Exact Recovery in the General Hypergraph Stochastic Block Model
Qiaosheng Zhang and Vincent Y. F. Tan, Senior Member, IEEE

Abstract—This paper investigates fundamental limits of exact recovery in the general $d$-uniform hypergraph stochastic block model ($d$-HSBM), wherein $n$ nodes are partitioned into $k$ disjoint communities with relative sizes $(p_1, \ldots, p_k)$. Each subset of nodes with cardinality $d$ is generated independently as an order-$d$ hyperedge with a certain probability that depends on the ground-truth communities that the $d$ nodes belong to. The goal is to exactly recover the $k$ hidden communities based on the observed hypergraph. We show that there exists a sharp threshold such that exact recovery is achievable above the threshold and impossible below the threshold (apart from a small regime of parameters that will be specified precisely). This threshold is represented in terms of a quantity which we term as the generalized Chernoff-Hellinger divergence between communities. Our result for this general model recovers prior results for the standard SBM and $d$-HSBM with two symmetric communities as special cases. En route to proving our achievability results, we develop a polynomial-time two-stage algorithm that meets the threshold. The first stage adopts a certain hypergraph spectral clustering method to obtain a coarse estimate of communities, and the second stage refines each node individually via local refinement steps to ensure exact recovery.

Index Terms—Community detection, hypergraph stochastic block model (HSBM), exact recovery, hypergraph spectral clustering methods.

I. INTRODUCTION

THE stochastic block model (SBM) [1] is a celebrated random graph model that has been widely studied for the community detection problem, and the objective therein is to partition $n$ nodes into $k$ disjoint communities (a.k.a., clusters) based on the randomly generated graph. The recent award-winning1 papers [2], [3] discovered a phase transition phenomenon for exact recovery (i.e., all the nodes are required to be classified correctly) in the SBM with two symmetric communities. That is, there is a sharp threshold such that exact recovery is achievable above the threshold, and impossible below the threshold. This phase transition phenomenon was later extended to the general SBM with $k \geq 2$ communities and without imposing symmetric structures [4], [5]. Popularized by these breakthroughs, community detection in the SBM and its variants have then received significant attention, and many works have progressively contributed to this field by considering the information-theoretic limits of some variants of the SBM [6], [7], [8], [9], efficient algorithms with theoretical guarantees [10], [11], [12], [13], [14], [15], [16], [17], the effect of side information [18], [19], [20], [21], [22], etc. We refer the readers to [23] for a comprehensive survey.

While most prior works focused on community detection on graphs, it is also of keen interest to study community detection on hypergraphs. This is because higher-order relational information among multiple nodes, which can naturally be captured by hypergraphs, is ubiquitous in many applications. For example, friendships between users in social networks can be captured by graphs, but chat groups are usually represented by hyperedges in hypergraphs. Authors in co-authorship networks can also be connected by hyperedges. There are also applications in computer vision (such as object recognition and image registration) that are concerned with the point-set matching problem [24], [25], which aims to find strongly connected components in a uniform hypergraph.

Motivated by these applications, in recent years some efforts have been expended to advance our understanding of community detection on hypergraphs. In particular, Ghoshdastidar and Dukkipati [26] first proposed a random hypergraph model called the $d$-uniform hypergraph stochastic block model ($d$-HSBM), in which each subset of nodes with cardinality $d$ is generated independently as an order-$d$ hyperedge with a certain probability that depends on the communities that the $d$ nodes belong to. Subsequent researchers further investigated the recovery limits of $d$-HSBMs by developing various hypergraph clustering algorithms (such as spectral clustering methods [27], [28], [29], [30], [31], [32], [33], semidefinite programming-based methods [34], [35], [36], tensor decomposition-based methods [37], approximate-message passing algorithms [38], [39], etc) with theoretical guarantees, and characterizing the minimax misclassification proportion [40], [41], [42] as well as the exact recovery criterion for the special case of two symmetric communities [34].2

1The Information Theory Society Paper Award, 2020.

2While this work mainly focuses on the $d$-HSBM, we are also aware that the clustering problem has also been explored in other hypergraph models, such as the non-uniform HSBM [27], [28], [29], the generalized censored block model [43], the sub-hypergraph models [44], [45], etc.0018-9448 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Although an information-theoretic limit for exact recovery has been derived in [34], their setting considering only two equal-sized communities and symmetric hyperedge generation probabilities is rather restrictive for real-world applications, and a general theory for exact recovery in the $d$-HSBM is still lacking. Motivated by this gap in the literature, in this work we consider the exact recovery criterion in the general $d$-HSBM. Our problem setting and the distinctions compared to other works are summarized as follows.

1) Nodes are partitioned into $k \geq 2$ non-overlapping communities, in which each node is assigned to one of these $k$ communities with probabilities $\{p_i\}_{i=1}^k$. This generalizes the setting in [34] and [30] in which $k = 2$ and $p_1 = p_2$, and the setting in [31], [40], [41], and [42] wherein the $k$ communities are of equal or approximately equal sizes (i.e., $p_i \approx p_j$ for all $1 \leq i < j \leq k$).

2) The probability that an order-$d$ hyperedge appears depends on the number of nodes in each of the $k$ communities (which is quantified by a length-$k$ vector $(T_1, T_2, \ldots, T_k)$ with $\sum_i T_i = d$, where $T_i$ is the number of nodes in community $i$). In contrast, many prior works with theoretical guarantees consider more restrictive assumptions on the probability that a hyperedge is present. For example, in [33], [30], [31], [34], and [35] the hyperedge probability can only take two values depending on whether all the $d$ nodes belong to the same community. Although other works such as [40], [41], [42] relax the restrictions in [33], [30], [31], [34], and [35], they are nevertheless particularizations of our general HSBM model. We are aware that the assumption on hyperedge probabilities made in [26], [27], [28], and [29] is similar to ours, but their focus is to characterize the performance of a hypergraph spectral clustering method contained therein in terms of the fraction of misclassified nodes, whereas we aim to quantify necessary and sufficient conditions for exact recovery (see Definition 1 for details).

3) Based on the observed $d$-uniform hypergraph, the learner is tasked to achieve exact recovery of the hidden partition, i.e., all the $n$ nodes should be assigned to the ground-truth communities that they belong to with high probability as the size of the hypergraph grows. We are also interested in deriving an algorithm-agnostic impossibility result that matches the performance guarantee of the learner’s algorithm.

A. Main Contributions

The main contributions and key technical challenges of this work are summarized as follows.

1) We establish a phase transition for the general $d$-HSBM (see Theorems 1 and 2), apart from a small subset of $d$-HSBMs that contains communities whose so-called second-order degree profiles are identical (to be specified in Section III-A). That is, there is a sharp threshold such that exact recovery is possible above the threshold, and impossible below the threshold. This threshold is represented in terms of a quantity which we term as the generalized Chernoff-Hellinger (GCH) divergence between different communities, and is a generalization of the CH-divergence discovered in [4] for the SBM. Our result also recovers the exact recovery criterions for the SBM [2], [3] and $d$-HSBM with two symmetric communities [34] as special cases. The techniques for proving the fundamental limits are inspired by [4]; however, dealing with hypergraphs requires us to carefully characterize different types of hyperedges that are induced by complicated community relations.

2) We develop a polynomial-time algorithm that meets the information-theoretic limits. This implies that there is no information-computation gap for exact recovery in the general $d$-HSBM (apart from the aforementioned small regime of parameters). Our two-stage algorithm consists of a hypergraph spectral clustering step in the first stage to ensure almost exact recovery$^3$ (see Definition 2). It then performs local refinement steps for each of the $n$ nodes in the second stage to ensure exact recovery. To circumvent the problem that conditioned on the success of the first stage certain a priori independent random variables become dependent, we adopt a hypergraph splitting technique to split the hypergraph into two sub-hypergraphs (see Section IV-A), such that the two stages can be run on the two independent sub-hypergraphs respectively, preserving the independence of the two stages to facilitate the analysis. Although this technique is not new, our analytical method is different from previous analyses (such as [4]). We prove that with high probability over splitting of the given hypergraph into two sub-hypergraphs, desirable properties of the resultant sub-hypergraphs are preserved, which further guarantees the success of the two stages. This new analytical method for analyzing multi-stage algorithms may be of independent interest. Algorithm 1 can also be improved to an agnostic algorithm that does not require the knowledge of model parameters (see Remark 4).

3) A main technical challenge lies in the development and analysis of an efficient algorithm that leads to almost exact recovery (for the first stage) for the general $d$-HSBM. To the best of our knowledge, such an algorithm with accompanying guarantees is lacking in the literature. Thus, the hypergraph spectral clustering method developed here and its analysis may be of independent interest. We are aware that various clustering algorithms have been developed. However, theoretical guarantees (on the fraction of misclassified nodes) are usually restricted to special classes of HSBMs and they do not readily apply to the general $d$-HSBMs. For example, the performance of the spectral clustering method in [40] depends on the $k$-th largest singular value of a specific matrix, but bounding this value turns out to be non-trivial for general $d$-HSBMs. The semidefinite programming-based method in [34] is only applicable to symmetric settings. Our new algorithm overcomes these stumbling blocks by

$^3$In the literature, “almost exact recovery” is sometimes also called “weak consistency”, and “exact recovery” is called “strong consistency”.

leveraging and judiciously combining various ideas from prior works for the SBM [9], [10] and the HSBM [40].
Our theoretical result (Theorem 3) shows that, with probability approaching one, all but a vanishing fraction of
the \( n \) nodes can be assigned to their true communities
(i.e., almost exact recovery is achieved) in the general
d-HSBM.

B. Organization

We describe the general \( d \)-HSBM and the exact recovery
criterion in Section II, and provide our main results (with
accompanying discussions) in Section III. Our computationally
efficient two-stage algorithm is introduced in Section IV, and
its theoretical guarantee is formally established in Section V.
The converse part is proved in Section VI. Section VII
concludes this work and proposes several directions that are
fertile avenues for future research.

II. PRELIMINARIES AND PROBLEM STATEMENT

A. Notation

For any integer \( m \geq 1 \), let \([m]\) represent the set of
integers \( \{1, 2, \ldots, m\} \), and \( S_m \) be the set of all permutations
from \([m]\) to \([m]\). Random variables and their realizations
are respectively denoted by upper-case and lower-case letters,
while vectors, matrices, and tensors are denoted by boldface
letters. For a length-\( n \) vector \( x \), let \( x_i \) denote its \( i \)-th element, and \( x_{-i} \) denote the length-\((n-1)\) sub-vector that excludes
\( x_i \). For a matrix \( A \), its operator norm and Frobenius norm
are respectively represented by \( \|A\|_{op} \) and \( \|A\|_F \), and its \( v \)-th column is denoted by \( A_v \).

B. The \( d \)-Uniform Hypergraph Stochastic
Block Model (\( d \)-HSBM)

Let \( n \in \mathbb{N} \) be the number of nodes, and \( k \geq 2 \) be the
number of non-overlapping communities. Each node \( v \in [n] \)
belongs to one of the \( k \) communities, and is associated with
a latent random variable \( Z_v \) on \([k]\) with prior distribution
\( p = (p_1, p_2, \ldots, p_k) \), where \( \sum_{i \in [k]} p_i = 1 \). That is, if node
\( v \) belongs to community \( i \), then \( Z_v = i \). The length-\( n \) vector
\( Z = (Z_1, Z_2, \ldots, Z_n) \) thus represents the ground-truth
community vector of the \( n \) nodes. Furthermore, we define
\( \mathcal{V}_i = \{ v \in [n] : Z_v = i \} \) as the collection of nodes that
belong to the community \( i \) (for \( i \in [k] \)).

Let \( d \geq 2 \) be the order of the hyperedges (i.e., the number of
nodes contained in each hyperedge), and \( \mathcal{W} \) be the set of all
order-\( d \) hyperedges on \([n]\) (where \( |\mathcal{W}| = \binom{n}{d} \)). It is assumed
that the hypergraph considered in this work only contains
order-\( d \) hyperedges; thus it is referred to as a \( d \)-uniform
hypergraph. The generation process (underlying statistical
model) of our random \( d \)-uniform hypergraph \( G \) is as follows. For each \( e \in \mathcal{W} \), the probability that it appears in the hypergraph \( G \) (i.e., \( e \in \mathcal{E} \)) depends on the number of nodes in each community \( \{\mathcal{V}_i\}_{i=1}^k \). Formally, let
\[
\mathcal{T} \triangleq \{ T \in \mathbb{N}^k : T_1 + T_2 + \cdots + T_k = d \}
\]
be the collection of length-\( k \) vectors such that each vector
\( T \in \mathcal{T} \) (with \( T_i \) representing the number of nodes in \( \mathcal{V}_i \))
represents a possible community assignment of \( n \) nodes, where
“community assignment” is referred to as the number of
nodes contained in each community. The generation of the
hyperedges in \( G \) is fully characterized by a set of numbers
\( \{Q_T\}_{T \in \mathcal{T}} \subset \mathbb{R}_+ \). The probability of a hyperedge \( e \in \mathcal{W} \) appearing is \( Q_{T(e)} \frac{\log n}{n^2} \), where \( T(e) \in \mathcal{T} \) denotes
the community assignment of the \( d \) nodes in hyperedge \( e \). That is, \( T(e) \) is the length-\( k \) vector whose \( i \)-th entry represents the
number of nodes in the hyperedge \( e \) that belongs to the \( i \)-th community.

Example 1: Suppose \( d = 3 \), \( k = 4 \), \( n = 8 \), and \( \mathcal{V}_1 = \{1, 2\}, \mathcal{V}_2 = \{3, 4\}, \mathcal{V}_3 = \{5, 6\}, \mathcal{V}_4 = \{7, 8\} \). We list three different order-\( d \) hyperedges \( e_1, e_2, e_3 \in \mathcal{W} \), as well as their community assignments in the table below. Although \( e_1 \neq e_2 \), the probabilities that \( e_1 \in \mathcal{E} \) and \( e_2 \in \mathcal{E} \) are the same since
they have the same community assignment. On the other hand,
the probability that \( e_3 \in \mathcal{E} \) is in general different from that for
\( e_1 \) and \( e_2 \).

| Hyperedges | Community assignment | Hyperedge probability |
|------------|----------------------|-----------------------|
| \( e_1 = (1, 4, 8) \) | \( T(e_1) = (1, 1, 0, 1) \) | \( \frac{Q_{(1,1,0,1)} \log n}{n^2} \) |
| \( e_2 = (1, 3, 7) \) | \( T(e_2) = (1, 1, 0, 1) \) | \( \frac{Q_{(1,1,0,1)} \log n}{n^2} \) |
| \( e_3 = (2, 5, 7) \) | \( T(e_3) = (1, 0, 1, 1) \) | \( \frac{Q_{(1,0,1,1)} \log n}{n^2} \) |

The reason why we consider the \( \Theta(\frac{\log n}{n^2}) \)-regime for the
connectivity probability is that it ensures the average degree
of each node is \( \Theta(\log n) \) and it was shown [33], [34], [40] that
phase transition for exact recovery occurs in this logarithmic
average degrees regime. Furthermore, we define \( Q_{\text{max}} \triangleq \max_{T \in \mathcal{T}} Q_T \) and \( Q_{\text{min}} \triangleq \min_{T \in \mathcal{T}} Q_T \), and it is assumed
that the parameters \( Q_{\text{max}}, Q_{\text{min}}, k, d \) and \( \{p_i\}_{i \in [k]} \) do not scale with \( n \). We also note that several related works [26], [29], [37]
allow the number of communities \( k \) to diverge as \( n \) grows.

Similar to the adjacency matrices for graphs, any \( d \)-uniform
hypergraph \( G \) can be represented by an order-\( d \) \( n \times \cdots \times n \)
adjacency tensor \( A = [A_{b}] \), where \( b = [b_1, \ldots, b_d] \in [n]^d \)
is the access index of the element in the tensor. Here, \( A_{b} \in \{0, 1\} \), and \( A_{b} = 1 \) means the presence of the hyperedge
corresponding to the \( d \) nodes in \( b \). In particular, \( A_{b} = 0 \) if the \( d \) elements in \( b \) are not distinct (since each hypergraph must contain \( d \) nodes), and \( A_{b} = A_{b'} \) if there exists a permutation \( \pi \) such that \( (b_{\pi(1)}, b_{\pi(2)}, \ldots, b_{\pi(d)}) = (b_1, b_2, \ldots, b_d) \).

