Hypergraph Spectral Clustering in the Weighted Stochastic Block Model

Kwangjun Ahn, Kangwook Lee, and Changho Suh, *Member, IEEE*

***Abstract*—Spectral clustering is a celebrated algorithm that partitions the objects based on pairwise similarity information. While this approach has been successfully applied to a variety of domains, it comes with limitations. The reason is that there are many other applications in which only *multi*way similarity measures are available. This motivates us to explore the multiway measurement setting. In this paper, we develop two algorithms intended for such setting: hypergraph spectral clustering (HSC) and hypergraph spectral clustering with local refinement (HSCLR). Our main contribution lies in performance analysis of the polytime algorithms under a random hypergraph model, which we name the weighted stochastic block model, in which objects and multiway measures are modeled as nodes and weights of hyperedges, respectively. Denoting by n the number of nodes, our analysis reveals the following: 1) HSC outputs a partition which is better than a random guess if the sum of edge weights (to be explained later) is $\Omega(n)$; 2) HSC outputs a partition which coincides with the hidden partition except for a vanishing fraction of nodes if the sum of edge weights is $\omega(n)$; and 3) HSCLR exactly recovers the hidden partition if the sum of edge weights is on the order of $n \log n$. Our results improve upon the state of the arts recently established under the model and they first settle the orderwise optimal results for the binary edge weight case. Moreover, we show that our results lead to efficient sketching algorithms for subspace clustering, a computer vision application. Finally, we show that HSCLR achieves the information-theoretic limits for a special yet practically relevant model, thereby showing no computational barrier for the case.**

***Index Terms*—Clustering, hypergraph clustering, information-theoretic limits, stochastic block model, subspace clustering.**

I. Introduction

The problem of clustering is prevalent in a variety of applications such as social network analysis, computer vision, and computational biology. Among many clustering algorithms, *spectral clustering* is one of the most prominent algorithms proposed by [2] in the context of image segmentation, viewing an image as a graph of pixel nodes, connected by weighted edges representing visual similarities between two adjacent pixel nodes. This approach has become popular, showing its wide applicability in numerous applications, and has been extensively analyzed under various models [3]–[5].

While the standard spectral clustering relies upon interactions between *pairs of two nodes*, there are many applications where interaction occurs across more than two nodes. One such application includes a social network with online social communities, called folksonomies, in which users attach tags to resources. In the example, a three-way interaction occurs across users, resources and annotations [6]. Another application is molecular biology, in which multi-way interactions between distinct systems capture molecular interactions [7]. See [8] and the list of applications therein. Hence, one natural follow-up research direction is to extend the celebrated framework of *graph* spectral clustering into a *hypergraph* setting in which edges reflect multi-way interactions.

As an effort, in this work, we consider a random weighted uniform hypergraph model which we call *the weighted stochastic block model*, which is a special case of that considered in [8]. An edge of size d is *homogeneous* if it consists of nodes from the same group, and is *heterogeneous* otherwise.[1] Given a hidden partition of n nodes into k groups, a weight is independently assigned to each edge of size d such that homogeneous edges tend to have higher weights than heterogeneous edges. More precisely, for some constants $p > q$, the expectation of homogeneous edges' weights is $p\alpha_n$ and that of heterogeneous edges' weights is $q\alpha_n$.[2] Here, α_n captures the sparsity level of the weights, which may decay in n. The task here is to recover the hidden partition from the weighted hypergraph. In particular, we aim to develop computationally efficient algorithms that provably find the hidden partition.

Our contributions: By generalizing the spectral clustering algorithms proposed for the graph clustering, we first propose two poly-time algorithms which we name *Hypergraph Spectral Clustering* (HSC) and *Hypergraph Spectral Clustering with Local Refinement* (HSCLR). We then analyze their performances, assuming that the size of hyperedges is d, the number of clusters k is constant, and the size of each group is linear in n. Our main results can be summarized as follows. For some constants c and c', which depend only on p, q, and k, the following statements hold with high probability:

- *Detection:* If $\binom{n}{d}\alpha_n \geq c \cdot n$, the output of HSC is more consistent with the hidden partition than a random guess;

Manuscript received December 8, 2017; revised March 27, 2018; accepted May 2, 2018. Date of publication May 17, 2018; date of current version September 27, 2018. This work was supported by the National Research Foundation of Korea funded by the Korea Government (MSIT) under Grant 2015R1C1A1A02036561. This paper was presented in part at the 2017 IEEE International Symposium on Information Theory, Aachen, Germany, Jun. 25–30. The guest editor coordinating the review of this paper and approving it for publication was Dr. Miguel Rodrigues. *(Corresponding author: Changho Suh.)*

K. Ahn is with the Department of Mathematical Sciences, KAIST, Daejeon 34141, South Korea (e-mail: kjahnkorea@kaist.ac.kr).

K. Lee and C. Suh are with the School of Electrical Engineering, KAIST, Daejeon 34141, South Korea (e-mail: kw1jjang@kaist.ac.kr; chsuh@kaist.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2018.2837638

[1] While edges of a graph are pairs of nodes, edges of a hypergraph (or hyperedges) are arbitrary sets of nodes. Further, the size of an edge is the number of nodes contained in the edge.

[2] For illustrative purpose, we focus on a symmetric setting. In Section V, we will extend our results (to be described later) to a more general setting.

- *Weak consistency:* If $\binom{n}{d}\alpha_n = \omega(n)$, HSC outputs a partition which coincides with the hidden partition except $o(n)$ number of nodes; and
- *Strong consistency:* If $\binom{n}{d}\alpha_n \geq c' \cdot n \log n$, HSCLR exactly recovers the hidden partition.

We remark that our main results are the first order-wise optimal results for the binary edge weight case (see Proposition 1).

A. Related Work

1) Graph Clustering: The problem of standard graph clustering, i.e., $d = 2$, has been studied in great generality. Here, we summarize some major developments, referring the readers to a recent survey by Abbe [13] for details. The detection problem, whose goal is to find a partition that is more consistent with the hidden partition than a random guess, has received a wide attention. A notable work by Decelle *et al.* [14] firstly observes phase transition and conjectures the transition limit. Further, they also conjecture that the computational gap exists for the case of $k \geq 4$. For the case of $k = 2$, the phase transition limit is fully settled jointly by [15] and [16], [17]: The impossibility of the detection below the conjectured threshold is established in [15], and it is proved that the conjectured threshold can be achieved via some efficient algorithms in [16], [17]. The limits for the case $k \geq 3$ have been studied in [18]–[21], and are settled in [22].

The weak/strong consistency problem aims at finding a cluster that is correct except a vanishing or zero fraction. The necessary and sufficient conditions for weak consistency have been studied in [23]–[27], and those for strong consistency in [25], [27]–[29]. In particular for strong consistency, both the fundamental limits and computationally efficient algorithms are investigated initially for $k = 2$ [25], [28], [29], and recently for general k [27]. While most of the works assume that the graph parameters such as p, q, k, and the size of clusters are fixed, one can also study the minimax scenario where the graph parameters are adversarially chosen against the clustering algorithm. In [30], the authors characterize the minimax-optimal rate. Further, [24] shows that the minimax-optimal rate can be achieved by an efficient algorithm.

2) Hypergraph Clustering: Compared to graph clustering, the study of hypergraph clustering is still in its infancy. In this section, we briefly summarize recent developments. For detection, analogous to the work by Decelle *et al.* [14], Angelini *et al.* [31] firstly conjecture phase transition thresholds. These conjectures have not been settled yet unlike the graph case. In [8], the authors study a specific spectral clustering algorithm, which can be shown to detect the hidden cluster if $\binom{n}{d}\alpha_n = \Omega(n(\log n)^2)$, while the conjectured threshold for detection is $\binom{n}{d}\alpha_n = c^\star n$ for some constant $c^\star$. Actually, this gap is due to the technical challenge that is specific to the hypergraph clustering problem: See Remark 7 for details. In [12], the authors study the bipartite stochastic block model, and as a byproduct of their results, they show that detection is possible under some specific model if $\binom{n}{d}\alpha_n = \Omega(n)$. While this guarantee is order-wise optimal, it holds only when edge weights are binary-valued and the size of two clusters are equal. Our detection guarantee, obtained by delicately resolving the technical challenges specific to hypergraphs, is also order-wise optimal but does not require such assumptions.

