) has time complexity $O(n^m)$ (see comments below Remark 1 of (4)). Hence, Theorem 2.3 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

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

2: **Step 1: random splitting.**

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

3: **Step 2: spectral algorithm.**

Apply the weak recovery algorithm HSC in (4) to $\tilde{A}$ and denote the community output as $\tilde{I}_+(\sigma)$, $\tilde{I}_-(\sigma)$.

4: **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.

5: **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 Semidefinite relaxation algorithm

In subsection 2.2, we show that spectral algorithm with refinement step can achieve exact recovery. It is also interesting to study whether a single spectral algorithm without refinement achieves the threshold. In graph case ($m = 2$), the answer is confirmative and the semidefinite relaxation algorithm and spectral algorithm are shown to succeed without refinement step ((17; 9)). In hyergraph case ($m \geq 3$), either censored or uncensored, it is still an open problem. In this subsection, we study the semidefinite 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 the semidefinite relaxation algorithm applied to the weighted graph can achieve exact recovery.

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

$$\widetilde{A}_{i_1 i_2 \ldots i_m} = \mathbb{1}[A_{i_1 i_2 \ldots i_m} = 1],$$

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

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

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

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

Define the semidefinite program problem (SDP) as

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

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

Theorem 2.4. 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 ((17)). However, for $m \ge 3$, it is unclear whether $\widehat{Y}$ succeeds or not in the range $I_m(p,q) < t < J_m(p,q)$. Similar gap exists in uncencored hypergraph case ((24)). The proof of Theorem 2.4 is in the Appendix.

3. Proof of Theorem 2.1

In this section, we provide proof of Theorem 2.1.

Proof of Theorem 2.1 : 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 (9), if there are labels η_t $(1 \leq t \leq k_n)$ with $k_n \to \infty$ such that $l(\eta_1) = l(\eta_2) = \cdots = l(\eta_{k_n}) = l(\sigma)$, then the maximum likelihood estimator (MLE) fails to exactly recover the true label with probability $1 - o(1)$. Our proof proceeds by constructing labels η_t $(1 \leq t \leq k_n)$ with $k_n \to \infty$ under the condition $t < I_m(p,q)$.

Firstly, we write down the 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

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

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

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

The maximum likelihood estimator(MLE) is obtained by maximizing L with respect to σ. The first product factor of L does not involve σ. Hence we only need to maximize the second product factor of L to get MLE. Denote

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

The log-likelihood function is equal to

$$\log L = R_n + l(\sigma), \tag{3.3}$$

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 $t < I_m(p,q)$. Since R_n is independent of σ. We only need to 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 $I_-(\sigma)$. 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 label of i_0 and j_0. Let $\tilde{\sigma}$ be labels obtained from σ by flipping the labels of i_0, j_0. We shall verify that $l(\sigma) = l(\tilde{\sigma})$. To prove this, let $T_1 = \log\left(\frac{p}{q}\right)$, $T_2 = \log\left(\frac{1-p}{1-q}\right)$, 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

$$\begin{aligned} l(\sigma) &= T_1 \sum_{\substack{i_1i_2\ldots i_m \in I_+(\sigma)\\ i_1i_2\ldots i_m \neq i_0}} \mathbb{1}[A_{i_1i_2\ldots i_m}=1] + T_1 \sum_{\substack{i_2\ldots i_m \in I_+(\sigma)\\ i_2\ldots i_m \neq i_0}} \mathbb{1}[A_{i_0i_2\ldots i_m}=1] \\ &+ T_2 \sum_{\substack{i_1i_2\ldots i_m \in I_+(\sigma)\\ i_1i_2\ldots i_m \neq i_0}} \mathbb{1}[A_{i_1i_2\ldots i_m}=0] + T_2 \sum_{\substack{i_2\ldots i_m \in I_+(\sigma)\\ i_2\ldots i_m \neq i_0}} \mathbb{1}[A_{i_0i_2\ldots i_m}=0] \\ &+ T_1 \sum_{\substack{i_1i_2\ldots i_m \in I_-(\sigma)\\ i_1i_2\ldots i_m \neq j_0}} \mathbb{1}[A_{i_1i_2\ldots i_m}=1] + T_1 \sum_{\substack{i_2\ldots i_m \in I_-(\sigma)\\ i_2\ldots i_m \neq j_0}} \mathbb{1}[A_{j_0i_2\ldots i_m}=1] \\ &+ T_2 \sum_{\substack{i_1i_2\ldots i_m \in I_-(\sigma)\\ i_1i_2\ldots i_m \neq j_0}} \mathbb{1}[A_{i_1i_2\ldots i_m}=0] + T_2 \sum_{\substack{i_2\ldots i_m \in I_-(\sigma)\\ i_2\ldots i_m \neq j_0}} \mathbb{1}[A_{j_0i_2\ldots i_m}=0], \end{aligned}$$