C. Objective

Given the observation of the hypergraph \( G \) (or the adjacency
tensor \( A \)), the learner aims to use an estimator \( \hat{\phi} = \phi(G) \) to
recover the partition of the \( n \) nodes into \( k \) communities. The
output of the estimator \( \phi \) is denoted by \( \hat{Z} = (\hat{Z}_1, \hat{Z}_2, \ldots, \hat{Z}_n) \).
We measure the accuracy of \( \hat{Z} \) in terms of the misclassification
propportion \( l(\hat{Z}, Z) \), which is defined as
\[
l(\hat{Z}, Z) \triangleq \min_{\pi \in \mathcal{S}_n} \frac{1}{n} \sum_{v \in [n]} \mathbb{1}\{ \hat{Z}_v \neq \pi(Z_v) \}.
\]

**Definition 1 (Exact Recovery):** An estimator \( \hat{\phi} \) is said to
achieve exact recovery if it ensures that with probability \( 1 - o(1) \), the misclassification proportion \( l(\hat{Z}, Z) = 0 \).
Definition 2 (Almost Exact Recovery): An estimator $\phi$ is said to achieve almost exact recovery if it ensures that with probability $1-o(1)$, the misclassification proportion $l(\hat{Z}, Z) \to 0$ as $n$ tends to infinity.

III. MAIN RESULTS AND DISCUSSIONS
We first introduce several notations that are useful for stating our main results. Let
\[ M \triangleq \{ \mathbf{m} \in \mathbb{N}^k : m_1 + m_2 + \cdots + m_k = d - 1 \} \tag{3} \]
be the collection of length-$k$ vectors such that the sum of the $k$ elements equals $d - 1$. Each element in $M$ represents one possible community assignment of $d - 1$ nodes. For each $\mathbf{m} \in M$, we define
\[ R_m \triangleq \prod_{s=1}^{k} \left( \frac{np_s}{m_s} \right) \quad \text{and} \quad R_m' \triangleq \frac{R_m}{n^{d-1}} \tag{4} \]
as the expected number (and normalized expected number) of combinations of $d - 1$ nodes that have community assignment $\mathbf{m}$. Note that $R_m = \Theta(n^{d-1})$ since $\left( \frac{np_s}{m_s} \right) = \Theta(n)$ and $\sum_{s \in [k]} m_s = d - 1$, and thus $R_m' = \Theta(1)$.

A. Separation Between Communities
1) Degree Profile: For each community $V_i$ (where $i \in [k]$), we define $\mu_{m \oplus i}$ as the criterion for the degree profile of community $i$. Intuitively, two communities are easier to be separated if the degree profiles of these two communities are further apart. The discrepancy between any two communities in the $d$-HSBM can be measured in terms of the generalized Chernoff-Hellinger divergence (GCH-divergence) between their degree profiles, which generalizes the CH-divergence for the SBBM that was first discovered by Abbe and Sandon [4, Eqn. (3)].

Definition 3 (GCH-Divergence): For any $i, j \in [k]$ such that $i \neq j$, we define the GCH-divergence between $i$ and $j$ as
\[ D_+(i, j) \triangleq \max_{t \in [0,1]} \sum_{\mathbf{m} \in M} t \mu_{m \oplus i} + (1-t) \mu_{m \oplus j} - \mu^t_{m \oplus i} \mu^{1-t}_{m \oplus j}, \tag{6} \]
where $D_+(i, j)$ is a function of $\mathbf{p}$ and $\{Q_T\}_{T \in \mathcal{T}}$.

Note that $D_+(i, j) = 0$ if and only if the degree profiles of communities $V_i$ and $V_j$ are exactly the same, in which case the two communities are statistically indistinguishable.

2) Second-Order Degree Profile: For each community $V_i$ (where $i \in [k]$), we define its second-order degree profile as $\{ \sum_{\mathbf{m} \in M: m_s \geq 1} m_s \mu_{m \oplus i} \}_{s \in [k]}$, where each element $\sum_{\mathbf{m} \in M: m_s \geq 1} m_s \mu_{m \oplus i}$ represents the normalized expected number of hyperedges that contain two fixed nodes belonging to $V_i$ and $V_s$, respectively. When the second-order degree profile of two communities are exactly the same, our analysis also shows that there may be some inherent difficulties in distinguishing them. Formally, we define $\Xi$ as the subset of model parameters $\{ \mathbf{p}, \{Q_T\}_{T \in \mathcal{T}} \}$ such that there exist two communities having the same second-order degree profiles, i.e.,
\[ \Xi \triangleq \left\{ \{ \mathbf{p}, \{Q_T\}_{T \in \mathcal{T}} \} : \exists i \neq j \text{ s.t. } \sum_{\mathbf{m} \in M: m_s \geq 1} m_s \mu_{m \oplus i} = \sum_{\mathbf{m} \in M: m_s \geq 1} m_s \mu_{m \oplus j} \text{ for all } s \in [k] \right\}. \tag{7} \]

B. Main Results
Theorem 1 (Converse): It is impossible to achieve exact recovery when the model parameters $\{ \mathbf{p}, \{Q_T\}_{T \in \mathcal{T}} \}$ satisfy
\[ \min_{i,j \in [k]: i \neq j} D_+(i, j) < 1. \tag{8} \]

Theorem 2 (Achievability): Assume that the model parameters $\{ \mathbf{p}, \{Q_T\}_{T \in \mathcal{T}} \} \notin \Xi$. Then the polynomial-time two-stage algorithm (Algorithm 1) achieves exact recovery when the model parameters $\{ \mathbf{p}, \{Q_T\}_{T \in \mathcal{T}} \}$ satisfy
\[ \min_{i,j \in [k]: i \neq j} D_+(i, j) > 1. \tag{9} \]

Some remarks on Theorems 1 and 2 are in order.
1) For community detection in the $d$-HSBM, most of the settings considered in prior works, such as the one that $\{Q_T\}_{T \in \mathcal{T}}$ can only take two values depending on whether $d$ nodes belong to the same community $[30, 31, 32, 33, 34, 35]$, satisfy $\{ \mathbf{p}, \{Q_T\}_{T \in \mathcal{T}} \} \notin \Xi$. Thus, our result is a strict generalization of these existing works. An example for the case that $\{ \mathbf{p}, \{Q_T\}_{T \in \mathcal{T}} \} \in \Xi$ is as follows. Suppose $k = 2, d = 3$, $\mathbf{p} = (p_1, p_2)$, and $\{Q_T\}_{T \in \mathcal{T}} = \{ Q_{(3,0)}, Q_{(2,1)}, Q_{(1,2)}, Q_{(0,3)} \}$. When $\{ \mathbf{p}, \{Q_T\}_{T \in \mathcal{T}} \}$ satisfies $p_1 = p_2 = 1/2$, $Q_{(3,0)} = Q_{(1,2)}$ and $Q_{(2,1)} = Q_{(0,3)}$, one can check that the two communities $V_1$ and $V_2$ have the same second-order degree profile, thus $\{ \mathbf{p}, \{Q_T\}_{T \in \mathcal{T}} \} \in \Xi$.
2) When $\{ \mathbf{p}, \{Q_T\}_{T \in \mathcal{T}} \} \notin \Xi$ and $\min_{i,j \in [k]: i \neq j} D_+(i, j) > 1$, it remains open whether exact recovery is possible. However, we would like to point out that this scenario does not apply to the SBBM (equivalently, the 2-HSBM), since the condition $\min_{i,j \in [k]: i \neq j} D_+(i, j) > 1$ immediately implies that $\{ \mathbf{p}, \{Q_T\}_{T \in \mathcal{T}} \} \notin \Xi$ in this setting. Our sharp threshold is applicable to all model parameters when $d = 2$.
3) When $\{ \mathbf{p}, \{Q_T\}_{T \in \mathcal{T}} \} \notin \Xi$, the first stage of Algorithm 1 ensures almost exact recovery via hypergraph spectral clustering (as shown in Theorem 3), and the condition $\min_{i,j \in [k]: i \neq j} D_+(i, j) > 1$ is the criterion for
Stage 2 (local refinement steps) to succeed. Roughly speaking, performing a local refinement step for each node is equivalent to performing a hypothesis test with independent but non-identically distributed samples. The corresponding error probability can be represented by a variant of the Chernoff information [46, Chapter 11.9], and this further reduces to $n^{-\min_{i,j\in[k]:i\neq j} D_{+}(i,j)}$ which is in the form of the GCH-divergence. Thus, when $\min_{i,j\in[k]:i\neq j} D_{+}(i,j) > 1$, taking a union bound over the $n$ nodes results in a vanishing error probability (i.e., exact recovery is achieved).

4) Our analysis of the first stage is not able to handle the case in which $\{p, \{Q_{T}\}_{T\in\mathcal{T}}\} \notin \Xi$ because the key step in the hypergraph spectral clustering method is to map the order-$d$ adjacency tensor $A$ to an $n \times n$ matrix $L$ (defined in (17) below), and the subsequent clustering algorithm critically relies on the discrepancy between the columns of $L$ (which corresponds to the second-order degree profiles of communities). We conjecture that this issue may be circumvented if one directly applies clustering algorithms on the adjacency tensor (such as the method proposed in [37]), and the exact recovery threshold $\min_{i,j\in[k]:i\neq j} D_{+}(i,j) = 1$ holds without the assumption that the second-order degree profiles of any two communities are distinct.

5) The algorithm performance also depends on the value of $d$. As $d$ increases, the computational complexity of constructing $L$ increases accordingly. If the hyperedge probabilities $\{Q_{T}\}_{T\in\mathcal{T}}$ were unknown a priori, a larger value of $d$ would also increase the difficulty of learning $\{Q_{T}\}_{T\in\mathcal{T}}$, since $|\mathcal{T}|$ increases exponentially with $d$.

C. Recovering Prior Results From Theorems 1 and 2

To the best of our knowledge, the sharp threshold established by Theorems 1 and 2 is the most general result for exact recovery in the SBM/HSBM literature. As discussed below, several problem settings investigated in prior works are subsumed by our result, and the thresholds derived in the prior works can be recovered from Theorems 1 and 2.

1) Exact Recovery in the SBM [4]: The SBM considered in [4] corresponds to our $d$-HSBM with $d = 2$. In [4], the prior distribution of each node is also $p = (p_{1}, p_{2}, \ldots, p_{k})$, and the edge probabilities are characterized by $\{Q_{ij} \log \frac{n}{n}\}_{i,j\in[k]}$, where $Q_{ij}$ corresponds to edges that contain nodes in $V_{i}$ and $V_{j}$. The authors of [4] showed that the threshold for exact recovery is

$$\min_{i,j\in[k]:i\neq j} \max_{t\in[0,1]} \sum_{s\in[k]} p_{s} \left[ (1-t)Q_{s,i} + tQ_{s,j} - Q_{s,i}^{1-t}Q_{s,j}^{1-t} \right] = 1. \tag{10}$$

In our setting with $d = 2$, the set $M = \{m_{1}, m_{2}, \ldots, m_{k}\}$ contains $k$ distinct length-$k$ vectors, where each $m_{s} \notin (0,0,0,1,0,0,\ldots,0)$ contains a single one which is at the $s$-th location. By noting that $\mu_{m_{s}} = p_{s}Q_{s,i}$ (resp. $\mu_{m_{s}} = p_{s}Q_{s,j}$) in [4] and $\{p, \{Q_{T}\}_{T\in\mathcal{T}}\} \notin \Xi$ when $\min_{i,j\in[k]:i\neq j} D_{+}(i,j) > 1$, we recover their threshold stated in Eqn. (10) from Theorems 1 and 2.

Remark 1: One major distinction between the algorithms in [4] and this work is the initialization step (Stage 1). We use the spectral clustering method while [4] uses the so-called sphere comparison algorithm. The main idea of the sphere comparison algorithm is to determine whether two nodes belong to the same community by counting the common neighbors at a large enough depth between them. While it works well for regular graphs, generalizing it to hypergraphs may be non-trivial.

2) Exact Recovery in the $d$-HSBM With Two Symmetric Communities [34]: The model considered in [34] is a special $d$-HSBM with two equal-sized communities that have symmetric structures. It corresponds to our general $d$-HSBM with $k = 2$, $p_{1} = p_{2} = 1/2$, and $\{Q_{T}\}_{T\in\mathcal{T}} = \{Q_{1}, Q_{2}\}$ (where each hyperedge appears with probability $Q_{1}$ when $d$ nodes are in the same communities, and $Q_{2}$ otherwise). Kim, Bandeira, and Goemans [34] showed that the threshold for exact recovery is

$$\frac{1}{2^{d-1}} \left( \sqrt{Q_{1}/(d-1)!} - \sqrt{Q_{2}/(d-1)!} \right)^{2} = 1. \tag{11}$$

Specializing our result to this symmetric setting, we note that the set $M = \{(d-1,0),(d-2,1),\ldots,(0,d-1)\}$ contains $d$ distinct length-2 vectors, and the values of $\mu_{m_{1}}$ and $\mu_{m_{2}}$ for all $m \in M$ can then be calculated. By noting that the model parameters $\{p, \{Q_{T}\}_{T\in\mathcal{T}}\} \notin \Xi$ and $t = 1/2$ maximizes the GCH-divergence in (6) for symmetric SBMs and HSBMs, we recover the threshold stated in Eqn. (11) from Theorems 1 and 2. Furthermore, we also recover the celebrated exact recovery threshold $\sqrt{Q_{1} - Q_{2}} = \sqrt{d}$ for the SBM with two symmetric communities [2], since it is a special case of [34] for $d = 2$.

D. Comparisons With the Results on the Misclassification Proportion in the HSBM [40]

The work [40] studied the fundamental limit of misclassification proportion in $d$-HSBMs. Their model assumes that there are $k$ approximately equal-sized communities, and the hyperedge probabilities depend only on the sorted histogram vector (in descending order) of the community assignment vector (e.g., the community assignment vectors $T = (d,0,\ldots,0)$ and $T' = (0,\ldots,0,d)$ correspond to a same sorted histogram vector $(d,0,\ldots,0)$). Thus, their model is a particularization of our general HSBM model. While the main focus of their work is to characterize the negative exponent of the misclassification proportion $l(Z, Z)$ (as defined in Eqn. (2)), their results can also be applied to finding the exact recovery threshold by setting the negative exponent to be greater than $\log n$ (which means $l(Z, Z) < 1/n$ and thus implies exact recovery). In the following, we show that the exact recovery thresholds derived in this work and [40, Theorem 3.1, Theorem 3.2] are exactly the same when $d = 2$ and $d = 3$. When $d \geq 4$, the expressions in both works become highly complicated (and moreover, [40] did not provide the precise value of their expression for $d \geq 5$), thus it is difficult to make comparisons; however, we conjecture that the thresholds should still be the same for $d \geq 4$ due to the evidence shown for $d = 2$ and $d = 3$. 
1) Comparison for \( d = 2 \): For a valid comparison, we assume that (i) there are \( k \) communities of equal sizes, and (ii) the hyperedge probabilities scale as \( \Theta(n) \). This implies that the misclassification proportion in [47], in which the negative exponent is dominated by \((n/k) \cdot I_{q_1q_2} \) where \( I_{q_1q_2} \triangleq -2 \log(\sqrt{q_1q_2} + \sqrt{1-q_1}\sqrt{1-q_2}) \) and can further be simplified as

\[
-2 \log(\sqrt{q_1q_2} + \sqrt{1-q_1}\sqrt{1-q_2}) = -2 \log \left\{ \sqrt{q_1q_2} + \left( \frac{1}{2}q_1 + O(q^2) \right) + \left( \frac{1}{2}q_2 + O(q^2) \right) \right\} \\
= -2 \log \left\{ 1 - \left( \frac{1}{2}(q_1 + q_2) - \sqrt{q_1q_2} + O(q_1q_2) \right) \right\} \\
= 2 \left( \frac{1}{2}(q_1 + q_2) - \sqrt{q_1q_2} \right) + O(q_1q_2), \tag{12}
\]

The first equality follows from \( \sqrt{1-x} = 1 - \frac{1}{2}x + O(x^2) \) for \( x \to 0 \). For sufficiently large \( n \), when the parameters satisfy

\[
\frac{(\sqrt{q_1} - \sqrt{q_2})^2}{k} > 1, \tag{13}
\]

the negative exponent of the misclassification proportion will be greater than \( \log n \) (i.e., the misclassification proportion will be less than \( 1/n \)), which implies exact recovery.

On the other hand, when \( d = 2 \), the theoretical guarantee of exact recovery in our work reduces to the threshold in [4], which is exactly the condition given in (13). This means that for \( d = 2 \), the exact recovery thresholds in [40] and this work are the same.