While several consistency results under various models are shown in [8]–[12], to the best of our knowledge, our consistency guarantees are the first order-wise optimal ones. We briefly overview the existing results below. In [9], [10], the authors derive consistency results for the case in which $\alpha_n = 1$ and weights are binary-valued. In [8], the authors investigate consistency results of a certain spectral clustering algorithm under a fairly general random hypergraph model, called the planted partition model in hypergraphs. Indeed, our hypergraph model is a special case of the planted partition model, and hence the algorithm proposed in [8] can be applied to our model as well. One can show that their algorithm is weakly consistent if $\binom{n}{d}\alpha_n = \Omega(n(\log n)^2)$ under our model. The case of non-uniform hypergraphs, in which the size of edges may vary, is studied in [11]. See Table I for a summary.

TABLE I
COMPARISON TO THE STATE OF THE ARTS: "WEAK" AND "STRONG" ARE FOR CONSISTENCY RESULTS

	Model assumption		Order of $\binom{n}{d}\alpha_n$ required for		
	Multi Groups	Weighted Edges	Detection	Weak	Strong
[9]	$\surd$	$\times$	NA	n^d	NA
[10]	$\surd$	$\times$	NA	n^d	NA
[11]	$\surd$	$\times$	NA	$\Omega(n(\log n)^2)$	NA
[8]	$\surd$	$\surd$	NA	$\Omega(n(\log n)^2)$	NA
[12]	$\times$	$\times$	$\Omega(n)$	NA	NA
Ours	$\surd$	$\surd$	$\Omega(n)$	$\omega(n)$	$\Omega(n \log n)$

While most of the existing works focus on analyzing the performance of certain clustering algorithms, some study the fundamental limits. In [1], [32], the information-theoretic limits are characterized for specific hypergraph models. In [33], the minimax optimal rates of error fraction are derived for the binary weighted edge case. However, it has not been clear whether or not a computationally efficient algorithm can achieve such limits. In this work, we show that HSCLR achieves the fundamental limit for the model considered in [1].

3) Main Innovation Relative to [1]: The new algorithms proposed in this work can be viewed as strict improvements over the algorithm proposed in our previous work [1]. First, the algorithm of [1] cannot handle the sparse-weight regime, i.e., $\binom{n}{d}\alpha_n = \Theta(n)$. In order to address this, we employ a preprocessing step prior to the spectral clustering step. It turns out this can handle the sparse regime; see Lemma 2 for details.

Another limitation of the original algorithm is related to its refinement step (to be detailed later). The original refinement step is tailored for a specific model, which assumes binary-valued weights and two clusters (see Definition 7). On the other hand, our new refinement step can be applied to the general case with weighted edges and k clusters. Further, the original refinement step involves iterative updates, and this is solely because our old proof holds only with such iterations. However, we observe via experiments that a single refinement step is always sufficient. By integrating a well-known sample splitting technique into our algorithm, we are able to prove that a single refinement step is indeed sufficient.

Apart from the improvements above, we also propose a sketching algorithm for subspace clustering based on our new algorithm, and we show that it outperforms existing schemes in terms of sample complexity as well as time complexity.

4) Computer Vision Applications: The weighted stochastic block model that we consider herein is well-fitted into computer vision applications such as geometric grouping and subspace clustering [34]–[36]. The goal of such problems is to cluster a union of groups of data points where points in the same group lie on a common low-dimensional affine space. In these applications, similarity between a fixed number of data points reflects how well the points can be approximated by a low-dimensional flat. By viewing these similarities as the weights of edges in a hypergraph, one can relate it to our model. Note that edges connecting the data points from the same low-dimensional affine space have larger weights than other edges: See Section VI for detailed discussion.

5) Connection With Low-Rank Tensor Completion: Our model bears strong resemblance to the low-rank tensor completion. To see this, consider the following model: for each $e = \{i_1, i_2, i_3\} \in \mathcal{E}$, edge weight of e is generated as $W_e = pX_e$ (where $X_e \sim \mathsf{Bern}(\alpha_n)$) if (i_1, i_2, i_3) are from the same cluster; $W_e = qX_e$ otherwise. This model generates a weighted hypergraph, whose weights are either p, q, or 0. Now, view each weight as an observation of an entry of a hidden tensor $\mathbf{T}$, whose entries $\mathbf{T}_{i_1 i_2 i_3} = p$ if (i_1, i_2, i_3) are from the same cluster; $\mathbf{T}_{i_1 i_2 i_3} = q$ otherwise. Here, 0 weight indicates that the entry is "unobserved". Then, the knowledge of hidden partition will directly lead to "completion" of unobserved entries. This way, one can draw a parallel between hypergraph clustering and the low-rank tensor completion.[3] This connection allows us to compare our results with the guarantee in the tensor completion literature. For instance, the sufficient condition for vanishing estimation error, i.e., weak consistency, derived in [38] reads $\binom{n}{d}\alpha_n = \omega(n^{3/2}\log^4 n)$, while ours reads $\binom{n}{d}\alpha_n = \omega(n)$. This favors our approach. Moreover, a more interesting implication arises in computational aspects. Notice that a naïve lower bound for tensor completion is[4] $\binom{n}{d}\alpha_n = \Omega(n)$, and the tensor completion guarantee comes with an additional $\Omega(n^{1/2})$ factor to the lower bound. Actually this gap has not been closed in the literature, raising a question whether this *information-computation gap* is fundamental. Interestingly, this gap does not appear in our result, hence hypergraph clustering can shed new light on the computational aspects of tensor completion. Recently, a similar observation has been made independently in [39] for spike-tensor-related models (see Section 4.3. therein).

B. Paper Organization

Section II introduces the considered model; in Section III, our main results are presented along with some implications; in Section IV, we provide the proofs of the main theorems; in Section V, we discuss as to how our results can be extended and adapted to other models; Section VI is devoted to practical applications relevant to our model, and presents the empirical performances of the proposed algorithms; and in Section VII, we conclude the paper with some future research directions.

C. Notations

Let $\mathbf{M}_{i*}$ ($\mathbf{M}_{*j}$) be the ith row (the jth column) of matrix $\mathbf{M}$. For a positive integer n, $[n] := \{1, 2, \ldots, n\}$. For a set A and an integer m, $\binom{A}{m} := \{B \subset A : |B| = m\}$. Let $\log(\cdot)$ denote the natural logarithm. Let $\mathbb{I}\{\cdot\}$ denote the indicator function. For a function $F : \mathcal{A} \to \mathcal{B}$ and $b \in \mathcal{B}$, $F^{-1}(b) := \{i \in \mathcal{A} : F(i) = b\}$.

II. The Weighted Stochastic Block Model

We first remark that our definition of the weighted SBM is a generalization of the original model for graphs [40], [41] to a hypergraph setting. For simplicity, we will focus on the following symmetric assortative model. In Section V, we generalized our results to a broader class of graph models.

A. Model

Let $\mathcal{V} = [n]$ be the indices of n nodes, and $\mathcal{E} := \binom{[n]}{d}$ be the set of all possible edges of size d for a fixed integer $d \geq 2$. Let $\Psi : \mathcal{V} \to [k]$ be the hidden partition function that maps n nodes into k groups for a fixed integer k. Equivalently, the membership function can be represented in a matrix form $\mathbf{Z} \in \{0, 1\}^{n \times k}$, which we call *the membership matrix*, whose (i, j)th entry takes 1 if $j = \Psi(i)$ and 0 otherwise. We denote by n_i the size of the jth group for $j = 1, 2, \ldots, k$, i.e., $n_j := |\Psi^{-1}(j)|$. Let $n_{\min} := \min_j n_j$ and $n_{\max} := \max_j n_j$. An edge $e = \{i_1, \ldots, i_d\}$ is *homogeneous* if $\Psi(i_1) = \Psi(i_2) = \cdots = \Psi(i_d)$ and *heterogeneous* otherwise. We now formally define the weighted SBM.