and

$$\begin{aligned} l(\tilde{\sigma}) &= T_1 \sum_{\substack{i_1i_2\ldots i_m \in I_+(\sigma)\\ i_1i_2\ldots i_m \neq j_0}} \mathbb{1}[A_{i_1i_2\ldots i_m}=1] + T_1 \sum_{\substack{i_2\ldots i_m \in I_+(\sigma)\\ i_2\ldots i_m \neq j_0}} \mathbb{1}[A_{j_0i_2\ldots i_m}=1] \\ &+ T_2 \sum_{\substack{i_1i_2\ldots i_m \in I_+(\sigma)\\ i_1i_2\ldots i_m \neq j_0}} \mathbb{1}[A_{i_1i_2\ldots i_m}=0] + T_2 \sum_{\substack{i_2\ldots i_m \in I_+(\sigma)\\ i_2\ldots i_m \neq j_0}} \mathbb{1}[A_{j_0i_2\ldots i_m}=0] \\ &+ T_1 \sum_{\substack{i_1i_2\ldots i_m \in I_-(\sigma)\\ i_1i_2\ldots i_m \neq i_0}} \mathbb{1}[A_{i_1i_2\ldots i_m}=1] + T_1 \sum_{\substack{i_2\ldots i_m \in I_-(\sigma)\\ i_2\ldots i_m \neq i_0}} \mathbb{1}[A_{i_0i_2\ldots i_m}=1] \\ &+ T_2 \sum_{\substack{i_1i_2\ldots i_m \in I_-(\sigma)\\ i_1i_2\ldots i_m \neq i_0}} \mathbb{1}[A_{i_1i_2\ldots i_m}=0] + T_2 \sum_{\substack{i_2\ldots i_m \in I_-(\sigma)\\ i_2\ldots i_m \neq i_0}} \mathbb{1}[A_{i_0i_2\ldots i_m}=0] \end{aligned}$$

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

Next we will show there are k_n ($k_n \to \infty$) such pairs. More specifically, we will show that there exists $i_1, i_2, \ldots, i_k \in I_+(\sigma)$ and $j_1, j_2, \ldots, j_k \in I_-(\sigma)$ with $k \gg 1$ such that the likelihood function remains unchanged if we flip the label of a pair (i_t, j_t) $t = 1, 2, \ldots, k$. 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 $1 \leq t \leq k \to \infty$.

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

$$S_0 = \left\{ i \in S | any\ i_2, \ldots, i_t \in S, i_{t+1}, \ldots, i_m \in S^c,\ s.t.\ A_{ii_2 \ldots i_t i_{t+1} \ldots i_m} = *,\ t \geq 2 \right\}.$$

For each node $i \in S_0$, hyperedge $A_{ii_2 \ldots i_m}$ is possibly revealed if and only if $\{i_2, \ldots, i_m\} \subset I_+(\sigma) - S$ or $\{i_2, \ldots, i_m\} \subset I_-(\sigma) - S$.