2) Comparison for \( d = 3 \): For a valid comparison, we assume that (i) there are \( k \) communities of equal sizes, and (ii) the hyperedge probabilities scale as \( \Theta(n) \), and depend only on the sorted histogram vector of the community assignment:

- A hyperedge appears with probability \( q_1 = Q_1 \log n \) if all three nodes belong to a same community;
- A hyperedge appears with probability \( q_2 = Q_2 \log n \) if only two nodes belong to a same community;
- A hyperedge appears with probability \( q_3 = Q_3 \log n \) if three nodes belong to three different communities.

We note that [40, Theorem 3.1] guarantees that the misclassification proportion between the true and estimated labels is at most \( \exp\{-n(1-\xi_n)\} \frac{n^2(k-2)}{2k^2} I_{q_1q_2} + \frac{n^2(k-2)}{2k^2} I_{q_1q_3} \) with high probability, where \( I_{q_1q_2} \triangleq -2 \log(\sqrt{q_1q_2} + \sqrt{1-q_1}\sqrt{1-q_2}) \) and \( \xi_n \to 0 \) as \( n \to \infty \). To ensure exact recovery, the negative exponent should satisfy

\[
\frac{n^2}{2k^2} I_{q_1q_2} + \frac{n^2(k-2)}{k^2} I_{q_1q_3} \log n > 0. \tag{14}
\]

Next, we figure out the condition under which (14) holds. Recalling from (12) that \( I_{q_1q_3} = (\sqrt{q_1} - \sqrt{q_3})^2 + O(q_1q_3) \), thus the LHS of (14) can be expressed as

\[
\frac{n^2}{2k^2} (\sqrt{q_1} - \sqrt{q_2})^2 + \frac{n^2(k-2)}{2k^2} (\sqrt{q_2} - \sqrt{q_3})^2 + O(n^2q_1q_2)
\]

For sufficiently large \( n \), when the model parameters satisfy

\[
\frac{(\sqrt{q_1} - \sqrt{q_2})^2}{k^2} + \frac{(k-2)(\sqrt{q_2} - \sqrt{q_3})^2}{k^2} > 1, \tag{15}
\]

the misclassification proportion will be less than \( 1/n \), which implies exact recovery.

Next, we specialize our results to the setting of interest. Note that \( M = \{ m \in \mathbb{N}^k : m_1 + m_2 + \cdots + m_k = 2 \} \), and \( R_m = 1/k^2 \) if \( \max_{|k|} m_l = 1 \), and \( R_m = 1/(2k^2) \) if \( \max_{|k|} m_l = 2 \). One can check that the second-order degree profile condition is satisfied. Without loss of generality, we focus on the first two communities: \( V_1 \) with degree profile \( \{\mu_m\}_{m \in M} \) and \( V_2 \) with degree profile \( \{\mu_m\}_{m \in M} \). In the following, we consider \( m \in M \) such that \( \mu_m = 1 \) and \( \mu_m = 2 \) are different:

- When \( m = (2,0,0,\ldots,0) \), we have \( \mu_m = Q_1/(2k^2) \) and \( \mu_m = Q_2/(2k^2) \);
- When \( m = (0,2,0,\ldots,0) \), we have \( \mu_m = Q_1/(2k^2) \) and \( \mu_m = Q_2/(2k^2) \);
- When \( m \) satisfies \( m_1 = 2, m_2 = 0 \), and there exists only one index \( l \in \{3,\ldots,k\} \) such that \( m_l = 1 \), we have \( \mu_m = Q_1/k^2 \) and \( \mu_m = Q_3/k^2 \);
- When \( m \) satisfies \( m_1 = 0, m_2 = 2 \), and there exists only one index \( l \in \{3,\ldots,k\} \) such that \( m_l = 1 \), we have \( \mu_m = Q_3/k^2 \) and \( \mu_m = Q_2/k^2 \).

Thus, the GCH-Divergence \( D_+(1,2) \) between the first two communities is

\[
\max_{t \in [0,1]} \left\{ \frac{Q_1}{2k^2} + (1-t) \frac{Q_2}{2k^2} - \left( \frac{Q_1}{2k^2} \right)^t \left( \frac{Q_2}{2k^2} \right)^{1-t} \right\}
+ \left\{ \frac{Q_2}{2k^2} + (1-t) \frac{Q_1}{2k^2} - \left( \frac{Q_2}{2k^2} \right)^t \left( \frac{Q_1}{2k^2} \right)^{1-t} \right\}
+ (k-2) \left\{ \frac{Q_3}{k^2} + (1-t) \frac{Q_2}{k^2} - \left( \frac{Q_3}{k^2} \right)^t \left( \frac{Q_2}{k^2} \right)^{1-t} \right\}
+ (k-2) \left\{ \frac{Q_3}{k^2} + (1-t) \frac{Q_1}{k^2} - \left( \frac{Q_3}{k^2} \right)^t \left( \frac{Q_1}{k^2} \right)^{1-t} \right\},
\]

where the minimum is obtained at \( t = 1/2 \), yielding that \( D_+(1,2) = \frac{(\sqrt{Q_1} - \sqrt{Q_2})^2}{2k^2} + \frac{(k-2)(\sqrt{Q_2} - \sqrt{Q_3})^2}{k^2} \). Finally, by symmetry one can show that \( D_+(i,j) = D_+(1,2) \) for other pairs of \( i,j \in [k] \). Therefore, the exact recovery threshold is

\[
\frac{(\sqrt{Q_1} - \sqrt{Q_2})^2}{2k^2} + \frac{(k-2)(\sqrt{Q_2} - \sqrt{Q_3})^2}{k^2} > 1, \tag{16}
\]

which is exactly the same as the threshold (15) derived in [40].
IV. THE TWO-STAGE ALGORITHM FOR EXACT RECOVERY

In this section, we present our polynomial-time algorithm that is used to achieve the information-theoretic limit shown in Theorem 2. As mentioned in Section III-B, our algorithm consists of two stages such that the first stage achieves almost exact recovery via the hypergraph spectral clustering method and the second stage achieves exact recovery via local refinement steps. This “from global to local” principle has been employed in many contexts, such as community detection in the SBM [4], [9], [10], [11], [48] and HSBM [33], [34], [40], [41], [42], matrix completion [49], [50], [51], etc. It is also worth noting that when analyzing two-stage algorithms, random variables that are initially independent may become dependent conditioned on the success of a preceding stage. To ameliorate this problem, we adopt the graph splitting technique (as described in Subsection IV-A) which is inspired by prior works on community detection [4], [11], [33], [52]. Our algorithm is described in detail in Algorithm 1.

Algorithm 1 THE TWO-STAGE ALGORITHM

1. **Input:** Hypergraphs \( G_1, G_2 \), \( \gamma_n = \sqrt{\log n} \), number of communities \( k \), radius \( r = \frac{\gamma_n^2}{\tau \log(\gamma_n)} \), \( \hat{V}_0^{(0)} \)

2. **Stage 1:** Hypergraph spectral clustering
   - \( L \leftarrow HH^T \otimes D; \Gamma \leftarrow \text{trim the rows and columns in } L \text{ that correspond to } v \notin \Gamma \)
   - \( L^{(k)}_\Gamma \leftarrow \text{rank-}k \text{ approximation of } L_\Gamma \)
   - \( \hat{V}^{(0)}_\Gamma \leftarrow \text{argmax }_{i,j} \{ \| (L^{(k)}_\Gamma)_{i,j} \|^2 \leq 1 \}, \forall v \in \Psi \)
   - \( v^*_t \leftarrow \text{argmax }_{v \in \Psi} \{ B_v \setminus \{ (L^{(k)}_\Gamma)_{j,v} \} \} \)
   - \( \hat{V}^{(0)}_{j,v} \leftarrow B_{v} \setminus \{ (L^{(k)}_\Gamma)_{j,v} \} \)

3. **Stage 2:** Local refinement steps
   - \( \{ \hat{V}^{(0)}_{i} \}_i \) \( i \in [k] \)
   - **Output:** Final estimate \( \hat{Z} = (\hat{Z}_1, \hat{Z}_2, \ldots, \hat{Z}_n) \)

A. Graph Splitting

Let \( F = (\{ n \}, \mathcal{W}) \) be the complete \( d \)-uniform hypergraph on node set \( \{ n \} \), and the hyperedge set \( \mathcal{W} \) contains all the \( \binom{n}{d} \) order-\( d \) hyperedges (as defined in Section II-B). We randomly split \( F \) into two sub-hypergraph \( F_1 = (\{ n \}, \mathcal{W}_1) \) and \( F_2 = (\{ n \}, \mathcal{W}_2) \). Each hyperedge in \( \mathcal{W} \) is sampled to \( \mathcal{W}_1 \) with probability \( \gamma_n / \log n \), and to \( \mathcal{W}_2 \) with probability \( 1 - (\gamma_n / \log n) \), where \( \gamma_n \) can be any value in \( \omega(1) \cap o(\log n) \). For concreteness we set \( \gamma_n = \sqrt{\log n} \). Note that \( \mathcal{W}_2 \) is the complement of \( \mathcal{W}_1 \). This splitting process is independent of the generation of the hypergraph \( G = (\{ n \}, \mathcal{E}) \) (which is generated according to \( p \) and \( \Omega_{T \in T} \mathcal{E} \)). We then define \( G_1 = (\{ n \}, \mathcal{E}_1) \) as the sub-HSBM that is generated on the hyperedge set \( \mathcal{W}_1 \) of \( F_1 \), where \( \mathcal{E}_1 = \mathcal{E} \cap \mathcal{W}_1 \) is the intersection of the hyperedge sets of the HSBM \( G \) and the sub-hypergraph \( F_1 \). Similarly, we define \( G_2 = (\{ n \}, \mathcal{E}_2) \) as the sub-HSBM that is generated on the hyperedge set \( \mathcal{W}_2 \) of \( F_2 \), where \( \mathcal{E}_2 = \mathcal{E} \cap \mathcal{W}_2 \).

B. Almost Exact Recovery via Hypergraph Spectral Clustering (Stage 1)

The main focus of this subsection is the sub-HSBM \( G_1 = (\{ n \}, \mathcal{E}_1) \). We apply a hypergraph spectral clustering method on \( G_1 \) to obtain an initial estimate of the ground-truth community vector \( Z_0 \), denoted by \( \tilde{Z}_0^{(k)} = (\tilde{Z}_0^{(k)}_1, \ldots, \tilde{Z}_0^{(k)}_n) \)

Let \( H = [H_{vw}] \) be the \( n \times (\binom{n}{d}) \) binary incidence matrix corresponding to \( G_1 \) such that each entry \( H_{vw} = 1 \) if the hyperedge \( e \in \mathcal{E}_1 \) and \( e \) contains node \( v \), and \( H_{vw} = 0 \) otherwise. Note that there is an one-to-one mapping between \( H \) and the observed adjacency tensor \( \Theta \), thus one can obtain \( H \) from \( \Theta \). For each node \( v \in \{ n \} \), its degree (in \( G_1 \)) is denoted by \( d_v = \sum_{e \in \mathcal{E}_1} H_{ev} \). Let \( D = \text{diag}(d_1, \ldots, d_n) \) be an \( n \times n \) diagonal matrix that represents the degrees of the \( n \) nodes. We then define the hypergraph Laplacian as

\[
L \triangleq HH^T - D,
\]

where \( L \) is an \( n \times n \) matrix and the \((i,j)-entry\) represents the number of hyperedges that contain both node \( i \) and node \( j \). To ensure a good performance of the hypergraph spectral clustering method, one typically needs to remove a small fraction of nodes that have significantly higher degrees [40] than the average. Thus, we define the set of “good” nodes that have degree no larger than a certain threshold \( \tau \) as

\[
\Gamma \triangleq \{ v \in [n] : d_v \leq \tau \},
\]

where \( \tau \) is set to be \( CQ_{\text{max}} \gamma_n \) for some large constant \( C > 0 \), such that \( \tau \) is much larger than the expected degree of every node.

We apply Stage 1 of Algorithm 1 (lines 2 – 16) to obtain an almost exact recovery of the \( k \) communities. Initially, we calculate the hypergraph Laplacian \( L \), and then “trim” the rows and columns in \( L \) that correspond to nodes that do not belong to \( \Gamma \). Specifically, for each of the \( n \) nodes \( v \in \{ n \} \), if \( v \notin \Gamma \), we replace all the entries in the \( v \)-th row and \( v \)-th column of \( L \) by zeros. This yields the trimmed hypergraph Laplacian \( L_\Gamma \). In addition, we also perform a singular value decomposition (SVD) on \( L_\Gamma \) to obtain the optimal rank-\( k \) approximation \( L_\Gamma^{(k)} \), i.e., \( L_\Gamma^{(k)} = \sum_{i=1}^{k} \sigma_i u_i v_i^T \) where \( \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \) are the largest \( k \) singular values, and \( u_i \) and \( v_i \) are the corresponding singular vectors of \( L_\Gamma \).

We then perform a clustering algorithm (lines 5 – 16) on the columns of \( L_\Gamma^{(k)} \), i.e., the set of column vectors \( \{ L_\Gamma^{(k)}_v \}_{v \in \Gamma} \). An example of our clustering algorithm is illustrated in Fig. 1. We first randomly select \( \lceil \log n \rceil \) nodes from \( \Gamma \) (with replacement) as reference nodes, and it can be
shown (in Lemma 5 below) that each community contains at least one reference node with high probability. This set of reference nodes is denoted by $\Psi$. For each node $v \in \Psi$, we construct a ball $B_v$ with center $v$ and radius $r \triangleq \gamma^2 n \log n$ which includes all the neighboring nodes (i.e., the nodes in $B_v$). Among $\{B_v \mid v \in \Psi\}$, we find the one that has the largest cardinality, declare $v^*_1 \triangleq \arg \max_{v \in \Psi} |B_v|$, and set the largest community $\hat{V}^{(0)}_1$ to be $B_{v^*_1}$. To find the second largest community, we remove all the nodes in $\hat{V}^{(0)}_1$ and then follow a similar procedure to find the ball with the largest cardinality. That is, we declare $v^*_2 \triangleq \arg \max_{v \in \Psi} |B_v \setminus \hat{V}^{(0)}_1|$, and set the second largest community $\hat{V}^{(0)}_2$ to be $B_{v^*_2} \setminus \hat{V}^{(0)}_1$. By repeating this procedure for $2 \leq j \leq k$, we obtain $k$ estimated communities $\hat{V}^{(0)}_1, \hat{V}^{(0)}_2, \ldots, \hat{V}^{(0)}_k$ (lines 7–10). Furthermore, we assign the nodes belonging to $\Gamma \setminus (\bigcup_{j \in [k]} \hat{V}^{(0)}_j)$ to their nearest communities (lines 11–14), and the nodes that do not belong to $\Gamma$ to each community randomly (line 15). Finally, for each node $v \in \hat{V}^{(0)}_j$ (for all $j \in [k]$), we set $\hat{Z}^{(0)}_v = j$.

The high-level intuition of the analysis of Stage 1 is as follows. Let $M \triangleq E(L_T)$ be the expected trimmed hypergraph Laplacian, where $M$ is identical to $E(L)$ except that the rows and columns corresponding to nodes that do not belong to $\Gamma$ are set to zeros. Note that $M$ is an $n \times n$ matrix of rank at most $k$ when $(p, \{Q_T\}_{T \in T}) \notin \Xi$. If nodes $u$ and $v$ are in the same cluster, we have $M_u = M_v$; otherwise they are far apart in the sense that $|M_u - M_v|_2^2 = \Omega(\gamma^2 n)$ (as shown in Lemma 3 below). On the other hand, the sum of the distances between each column $(L_T^{(k)})_v$ and its expectation $M_{v}^{(k)}$ with high probability, where (a) follows from the Eckart-Young-Mirsky theorem and Weyl’s inequality, and (b) is proved by leveraging the random matrix theory (see Lemma 4 for details). By a careful analysis of Stage 1, one can show that each node $v$ will be misclassified only if the distance $(L_T^{(k)})_v - M_v^{(k)}_v$ is large enough, one can use the reference node in each community (denoted by $v$) to find most of its community members, via the ball $B_v$, with center $v$ and radius $r$. As a result, we only need to compute $\Theta(n \log n)$ pairwise distances $\|L_T^{(k)} - L_T^{(k)}\|^2_2$, while some related works [40] use all the $n$ nodes as reference nodes and thus $\Theta(n^2)$ pairwise distances need to be computed.

C. Local Refinements (Stage 2)

After obtaining the initial estimate $\hat{Z}^{(0)}$, we refine the label of each node $v \in [n]$ based on the observation of the hypergraph $G_2 = g_2$ as well as the estimated labels $\{\hat{Z}^{(0)}_v\}$ for the remaining nodes. For each node $v \in [n]$, we perform a local maximum a posteriori (MAP) estimation as follows:

$$\tilde{Z}_v(g_2, \hat{Z}^{(0)}_v) \triangleq \arg \max_{i \in [k]} \mathbb{P}(Z_v = i \mid G_2 = g_2, \hat{Z}^{(0)}_v = \tilde{Z}^{(0)}_v).$$

(19)

This leads to the final estimate $\tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_n)$ of the ground-truth community vector. A detailed analysis of Stage 2 is provided in Section V-B.