Definition 1 (The weighted SBM(p, q, α_n)): A random weight $W_e \in [0, 1]$ is assigned to each edge e independently[5]: for homogeneous edges, $\mathbb{E}[W_e] = p\alpha_n$; and for heterogeneous edges, $\mathbb{E}[W_e] = q\alpha_n$.

Note that the weighted SBM does not assume a specific edge weight distribution but only specifies the expected values. For instance, it can capture the case with a single location family distribution with different parameters as well as the case with two completely different weight distributions.

Example 1 (The unweighted hypergraph case): For homogeneous edges, $W_e \sim \mathsf{Bern}(p\alpha_n)$; and for heterogeneous edges, $W_e \sim \mathsf{Bern}(q\alpha_n)$. This is an instance of the weighted SBM(p, q, α_n). When $d = 2$, it captures the standard models such as planted multisection [3] and the SBM [42].

Example 2 (The weighted hypergraph case): For homogeneous edges, $W_e \sim \mathsf{Bern}(0.75)$; and for heterogeneous edges, $W_e \sim \mathsf{Unif}[0, 1]$, a uniform distribution on $[0, 1]$. This model can be seen as an instance of the weighted SBM$(0.75, 0.5, 1)$.

B. Performance Metric

Given $\{W_e\}_{e \in \mathcal{E}}$ and the number of clusters k, we intend to recover a hidden partition Ψ up to a permutation. Formally, for any estimator $\Phi : [n] \to [k]$, we define the error fraction as $err(\Phi) := \frac{1}{n}\min_{\Pi \in \mathcal{P}} |\{i : \Psi(i) \neq \Pi(\Phi(i))\}|$, where $\mathcal{P}$ is the collection of all permutations of $[k]$. We study three types of consistency guarantees [13], [43].

Definition 2 (Recovery types): An estimator Φ is

- *strongly consistent* if $\lim_{n\to\infty} \Pr(err(\Phi) = 0) = 1$;
- *weakly consistent* if $\lim_{n\to\infty} err(\Phi) = 0$ in prob.; and
- is solving *detection* if it outputs a partition which is more consistent relative to a random guess.[6]

[3] Here, $\mathbf{T}$ is of rank at most k since it admits a CP-decomposition [37] $\mathbf{T} = q\mathbf{1}^{\otimes 3} + \sum_{i=1}^{k}(p - q)(\mathbf{Z}_{*i})^{\otimes 3}$.

[4] The number of free parameters defining a rank k, d-th order, n-dimensional tensor is ndk, which scales like $\Theta(n)$ when d and k are fixed.

[5] Our results hold as long as the weights are upper bounded by any fixed positive constant since one can always normalize the edge weights such that they are within $[0, 1]$. The global upper bound on the edge weights are required for deriving our large deviation results (Lemmas 3 and 5) in the proof.

[6] Here we provide an informal definition for simplicity. See Definition 7 in [13] for the formal definition.

Algorithm 1: HSC.

```
Input: A weighted hypergraph $\mathcal{H} = ([n], \{W_e\}_{e\in\mathcal{E}})$, the number of clusters $k$.
Compute the processed similarity matrix $\mathbf{A}^0$:
    Compute the similarity matrix $\mathbf{A}$ where $\mathbf{A}_{ij} = \sum_{e:\{i,j\}\subset e} W_e$ if $i \neq j$; and $\mathbf{A}_{ij} = 0$ otherwise. Then, obtain $\mathbf{A}^0$ by zeroing-out row $i$ (and the corresponding column) if $\sum_j \mathbf{A}_{ij} > c_{\text{thr}} \frac{1}{n} \sum_{i,j} \mathbf{A}_{ij}$, where $c_{\text{thr}} > 0$ is a constant depending only on $d$ (e.g., $c_{\text{thr}} = 6$ when $d = 2$).
Apply spectral clustering to $\mathbf{A}^0$: Find $k$ largest eigenvectors of $\mathbf{A}^0$, stack them side by side to obtain $\mathbf{U}^0 \in \mathbb{R}^{n\times k}$, and cluster the rows of $\mathbf{U}^0$ using the approximate geometric $k$-clustering [48] with an approximation rate $\epsilon > 0$.
Output: $\Phi_{\text{HSC}}(i)$ = cluster index of the $i$th row.
```

III. Main Results

A. Hypergraph Spectral Clustering

Hypergraph Spectral Clustering (HSC) is built upon the spectral relaxation technique [10] and the spectral algorithms [5], [24], [26], [44]–[47]. The first step of the algorithm is to **compute the processed similarity matrix** whose entries represent similarities between pairs. To this end, we first compute *the similarity matrix* $\mathbf{A}$, where $\mathbf{A}_{ij} = \sum_{e:\{i,j\}\subset e} W_e$ if $i \neq j$; $\mathbf{A}_{ij} = 0$ if $i = j$. This is inspired by the spectral relaxation technique in [10]. Next, we zero-out every row and column whose sum is larger than a certain threshold, constructing an output $\mathbf{A}^0$, which we call *the processed similarity matrix*. We then apply **spectral clustering** to the processed similarity matrix. That is, we first find the k largest eigenvectors $\mathbf{U}^0 \in \mathbb{R}^{n\times k}$ of $\mathbf{A}^0$, and cluster n rows of $\mathbf{U}^0$ using the approximate geometric k-clustering [48]. Note that HSC is non-parametric, i.e., it does not require the knowledge of model parameters. See Algorithm 1 for the detailed procedure.

Remark 1: The zeroing-out procedure, proposed in [44] (see Section 3 therein), is used to remove outlier rows whose sums are much larger than the average. This is necessary since if such outliers exist, the eigenvector estimate will be biased, and hence the spectral clustering will also fail. Note that this technique is widely adopted in various graph clustering algorithms [26], [45], [49].

The time complexity of HSC is $O\left(n^d\right)$. As each edge appears $2\binom{d}{2}$ times during the construction of the similarity matrix, this step requires $2\binom{d}{2}|\mathcal{E}| = O\left(n^d\right)$ time. The first k eigenvectors can be computed via power iterations, which can be done within $O(kn^2 \log n)$ time [50]. Geometric k-clustering can be done in time $O(n(\log n)^k)$ [48].

B. Hypergraph Spectral Clustering With Local Refinement

Our second algorithm consists of two stages: HSC and local refinement. The HSCLR algorithm is inspired by a similar refinement procedure, which has been proposed for the graph case [27], [28]. The algorithm begins with randomly splitting edges into two sets $\mathcal{E}_1$ and $\mathcal{E}_2$. For small $\beta > 0$, we assign each edge to $\mathcal{E}_1$ independently with probability β. $\mathcal{E}_2$ is the complement of $\mathcal{E}_1$. Then, we run HSC on $\mathcal{H}_1 = ([n], \{W_e\}_{e\in\mathcal{E}_1})$. Next, we do local refinement with $\mathcal{E}_2$. For $i \in [n]$ and $j \in [k]$, define $\mathcal{E}^{(i)}(j)$ to be the set of edges ($\in \mathcal{E}_2$) which connect node i with $d-1$ nodes from $\Phi_{\text{HSC}}^{-1}(j)$, i.e., $\mathcal{E}^{(i)}(j) := \left\{e \in \mathcal{E}_2 : i \in e, (e \setminus \{i\}) \subset \Phi_{\text{HSC}}^{-1}(j)\right\}$. Then, for each $i \in [n]$, we update $\Phi_{\text{HSC}}(i)$ with

$$\arg\max_{j\in[k]} \frac{1}{|\mathcal{E}^{(i)}(j)|} \sum_{e\in\mathcal{E}^{(i)}(j)} W_e. \tag{1}$$

That is, the refinement step first measures the *fitness* of each node with respect to different clusters, and updates the cluster assignment of each node accordingly. Note that HSCLR is also non-parametric. See Algorithm 2 for the detailed procedure.