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

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

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 Markov inequality we have

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

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

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

absent hyperedges in $I_+(\sigma)$ and $I_-(\sigma)$ respectively is,

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

If $t<\frac{2^{m-1}(m-1)!}{(\sqrt{p}-\sqrt{q})^2+(\sqrt{1-p}-\sqrt{1-q})^2}$, then $p_0\gg\frac{n^{1-\epsilon}}{n}$ 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 $I_-(\sigma)$ respectively. Define two random variables

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

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

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

Since $p_0 \gg \frac{n^{1-\epsilon}}{n}$ for some $\epsilon > 0$ and $|S_0 \cap S_+| \geq \frac{2n}{\log^2 n}(1 - o(1))$. Then $X \geq |S_0 \cap S_+|p_0 \to +\infty$ with probability $1 - o(1)$. Similarly $Y \geq |S_0 \cap S_+|p_0 \to +\infty$ with probability $1 - o(1)$. As a result, we have pairs (i_t, j_t) $(1 \leq t \leq k \to \infty)$. For each t, the likelihood is constant by flipping the labels of i_t and j_t. The proof is complete by Proposition 4.1 in (9).

4. Proof of Theorem 2.2

Proof of Theorem 2.2 : Let σ be the maximum likelihood estimator(MLE). Recall the log-likelihood function in (3.3). The MLE fails to exactly recover the true label if there exists a label η such that $l(\eta) \geq l(\sigma)$ with probability δ for some constant $\delta > 0$. Our proof proceeds by showing that the probability MLE fails is $o(1)$.

The maximum likelihood estimator(MLE) is obtained by maximizing $\log L$ in (3.3) with respect to σ. The first term of $\log L$ does not involve σ. Hence we only need to maximize the second term of $\log L$ to get 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 MLE fails is $o(1)$.

Let k be an even number and $1 \leq k \leq \frac{n}{2}$. Define the Hamming distance between two labels σ, 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 log-likelihood difference at η and σ is

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

We will show

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

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

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

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

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

It is easy to verify that there are $n_k = 2\left[\binom{\frac{n}{2}}{m} - \binom{\frac{k}{2}}{m} - \binom{\frac{n-k}{2}}{m}\right]$ hyperedges $\{i_1, \ldots, i_m\}$ such that $\{i_1 \ldots i_m\} \subset \mathbb{1}_+(\eta) \textit{ or } \mathbb{1}_-(\eta), \{i_1 \ldots i_m\} \not\subset \mathbb{1}_+(\sigma), \mathbb{1}_-(\sigma)$.

For convenience, define random variables X, Y as

$$\mathbb{P}(X = 1) = \alpha p, \quad \mathbb{P}(X = 0) = \alpha(1-p), \quad \mathbb{P}(X = -1) = 1 - \alpha.$$

$$\mathbb{P}(Y = 1) = \alpha q, \quad \mathbb{P}(Y = 0) = \alpha(1-q), \quad \mathbb{P}(Y = -1) = 1 - \alpha.$$

Let X_i, Y_i be $i.i.d$ copies of X, Y respectively and

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

For any $r > 0$, by 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 the explicit expression of expectations $\mathbb{E}\left(e^{-rW_1}\right)$ and $\mathbb{E}\left(e^{rV_1}\right)$.

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

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

Taking $r = \frac{1}{2}$ 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,

$$\begin{aligned}
\log \mathbb{P}\left(l(\eta) - l(\sigma) \geq 0\right) &\leq n_k \log \mathbb{E}[e^{-rW_1}] + n_k \log \mathbb{E}[e^{rV_1}] \\
&\leq n_k[2\alpha(\sqrt{pq} + \sqrt{(1-p)(1-q)} - 1)] \\
&= n_k\alpha\left[(-1)\left\{(\sqrt{p} - \sqrt{q})^2 + (\sqrt{1-p} - \sqrt{1-q})^2\right\}\right] \\
&= -n_k\alpha\left[(\sqrt{p} - \sqrt{q})^2 + (\sqrt{1-p} - \sqrt{1-q})^2\right]. \quad (4.4)
\end{aligned}$$

For $k \geq \frac{n}{\log\log n}$, it is easy to check $n_k \geq \frac{1}{2^{m-1}} \frac{n}{\log\log n}\binom{n-1}{m-1}$. Hence by (4.4), we get