Remark 3: Instead of computing the posterior probability $\mathbb{P}(Z_v = i \mid G_2 = g_2, \hat{Z}^{(0)}_v = \tilde{Z}^{(0)}_v)$ directly, one can compute the probability $\mathbb{P}(G_2 = g_2 \mid Z_v = i, \hat{Z}^{(0)}_v = \tilde{Z}^{(0)}_v) \cdot p_i$ which is proportional to the posterior probability. Note that the sub-HSBM $G_2$ is generated on the hyperedge set $\mathcal{W}_2$ of the sub-hypergraph $F_2$ (as defined in Section IV-A due to graph splitting). The presence or absence of each hyperedge in $\mathcal{W}_2$ can be modeled by a Bernoulli random variable whose success probability is governed by $Z_v$ and $\hat{Z}^{(0)}_v$. Hence, the probability $\mathbb{P}(G_2 = g_2 \mid Z_v = i, \hat{Z}^{(0)}_v = \tilde{Z}^{(0)}_v)$ is essentially a product of $|\mathcal{W}_2|$ terms corresponding to the presence or not of hyperedges in $\mathcal{W}_2$, and each term equals either the “success probability” or “one minus the success probability” depending on whether the hyperedge appears in $g_2$.

Remark 4: It is straightforward to improve Algorithm 1 to an agnostic algorithm that does not require the knowledge of model parameters. Before performing the local refinement steps, one can estimate the distribution of communities $\tilde{p} = (p_1, \ldots, p_k)$ based on the estimated community vector $\hat{Z}^{(0)}$, and the hyperedge probabilities $\{Q_T\}_{T \in T}$ based on both $\hat{Z}^{(0)}$ and the hyperedges in the sub-hypergraph $G_2$. Due to the law of large numbers and the fact that $l(Z, \hat{Z}^{(0)}) = o(1)$, these estimates $\tilde{p}$ and $\{\tilde{Q}_T\}_{T \in T}$ are expected to be close to the true values $p$ and $\{Q_T\}_{T \in T}$ respectively. As a result, it can be shown that running the local refinement steps in Stage 2 still yields exact recovery. Furthermore, even if the number of communities $k$ is not given a priori, one can still apply a singular value thresholding method (as employed in [9]) to the hypergraph Laplacian $L_T$ to estimate the value of $k$ in Stage 1.

Remark 5: We note that the algorithm in [40] also relies on a hypergraph spectral clustering step plus a local refinement step. However, their hypergraph spectral clustering method is different from ours, with main distinctions described as follows.

- In our work, we use the entire rank-$k$ approximation of the trimmed hypergraph Laplacian $L_T^{(k)} = \sum_{i=1}^k \sigma_i u_i v_i^T$ as the input of our subsequent clustering step (lines 5-15 in Algorithm 1).
- The algorithm in [40] applies a singular value decomposition to $L_T$ to obtain the $k$ leading singular vectors $[u_1, \ldots, u_k]$, and then apply their subsequent clustering.

Note that it suffices to focus on hyperedges in $\mathcal{W}_2$ that contains node $v$ only, since the presence or absence of other hyperedges does not depend on which community node $v$ belongs to, thus has no influence on the decision rule in (19).
step by representing each of the $n$ node by a reduced $k$-dimensional vector.

The advantage of our algorithm is that its accompanying analysis does not involve the $k$-th largest singular value of $E(L)$, while the theoretical guarantee of the spectral clustering step in [40] depends on the $k$-th largest singular value. In the setting of [40] (described in Section III-D), their algorithm works only when the $k$-th largest singular value is reasonably large (as shown in [40, Lemma 5.1]). However, in our more general setting, the $k$-th largest singular value is not always large enough, which prevents the applicability of the algorithm in [40]. In contrast, our algorithm is applicable to a larger set of parameters, i.e., as long as the second-order degree profile condition is satisfied.

V. THEORETICAL GUARANTEES OF ALGORITHM 1

In this section, we prove that as long as 

$$\min_{n, j \in [k]: i \neq j} D_+(i, j) > 1$$

and 

$$(p, \{Q_T\}_{T \in \mathcal{T}}) \notin \Xi,$$ 

applying Algorithm 1 on the observed hypergraph $G$ ensures $l(Z, \tilde{Z}) = 0$ with high probability for sufficiently large $n$ (i.e., exact recovery).

First, we note that with high probability, the size of each community $V_j = \{v \in [n]: Z_v = j\}$ is close to $n p_j$, for all $j \in [k]$. This is stated in Lemma 1 below and can be proved by applying the Chernoff bound.

**Lemma 1:** Fix a constant $\delta \in (0, 1)$ which can be chosen to be arbitrarily small. We say that the length-$n$ vector $Z \in A_Z$ (where $A_Z$ is referred to as the typical set of $Z$) if the communities $\{V_j\}_{j \in [k]}$ associated with $Z$ satisfy

$$\left(1 - n^{-\frac{\delta}{2} + \frac{1}{2}}\right) p_j n \leq |V_j| \leq \left(1 + n^{-\frac{\delta}{2} + \frac{1}{2}}\right) p_j n, \quad \forall j \in [k].$$

(20)

Then, we have $P(Z \in A_Z) \geq 1 - \exp(-\Theta(n^\delta))$.

Therefore, one can focus on typical ground-truth community vectors $Z \in A_Z$ in the following analysis.

A. Theoretical Guarantees of Stage 1

For a fixed ground-truth community vector $Z \in A_Z$, we first introduce an artificial $d$-HSBM $G_1$ which is generated with respect to the ground-truth community vector $Z$ and hyperedge probabilities $\{\frac{Q_T \gamma_T}{\tau}\}_{T \in \mathcal{T}}$. Note that the generation process of $G_1$ is equivalent to first generating a sub-hypergraph $F_1$ (with splitting parameter $\gamma_n / \log n$) and then generating a sub-HSBM $G_1$ on the hyperedge set of $F_1$ (with hyperedge probabilities $\{\frac{Q_T \log n}{\tau}\}_{T \in \mathcal{T}}$). Thus, we investigate the misclassification proportion $l(z, \hat{Z}^{(0)})$ based on the random hypergraph $G_1$.

**Theorem 3 (Theoretical Guarantee of Stage 1):** Suppose the model parameters $(p, \{Q_T\}_{T \in \mathcal{T}}) \notin \Xi$. For any fixed $z \in A_Z$, there exist vanishing sequences $\{\epsilon_n\}$ and $\{\eta_n\}$ (which depend on $\{\gamma_n\}$) such that with probability at least $1 - \epsilon_n$ over the generation of $G_1$, running Stage 1 of Algorithm 1 ensures that $l(z, \hat{Z}^{(0)}) \leq \eta_n$, i.e., almost exact recovery is achieved.

Theorem 3 is proved in the rest of this section, and in the following we assume that $(p, \{Q_T\}_{T \in \mathcal{T}}) \notin \Xi$.

Note that with high probability over the generation of $G_1$, the degrees of most nodes in $[n]$ are smaller than the threshold $\tau$, thus only a vanishing fraction of nodes (at most $O(n^{1/\gamma_n})$ nodes) is trimmed. This result is adapted from [40, Lemma D.3] and stated below.

**Lemma 2 (Adapted From Lemma D.3 in [40]):** There exists a large constant $C > 0$ such that if we set $\tau = C J_{\max} \gamma_n$, then with probability at least $1 - \exp(-C' n)$ (for some constant $C' > 0$), the cardinality of the set $\Gamma$ satisfies $|\Gamma| \geq n(1 - (1/\tau))$.

Furthermore, let $V_j^\Gamma \triangleq V_j \cap \Gamma$ for all $j \in [k]$. Lemma 2 also implies that with high probability,

$$(1 - o(1)) p_j n \leq |V_j^\Gamma| \leq (1 + o(1)) p_j n, \quad \forall j \in [k].$$

We then focus on the remaining nodes in the set $\Gamma$. First recall that $M = E(L_T)$ is a matrix of rank at most $k$, and satisfies $M_u = M_v$ if $u$ and $v$ belong to the same community (where $u, v \in \Gamma$). Lemma 3 below shows that the distance between the two columns $M_u$ and $M_v$ scales as $\Omega(\gamma_n^2 / n)$ if $u$ and $v$ belong to different communities.

**Lemma 3:** Suppose $u, v \in \Gamma$. When the second-order degree profile condition is satisfied (i.e., $(p, \{Q_T\}_{T \in \mathcal{T}}) \notin \Xi$),

$$\|M_u - M_v\|_2 = 0 \text{ if } u \text{ and } v \text{ belong to the same community}, \text{ and } \|M_u - M_v\|_2 \geq \Omega(\gamma_n^2 / n) \text{ otherwise}.$$

The proof of Lemma 3 can be found in Appendix A. Lemma 4 below shows that with high probability, both the
operator norm and the Frobenius norm of the difference between $L_r$ and $M$ can be appropriately upper bounded.

**Lemma 4:** For any constant $C_1 > 0$, there exists some constant $C_2 > 0$ such that with probability at least $1 - n^{-C_1}$,

$$\left\| L_r - M \right\|_{op} \leq C_1 \cdot \left[ \sqrt{\gamma_n Q_{\max}^2 + \sqrt{\sigma_{k+1}}} + \frac{\gamma_n Q_{\max}}{\sqrt{\gamma_n Q_{\max} + 4 \sqrt{\sigma_{k+1}}}} \right].$$

As a result, the Frobenius norm between $M$ and $L_r^{(k)}$ (the rank-\(k\) approximation of $L_r$) can be upper-bounded as

$$\left\| L_r^{(k)} - M \right\|_F^2 \leq 8kC_2^2 \left[ \sqrt{\gamma_n Q_{\max}^2 + \sqrt{\sigma_{k+1}}} + \frac{\gamma_n Q_{\max}}{\sqrt{\gamma_n Q_{\max} + 4 \sqrt{\sigma_{k+1}}}} \right]^2 = O(\gamma_n).$$

**Proof of Lemma 4:** The proof of (21) can be adapted from [40, Lemma D.4], so our main focus is on the proof of (22). Note that

$$\left\| L_r^{(k)} - M \right\|_F^2 \leq 2k \cdot \left\| L_r^{(k)} - M \right\|_{op}^2 \leq 4k \cdot \left\| L_r^{(k)} - L_r \right\|_{op}^2 + 4k \cdot \left\| L_r - M \right\|_{op}^2$$

$$\leq 4k \cdot \sigma_{k+1}^2 + 4k \cdot \left\| L_r - M \right\|_{op}^2 \leq 8k \cdot \left\| L_r - M \right\|_{op}^2,$$

where $\sigma_{k+1}$ is the $(k + 1)$-th largest singular value of $L_r$. Eqn. (23) holds since the rank of $L_r^{(k)} - M$ is at most $2k$, and $\left\| X \right\|_{op} \leq r \cdot \left\| X \right\|_F$ for any matrix $X$ of rank $r$. Eqn. (24) holds since $\left\| L_r^{(k)} - M \right\|_{op} \leq \left\| L_r^{(k)} - L_r \right\|_{op} + \left\| L_r - M \right\|_{op}$ and $\left\| L_r^{(k)} - L_r \right\|_{op} \leq \left\| L_r^{(k)} - L_r \right\|_{op} + \left\| L_r - M \right\|_{op}$. Eqn. (25) follows from the Eckart-Young-Mirsky theorem [53], while Eqn. (26) is due to Weyl’s inequality [54]. Combining Eqns (21) and (26), it is then clear that

$$\left\| L_r^{(k)} - M \right\|_F^2 \leq 2k \cdot \left\| L_r^{(k)} - M \right\|_{op}^2 \leq 8kC_2^2 \left[ \sqrt{\gamma_n Q_{\max}^2 + \sqrt{\sigma_{k+1}}} + \frac{\gamma_n Q_{\max}}{\sqrt{\gamma_n Q_{\max} + 4 \sqrt{\sigma_{k+1}}}} \right]^2 = O(\gamma_n).$$

This completes the proof of Lemma 4. □

1) **Analysis of Stage 1:** In the following, we show that when Lemmas 2–4 hold, running Stage 1 ensures that \( I(\mathbf{z}, \mathbf{Z}^{(0)}) \to 0 \) with high probability. Our analysis is inspired by [9] for SBMs, but is adapted to our general \(d\)-HSBM problem. We first partition the nodes in $\Gamma$ as follows. Recall that the radius $r$ is set to be $\gamma_n^2/(n \log(\gamma_n))$, and let

$$T_j^{in} \triangleq \{ v \in V_j^* : \left\| (L_r^{(k)})_v - M_v \right\|_2^2 \leq r/4 \}, \ \forall j \in [k],$$

$$T_j^{out} \triangleq \{ v \in V_j^* : \left\| (L_r^{(k)})_v - M_v \right\|_2^2 \leq 4r \}, \ \forall j \in [k],$$

and the remaining nodes in $\Gamma$ belong to $U \triangleq \Gamma \setminus \bigcup_{j \in [k]} T_j^{out}$. Note that these sets have the following properties:

3Let $A$ and $B$ be $n \times n$ matrices, and their singular values are respectively denoted by $\{\sigma_i(A)\}_{i \in [n]}$ and $\{\sigma_i(B)\}_{i \in [n]}$, both in decreasing orders. Weyl’s inequality states that for every $i \in [n]$, $|\sigma_i(A) - \sigma_i(B)| \leq \|A - B\|_{op}$.

(i) For all $j \in [k]$, most of the nodes $v \in V_j^*$ are such that $(L_r^{(k)})_v$ is $\frac{r}{16}$-close to its expectation $M_v$, since

$$|V_j^* \setminus T_j^{in}| \leq \frac{\left\| (L_r^{(k)} - M) \right\|_F^2}{r/4} = O\left( \frac{\log(\gamma_n)}{\gamma_n^2 n} \right) = o(n).$$

Thus, $|T_j^{in}| \geq |V_j^* \cap T_j^{in}| = |V_j^*| - |V_j^* \setminus T_j^{in}| \geq p_j n(1 - o(1))$.

(ii) Similar to (i), we have

$$|U| \leq |\Gamma| \setminus \bigcup_{j \in [k]} T_j^{in} | \leq \|L_r^{(k)} - M\|_F^2/(r/4) = O\left( \frac{\log(\gamma_n)}{\gamma_n^2 n} \right) = o(n).$$

(iii) If node $v \in U \cap \Psi$, then $B_v \cap (\cup j \in [k] T_j^{in}) = \emptyset$, and as a result, we have $|B_v| = O\left( \frac{\log(\gamma_n)}{\gamma_n^2 n} \right) = o(n)$.

(iv) If node $v \in T_j^{in} \cap \Psi$, then $T_j^{in} \subseteq B_v$.

(v) If nodes $u \in T_i^{out}$ and $v \in T_j^{out}$ such that $i \neq j$, then $T_u \cap B_v = \emptyset$.

(vi) For all $j \in [k]$, $|T_j^{out}| \leq p_j (n(1 - o(1)))$. This is because

$$|T_j^{out}| \leq n - |\Gamma| - \sum_{j' \neq j} |T_j^{out}'| \leq n - \sum_{j' \neq j} |T_j^{out}'| \leq n - (p_j n (1 - o(1))) \leq p_j n (1 + o(1)).$$

Lemma 5 below states that with high probability, each community contains at least one reference node.

**Lemma 5:** With probability $1 - o(1)$ over the selection of $\Psi$, we have $|T_j^{in} \cap \Psi| \geq 1$ for all $j \in [k]$.

**Proof:** For any $j \in [k]$, the probability that a randomly selected reference node does not belong to $T_j^{in}$ is $1 - (|T_j^{in}|/n)$, thus the probability that there exists at least one reference node belongs to $T_j^{in}$ is

$$1 - \left( 1 - (|T_j^{in}|/n) \right)^{n} \log n = 1 - \exp(-\Theta(n/\log n)).$$

Taking a union bound over all $j \in [k]$, we complete the proof. □

With these properties, we are able to show that running Stage 1 yields almost exact recovery. For ease of presentation, in the following we focus mainly on the case when $p_1 > p_2 > \ldots > p_k$ (such that the community sizes satisfy $|V_1^*| > |V_2^*| > \ldots > |V_k^*|$ with high probability). The analysis can be easily generalized to the case when $p_i = p_j$ for some $i \neq j$, but it requires more cumbersome notations (such as permutations) which makes the subsequent analysis more difficult to understand.