Algorithm 2: HSCLR.

```
Input: A weighted hypergraph $\mathcal{H} = ([n], \{W_e\}_{e\in\mathcal{E}})$, the number of clusters $k$, and sample splitting rate $\beta > 0$.
Randomly split $\mathcal{E}$: for small enough $\beta > 0$, include each edge of $\mathcal{E}$ in $\mathcal{E}_1$ independently with probability $\beta$. Denote by $\mathcal{E}_2$ the complement of $\mathcal{E}_1$.
Apply Hypergraph Spectral Clustering to $\mathcal{H}_1 = ([n], \{W_e\}_{e\in\mathcal{E}_1})$ to yield an estimate $\Phi_{\text{HSC}}$.
Local refinement: for $i = 1, 2, \ldots, n$, $\Phi_{\text{HSCLR}}(i) = \arg\max_{j\in[k]} \frac{1}{|\mathcal{E}^{(i)}(j)|} \sum_{e\in\mathcal{E}^{(i)}(j)} W_e$.
Output: $\Phi_{\text{HSCLR}}$.
```

The time complexity of HSCLR is $O\left(n^d\right)$. For each node i, the local refinement requires $\sum_{j=1}^{k} |\mathcal{E}^{(i)}(j)|$ flops, which is bounded by $k|\mathcal{E}_i|$, where $|\mathcal{E}_i|$ is the number of edges containing node i. As $\sum_i |\mathcal{E}_i| = d|\mathcal{E}|$, the local refinement step can be done within $O\left(|\mathcal{E}|\right)$ time.

Remark 2: HSCLR is inspired by the recent paradigm of solving non-convex problems, which first approximately estimates the solution, followed by some local refinement. This two-stage approach has been applied to a variety of contexts, including matrix completion [51], [52], phase retrieval [53], [54], robust PCA [55], community recovery [28], [56], EM-algorithm [57], and rank aggregation [58].

C. Theoretical Guarantees

Theorem 1: Let Φ_{HSC} be the output of HSC. Suppose that $\frac{n_{\max}}{n_{\min}} = O(1)$. Then, there exist constants $c_0, c_1 > 0$ (where c_1 depends on p and q) such that if $\binom{n}{d}\alpha_n \geq c_0 n$, then,

$$err(\Phi_{\text{HSC}}) \leq c_1 k \frac{n^3}{d(d-1)n_{\min}^2 \binom{n}{d}\alpha_n} \tag{2}$$

w.p. $1 - O(n^{-1})$, provided that $c_1 k \frac{n^3}{d(d-1)n_{\min}^2 \binom{n}{d}\alpha_n} < 1$.

Proof: See Section IV-A. ■

Note that when $d = 2$, Theorem 1 recovers [24, Th. 6].

Remark 3: We remark a technical challenge that arises in proving Theorem 1 relative to the graph case. Actually, the key step in the proof is to derive the sharp concentration bound on a certain matrix spectral norm (to be detailed later). But the bounding technique employed in the graph case does not carry over to the hypergraph case, as the matrix has strong dependencies across entries. We address this challenge by developing a delicate analysis that carefully handles such dependencies. See Remark 7 in Section IV for details.

Corollary 1 (Detection): Suppose that $\frac{n_{\max}}{n_{\min}} = O(1)$. There exists a constant c_2 depending on p, q and k such that HSC solves detection if $\binom{n}{d}\alpha_n \geq c_2 \cdot n$.

Proof: In Theorem 1, $err(\Phi_{\text{HSC}}) = O(1/c)$ when α_n satisfies $\binom{n}{d}\alpha_n \geq cn$ for sufficiently large $c > 0$.

Remark 4: We compare our algorithm to the one proposed in [31]. To compare, we first note that in the graph case, the threshold for detection [14] is achieved by new methods based on the *non-backtracking operator* [17], [22], [59]. In [59], the spectral analysis based on a plain adjacency matrix is shown to fail, while the one based on the non-backtracking operator succeeds. Recently, it is shown that the non-backtracking based approach can be extended to the hypergraph case, and it is empirically observed to outperform a spectral method that is similar to HSC except the preprocessing step [31].

Corollary 2 (Weak consistency): Suppose that $\frac{n_{\max}}{n_{\min}} = O(1)$. HSC is weakly consistent if $\binom{n}{d}\alpha_n = \omega(n)$.

Proof: By (2), $\binom{n}{d}\alpha_n = \omega(n) \Rightarrow err(\Phi_{\text{HSC}}) = o(1)$. ■

Remark 5: When specialized to weighted stochastic block model, the weak consistency guarantee of [8] becomes $\binom{n}{d}\alpha_n = \Omega(n(\log n)^2)$, which comes with an extra poly-logarithmic factor gap to ours.

The following theorem provides the theoretical guarantee of HSCLR. See Section IV-B for the proof.

Theorem 2 (Strong consistency): Suppose that $\frac{n_{\max}}{n_{\min}} = O(1)$. Then, HSCLR with sampling rate[7] $\beta = \frac{\log\log n}{\log n}$ is strongly consistent provided that for any $\epsilon > 0$,

$$\frac{(p-q)^2}{p}\binom{n}{d}\alpha_n \geq (8+\epsilon)\frac{(n/n_{\min})^{d-1}}{d} n \log n \,. \quad (3)$$

Remark 6: We remark that Theorem 2 characterizes the performance of our *non-parametric* algorithm for any hypergraphs with (bounded) *real-valued* weights. Hence, one may obtain a tighter threshold and a *parametric* algorithm by focusing on a more specific hypergraph model. For instance, in [60], Chien *et al.* derive a tighter bound for the binary weight case. As a concrete example, when $d = 3$ and $k = 2$ with two equal-sized clusters, the sufficient condition of [60, Th 4.1] reads $(\sqrt{p} - \sqrt{q})^2\binom{n}{3}\alpha_n \geq \frac{1}{12} n \log n$, while that of Theorem 2 reads $\frac{(p-q)^2}{p}\binom{n}{3}\alpha_n \geq \frac{32}{3} n \log n$.

Indeed, by leveraging recent works on phase transition of random hypergraphs [61], [62], we can prove the order-wise optimality of our algorithms for the binary-valued edge case.

Proposition 1: For the binary-valued edge case, there is no estimator which

- solves detection when $\binom{n}{d}\alpha_n = o(n)$
- is weakly consistent when $\binom{n}{d}\alpha_n = O(n)$; and
- is strongly consistent when $\binom{n}{d}\alpha_n = o(n \log n)$.

Proof: If $\binom{n}{d}\alpha_n = o(n)$, the fraction of isolated nodes approaches 1, hence detection is infeasible. In [61], the authors show that if $\binom{n}{d}\alpha_n = \Theta(n)$, there is no connected component of size $(1 - o(1))n$, implying that weak consistency is infeasible. Lastly, [62] shows that $\binom{n}{d}\alpha_n > c \cdot n \log n$ for some constant $c > 0$ is required for connectivity, a necessary condition for strong consistency.

IV. Proofs

A. *Proof of Theorem 1*

We first outline the proof. Proposition 2 asserts that spectral clustering finds the exact clustering if $\mathbb{E}[\mathbf{A}]$ is available instead of $\mathbf{A}^0$. We then make use of Lemma 1 to bound the error fraction in terms of $\|\mathbf{A}^0 - \mathbb{E}[\mathbf{A}]\|$. Finally, we derive a new concentration bound for the above spectral norm, and combine it with Lemma 1 to prove the theorem.