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

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

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

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

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

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

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

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

5. Proof of Theorem 2.3

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

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

Then $\tilde{A}$ has the same community structure as the original hypergraph A and in $\tilde{A}$, the order of hyperedge probability is $\frac{\log\log n}{n^{m-1}}$. With probability $1 - o(1)$, the weak recovery algorithm in (4) will recover the true labels of $(1-\delta)n$ nodes of $\tilde{A}$ with $\delta = o(1)$. Denote the communities as $\tilde{I}_+(\sigma)$, $\tilde{I}_-(\sigma)$. Hence, with probability $1 - o(1)$, there are $\frac{\delta}{2}n$ nodes in $\tilde{I}_+(\sigma)$ and $\tilde{I}_-(\sigma)$ that are mislabelled. By the refinement step, a node i among the correctly labelled $\frac{1-\delta}{2}n$ nodes in $\tilde{I}_+(\sigma)$ is mislabelled if

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

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

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

Similar result holds for nodes in $\tilde{I}_-(\sigma)$. Let X_i, Y_i, W_i, V_i be defined as in the proof of Theorem 2.2 and W_i', V_i' be $i.i.d.$ copies of W_i, V_i. Then a node i is mislabelled is equivalent to

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

We are going to bound the probability that node i is mislabelled and then apply the union bound. Let $r = \frac{1}{\delta\sqrt{\log(\frac{1}{\delta})}}$. Then we have

$$\begin{aligned}
p_i &= \mathbb{P}(\text{node } i \text{ is mislabelled}) \\
&= \mathbb{P}\left[\sum_{i=1}^{\binom{\frac{\delta}{2}n}{m-1}} W_i + \sum_{i=1}^{\binom{\frac{n}{2}}{m-1}-\binom{\frac{\delta}{2}n}{m-1}} V_i \geq \sum_{i=1}^{\binom{(1-\delta)\frac{n}{2}}{m-1}} W_i' + \sum_{i=1}^{\binom{\frac{n}{2}}{m-1}-\binom{(1-\delta)\frac{n}{2}}{m-1}} V_i'\right] \\
&= \mathbb{P}\left[\sum_{i=1}^{\binom{\frac{n}{2}}{m-1}-\binom{\frac{\delta}{2}n}{m-1}} (V_i - W_i') + \sum_{i=1}^{\binom{\frac{\delta}{2}n}{m-1}} W_i \geq \sum_{i=1}^{\binom{\frac{n}{2}}{m-1}-\binom{(1-\delta)\frac{n}{2}}{m-1}} V_i' - \sum_{i=1}^{\binom{\frac{n}{2}}{m-1}-\binom{\frac{\delta n}{2}}{m-1}-\binom{(1-\delta)\frac{n}{2}}{m-1}} W_i'\right] \\
&\leq \mathbb{P}\left[\sum_{i=1}^{\binom{\frac{n}{2}}{m-1}-\binom{\frac{\delta}{2}n}{m-1}} (V_i - W_i') \geq -r\delta\log n\right] + \\
&\quad \mathbb{P}\left[\sum_{i=1}^{\binom{\frac{\delta}{2}n}{m-1}} W_i + \sum_{i=1}^{\binom{\frac{n}{2}}{m-1}-\binom{\frac{\delta n}{2}}{m-1}-\binom{(1-\delta)\frac{n}{2}}{m-1}} W_i' - \sum_{i=1}^{\binom{\frac{n}{2}}{m-1}-\binom{(1-\delta)\frac{n}{2}}{m-1}} V_i' \geq r\delta\log n\right] \\
&= (I) + (II).
\end{aligned}$$

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

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

Since $p > q > 0$, 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, one has

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

Similarly, we get

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

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

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

Then we conclude that $(II) = O\left(n^{-2}\right)$.

Next we bound (I). Note that $\binom{\frac{n}{2}}{m-1} - \binom{\frac{\delta}{2}n}{m-1} = \frac{n^{m-1}}{2^{m-1}(m-1)!}(1+o(1))$.