The “**for Loop**” in Lines 7–10: When $j = 1$, we will show that $v_1^* \in T_1^{out}$ and $|B_{v_1^*}| \geq p_1 (1 - o(1))$. From Lemma 5, there exists a node $v_1 \in T_1^{in} \cap \Psi$, and its corresponding set $B_{v_1}$ is a superset of $T_1^{in}$ (due to (iv)). Thus, we have $|B_{v_1}| \geq |T_1^{in}| \geq p_1 n (1 - o(1))$, where the last inequality is due to (i). As $|B_{v_1}|$ is at least $|B_{v_1}|$, we obtain that $|B_{v_1}| \geq p_1 n (1 - o(1))$.

To prove that $v_1^* \in T_1^{out}$, one can verify that

1) For any $v \in U \cap \Psi$, $|B_v| = o(n) < |B_{v_1^*}|$ by (iii);
2) for any \( v \in T_j^\text{out} \cap \Psi \) where \( j \neq 1 \), \(|B_v| \leq |T_j^\text{out}| + |U| \leq np_j(1 + o(1)) + o(n) < |B_v^*| \) (due to (vi) and the fact that \( p_1 > p_j \)).

For \( 2 \leq j \leq k \), we will show that \( v_j^* \in T_j^\text{out} \) and \(|B_v^*| \geq np_j(1 - o(1)) \). Similar to the analysis above, there exists a node \( v_j \in T_j^\text{in} \cap \Psi \) and \( B_v \geq \sum_{l=0}^{1} V_l(0) \). From (v) we know that \(|B_v^*| \setminus (\cup_{l=0}^{1} V_l(0)) = |B_v|\), thus \(|B_v^*| \setminus (\cup_{l=0}^{1} V_l(0)) \geq |T_j^\text{in}| \geq np_j(1 - o(1)) \). As \(|B_v^*| \setminus (\cup_{l=0}^{1} V_l(0)) \) is at least \(|B_v^*| \setminus (\cup_{l=0}^{1} V_l(0)) \), we obtain that \(|B_v^*| \geq np_j(1 - o(1)) \).

To prove that \( v_j^* \in T_j^\text{out} \), one can verify that

1) for any \( v \in U \cap \Psi \), \(|B_v \setminus (\cup_{j=0}^{1} V_j(0))| = o(n);
2) for any \( v \in T_j^\text{out} \cap \Psi \) where \( j > j' \), \(|B_v \setminus (\cup_{l=0}^{1} V_l(0))| \leq |B_v| \leq |T_j^\text{out}| + |U| \leq np_j(1 + o(1)) + o(n) < |B_v^*| \) (due to (vi) and the fact that \( p_j > p_j' \));
3) for any \( v \in T_j^\text{out} \cap \Psi \) where \( j < j' \), we have \(|B_v \setminus (\cup_{l=0}^{1} V_l(0))| \leq |B_v| \leq |T_j^\text{out}| + |U| \leq np_j(1 - o(1)) \).

Therefore, under the assumption \( p_1 > p_2 > \cdots > p_k \), we have \( v_j^* \in T_j^\text{out} \) for all \( j \in \{1, \ldots, k\} \). This implies that all the elements in \( \{v_j^*\}_{j \in [k]} \) are far from each other—specifically, every pair of nodes \((v_j^*, v_{j'}^*)\) satisfies

\[
\| (L_{j'}(k))_{v_j} - (L_{j'}(k))_{v_{j'}} \|^2 \\
\geq \frac{1}{2} \| (M_{j'} - M_{j'})_{v_j} \|^2 - k \| M_{j'} - (L_{j'}(k))_{v_{j'}} \|^2 \\
\geq \frac{1}{2} \left\{ \frac{1}{2} \| M_{j'} - M_{j'} \|^2 - \| M_{j'} - (L_{j'}(k))_{v_{j'}} \|^2 \right\} \\
- \| M_{j'} - (L_{j'}(k))_{v_{j'}} \|^2 = \Omega(\gamma_n^2/n),
\]

since \( \| (L_{j'}(k))_{v_j} - M_{j'} \|^2 \leq \epsilon_r \| (L_{j'}(k))_{v_j} - M_{j'} \|^2 \leq 4 \epsilon_r \), \(|L_{j'}(k))_{v_j} - M_{j'} \|^2 \leq 4 \epsilon_r \) and \(|L_{j'}(k))_{v_{j'}} - M_{j'} \|^2 \leq \Omega(\gamma_n^2/n) \) by Lemma 3. If a node \( v \in \cap_j V_j^*(0) \) is misclassified to \( \cap_j V_j^*(0) \) (where \( i \neq j \)) in the first “for loop”, then it must be close to the center of \( \cap_j V_j^*(0) \), i.e., \( \| (L_j(0))_{v} - (L_j(0))_{v} \|^2 \leq r \). Thus, we have

\[
\| (L_j(0))_{v} - M_{v} \|^2 \\
\geq \frac{1}{2} \| M_{v} - M_{v} \|^2 - \| (L_j(0))_{v} - M_{v} \|^2 \\
\geq \frac{1}{2} \| M_{v} - M_{v} \|^2 - 2 \| (L_j(0))_{v} - (L_j(0))_{v} \|^2 \\
- 2 \| (L_j(0))_{v} - M_{v} \|^2 = \Omega(\gamma_n^2/n).
\]

The “for Loop” in Lines 11 – 14: Consider a specific node \( v \in \cap_j V_j^*(0) \). If \( v \in \cap_j V_j^*(0) \) is misclassified to \( \cap_j V_j^*(0) \) (where \( i \neq j \)) in the second “for loop”, then it must be closer to the center of \( \cap_j V_j^*(0) \) than the center of \( \cap_j V_j^*(0) \), i.e., \( \| (L_j(0))_{v} - (L_j(0))_{v} \|^2 \leq \| (L_j(0))_{v} - (L_j(0))_{v} \|^2 \). Since the two centers are far from each other, one can show that \( v \) is far from the center of \( \cap_j V_j^*(0) \), i.e.,

\[
\Omega(\gamma_n^2/n) \leq \| (L_j(k))_{v} - (L_j(k))_{v} \|^2 \\
\leq 2 \| (L_j(k))_{v} - (L_j(k))_{v} \|^2 + 2 \| (L_j(k))_{v} - (L_j(k))_{v} \|^2 \\
\leq 4 \| (L_j(k))_{v} - (L_j(k))_{v} \|^2.
\]

Thus, node \( v \) must satisfy

\[
\| (L_j(k))_{v} - M_{v} \|^2 \\
\geq \frac{1}{2} \| (L_j(k))_{v} - (L_j(k))_{v} \|^2 - \| M_{v} - (L_j(k))_{v} \|^2 = \Omega(\gamma_n^2/n).
\]

Combining Eqs. (28) and (29), we conclude that for any node \( v \in \Gamma \), if it is misclassified to another cluster, it must satisfy \( \| (L_j(k))_{v} - M_{v} \|^2 = \Omega(\gamma_n^2/n) \). Since \( \sum_{v \in \Gamma} \| (L_j(k))_{v} - M_{v} \|^2 = \| L_j(k) - M \|^2 = \Omega(n) \) (by Lemma 4), we know that the number of misclassified nodes in \( \Gamma \) is at most \( \Omega(n/\gamma_n) \). Taking into account the number of nodes that do not belong to \( \Gamma \) (which also scales as \( \Omega(n/\gamma_n) \) by Lemma 2), we complete the proof of Theorem 3.

B. Theoretical Guarantees of Stage 2

From Theorem 3 we know that for a fixed ground-truth community vector \( z \in A_Z \), running Stage 1 on \( G_1 \) ensures that \( l(z, \tilde{Z}^{(0)}) \leq \eta_n \) with probability at least \( 1 - \epsilon_n \). In the following, we show that the hypergraph spectral clustering method does not only work well on \( G_1 \), but also works well on the sub-HSBM \( G_1 \) (which is generated on the fixed sub-hypergraph \( F_1 \)) with high probability over the graph splitting process.

Definition 4: Let \((f_1, f_2)\) be the realizations of the sub-HSBMs \((F_1, F_2)\), and consider a fixed ground-truth community vector \( z \).

- We say \( f_1 \) is a good realization of the first sub-hypergraph with respect to \( z \) (denoted by \( f_1 \in G_1^* \)) if the probability that “running Stage 1 on \( G_1 \) (which depends on \( f_1 \)) ensures \( l(z, \tilde{Z}^{(0)}) \leq \eta_n \)” at least \( 1 - \sqrt{\epsilon_n} \), i.e.,

\( \mathbb{P}_{G_1}(l(z, \tilde{Z}^{(0)}) \leq \eta_n) \geq 1 - \sqrt{\epsilon_n} \).

- We say \( f_2 \) is a good realization of the second sub-hypergraph with respect to \( z \) (denoted by \( f_2 \in G_2^* \)) if for every node \( v \in [n] \) and every community assignment \( m \in M \), the number of hyperedges in \( f_2 \) that includes node \( v \) and other \( d - 1 \) nodes with community assignment \( m \) (denoted by \( D_{v,m} \)) satisfies

\( 1 - (3\gamma_n/\log n) \leq R_{v,m} \leq 1 + (\frac{1 - \gamma_n}{\epsilon_2 \gamma_n})R_{v,m} \).

Lemma 6: Suppose \( z \in A_Z \). With probability at least \( 1 - 2\sqrt{\epsilon_n} \), the randomly generated sub-hypergraphs \((F_1, F_2)\) satisfy \( F_1 \in G_1^* \) and \( F_2 \in G_2^* \) simultaneously.

Proof: See Appendix B.

Due to Lemma 6, one can then focus on a specific type of ground-truth community vector \( z \in A_Z \) and good realizations of the sub-hypergraphs \( f_1 \in G_1^* \) and \( f_2 \in G_2^* \) in the following. Also, one can suppose that the initial estimate \( \tilde{Z}^{(0)} \) (after Stage 1) satisfies \( l(z, \tilde{Z}^{(0)}) \leq \eta_n \), since running Stage 1 on \( G_1 \) (which depends on the good realization \( f_1 \)) ensures \( l(z, \tilde{Z}^{(0)}) \leq \eta_n \) with high probability. Without loss
of generality, we further assume that the minimum value of
\( l(z, \tilde{z}^{(0)}) = \min_{z \epsilon S_k} \sum_{v \epsilon [n]} 1 \{ z_v = \pi(\tilde{z}_v) \} \) is achieved by the identity permutation in \( S_k \), and this assumption helps
us to remove the presence of the permutation in the following analysis.

Note that the local MAP estimation in (19) for each node
\( v \epsilon [n] \) can be alternatively represented as

\[
\tilde{z}_v(g_2, \tilde{z}^{(0)}_{-v}) = \arg \max_{i \epsilon [k]} P \left( Z_v = i | G_2 = g_2, \tilde{z}^{(0)}_{-v} = \tilde{z}^{(0)}_{-v} \right) = \arg \max_{i \epsilon [k]} P \left( G_2 = g_2 | Z_v = i, \tilde{z}^{(0)}_{-v} = \tilde{z}^{(0)}_{-v} \right) \cdot p_i.
\]

Thus, for a fixed node \( v \epsilon V \) (i.e., \( z_v = i \) for some \( i \epsilon [k] \),
the probability that it is misclassified to a different community
can be bounded as follows:

\[
P \left( \tilde{z}_v(G_2; \tilde{z}^{(0)}_{-v}) \neq z_v \right) = P_{G_2} \left( \arg \max_{i \epsilon [k]} P \left( G_2 | Z_v = i, \tilde{z}^{(0)}_{-v} \right) \cdot p_i \neq i \right)
\]

\[
= P_{G_2} \left[ \exists j \neq i : P \left( G_2 | Z_v = j, \tilde{z}^{(0)}_{-v} \right) \sum_{z \epsilon M \in G} P \left( Z_v = i, \tilde{z}^{(0)}_{-v} \right) \right]
\]

\[
\leq \sum_{j \neq i} \sum_{z \epsilon M} P \left( G_2 | Z_v = j, \tilde{z}^{(0)}_{-v} \right) \cdot p_j \left( G_2 | Z_v = i, \tilde{z}^{(0)}_{-v} \right) \]

\[
\times \left\{ \frac{P \left( G_2 | Z_v = j, \tilde{z}^{(0)}_{-v} \right) \cdot p_j \geq P \left( G_2 | Z_v = i, \tilde{z}^{(0)}_{-v} \right) \cdot p_i \right\} \right].
\]

Note that the above error probability depends only on the
hyperedges in \( G_2 \) (which is generated on the sub-hypergraph
\( f_2 \)) that contains node \( v \). Recall that the number of hyperedges
in \( f_2 \) that comprise node \( v \) and other \( d - 1 \) nodes with
community assignment \( m \) is \( D_{v,m} \) (as defined in Definition 4),
thus these hyperedges can be represented by Bernoulli random
variables \( \{ X_m \}_{i=1}^{D_{v,m}} \), where \( X_m = 1 \) means the presence
of this hyperedge in \( G_2 \). Taking all possible \( m \epsilon M \) into
account, we know that the hyperedges in \( f_2 \) that contains
node \( v \) can be represented by the collection of Bernoulli
random variables \( \{ \{ X_m \}_{i=1}^{D_{v,m}} \}_{m \epsilon M} \). We also denote \( D_v = \sum_{m \epsilon M} D_{v,m} \) as the total number of hyperedges in \( f_2 \) that contains
\( v \), where \( D_v \) scales as \( \Theta(n^{d-1}) \) by Lemma 6.

To further upper bound (30), we would like to substitute the
terms \( P \left( g_2 | Z_v = i, \tilde{z}^{(0)}_{-v} \right) \) and \( P \left( g_2 | Z_v = j, \tilde{z}^{(0)}_{-v} \right) \) in the indicator
function by \( P \left( g_2 | Z_v = i, z_{-v} \right) \) and \( P \left( g_2 | Z_v = j, z_{-v} \right) \)
respectively; thus we can get rid of the fact that \( \tilde{z}^{(0)}_{-v} \)
not exactly the same as \( z_{-v} \). We now consider the ratio
between \( P \left( g_2 | Z_v = j, \tilde{z}^{(0)}_{-v} \right) \) and \( P \left( g_2 | Z_v = j, z_{-v} \right) \) for an
arbitrary \( j \epsilon [k] \). Due to the independence of hyperedges
\( \{ X_m \}_{i=1}^{D_{v,m}} \) in \( G_2 \), we have

\[
P \left( G_2 | Z_v = j, \tilde{z}^{(0)}_{-v} \right) = \prod_{m \epsilon M} \prod_{i=1}^{D_{v,m}} \left[ P \left( X_m | Z_v = j, \tilde{z}^{(0)}_{-v} \right) \right]
\]

\[
P \left( G_2 | Z_v = j, z_{-v} \right) = \prod_{m \epsilon M} \prod_{i=1}^{D_{v,m}} \left[ P \left( X_m | Z_v = j, z_{-v} \right) \right]
\]

\[
= \prod_{m \epsilon M} \prod_{i=1}^{D_{v,m}} \left[ P \left( X_m | Z_v = j, \tilde{z}^{(0)}_{-v} \right) \right]
\]

\[
\left\{ P \left( g_2 | Z_v = j, \tilde{z}^{(0)}_{-v} \right) \geq P \left( g_2 | Z_v = j, z_{-v} \right) \right\}
\]

\[
\sum_{g_2 \epsilon G_v} \left\{ P \left( g_2 | Z_v = j, z_{-v} \right) \right\} \leq n^{-\omega(1)}.
\]

Conditioned on \( Z_v = j \) and the ground truth \( z_{-v} \), the
hyperedge \( X_m \) includes node \( v \epsilon V_j \) and \( d - 1 \) other nodes with
community assignment \( m \), thus \( P \left( X_m = 1 | Z_v = j, z_{-v} \right) =
\frac{Q_{m,j} \log n}{n^{d-1}} \). On the other hand, conditioned on \( Z_v = j \) and
the estimated labels \( \tilde{z}^{(0)}_{-v} \), the hyperedge \( X_m \) is considered to
include node \( v \epsilon V_j \), but the community assignment of the
other \( d - 1 \) nodes, denoted by \( m' \epsilon M \), may not be equal to \( m \),
and we have \( P \left( X_m = 1 | Z_v = j, \tilde{z}^{(0)}_{-v} \right) = \frac{Q_{m',j} \log n}{n^{d-1}} \). Here,
\( m' \neq m \) if all the \( d - 1 \) nodes (except for \( v \) in this hyperedge)
are not misclassified in \( \tilde{z}^{(0)}_{-v} \), while \( m' = m \) otherwise.
For hyperedges such that their community assignments \( m' \)
(under \( \tilde{z}^{(0)}_{-v} \)) equal \( m \), they are cancelled out in (31) since the
distances are invariant conditioned on either \( z_{-v} \) or \( Z_{-v} \).
It then remains to focus on the hyperedges that contain at
least one misclassified node in \( z_{-v} \). Since \( l(z, \tilde{z}) \leq \eta_n \), the
number of such hyperedges is at most \( \eta_n n^{d-1} \). For ease of presentation,
these \( D' \) Bernoulli random variables in \( \{ \{ X_m \}_{i=1}^{D_{v,m}} \}_{m \epsilon M} \)
are alternatively relabelled as \( \{ Y_{a} \}_{a=1}^{D'} \).