Consider two off-diagonal entries $\mathbf{A}_{i,j}$ and $\mathbf{A}_{i',j'}$ such that $\Psi(i) = \Psi(i')$ and $\Psi(j) = \Psi(j')$. One can see from the definition that $\mathbf{A}_{i,j}$ is statistically identical to $\mathbf{A}_{i'j'}$, so $\mathbb{E}[\mathbf{A}_{ij}] = \mathbb{E}[\mathbf{A}_{i'j'}]$. Hence, by defining a $k \times k$ matrix $\mathbf{B}$ such that $\mathbf{B}_{\ell,m} = \mathbb{E}[\mathbf{A}_{ij}]$, where $i \in \Psi^{-1}(\ell)$, $j \in \Psi^{-1}(m)$, for some $i \neq j$, one can verify that $\mathbf{P} := \mathbf{Z}\mathbf{B}\mathbf{Z}^T$ coincides with $\mathbb{E}[\mathbf{A}]$ except for the diagonal entries. Our model implies that the diagonal entries of $\mathbf{B}$ are strictly larger than its off-diagonal entries, so $\mathbf{B}$ is of full rank.

Proposition 2: ([5, Lemma 2.1]) Consider $\mathbf{B} \in \mathbb{R}^{k\times k}$ of full rank and the membership matrix $\mathbf{Z} \in \{0,1\}^{n\times k}$. Let $\mathbf{P} = \mathbf{Z}\mathbf{B}\mathbf{Z}^T$. Then the matrix $\mathbf{U} \in \mathbb{R}^{n\times k}$ whose columns are the first k eigenvectors of $\mathbf{P}$ satisfies:$\mathbf{U}_{i*} = \mathbf{U}_{j*}$ whenever $\Psi(i) = \Psi(j)$; $\mathbf{U}_{i*}$ and $\mathbf{U}_{j*}$ are orthogonal whenever $\Psi(i) \neq \Psi(j)$. In particular, a clustering algorithm on the rows of $\mathbf{U}$ will exactly output the hidden partition.

Proposition 2 suggests that spectral clustering successfully finds Ψ if $\mathbf{P}$ is available. We now turn to the case where $\mathbf{A}^0$ is available instead of $\mathbf{P}$. It is developed in [5] a general scheme to prove error bounds for spectral clustering under an assumption that k-clustering step outputs a "good" solution. To clarify the meaning of "goodness", we formally describe the k-means clustering problem.

Definition 3 (k-means clustering problem): The goal is to cluster the rows of an $n \times k$ matrix $\mathbf{U}$. Define the cost function of a partition $\Phi : [n] \to [k]$ as $\text{cost}(\Phi) = \sum_{j=1}^{k} \mathbf{Var}(\Phi^{-1}(j))$, where $\mathbf{Var}(\mathcal{A}) = \sum_{i\in\mathcal{A}} \|\mathbf{U}_{i*} - \frac{1}{|\mathcal{A}|}\sum_{\ell\in\mathcal{A}} \mathbf{U}_\ell\|^2$. We say Φ is $(1+\epsilon)$-approximate if $\text{cost}(\Phi) \leq (1+\epsilon)\min_{\Phi':[n]\to[k]} \text{cost}(\Phi')$.

We now introduce the general scheme to prove error bounds, formally stated in the following lemma.

Lemma 1: Assume that $\mathbf{P}$ is defined as in Proposition 2 and $\sigma_{\min}(\mathbf{P})$ is the smallest non-zero singular value of $\mathbf{P}$. Let $\mathbf{M}$ be any symmetric matrix and $\mathbf{U} \in \mathbb{R}^{n\times k}$ be the k largest eigenvectors of $\mathbf{M}$. Suppose a $(1+\epsilon)$-approximate solution Φ for a constant $\epsilon > 0$. Then, for some $c_3 > 0$, $err(\Phi) \leq c_3 k(1+\epsilon)\frac{\|\mathbf{M}-\mathbf{P}\|^2}{\sigma_{\min}(\mathbf{P})^2}$, provided that $c_3 k(1+\epsilon)\frac{\|\mathbf{M}-\mathbf{P}\|^2}{\sigma_{\min}(\mathbf{P})^2} \leq 1$.

Proof: We refer to [5] for the proof. ■

Theorem 1.2. in [48] implies that a $(1+\epsilon)$-approximate solution can be found using the approximate geometric k-clustering.[8] Hence, the above lemma implies that one needs to bound $\|\mathbf{A}^0 - \mathbf{P}\|$ in order to analyze the error fraction of the spectral clustering. Our technical contribution lies mainly in deriving such concentration bound, formally stated below.

Lemma 2: There exist constants c_{thr} (depending only on d), $c_4, c_5 > 0$ such that the processed similarity matrix $\mathbf{A}^0$ with constant c_{thr} (see Algorithm 1) satisfies $\|\mathbf{A}^0 - \mathbf{P}\| \leq c_4\sqrt{n\binom{n-2}{d-2}\alpha_n}$ with probability exceeding $1 - O(n^{-1})$, provided that $n\binom{n-2}{d-2}p\alpha_n \geq c_5$.

[7] We note that β can be chosen arbitrarily as long as $\beta = o(1)$ and $\beta = \omega(1/\log n)$. See Section IV-B for detail.

[8] Note that this result holds only for a fixed k [48].

Proof: See Appendix A. ■

Note that this lemma holds for a fixed d. We now conclude the proof with these lemmas. Let $\mu := \binom{n-2}{d-2}\alpha_n$. We first estimate $\sigma_{\min}(\mathbf{P})$.

Claim 1: $\sigma_{\min}(\mathbf{P}) \le c_6 n_{\min}\mu$ for some constant $c_6 > 0$.

Proof: By definition, $\mathbf{P} = (\mathbf{Z}\Delta^{-1})\Delta\mathbf{B}\Delta^T(\mathbf{Z}\Delta^{-1})^T$, where $\Delta = \operatorname{diag}(\sqrt{n_1}, \sqrt{n_2}, \ldots, \sqrt{n_k})$. Since the columns of $\mathbf{Z}\Delta^{-1}$ are orthonormal, $\sigma_{\min}(\mathbf{P}) = \sigma_{\min}(\Delta\mathbf{B}\Delta^T)$. One can show that $\sigma_{\min}(\Delta\mathbf{B}\Delta^T) \ge \sigma_{\min}(\Delta)^2\sigma_{\min}(\mathbf{B})$. Hence, $\sigma_{\min}(\mathbf{P}) \ge \sigma_{\min}(\Delta)^2\sigma_{\min}(\mathbf{B}) = n_{\min}\sigma_{\min}(\mathbf{B})$. Hence, we calculate $\sigma_{\min}(\mathbf{B})$. By the definition of $\mathbf{B}$,

$$\mathbf{B}_{\ell m} = \begin{cases} p\alpha_n\binom{n_\ell - 2}{d-2} + q\alpha_n\left[\binom{n-2}{d-2} - \binom{n_\ell - 2}{d-2}\right] & \text{if } \ell = m, \\ q\alpha_n\binom{n-2}{d-2} & \text{if } \ell \ne m \end{cases}$$

$$= \mu \cdot \begin{cases} \frac{\binom{n_\ell - 2}{d-2}}{\binom{n-2}{d-2}}(p-q) + q & \text{if } \ell = m, \\ q & \text{if } \ell \ne m. \end{cases}$$

Thus, $\mathbf{B} = \mu \cdot q\mathbf{1}\mathbf{1}^T + \mu \cdot \operatorname{diag}(f_1, f_2, \ldots, f_\ell)$, where $f_\ell := \frac{\binom{n_\ell - 2}{d-2}}{\binom{n-2}{d-2}}(p-q)$. As $n_{\max}/n_{\min} = O(1)$, each f_ℓ converges to a positive constant, implying that $\sigma_{\min}(\mathbf{B}) = \Theta(\mu)$. ■

By Lemma 2 and the above claim, $c_3 k(1+\epsilon)\frac{\|\mathbf{A}^0 - \mathbf{P}\|^2}{\sigma_{\min}(\mathbf{P})^2} \le \frac{c_3 k(1+\epsilon)c_4^2}{c_6^2}\frac{n}{n_{\min}^2\mu}$ holds w.p. $1 - O(n^{-1})$ for $n\mu \ge c_5$. Choosing $c_0 = \frac{c_5}{d(d-1)}$, $c_1 = \frac{c_3 k(1+\epsilon)c_4^2}{c_6^2}$ completes the proof.