By Markov's inequality, one has

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

Taking logrithm on both side yields

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

Hence,

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

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

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

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

Acknowledgement

The authors are grateful to Co-Editor, Associate Editor and reviewers for valuable comments that significantly improve the manuscript.

References

[1] Abbe, E., Community Detection and Stochastic Block Models: Recent Developments. *Journal of Machine Learning Research* **2018**, *18*, 1-86.

[2] Abbe, E, Bandeira, A.S. and Hall, G. (2016). Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1): 471-487.

[3] 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(1), 10-22.

[4] 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(5), 2018.

[5] Ahn, K., Lee, K. and Suh, C. (2019). Community Recovery in Hypergraphs. *IEEE Transactions on Information Theory*, 12(5), 6561-6578.

[6] 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.

[7] 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.

[8] Chen, J. and Yuan, B. (2006). Detecting functional modules in the yeast proteinprotein interaction network. *Bioinformatics*, **22(18)**, 2283-2290.

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

[10] Estrada, E. and Rodriguez-velasquez, J. (2005). Complex networks as hypergraphs. `https:`

//arxiv.org/ftp/physics/papers/0505/0505137.pdf

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

[12] Ghoshdastidar, D. and Dukkipati A. Consistency of spectral hypergraph partitioning under planted partition model. *The Annals of Statistics* **2017**, **45(1)**, 289-315.

[13] 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.

[14] Gao, C., Ma, Z., Zhang, A.Y. and Zhou, H. H. (2018). Community detection in degree-corrected block models. *The Annals of Statistics.* (46)(5), 2153-2185

[15] 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*, **2**, 129-233.

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

[17] Hajek, B., Wu, Y. and Xu, J. (2016). Achieving exact cluster recovery threhold via semidefinite programming. *IEEE Transaction on Information Theory*, 62(5): 2788-2796.

[18] 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, 2: 325-352.

[19] Hu,F. and Shi, J.(2015). Neighborhood Hypergraph Based Classification Algorithm for Incomplete Information System. *Mathematical Problems in Engineering*, Article ID 735014.

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

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

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

[23] Kim, S. 2011. Higher-order correlation clustering for image segmentation. *Advances in Neural Information Processing Systems* 1530–8.

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

[25] Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, **43(1)**, 215-237.

[26] 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.

[27] 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.

[28] Luo, Y. and Zhang, A. (2020). Open problem: average-case hardness of hypergraphic planted clique detection. *Proceedings of Machine Learning Research*, 1-4, 2020.

[29] Liu, H., Jan, L. and Yan, S.(2015). Dense subgraph partition of positive hypergraphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,37,3:541-554.

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

[31] Mossel, E., Neeman, J. and Sly, A. (2017). A proof of the block model threshold conjecture. *Combinatorica*, 1-44.

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

[33] Ouvrard,X., Goff,J. and Marchand-Maillet, S. (2017). Networks of collaborations: hypergraphmodeling and visualisation.https://arxiv.org/pdf/1707.00115.pdf

[34] Ramasco, J., Dorogovtsev, S. N. and Pastor-Satorras, R. (2004). Self-organization of collaboration networks, *Phys. Rev. E* **70**, 036-106.

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

[36] Shi, J. and Malik, J. (1997). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22(8)**, 888-905.

[37] Weng, H. and Feng, Y.(2021). Community detection with nodal information: likelihood

and its variational approximation. *Stat*, e428.

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

[39] Yuan, M. and Shang, Z. (2021). Informatin Limits for Detection a Subhypergraph. *STAT*, e407.

[40] Yuan, M. and Shang, Z. Sharp detection boundaries on testing dense subhypergraph. *Bernoulli* **2021**, to appear.

[41] Zhao, Y., Levina, E. and Zhu., J.(2011). Community extraction for social networks. *Proc. Natn. Acad. Sci. USA*, **108**, 7321-7326.

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

Department of Statistics, North Dakota State University, USA.

E-mail: mingao.yuan@ndsu.edu

Department of Statistics, North Dakota State University, USA.

E-mail: bin.zhao@ndsu.edu

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

E-mail: zxfstat@ncwu.edu.cn