Definition 5: Let \( \{ c_n \} \) be a vanishing sequence that satisfy
\( c_n \log(c_n / \eta_n) = \omega(1) \). We say \( G_2 \epsilon G_v \) if the random
variables \( \{ Y_{a} \}_{a=1}^{D'} \) satisfies \( \sum_{a=1}^{D'} Y_a \leq c_n \log n \), and \( G_2 \neq G_v \)
otherwise.

Lemma 7 below states that \( G_2 \epsilon G_v \) with high probability
(conditioned on the ground truth \( Z_v = i \) and \( z_{-v} \)), and for
every realization \( g_2 \epsilon G_v \), the ratio between \( P \left( g_2 | Z_v = j, \tilde{z}^{(0)}_{-v} \right) \)
and \( P \left( g_2 | Z_v = j, z_{-v} \right) \) is bounded for any \( j \epsilon [k] \).

Lemma 7: Suppose the initial estimate \( \tilde{z}^{(0)} \) satisfies
\( l(z, \tilde{z}^{(0)}) \leq \eta_n \). The probability that \( G_2 \) does not belong to
\( G_v \) is at most \( n^{-\omega(1)} \), i.e.,

\[
P \left( G_2 \neq G_v | Z_v = i, z_{-v} \right)
\]

\[
= \frac{\sum_{a=1}^{D'} Y_a > c_n \log n | Z_v = i, z_{-v}}{P \left( g_2 | Z_v = i, z_{-v} \right)} \leq n^{-\omega(1)}.
\]

Furthermore, there exist \( L_h = n^{o(1)} \) and \( L_1 = n^{o(1)} \) such that for all \( g_2 \epsilon G_v \) and \( j \epsilon [k] \),

\[
L_h \leq \frac{P \left( g_2 | Z_v = j, \tilde{z}^{(0)}_{-v} \right)}{P \left( g_2 | Z_v = j, z_{-v} \right)} \leq L_h.
\]

Proof: See Appendix C.
Lemma 8 below bounds the term
\[
\sum_{g_2} \min \left\{ \mathbb{P} \left( g_2 \mid Z_v = i, z_{\sim v} \right), \mathbb{P} \left( g_2 \mid Z_v = j, z_{\sim v} \right) \right\}
\]
for any \( i \neq j \) in terms of the GCH-divergence \( D_+(i,j) \) between communities \( V_i \) and \( V_j \).

**Lemma 8:** When the ground-truth community vector \( z \in A_Z \) and the sub-hypergraphs \( f_1 \in G^z_1 \) and \( f_2 \in G^z_2 \), we have
\[
\sum_{g_2} \min \left\{ \mathbb{P} \left( g_2 \mid Z_v = i, z_{\sim v} \right), \mathbb{P} \left( g_2 \mid Z_v = j, z_{\sim v} \right) \right\} 
\leq n^{-D_+(i,j)+o(1)},
\]
where \( D_+(i,j) = \max_{t \in [0,1]} \sum_{m \in M} t \mu_{m \oplus i} + (1-t) \mu_{m \oplus j} - \mu_{m \oplus i} 1^{t-1} \mu_{m \oplus j} \).

**Proof of Lemma 8:** As \( G^z \) is equivalent to the collection of random variables \( \{X^m_{a_1} \}_{a_1 \in M} \), we have
\[
\sum_{g_2} \min \left\{ \mathbb{P} \left( g_2 \mid Z_v = i, z_{\sim v} \right), \mathbb{P} \left( g_2 \mid Z_v = j, z_{\sim v} \right) \right\} 
= \sum_{m \in M} \sum_{a_1 \in \{0,1\}} \min \left\{ \prod_{a|m} \mathbb{P} \left( x^m_{a_1} \mid Z_v = i, z_{\sim v} \right), \prod_{a|m} \mathbb{P} \left( x^m_{a_1} \mid Z_v = j, z_{\sim v} \right) \right\} 
\leq \sum_{m \in M} \sum_{a_1 \in \{0,1\}} \prod_{a|m} \mathbb{P} \left( x^m_{a_1} \mid Z_v = i, z_{\sim v} \right) 
\times \left( \prod_{a|m} \mathbb{P} \left( x^m_{a_1} \mid Z_v = j, z_{\sim v} \right) \right)^{1-t} \tag{32}
\]
\[
= \prod_{m \in M} \sum_{a_1 \in \{0,1\}} \mathbb{P} \left( x^m_{a_1} \mid Z_v = i, z_{\sim v} \right)^t 
\times \mathbb{P} \left( x^m_{a_1} \mid Z_v = j, z_{\sim v} \right)^{1-t} \tag{33}
\]
\[
= \prod_{m \in M} \left[ \left( \frac{Q_{m \oplus i} \log n}{n^{d-1}} \right)^t \left( \frac{Q_{m \oplus j} \log n}{n^{d-1}} \right)^{1-t} 
+ \left( 1 - \frac{Q_{m \oplus i} \log n}{n^{d-1}} \right)^t \left( 1 - \frac{Q_{m \oplus j} \log n}{n^{d-1}} \right)^{1-t} \right] \tag{34}
\]
where (32)-(34) hold for any \( t \in [0,1] \). By applying a Taylor series expansion, we have
\[
\exp \left\{ \sum_{m \in M} D_{v,m} \cdot \log \left[ 1 - t \frac{Q_{m \oplus i} \log n}{n^{d-1}} \right] - (1-t) \frac{Q_{m \oplus j} \log n}{n^{d-1}} \right\} + \left( \frac{Q_{m \oplus i} \log n}{n^{d-1}} \right)^t \left( \frac{Q_{m \oplus j} \log n}{n^{d-1}} \right)^{1-t} \tag{36}
\]
\[
= \exp \left\{ - (\log n) \sum_{m \in M} \frac{D_{v,m}}{n^{d-1}} \cdot \left( t Q_{m \oplus i} + (1-t) Q_{m \oplus j} - Q_{m \oplus i} 1^{t-1} Q_{m \oplus j} + O \left( \frac{(\log n)^2}{n^{2d-2}} \right) \right) \right\} \tag{37}
\]
\[
\leq n^{-R_m (t Q_{m \oplus i} + (1-t) Q_{m \oplus j} - Q_{m \oplus i} 1^{t-1} Q_{m \oplus j}) + O(\gamma_n/\log n)}, \tag{38}
\]
which (35) one needs to be aware that the lower order term has a negative coefficient. Thus, one can rewrite (34) as (36)-(38) shown at the bottom of the page, where (38) is due to the fact that \( (1-(3\gamma_n/\log n)) R_m \leq D_{v,m} \leq (1+n^{-\frac{1}{2}+\delta}) R_m \). Since (38) is valid for any \( t \in [0,1] \) and recall that \( \mu_{m \oplus i} = R_{m \oplus i} Q_{m \oplus i} \), we eventually obtain that
\[
\sum_{g_2} \min \left\{ \mathbb{P} \left( g_2 \mid Z_v = i, z_{\sim v} \right), \mathbb{P} \left( g_2 \mid Z_v = j, z_{\sim v} \right) \right\} 
\leq n^{-\max_{t \in [0,1]} \sum_{m \in M} t \mu_{m \oplus i} + (1-t) \mu_{m \oplus j} - \mu_{m \oplus i} 1^{t-1} \mu_{m \oplus j} + o(1)} \tag{39}
\]
which completes the proof of Lemma 8. \( \square \)

Since \( \min_{i, j \in [k]: i \neq j} D_+ (i,j) > 1 \), there exists an \( \varepsilon > 0 \) such that \( D_+ (i,j) > 1 + \varepsilon \) for all \( i \neq j \). By also noting that \( L_u/L_i = n^{o(1)} \), we have that
\[
\sum_{g_2} \frac{p_i L_u}{p_i L_i} \sum_{g_2} \min \{ \mathbb{P} \left( g_2 \mid Z_v = i, z_{\sim v} \right), \mathbb{P} \left( g_2 \mid Z_v = j, z_{\sim v} \right) \} 
\leq n^{-1+\varepsilon} + o(1) \tag{40}
\]
Thus, one can bound the error probability for node \( v \) from above as
\[
\mathbb{P} \left( \tilde{Z}_v \neq z_v \right) \leq n^{-(1+\varepsilon)+o(1)} + n^{-w(1)}. \tag{41}
\]
Note that the above analysis is also valid for nodes that belong to any other communities (not necessarily community $V_j$), thus one can take a union bound over all the $n$ nodes to obtain that
\[
P\left(\exists v \in [n] : \tilde{Z}_v \neq z_v\right) = P\left(\tilde{Z} \neq z\right) \leq n^{-\epsilon/2}
\]
when $n$ is sufficiently large. This means that all the nodes can be recovered correctly with probability at least $1 - n^{-\epsilon/2}$.

C. The Overall Success Probability

Let $\mathcal{E}_{\text{suc}}$ be the event that $l(z, \tilde{Z}) = 0$. From the analysis of Stage 2, we know that for all $z \in A_Z$, $f_1 \in G_{f_1}^1$, $f_2 \in G_{f_2}^1$, and $\tilde{z}^{(0)}$ satisfying $l(z, \tilde{Z}) \leq \eta_n$,
\[
P(\mathcal{E}_{\text{suc}}|z, f_1, f_2, \tilde{z}^{(0)}) \geq 1 - n^{-\epsilon/2}. \tag{39}
\]
Therefore, the overall success probability is
\[
P\left(l(Z, \tilde{Z}) = 0\right) = \sum_z P(z) \sum_{f_1, f_2} P(f_1, f_2) \sum_{\tilde{z}^{(0)}} P(\mathcal{E}_{\text{suc}}|z, f_1, f_2, \tilde{z}^{(0)}) \geq \sum_z P(z) \sum_{f_1, f_2} P(f_1, f_2) \sum_{\tilde{z}^{(0)}} P(\mathcal{E}_{\text{suc}}|z, f_1, f_2, \tilde{z}^{(0)}) \times \sum_{\tilde{z}^{(0)}: l(\tilde{z}^{(0)}) \leq \eta_n} \sum_z P(z), f_1, f_2, \tilde{z}^{(0)}) \geq (1 - \exp(-\Theta(n^{\delta})))(1 - 2\sqrt{e}) \cdot (1 - \sqrt{\epsilon_n}) \cdot (1 - n^{-\epsilon/2}). \tag{40}
\]
where inequality (40) follows from Lemma 1, Lemma 6, the definition of good realization $f_1$ in Definition 4, and Eqn. (39). This means that exact recovery is achievable.

VI. PROOF OF CONVERSE (THEOREM 1)

In this section, we show that when the model parameters $(p, (Q_T)_{T \in \mathcal{G}})$ satisfy $\min_{i,j \in [k]: i \neq j} D_{\star}(i, j) < 1$, exact recovery is impossible. This converse proof is inspired by that for the SBM [4], but is adapted to the $d$-HSBM setting.

First, we recall from Lemma 1 that with high probability the number of nodes in each community $V_j$ is tightly concentrated around the expectation $n_{ij}$ (i.e., the ground-truth community vector $Z \in A_Z$). Hence, we consider a fixed $z \in A_Z$ from now on. Let $S$ be a random set that contains $n/(\log n)^3$ randomly selected nodes from $[n]$. By applying the Chernoff bound, we can show that with probability $1 - \exp(-\Theta(n^{\delta}))$ over the selection process, the number of nodes in both $V_j$ and $S$, denoted by $|V_j|$ and $S_j$, satisfies that $V_j \in [k],$
\[
(1 - n^{-\epsilon/2})^2 \frac{n_p}{(\log n)^3} \leq |V_j| \leq (1 + n^{-\epsilon/2})^2 \frac{n_p}{(\log n)^3}, \tag{41}
\]
and thus the number of nodes in both $V_j$ and $S_j$, denoted by $V_j^{S_j}$, satisfies that $\forall j \in [k],$
\[
np_j \left(1 - \frac{1}{(\log n)^3} - 2n^{-\epsilon/2}\right) \leq |V_j^{S_j}| \leq np_j \left(1 - \frac{1}{(\log n)^3} + 2n^{-\epsilon/2}\right), \tag{42}
\]
for sufficiently large $n$. We then consider a fixed set $S$ that satisfies (41) and (42). Let
\[
f_{i, j}(t) \triangleq \sum_{\mu_{m^{\oplus}i}} (1 - t)\mu_{m^{\oplus}j} - \mu_{m^{\oplus}i} - \mu_{m^{\oplus}j},
\]
and note that $D_+(i, j) = \max_{i \in [n]} f_{i, j}(t)$. To obtain the maximizer of $f_{i, j}(t)$, we set $f_{i, j}(t) = 0$ and this implies
\[
\sum_{t \in M} \mu_{m^{\oplus}i} |1 - t| \mu_{m^{\oplus}j} \log \left(\frac{\mu_{m^{\oplus}i}}{\mu_{m^{\oplus}j}}\right) = \sum_{t \in M} \mu_{m^{\oplus}i} - \mu_{m^{\oplus}j}, \tag{43}
\]
Let $\tau_{m}^{i, j} \triangleq |\mu_{m^{\oplus}i}^{1 - t} - \mu_{m^{\oplus}j}^{1 - t}| \log n$ for each $m \in M$, where $t^*$ is set to satisfy (43).

Definition 6: For each node $v \in S$, let $N_v$ denote the number of hyperedges that contains $v$ and other $d - 1$ nodes from $[n]\setminus S$ that have community assignment $m \in M$. A node $v \in S$ is said to be ambiguous if $N_v = \tau_{m}^{i, j}$ for all $m \in M$.

In the following, we show that if $D_{\star}(i, j) < 1$, there is at least one ambiguous node in $V_j$ and one ambiguous node in $S_j$. For a node $v \in V_j$, $N_v$ equals the sum of $R_m$ i.i.d. Bernoulli random variables $\{B_m[i]\}_{i \in [1]}$ with expectation $Q_m^{i, j} \log n$, where $R_m = \sum_{i \in [1]} (1/^{m^{i, j}}_{m^{i, j}})$. Due to (41) and (42), one can show that
\[
\left(1 - \frac{1}{(\log n)^3} - 2n^{-\epsilon/2}\right) \leq \frac{R_m}{R_m} \leq \left(1 - \frac{1}{(\log n)^3} + 2n^{-\epsilon/2}\right). \tag{44}
\]

Following [55, Exercise 2.2], one can show that the probability of $N_v = \tau_{m}^{i, j}$ is
\[
P\left(\sum_{r=1}^{R_m} B_m[i] = \tau_{m}^{i, j}\right) = \exp\left(-\frac{1}{2\log(2\pi R_m)} \log \left(1 - \frac{R_m Q_m^{i, j} \log n}{R_m} n^{-d-1}\right) + \frac{1}{2\log(2\pi R_m)} \log \left(1 - \frac{R_m Q_m^{i, j} \log n}{R_m} n^{-d-1}\right) - C_2\right). \tag{45}
\]
where $C_2 > 0$ is a constant. By using a Taylor series expansion and the fact that $R_m/R_m$ is bounded (as shown in (44)), we have
\[
\tilde{R}_m \left(\frac{R_m Q_m^{i, j} \log n}{R_m} n^{-d-1}\right) = R_m \left(\frac{R_m Q_m^{i, j} \log n}{R_m} n^{-d-1}\right) + \tilde{R}_m \left(1 - \frac{R_m Q_m^{i, j} \log n}{R_m} n^{-d-1}\right) \times \log \left(1 + \frac{R_m Q_m^{i, j} \log n}{R_m} n^{-d-1} - Q_m^{i, j} \log n\right) \log \left(1 + \frac{R_m Q_m^{i, j} \log n}{R_m} n^{-d-1} - Q_m^{i, j} \log n\right),
\]

show that for a node

\[ v \]

Combining (45), (46), and (47), we then have

\[
\begin{align*}
\mathbb{P}(N_v^m &= t^{i,j}_{m} \\
&= n - \left(1 - t^*\right)\mu_{m^{i,j}}^{1-t} + \mu_{m^{i,j}}^{1-t} + o(1) \\
&= n^{-[D_+(i,j) + o(1)]},
\end{align*}
\]

where (48) is due to the independence of \( \{N_v^m\}_{m \in \mathcal{M}} \) and (49) holds since \( t^* \) satisfies (43). Similarly, one can also show that for a node \( v \in \mathcal{V}_S \),

\[
\mathbb{P}(v \in \mathcal{V}_S^j \text{ is ambiguous})
\]

\[
\begin{align*}
&= \mathbb{P}(\gamma_m \in \mathcal{M} : N_v^m = t^{i,j}_{m}) \\
&= n^{-[D_+(i,j) + o(1)]}.
\end{align*}
\]

By noting that \( D_+(i,j) < 1 \) and the cardinalities of both \( \mathcal{V}_S \) and \( \mathcal{V}_S^j \) scale as \( \Theta(n/(\log n)^3) \), one can show that with probability \( 1 - o(1) \), there is at least one ambiguous node in \( \mathcal{V}_S \) (denoted by \( v_1 \)) and also one ambiguous node in \( \mathcal{V}_S^j \) (denoted by \( v_2 \)).