Remark 7: (Technical novelty relative to the graph case): Indeed, proving the sharp concentration of a spectral norm has been a key challenge in the spectral analysis [44], [63]. While most bounds developed hinge upon the independence between entries[9], the matrix $\mathbf{A}$ in HSC has strong dependencies across entries due to its construction. For instance, the entries $\mathbf{A}_{12}$ and $\mathbf{A}_{13}$ both have a term W_e for any edge e of the form $\{1, 2, 3, j_4, j_5, \ldots, j_d\}$, hence sharing $\binom{n-3}{d-3}$ many terms.

One approach to handle this dependency is to use matrix Bernstein inequality [65] on the decomposition $\mathbf{A} = \sum_{e\in\mathcal{E}} W_e\mathbf{S}_e$, where $S_e := \sum_{\substack{i,j\in e \\ i\ne j}} \mathbf{e}_i\mathbf{e}_j^T$. See [8], [11]. However, this approach provides a bound which comes with an extra $\sqrt{\log n}$ factor relative to the bound in Lemma 2, resulting in a suboptimal consistency guarantee as described in Section I-A.

Another approach is a combinatorial method [63], which counts the number of edges between subsets. The rationale behind this method is as follows. From the definition of the spectral norm, one needs to bound the quantity $\mathbf{x}^T(\mathbf{A}^0 - \mathbf{P})\mathbf{x}$ for any vector $\mathbf{x}$. It turns out that this quantity has a close connection to the number of (hyper)edges between two subsets in a random (hyper)graph. For instance, $\mathbf{1}_A^T\mathbf{A}\mathbf{1}_B$ is precisely the number of edges between A and B.

Indeed, a technique for estimating the number of hyperedges between two arbitrary subsets is developed in [66]. Using this method, however, one may only obtain a suboptimal guarantee, which is $\binom{n}{d}\alpha_n = \Omega(n^{1.5})$. On the other hand, we show via our analysis that the order-optimal guarantee can be obtained by improving the standard combinatorial method. See Appendix A.

B. Proof of Theorem 2

We first outline the proof. Using the union bound, we show that it is sufficient to prove $\Pr(\Phi_{\text{HSCLR}}(i) = j) = o(n^{-1})$ for all $1 \le i \le n$ and $j \ne \Psi(i)$. We then consider the following events to bound this error probability. The first event is that the average edge weight of the edges between the true community $\Psi(i)$ and node i is less than a certain threshold, and the other one is that the average edge weight of the edges between the wrong community j and node i is greater than the certain threshold. We will first show that if the misclassification event occurs, at least one of these two events must occur. Thus, we bound the error probability by bounding those of these two events using Lemma 3 and Lemma 4, respectively.

We consider the boundary case $\binom{n}{d}\alpha_n = \Theta(n\log n)$. As $\beta\binom{n}{d}\alpha_n = \Theta(n\log\log n) = \omega(n)$, Corollary 2 guarantees that Φ_{HSC} is weakly consistent. Without loss of generality, assume that the identity permutation is equal to $\arg\min_{\Pi\in\mathcal{P}}|\{i : \Psi(i) \ne \Pi(\Phi_{\text{HSC}}(i))\}|$. Then, $\frac{|\Phi_{\text{HSC}}^{-1}(j)\cap\Psi^{-1}(j)|}{|\Phi_{\text{HSC}}^{-1}(j)|} > 1-\gamma$, i.e., at least $1-\gamma$ fraction of the nodes that are classified as in community j are correctly classified. The second stage of HSCLR refines the output of the first stage Φ_{HSC}, resulting in Φ_{HSCLR}. By the union bound, we have $\Pr(err(\Phi_{\text{HSCLR}}) \ne 0) \le \sum_{i=1}^n\sum_{j\ne\Psi(i)}\Pr(\Phi_{\text{HSCLR}}(i) = j)$. Since the total number of summands is $\Theta(n)$, if $\Pr(\Phi_{\text{HSCLR}}(i) = j) = o(n^{-1})$ for all $1 \le i \le n$ and $j \ne \Psi(i)$, then $\Pr(err(\Phi_{\text{HSCLR}}) \ne 0) = o(1)$.

By the refinement rule (1), $\Pr(\Phi_{\text{HSCLR}}(i) = j) \le \Pr\left(\frac{\sum_{e\in\mathcal{E}^{(i)}(\Psi(i))}W_e}{|\mathcal{E}^{(i)}(\Psi(i))|} < \frac{\sum_{e\in\mathcal{E}^{(i)}(j)}W_e}{|\mathcal{E}^{(i)}(j)|}\right)$. For any real numbers (a, b, t), $[a \ge b] \supset [a \ge t] \cap [t \ge b]$ holds. By taking complements of both sides, we have $[a < b] \subset [a < t] \cup [t < b]$. Therefore, by the union bound, $P(a < b) \le P(a < t) + P(t < b)$ holds for any (a, b, t). Applying this bound, we have

$$\Pr\left(\frac{\sum_{e\in\mathcal{E}^{(i)}(\Psi(i))}W_e}{|\mathcal{E}^{(i)}(\Psi(i))|} < \frac{\sum_{e\in\mathcal{E}^{(i)}(j)}W_e}{|\mathcal{E}^{(i)}(j)|}\right) \tag{4}$$

$$\le \underbrace{\Pr\left(\frac{\sum_{e\in\mathcal{E}^{(i)}(\Psi(i))}W_e}{|\mathcal{E}^{(i)}(\Psi(i))|} < \frac{p+q}{2}\alpha_n\right)}_{R_1} \tag{5}$$

$$+ \underbrace{\Pr\left(\frac{p+q}{2}\alpha_n < \frac{\sum_{e\in\mathcal{E}^{(i)}(j)}W_e}{|\mathcal{E}^{(i)}(j)|}\right)}_{R_2}. \tag{6}$$

We first interpret R_1 and R_2. For illustration, assume that Φ_{HSCLR} coincides with Ψ. Under this assumption, observe that $\frac{\sum_{e\in\mathcal{E}^{(i)}(\Psi(i))}W_e}{|\mathcal{E}^{(i)}(\Psi(i))|}$ is equal to the average edge weight of the homogeneous edges within community $\Psi(i)$. Since the expected value of this term is $\frac{p}{2}\alpha_n$, one can show that the term R_1 vanishes. Similarly, $\frac{\sum_{e\in\mathcal{E}^{(i)}(j)}W_e}{|\mathcal{E}^{(i)}(j)|}$ is the average weight of the edges connecting i and the other nodes in community j. Since these edges are heterogeneous, R_2 also vanishes.

Indeed, as $err(\Phi_{\text{HSC}})$ is not exactly zero, but an arbitrarily small constant, the above interpretation is not precise. In what follows, we show that R_1 and R_2 vanish as well for the case.

We begin with bounding R_1. Denote by $\mathcal{E}_h$ the set of all homogeneous edges. Recall that edges in $\mathcal{E}^{(i)}(\Psi(i))$, except $O(\gamma)$ fraction, are homogeneous, so $|\mathcal{E}^{(i)}(\Psi(i)) \cap \mathcal{E}_h| =$

[9] For instance, the most studied model, called the Wigner matrix, assumes independence among entries. See [64] for more details.