In addition, we prove that with high probability, node \( v_1 \) (resp. \( v_2 \)) is not connected to any node in \( S \). This is because the number of hyperedges that contains \( v_1 \) (resp. \( v_2 \)) and another node in \( S \) is at most \( |S| \binom{d-2}{d-2} \leq \frac{n^{d-1}}{(\log n)^3} \), and the probability of each hyperedge is at most \( Q_{\max}(\log n)/n^{d-1} \),

Thus the probability that \( v_1 \) (resp. \( v_2 \)) does not have any connection with other nodes in \( S \) is at least

\[
\left(1 - Q_{\max}\frac{\log n}{n^{d-1}}\right) \geq e^{-\frac{2Q_{\max}(\log n)^2}{n^{d-1}}} \geq 1 - 4Q_{\max}\frac{\log n}{n^{d-1}},
\]

for sufficiently large \( n \). Finally, note that both \( v_1 \) and \( v_2 \) are not connected to any node in \( S \), and both of them are ambiguous (i.e., have the same number of hyperedges \( \{N_v^m\}_{m \in \mathcal{M}} \) outside \( S \)), thus it is impossible to distinguish them and to achieve exact recovery.

VII. Conclusion, Discussions, and Future Directions

This paper establishes a sharp phase transition for exact recovery in the general \( d \)-HSBM, apart from a small subset of generative distributions such that there exists two communities with the same second-order degree profiles. We also develop a polynomial-time algorithm (with theoretical guarantees) that achieves the information-theoretic limit, showing that there is no information-computation gap. Our two-stage algorithm is based on hypergraph spectral clustering and local refinement steps.

Next, we discuss some connections between our results and related works.

1) The second-order degree profile condition for our algorithm to succeed is milder than the conditions of several existing hypergraph spectral clustering methods, e.g., [29], [40], which typically require the \( k \)-th largest singular value of the expected hypergraph Laplacian \( \mathbb{E}(L) \) to be sufficiently large (referred to as the singular value condition below). Thus, our achievability result (Theorem 2) is applicable to a larger set of parameters.

To be specific:

- When the second-order degree profile condition is violated, there must exist two communities having the same second-order degree profile, which implies that the columns corresponding to these two communities in \( \mathbb{E}(L) \) are the same. Thus, the rank of \( \mathbb{E}(L) \) is less than \( k \) and the \( k \)-th largest singular value equals zero. This means that the singular value condition is also violated.

- When the singular value condition is violated, it does not necessarily imply that the second-order degree profile condition is violated. For example, suppose node \( u \in \mathcal{V}_1 \), node \( v \in \mathcal{V}_2 \), and their corresponding columns in \( \mathbb{E}(L) \) satisfy \( \mathbb{E}(L)_{uv} = 2\mathbb{E}(L)_{ee} \), then the rank of \( \mathbb{E}(L) \) is less than \( k \) and the \( k \)-th largest singular value is zero (i.e., the singular value condition is violated). However, since the columns corresponding to \( V_1 \) and \( V_2 \) are different (though they are linearly dependent), the second-order degree profiles of \( V_1 \) and \( V_2 \) are different, which does not imply that the second-order degree profile condition is violated.

2) Another work that is closely related to ours is [56], which considered the fundamental question of whether communities exist or not in a hypergraph. They characterized the condition under which a hypergraph generated
according to a $d$-HSBM can be successfully distinguished from a hypergraph generated according to an Erdős-Rényi hypergraph model, where the $d$-HSBM contains $k$ equal-sized communities and the hyperedge probabilities are assumed to be either $a_n$ or $b_n$ (depending on whether all $d$ nodes belong to a same community). Their main messages are that (i) when $a_n,b_n \in o(n^{-d+1})$, these two models are indistinguishable; (ii) when $a_n,b_n \in \omega(n^{-d+1})$, their proposed test ensures the two models to be distinguishable with probability approaching one; (iii) when $a_n,b_n \in \Theta(n^{-d+1})$, the two models are distinguishable if the so-called SNR is greater than a certain threshold, while indistinguishable if the SNR is below another threshold. Comparing [56] with our work, it is interesting to note that the phase transition occurs in the constant average degrees regime for detecting the existence of communities [56], while the phase transition occurs in the logarithmic average degrees regime for exactly recovering communities.

3) Community detection in hypergraphs is also related to the planted $k$-SAT problem [57], in which the objective is to identify a planted assignment $\sigma \in \{\pm 1\}^n$ of $n$ Boolean variables $\{x_1,x_2,\ldots,x_n\}$ given a sequence of randomly generated $k$-clauses, where each $k$-clause is a collection of $k$ distinct elements chosen from $\{x_1,x_2,\ldots,x_n\}$ and their negations $\{\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_n\}$. Let $X_k$ be the set of all $k$-clauses (with $|X_k| = \binom{n}{k}$), and $Q : \{\pm 1\}^{k} \rightarrow [0,1]$ be a probability distribution on $\{\pm 1\}^k$ such that $\sum_{\mathbf{y} \in \{\pm 1\}^k} Q(\mathbf{y}) = 1$. At each time when we generate a $k$-clause, the probability of a $k$-clause $c = [c_1,\ldots,c_k]$ (where $c_i \in \{x_1,x_2,\ldots,x_n\} \cup \{\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_n\}$) being selected is $P(c \text{ is selected}) = \frac{Q(\sigma(c))}{\sum_{c' \in \{\pm 1\}^k} Q(\sigma(c'))}$, where $\sigma(c) \in \{\pm 1\}^k$ is the assignment of the $k$ elements in $c$ under the assignment $\sigma$. In the planted $k$-SAT problem, we generate $M$ independent $k$-clauses, and the question of interest is to find how many clauses $M$ are required for successful recovery of the assignment $\sigma$ with high probability. This planted $k$-SAT problem can be viewed as a random hypergraph $(\mathcal{V},\mathcal{E})$, where the node set $\mathcal{V} = \{x_1,x_2,\ldots,x_n\} \cup \{\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_n\}$ is of size $2n$, and the edge set $\mathcal{E}$ contains $M$ $k$-uniform hyperedges with each one corresponding to a randomly generated $k$-clause. The nodes in $\mathcal{V}$ are partitioned into two communities that correspond to ‘+1’ and ‘−1’, where the two communities are of exactly equal sizes by construction. While the planted $k$-SAT problem can be approximately viewed as the HSBM problem studied in this work, there are also several notable differences. First, the generation process of $k$-clauses is different from the generation process of hyperedges in the HSBM—the former allows each $k$-clause to be selected for multiple times, while the latter only allows each $k$-uniform hyperedge to be selected once. Second, the assignments of nodes in the planted $k$-SAT problem are strongly correlated, e.g., the signs of $x_i$ and $\bar{x}_i$ must be different, while there is no such restriction in the HSBM.

Despite the differences, our algorithm is applicable to the planted $k$-SAT problem. One can first convert the $k$-SAT problem to a hypergraph with $2n$ nodes and $M$ $k$-uniform hyperedges, construct the corresponding trimmed hypergraph Laplacian, and then apply our spectral clustering method (lines 2-16 in Algorithm 1) to obtain an initial assignment $\hat{\sigma}^{(0)}$ of nodes $\{x_1,x_2,\ldots,x_n\} \cup \{\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_n\}$. It is expected that this stage leads to an almost exact recovery of the true assignment $\sigma$, as long as $M = \omega(n)$ (corresponding to hyperedge probabilities being $\omega(1/n^{-k-1})$ in the HSBM). In the second stage, one can use the local MAP estimation for each of the $2n$ nodes (lines 17-20 in Algorithm 1) to refine the assignments. However, since $x_i$ and $\bar{x}_i$ in the planted $k$-SAT problem are of different signs, one can instead choose to refine each pair $(x_i,\bar{x}_i)$ jointly via the local MAP estimation, which may lead to a better performance.

It is expected that when the probability distribution $Q : \{\pm 1\}^k \rightarrow [0,1]$ in the planted $k$-SAT problem is specialized to a simple function whose values depend only on the number of ‘+1’ in the input (in which case the distribution $Q$ is equivalent to the hyperedge probabilities $Q(\mathbf{X}) \mid \mathbf{T} \in \mathcal{T}$ in the HSBM), the second stage leads to exact recovery of the true assignment with high probability if $M \geq \Theta(n \log n)$ (corresponding to the logarithmic average degrees regime in the HSBM). We also expect that the GCH-divergence plays a role in the minimum pre-constant of $\Theta(n \log n)$; however, this pre-constant may not be obtained as a direct consequence of our result due to the several important differences between the planted $k$-SAT problem and HSBM.

Finally, we put forth two promising directions for future work.

1) Our algorithm fails if the parameters belong to $\Xi$ because we apply the hypergraph spectral clustering method to the processed hypergraph Laplacian $L$ (rather than the observed adjacency tensor $A$). This pre-processing step from $A$ to $L$ annihilates some salient information for distinguishing two communities with the same second-order degree profile. Thus, any clustering algorithms that rely merely on $L$ must be restricted to this second-order degree profile condition. On the other hand, we conjecture that the second-order degree profile condition is not necessary, and this issue may be circumvented if one directly applies clustering algorithms to the adjacency tensor $A$ (such as the tensor-based method proposed in [37]). As shown empirically in [37, Section 3.4] (particularly in Figure 3), their method avoids unwanted information loss caused by projecting hypergraphs to weighted graphs under a variety of parameter settings. Unfortunately, their concentration tools for random tensors are only applicable when the average degree is $\omega(\log^2(n))$, and thus are not powerful enough for the logarithmic average degrees regime considered in this paper. The analysis in [37] of tensor concentration relies on the notion of the incoherent
tensor operator norm}, the properties of the tensor, as well as concentration inequalities such as the Bernstein’s inequality and Chernoff bound. In contrast, the concentration of random matrices is relatively well understood, and the analysis in this paper relies mainly on techniques from random matrix theory. In future work, it is interesting to investigate whether tensor-based methods can be applied to hypergraphs with logarithmic average degrees, and to validate whether our conjecture that the exact recovery threshold \( \min_{i,j \in [k]: i \neq j} D_+(i, j) = 1 \) holds even without the condition on the second-order degree profile discussed in Section III-A.

2) It would also be interesting to extend our theory to even more general settings and other variants of the HSBM, such as the non-uniform HSBM (as proposed in [27], [28], and [29]), HSBM with overlapping communities, weighted or labelled HSMBs, HSBM with side information, etc.

**APPENDIX A**

**PROOF OF LEMMA 3**

Without loss of generality, we assume nodes \( u \) and \( v \) respectively belong to communities \( \mathcal{V}_i \) and \( \mathcal{V}_j \). Since we require \( (p, \{Q_T\}_{T \in T}) \notin \Xi \), there must exist a \( s \in [k] \) such that

\[
\sum_{m \in M: m_{s,i} \geq 1} m_{s,i} \mu_{m_{s,i}+1}^{m_{s,i}} \neq \sum_{m \in M: m_{s,j} \geq 1} m_{s,j} \mu_{m_{s,j}+1}^{m_{s,j}}.
\]  

(51)

Let \( R_{m,s} \triangleq \left( \frac{np_s}{m_{s-1}} \right) \cdot \prod_{a \in [k]: \{s\}} \left( \frac{np_a}{m_a} \right) \). Thus, we have

\[
\| M_u - M_v \|^2 \geq \sum_{w \in \mathcal{V}_i} (M_{w,u} - M_{w,v})^2
\]

\[
= \sum_{w \in \mathcal{V}_i} \left( \sum_{m \in M: m_{s,i} \geq 1} R_{m,s} Q_{m_{s,i}+1} \frac{\gamma_n}{n^{d-1}} - \sum_{m \in M: m_{s,j} \geq 1} R_{m,s} Q_{m_{s,j}+1} \frac{\gamma_n}{n^{d-1}} \right)^2
\]

\[
= \sum_{w \in \mathcal{V}_i} \left( \frac{\gamma_n}{np_s} \left[ \sum_{m \in M: m_{s,i} \geq 1} m_{s,i} \mu_{m_{s,i}+1}^{m_{s,i}} - m_{s,j} \mu_{m_{s,j}+1}^{m_{s,j}} \right] \right)^2
\]

\[
= \Omega \left( \frac{\gamma_n^2}{n} \right).
\]

(52)

(53)

where (52) holds since \( R_{m,s} = m_s R_m/(np_s) \) and \( R_m Q_{m_{s,i}+1}/n^{d-1} = \mu_{m_{s,i}+1} \), and (53) follows from (51) and \( |\mathcal{V}_i| \approx np_s \).

**APPENDIX B**

**PROOF OF LEMMA 6**

For any realization \( F_1 = f_1 \), let \( P_{\text{suc}}(f_1, z) \) be the probability that running a hypergraph spectral clustering method on \( G_T \) (which depends on \( f_1 \)) ensures \( I(z, Z^{(0)}) \leq \eta_n \). From Theorem 3, we have

\[
\sum_{f_1} P(F_1 = f_1) P_{\text{suc}}(f_1, z) \geq 1 - \epsilon_n.
\]  

(54)

We now prove Lemma 6 by contradiction. Suppose the probability that \( F_1 \in \mathcal{G}_T \) is less than \( 1 - \sqrt{\epsilon_n} \), then we have

\[
\sum_{f_1} \mathbb{P}(F_1 = f_1) P_{\text{suc}}(f_1, z)
\]

\[
< \sum_{f_1 \in \mathcal{G}_T} \mathbb{P}(F_1 = f_1) + \sum_{f_1 \notin \mathcal{G}_T} \mathbb{P}(F_1 = f_1)(1 - \sqrt{\epsilon_n})
\]

\[
= \sum_{f_1 \in \mathcal{G}_T} \mathbb{P}(F_1 = f_1) + (1 - \sqrt{\epsilon_n}) \left( 1 - \sum_{f_1 \notin \mathcal{G}_T} \mathbb{P}(F_1 = f_1) \right)
\]

\[
< 1 - \sqrt{\epsilon_n} + (1 - \sqrt{\epsilon_n}) \cdot \sqrt{\epsilon_n}
\]

\[
= 1 - \epsilon_n,
\]

(55)

(56)

(57)

where (55) follows from the fact that \( P_{\text{suc}}(f_1, z) < 1 - \sqrt{\epsilon_n} \) for \( f_1 \notin \mathcal{G}_T \) (see Definition 4), and (56) is due to our assumption. Since Eqs. (55)-(57) contradict the fact in (54), we obtain that \( \mathbb{P}(F_1 \in \mathcal{G}_T^*) \geq 1 - \sqrt{\epsilon_n} \).

Let \( N_m \triangleq \prod_{a=1}^k \binom{\{v_a\}}{\{m_a\}} \) for each \( m \in \mathcal{M} \). For each node \( v \in [n] \), the expected number of hyperedges in \( F_2 \) that contain node \( v \) and other \( d-1 \) nodes with community assignment \( m \) is \( \mathbb{E}(D_{v,m}) = N_m \left( 1 - \frac{\gamma_n^2}{n \log n} \right) \). By applying the Chernoff bound, we have

\[
\mathbb{P} \left( D_{v,m} \leq N_m \left( 1 - \frac{2\gamma_n}{\log n} \right) \right)
\]

\[
\leq \mathbb{P} \left( D_{v,m} \leq \left( 1 - \frac{\gamma_n^2}{n \log n} \right) \mathbb{E}(D_{v,m}) \right)
\]

\[
\leq \exp \left( - \frac{\gamma_n^2}{3 (\log n)^2} \mathbb{E}(D_{v,m}) \right)
\]

\[
= \exp \left( - \Theta(n^{-1} \gamma_n^2 / (\log n)^2) \right).
\]

(58)

Taking a union bound over all \( m \in \mathcal{M} \) and all the \( n \) nodes, we have that with probability at least \( 1 - \exp \left( - \Theta(n^{-1} \gamma_n^2 / (\log n)^2) \right) \), every node \( v \in [n] \) satisfies

\[
(1 - (2\gamma_n / \log n)) N_m \leq D_{v,m} \leq N_m, \forall m \in \mathcal{M}.
\]

Combining the fact that \( (1 - n^{-\frac{1}{2} + \frac{1}{2}}) R_m \leq N_m \leq (1 + n^{-\frac{1}{2} + \frac{1}{2}}) R_m \) (since \( z \in \mathcal{A}_Z \)), we have

\[
(1 - \frac{3\gamma_n}{\log n}) R_m \leq D_{v,m} \leq (1 + n^{-\frac{1}{2} + \frac{1}{2}}) R_m, \forall m \in \mathcal{M}.
\]

Thus, \( \mathbb{P}(F_2 \in \mathcal{G}_T^*) \geq 1 - \exp \left( - \Theta(n^{-1} \gamma_n^2 / (\log n)^2) \right) \). This completes the proof.

**APPENDIX C**

**PROOF OF LEMMA 7**

Note that \( \mathbb{E}(\sum_{a=1}^d Y_a) \leq D' \cdot \mathcal{Q}_{\max}(\log n) / n^{d-1} = O(\eta_n \log n) \), since the success probability of each Bernoulli random variable is at most \( \mathcal{Q}_{\max}(\log n) / n^{d-1} \). In the following, we show that the probability that \( \sum_{a=1}^d Y_a \geq c_n \log n \) is at most \( n^{-1} \), where \( c_n \log n = \omega(\eta_n \log n) \). Note that

\[
\mathbb{P} \left( \sum_{a=1}^d Y_a \geq c_n \log n \right)
\]

\[
= \sum_{\theta = c_n \log n} \mathbb{P} \left( \sum_{a=1}^d Y_a = \theta \right)
\]
where (59) follows from the facts that \( \binom{n}{k} \leq \left( \frac{en}{k} \right)^k \) and \( \theta = c_n \log n \) maximizes the terms in (58).

We then prove the second part. For random variables \( \{ Y_a \}_{a=1}^{D_n} \), we have

\[
\frac{Q_{\min}}{Q_{\max}} \leq \frac{\mathbb{P}(Y_a = 1|Z_v = j, \hat{Z}^{(0)}_{\mathrm{v}})}{\mathbb{P}(Y_a = 0|Z_v = j, \hat{Z}^{(0)}_{\mathrm{v}})} \leq \frac{Q_{\max}}{Q_{\min}}
\]

and

\[
1 - \frac{(Q_{\max} - Q_{\min}) \log n}{n^{d-1}} \leq \mathbb{P}(Y_a = 0|Z_v = j, \hat{Z}^{(0)}_{\mathrm{v}}) \leq 1 + \frac{(Q_{\max} - Q_{\min}) \log n}{n^{d-1}}.
\]

Since \( \sum_{a=1}^{D_n} y_a \leq c_n \log n \) for \( g_2 \in \mathcal{G}_v \), we have

\[
\frac{\mathbb{P}(g_2|Z_v = j, \hat{Z}^{(0)}_{\mathrm{v}})}{\mathbb{P}(g_2|Z_v = j, \hat{Z}^{(0)}_{\mathrm{v}})} = \frac{\prod_{a=1}^{D_n} \mathbb{P}(y_a|Z_v = j, \hat{Z}^{(0)}_{\mathrm{v}})}{\prod_{a=1}^{D_n} \mathbb{P}(y_a|Z_v = j, \hat{Z}^{(0)}_{\mathrm{v}})} \leq \frac{Q_{\max}}{Q_{\min}} \log n \left( 1 + \frac{(Q_{\max} - Q_{\min}) \log n}{n^{d-1}} \right) D_n c_n \log n
\]

\[ \triangleq L_h \]

and

\[
\frac{\mathbb{P}(g_2|Z_v = j, \hat{Z}^{(0)}_{\mathrm{v}})}{\mathbb{P}(g_2|Z_v = j, \hat{Z}^{(0)}_{\mathrm{v}})} = \frac{\prod_{a=1}^{D_n} \mathbb{P}(y_a|Z_v = j, \hat{Z}^{(0)}_{\mathrm{v}})}{\prod_{a=1}^{D_n} \mathbb{P}(y_a|Z_v = j, \hat{Z}^{(0)}_{\mathrm{v}})} \geq \frac{Q_{\min}}{Q_{\max}} c_n \log n \left( 1 - \frac{(Q_{\max} - Q_{\min}) \log n}{n^{d-1}} \right) D_n c_n \log n
\]

\[ \triangleq L_l \]

and note that \( L_h = n^{\alpha(1)} \) and \( L_l = n^{\alpha(1)} \).

REFERENCES

[1] P. W. Holland, K. B. Laskey, and S. Leinhardt, “Stochastic blockmodels: First steps,” Social Netw., vol. 5, no. 2, pp. 109–137, 1983.
[2] E. Abbe, A. S. Bandeira, and G. Hall, “Exact recovery in the stochastic block model,” IEEE Trans. Inf. Theory, vol. 62, no. 1, pp. 471–487, Jan. 2016.
[3] E. Mossel, J. Neeman, and A. Sly, “Consistency thresholds for the planted bisection model,” in Proc. 47th Annu. ACM Symp. Theory Comput., 2015, pp. 69–75.
[4] E. Abbe and C. Sandon, “Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery,” in Proc. IEEE 56th Annu. Symp. Found. Comput. Sci., Oct. 2015, pp. 670–688.
[5] E. Abbe and C. Sandon, “Recovering communities in the general stochastic block model without knowing the parameters,” 2015, arXiv:1506.03729.
[6] B. Hajek, Y. Wu, and J. Xu, “Information limits for recovering a hidden community,” IEEE Trans. Inf. Theory, vol. 63, no. 8, pp. 4729–4745, Aug. 2017.
[7] G. Reeves, V. Mayya, and A. Volfovsky, “The geometry of community detection via the MMSE matrix,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jul. 2019, pp. 400–404.
[8] V. Jog and P.-L. Loh, “Information-theoretic bounds for exact recovery in weighted stochastic block models using the Renyi divergence,” 2015, arXiv:1509.06418.
[9] S.-Y. Yun and A. Proustiere, “Optimal cluster recovery in the labeled stochastic block model,” in Proc. Adv. Neural Inf. Process. Syst., vol. 29, 2016, pp. 965–973.
[10] S.-Y. Yun and A. Proustiere, “Accurate community detection in the stochastic block model via spectral algorithms,” 2014, arXiv:1412.7335.
[11] P. Chin, A. Rao, and V. Vu, “Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery,” in Proc. Conf. Learn. Theory, 2015, pp. 391–423.
[12] B. Hajek, Y. Wu, and J. Xu, “Achieving exact cluster recovery threshold via semidefinite programming,” IEEE Trans. Inf. Theory, vol. 62, no. 5, pp. 2788–2797, May 2016.
[13] B. Hajek, Y. Wu, and J. Xu, “Achieving exact cluster recovery threshold via semidefinite programming: Extensions,” IEEE Trans. Inf. Theory, vol. 62, no. 10, pp. 5918–5937, Oct. 2016.
[14] A. Montanari and S. Sen, “Semidefinite programs on sparse random graphs and their application to community detection,” in Proc. 48th Annu. ACM Symp. Theory Comput., Jun. 2016, pp. 814–827.
[15] F. Caldogonale, M. Lelarge, and L. Miolane, “Recovering asymmetric communities in the stochastic block model,” IEEE Trans. Netw. Sci. Eng., vol. 5, no. 3, pp. 237–246, Jul. 2018.
[16] N. Agarwal, A. S. Bandeira, K. Koiliaris, and A. Kolla, “Multisection in the stochastic block model using semidefinite programming,” in Compressed Sensing and its Applications (Applied and Numerical Harmonic Analysis), Cham, Switzerland: Springer, 2017, pp. 125–156.
[17] A. Perry and A. S. Wein, “A semidefinite program for unbalanced multisection in the stochastic block model,” in Proc. Int. Conf. Sampling Theory Appl. (SampTA), Jul. 2017, pp. 64–67.
[18] A. R. Asadi, E. Abbe, and S. Verdú, “Compressing data on graphs with clusters,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jun. 2017, pp. 1583–1587.
[19] H. Saad and A. Nosratinia, “Community detection with side information: Exact recovery under the stochastic block model,” IEEE J. Sel. Topics Signal Process., vol. 12, no. 5, pp. 944–958, Oct. 2018.
[20] H. Saad and A. Nosratinia, “Recovering a single community with side information,” IEEE Trans. Inf. Theory, vol. 66, no. 12, pp. 7939–7966, Dec. 2020.
[21] H. Saad and A. Nosratinia, “Exact recovery in community detection with continuous-valued side information,” IEEE Signal Process. Lett., vol. 26, no. 2, pp. 332–336, Feb. 2019.
[22] V. Mayya and G. Reeves, “Mutual information in community detection with covariate information and correlated networks,” in Proc. 57th Annu. Allerton Conf. Commun., Control, Comput. (Allerton), Sep. 2019, pp. 602–607.
[23] E. Abbe, “Community detection and stochastic block models: Recent developments,” J. Mach. Learn. Res., vol. 18, no. 1, pp. 6446–6531, Jan. 2017.
[24] M. Chertok and Y. Keller, “Efficient high order matching,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 12, pp. 2205–2215, Dec. 2010.
[25] H. Liu, L. Latecki, and S. Yan, “Robust clustering as ensembles of affinity relations,” in Proc. Adv. Neural Inf. Process. Syst., vol. 23, 2010, pp. 1414–1422.
[26] D. Ghoshdastidar and A. Dukkipati, “Consistency of spectral partitioning of uniform hypergraphs under planted partition model,” in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 397–405.
[27] D. Ghoshdastidar and A. Dukkipati, “A provable generalized tensor spectral method for uniform hypergraph partitioning,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 400–409.
[28] D. Ghoshdastidar and A. Dukkipati, “Spectral clustering using multilinear SVD: Analysis, approximations and applications,” in Proc. AAAI Conf. Artif. Intell., 2013, vol. 29, no. 1, pp. 1025–1032.
[29] D. Ghoshdastidar and A. Dukkipati, “Consistency of spectral hypergraph partitioning under planted partition model,” Ann. Statist., vol. 45, no. 1, pp. 289–315, 2017.
[30] S. Pal and Y. Zhu, “Community detection in the sparse hypergraph stochastic block model,” *Random Struct. Algorithms*, vol. 59, no. 3, pp. 407–463, Oct. 2021.

[31] S. Cole and Y. Zhu, “Exact recovery in the hypergraph stochastic block model: A spectral algorithm,” *Linear Algebra Appl.*, vol. 593, pp. 45–73, May 2020.

[32] K. Ahn, K. Lee, and C. Suh, “Information-theoretic limits of subspace clustering,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2473–2477.

[33] K. Ahn, K. Lee, and C. Suh, “Hypergraph spectral clustering in the weighted stochastic block model,” *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 5, pp. 959–974, Oct. 2018.

[34] C. Kim, A. S. Bandeira, and M. X. Goemans, “Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach,” 2018, arXiv:1807.02884.

[35] J. Lee, D. Kim, and H. W. Chung, “Robust hypergraph clustering via convex relaxation of truncated MLE,” *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 3, pp. 613–631, Nov. 2020.

[36] C. Kim, A. S. Bandeira, and M. X. Goemans, “Community detection in hypergraphs, spiked tensor models, and sum-of-squares,” in *Proc. Int. Conf. Sampling Theory Appl. (SampTA)*, Jul. 2017, pp. 124–128.

[37] Z. Tracy Ke, F. Shi, and D. Xia, “Community detection for hypergraph networks via regularized tensor power iteration,” 2019, arXiv:1909.05053.

[38] M. C. Angelini, F. Caltagirone, F. Krzakala, and L. Zdeborova, “Spectral detection on sparse hypergraphs,” in *Proc. 53rd Ann. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2015, pp. 66–73.

[39] T. Lesieur, L. Miolane, M. Lelarge, F. Krzakala, and L. Zdeborova, “Statistical and computational phase transitions in spiked tensor estimation,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 511–515.

[40] I. E. Chien, C.-Y. Lin, and I.-H. Wang, “On the minimax misclassification ratio of hypergraph community detection,” *IEEE Trans. Inf. Theory*, vol. 65, no. 12, pp. 8095–8118, Dec. 2019.

[41] C.-Y. Lin, I. E. Chien, and I.-H. Wang, “On the fundamental statistical limit of community detection in random hypergraphs,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2178–2182.

[42] I. Chien, C.-Y. Lin, and I.-H. Wang, “Community detection in hypergraphs: Optimal statistical limit and efficient algorithms,” in *Proc. Int. Conf. Artif. Intell. Statist.*, 2018, pp. 871–879.

[43] K. Ahn, K. Lee, and C. Suh, “Community recovery in hypergraphs,” *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6561–6579, Oct. 2019.

[44] J. Liang, C. Ke, and J. Honorio, “Information theoretic limits of exact recovery in sub-hypergraph models for hypernetworks,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 2578–2583.

[45] M. Yuan and Z. Shang, “Information limits for detecting a subhypergraph,” *Statistics*, vol. 10, no. 1, p. e407, Dec. 2021.

[46] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

[47] A. Y. Zhang and H. H. Zhou, “Minimax rates of community detection in stochastic block models,” *Ann. Statist.*, vol. 44, no. 5, pp. 2252–2260, 2016.

[48] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou, “Achieving optimal misclassification proportion in stochastic block models,” *J. Mach. Learn. Res.*, vol. 18, no. 60, pp. 1980–2024, Jan. 2017.

[49] P. Jain, P. Netrapalli, and S. Sanghavi, “Low-rank matrix completion using alternating minimization,” in *Proc. 45th Ann. ACM Symp. Symp. Theory Comput. (STOC)*, 2013, pp. 665–674.

[50] R. Keshavan, A. Montanari, and S. Oh, “Matrix completion from a few entries,” *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2009.

[51] Q. Zhang, G. Suh, C. Suh, and V. Y. F. Tan, “MC2G: An efficient algorithm for matrix completion with social and item similarity graphs,” *IEEE Trans. Signal Process.*, vol. 70, pp. 2681–2697, 2022.

[52] V. Vu, “A simple SVD algorithm for finding hidden partitions,” 2014, arXiv:1404.3918.

[53] C. Eckart and G. Young, “The approximation of one matrix by another of lower rank,” *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[54] H. Weyl, “Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung),” *Mathematische Annalen*, vol. 71, no. 4, pp. 441–479, 1912.

[55] I. Csiszár and J. Körner, *Information Theory: Coding Theorems For Discrete Memoryless Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[56] M. Yuan, R. Liu, Y. Feng, and Z. Shang, “Testing community structure for hypergraphs,” *Ann. Statist.*, vol. 50, no. 1, pp. 147–169, Feb. 2022.

[57] V. Feldman, W. Perkins, and S. Vempala, “On the complexity of random satisfiability problems with planted solutions,” *SIAM J. Comput.*, vol. 47, no. 4, pp. 1294–1338, Jan. 2018.

Qiaosheng Zhang received the B.Eng. (Hons.) and Ph.D. degrees from the Department of Information Engineering, The Chinese University of Hong Kong (CUHK), in 2015 and 2019, respectively. From 2019 to 2022, he was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore (NUS). He is currently a Researcher at the Shanghai Artificial Intelligence Laboratory. His research interests include machine learning theory, information theory, and covert communication.

Vincent Y. F. Tan (Senior Member, IEEE) was born in Singapore, in 1981. He received the B.A. and M.Eng. degrees in electrical and information science from Cambridge University in 2005, and the Ph.D. degree in electrical engineering and computer science (EECS) from the Massachusetts Institute of Technology (MIT) in 2011. He is currently a Dean’s Chair Associate Professor with the Department of Mathematics and the Department of Electrical and Computer Engineering, National University of Singapore (NUS). His research interests include information theory, machine learning, and statistical signal processing. He is a member of the IEEE Information Theory Society Board of Governors. He received the MIT EECS Jin-Au Kong Outstanding Doctoral Thesis Prize in 2011, the NUS Young Investigator Award in 2014, the Singapore National Research Foundation (NRF) Fellowship (Class of 2018), and the NUS Young Researcher Award in 2019. He is also serving as a Senior Area Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY. He was an IEEE Information Theory Society Distinguished Lecturer from 2018 to 2019.