$(1-O(\gamma))|\mathcal{E}^{(i)}(\Psi(i))|$. By restricting the range of summation, $R_1 \le \Pr(\sum_{e\in\mathcal{E}^{(i)}(\Psi(i))\cap\mathcal{E}_h} W_e < \frac{p+q}{2}\alpha_n|\mathcal{E}^{(i)}(\Psi(i))|)$. Note that W_e's are not restricted to Bernoulli random variables. By tweaking the proof of conventional large deviation results [67] for Bernoulli variables, we obtain the following:

Lemma 3: Let S be the sum of m mutually independent random variables taking values in $[0,1]$. For any $\delta > 0$, we have $\Pr(S > (1+\delta)\mathbb{E}[S]) \le \exp(-\frac{\delta^2}{2+\delta}\mathbb{E}[S])$ and $\Pr(S < (1-\delta)\mathbb{E}[S]) \le \exp(-\frac{\delta^2}{2}\mathbb{E}[S])$.

Proof: See Appendix D-F. ■

As $\mathbb{E}[\sum_{e\in\mathcal{E}^{(i)}(\Psi(i))\cap\mathcal{E}_h} W_e] = (1-O(\gamma))p\alpha_n|\mathcal{E}^{(i)}(\Psi(i))|$,

$$\frac{\frac{p+q}{2}\alpha_n|\mathcal{E}^{(i)}(\Psi(i))|}{\mathbb{E}[\sum_{e\in\mathcal{E}^{(i)}(\Psi(i))\cap\mathcal{E}_h} W_e]} - 1 = (1+O(\gamma))\frac{q-p}{2p},$$

so Lemma 3$-$2) with $\delta = (1+O(\gamma))\frac{p-q}{2p}$ gives

$$R_1 \le \exp\left(-\frac{(p-q)^2}{8p}\alpha_n(1+O(\gamma))|\mathcal{E}^{(i)}(\Psi(i))|\right). \tag{7}$$

Next we consider R_2. Again, edges in $\mathcal{E}^{(i)}(j)$, except $O(\gamma)$ fraction, are heterogeneous, so $|\mathcal{E}^{(i)}(j)\cap\mathcal{E}_h^c| = (1-O(\gamma))|\mathcal{E}^{(i)}(j)|$. The following lemma says that the contribution due to the $O(\gamma)$ fraction of edges is marginal:

Lemma 4: For sufficiently small $\gamma > 0$,

$$\Pr\left(\sum_{e\in\mathcal{E}^{(i)}(j)\cap\mathcal{E}_h} W_e > \frac{p\alpha_n|\mathcal{E}^{(i)}(j)|}{\sqrt{\log(1/\gamma)}}\right) = o(n^{-1}). \tag{8}$$

Proof: See Appendix D-B. ■

Hence, we focus on heterogeneous edges only. Making a similar argument as above, the bound in Lemma 3 becomes

$$R_2 \le \exp\left(-\frac{\frac{(p-q)^2}{4q^2}}{2+\frac{p-q}{2q}}q\alpha_n(1+O(\gamma))|\mathcal{E}^{(i)}(j)|\right) \tag{9}$$

$$\overset{(a)}{\le} \exp\left(-\frac{1}{8}\frac{(p-q)^2}{p}\alpha_n(1+O(\gamma))|\mathcal{E}^{(i)}(j)|\right), \tag{10}$$

where (a) follows since $\frac{1}{2+\frac{p-q}{2q}} = \frac{1}{\frac{3}{2}+\frac{p}{2q}} = \frac{1}{\left(\frac{3}{2}\frac{q}{p}+\frac{1}{2}\right)\frac{p}{q}} \ge \frac{1}{2\frac{p}{q}}$.

Since $\frac{(p-q)^2}{p}\binom{n}{d}\alpha_n \ge (8+\epsilon)\frac{(n/n_{\min})^{d-1}}{d}n\log n$, a straightforward calculation yields $\frac{1}{8}\frac{(p-q)^2}{p}\alpha_n(1+O(\gamma))\binom{n_{\min}-1}{d-1} \ge \left(1+\frac{1}{16}\epsilon\right)\log n$, for sufficiently large n. Thus, R_1 and R_2 are both $o(n^{-1})$ from (7) and (10).

V. Discussion

We have shown that our algorithms can achieve the order-optimal sample complexity for all different recovery guarantees under a symmetric block model. In this section, we show that our main results indeed hold for a broader class of block models. We also show that HSCLR can achieve the sharp recovery threshold for a certain SBM model.

A. Extensions

For the graph case [13], a fairly general model, which subsumes as a special case the asymmetric SBM, has been investigated. Here we extend our model to one such model but in the context of hypergraphs. Specifically, we consider the following asymmetric weighted SBM.

Definition 4 (The asymmetric weighted SBM): Let $\{p_e\}_{e\in\mathcal{E}}$ be constants such that $p_e > p_e$ holds for any homogeneous edge e and heterogeneous edge e. A random weight is assigned to each edge independently as follows: For each edge $e\in\mathcal{E}$, $\mathbb{E}[W_e] = p_e\alpha_n$. Notice that this reduces to the condition of $p > q$ in the symmetric setting.

We find that our main results stated in Theorems 1 and 2 readily carry over the above asymmetric setting. The key rationale behind this is that our spectral clustering guarantee hinges only upon the full-rank condition on $\mathbf{B}$ (see Section III-A for the definition). Here, what one can easily verify is that the condition above implies the full-rank condition, and hence our results hold even for the asymmetric setting. The only distinction here is that the constants that appear in the theorems depend now on p_e's. Similarly, our technique can cover *disassortative* SBM in which heterogeneous edges have larger weights than homogeneous edges.

Definition 5 (The symmetric disassortative weighted SBM): In Definition 1, we assume instead that $0 < p < q < 1$.

Another prominent instance is the planted clique model.

Definition 6 (The planted clique model): Fix s-subset of nodes C $(s \le n)$. Consider a random hypergraph in which every d-regular edge $e = \{i_1, i_2, \ldots, i_d\}$ appears with probability 1 if $e \subseteq C$ or $\frac{1}{2}$ otherwise.

In this model, one wishes to detect the hidden subset C, which is called the *clique*. Following a similar analysis with a different notion of error fraction, one can show that the clique can be detected if $s \ge c^* \cdot \sqrt{n}$ for some constant c^*, which is consistent with the well-known result for $d = 2$ [68].

B. Sharpness

Recently, sharp thresholds on the fundamental limits are characterized in the graph case [22], [24], [27], [28], [30]. In contrast, such a tight result has been widely open in the hypergraph case. A notable exception is our companion paper [32] which studies a special case of the weighted SBM (considered herein), in which weights are *binary*-valued.

Definition 7 (Generalized Censored Block Model with Homogeneity Measurements [32]): Let $\theta \in (0, 1/2)$ be a fixed constant. Assume that $k = 2$ and denote erasure by x. If the edge e is homogeneous, $W_e = 1$ w.p. $\alpha_n(1-\theta)$, $W_e = 0$ w.p. $\alpha_n\theta$, and $W_e = \mathsf{x}$ w.p. $1-\alpha_n$. Otherwise, $W_e = 1$ w.p. $\alpha_n\theta$, $W_e = 0$ w.p. $\alpha_n(1-\theta)$, and $W_e = \mathsf{x}$ w.p. $1-\alpha_n$.

The information-theoretic limit for strong consistency has been characterized under this model, formally stated below.

Proposition 3: ([32, Th. 1]) Under the model in Definition 7, the maximum likelihood estimator is strongly consistent for any given hidden partition Ψ if $\binom{n}{d}\alpha_n \ge (1+\epsilon)\frac{2^{d-2}}{d}\frac{n\log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2}$ for any constant $\epsilon > 0$. Conversely, if $\binom{n}{d}\alpha_n \le (1-\epsilon)\frac{2^{d-2}}{d}\frac{n\log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2}$, no algorithm can be strongly consistent for any given hidden partition Ψ.

Using our results, we can show that there is no computational barrier under this model. We now state the theorem, deferring the proof to Appendix B.

Theorem 3: HSCLR[10] achieves the information-theoretic limits characterized in Proposition 3.

[10] Indeed, there should be some minor tweaks to make HSCLR better adapted to this model. See Appendix B for details.

VI. Application

In this section, based on our algorithms, we design a sketching algorithm for subspace clustering.

A. Subspace Clustering

It is well known that hypergraph clustering is closely related to computer vision applications such as subspace clustering [35]. In the subspace clustering problem, one is given with n data points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^{\ell}$ in a high dimensional ambient space. The n data points are partitioned into k groups, and data points in the same group approximately lie on the same subspace, each of dimension at most $m < \ell$. The goal is to recover the hidden partition of the n data points based on certain measurements. Among various approaches, tensor-based algorithms measure *similarities between the data points* to recover the cluster [34]–[36]. More specifically, they construct a weighted d-uniform ($d \geq m + 2$) hypergraph $([n], \{W_e\}_{e \in \mathcal{E}})$, in which each edge weight represents the similarity of the corresponding d points. One typical approach to measure the similarity between d data points is based on the hyperplane fitting. More specifically, denoting by $\text{fit}(\cdot)$ the error of fitting an m-dimensional affine subspace to d data points, one may set $W_e = \exp\left(-\text{fit}(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_d})\right)$ for $e = \{i_1, i_2, \ldots, i_d\}$. Note that $W_e \simeq 1$ if the d data points are approximately on the same subspace, and $W_e \simeq 0$ if the data points cannot be fit on a single subspace.

Consider a set of d data points of the same cluster, which approximately lie on the same subspace by definition. The edge weight corresponding to these d data points will be approximately 1, and one may model the edge weight as a random value whose expected value is close to 1. Similarly, one may model the edge weights of heterogeneous edges by a random variable whose expected value is close to 0.[11]

Clearly, our weighted SBM can precisely capture the above hypergraph model since our model only assumes that the average weights of homogeneous edges are larger than those of heterogeneous edges. We verify this claim using a real data set. Hopkins 155 is the most widely used dataset for the subspace clustering problem [69]. We first set $d = 8$ and $m = 3$, and then randomly sample 10000 homogeneous edges and 10000 heterogeneous edges. The empirical distributions of edge weights are shown in Fig. 1(a). We can see that the homogeneous edges have larger weights on average than the heterogeneous edges, well respecting the weighted SBM.

B. Sketching Algorithms for Subspace Clustering

Modern subspace clustering algorithms involve a large number of data points lying on a high-dimensional space, i.e., n and ℓ are very large. Hence, storing the entire raw data points is prohibitive, and one may have to resort to the *sketch* of the data set. A sketch can be viewed as a summary of the dataset, containing sufficient information of the data set.

As evidenced by the preceding section, we assume that the weighted hypergraph constructed from the data points follows the model in Section II. Under this assumption, subspace clustering can be done by clustering nodes of the weighted hypergraph. The following corollary asserts that one can exactly solve the subspace clustering problem with a sketch consisting of the weights of randomly chosen hyperedges.[12] We now state a corollary, a consequence of Theorems 1 and 2.

Corollary 3: Suppose that $n_{\max}/n_{\min} = O(1)$ and $\alpha_n = 1$. Then, HSC is weakly consistent if $\binom{n}{d} s_n = \omega(n)$, and HSCLR is strongly consistent if $\binom{n}{d} s_n \geq c_8 \cdot n \log n$ for some constant $c_8 > 0$. Moreover, the computational complexities of HSC, HSCLR reduce to $\max\{\binom{n}{d} s_n, n(\log n)^k\}$.

Remark 8: One can sketch data more aggressively if the subspaces are not similar to each other [70]. This is also captured in Corollary 3 as follows. As a concrete example, consider two subspaces of dimension m and a heterogeneous edge e. When the two subspaces are moving farther away from each other, the fitting error of e increases. Thus, W_e approaches 0, and hence $p - q$ increases. Since the sample complexity is inversely proportional to $p - q$,[13] one can sketch more aggressively.

Corollary 3 implies that our sketching method can reduce the storage overhead from $O(n\ell)$ to $O(n \log n)$. We now evaluate our sketching algorithm. The relevant parameters are n, k, ℓ, m, d, and s_n: in an ambient dimension of $\ell = 50$, we randomly generate k subspaces each being of dimension of $m = 3$; for each subspace, we randomly sample n/k points and perturb every point with Gaussian noise of variance σ^2; we set edge size $d = 5$ and sampling probability s_n. We first implement HSCLR in MATLAB[14]. We then compare HSCLR with other prior algorithm[15], adopting the experimental setups from [75] and [8].

We first measures the performance of various algorithms. We set $k = 3$ and $\binom{n}{d} s_n = 5k^{d-1} n \log n / d$, and report the average fractional errors of each algorithm over 20 trials for $(n/k, \sigma) \in \{300, 400, 500, 600\} \times \{0, 0.05, 0.1, 0.15\}$ in Fig. 1(b). Observe that our algorithm matches the state-of-the-art performance. We also measures the run time of the algorithms. We set $k = 2$, $\sigma = 0.025$, $n \in \{750, 1500, 3000, 6000\}$, $\binom{n}{d} s_n = 5k^{d-1} n \log n / d$, and report the average run time over 10 trials. Fig. 1(c) shows that the runtime of our proposed algorithm scales nearly linearly in n.

C. Other Applications

Apart from subspace clustering, there are many applications in which d-wise similarities can carry more information than pairwise ones. Those include other computer vision applications (such as geometric grouping [34], [36] and high-order matching [77]), tagged social networks [6], biological networks [7]

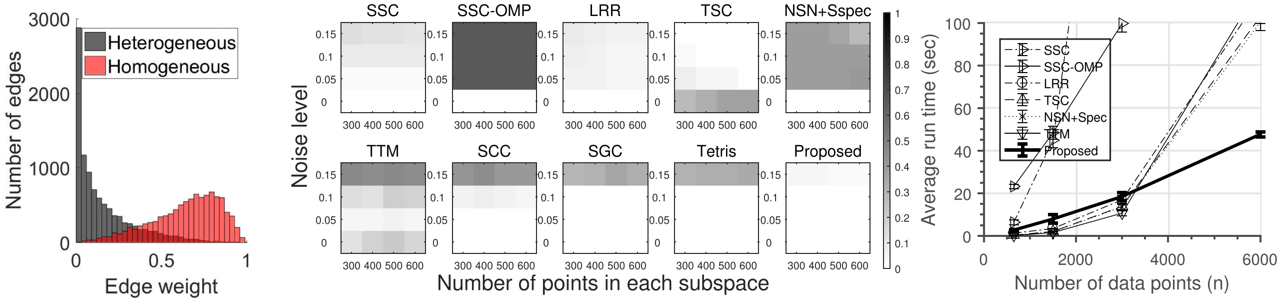
[11] Indeed, the edge weight of a heterogeneous edge can be very close to 1, i.e., the fitting error can be close to 0. This may happen when d data points, which are from different subspaces, are well aligned with another single subspace. Such a coincidence, however, happens with very low probability, and hence we simply treat these atypical events as statistical noise.

[12] We note that one may carefully choose similarity entries in order to achieve a more informative sketch than our random one, at the cost of increased computational complexity for sketch construction.

[13] The sufficient condition in Theorem 2 reads $\frac{(p-q)^2}{p} \binom{n}{d} s_n \geq C$ for some quantity C.

[14] We observe a large constant in the computational complexity of the geometric k-clustering, and hence we implement HSCLR with an efficient k-means algorithms for the experiments.

[15] Sparse subspace clustering (SSC) [71], a variant of SSC using OMP (SSC-OMP) [72], subspace clustering using low-rank representation (LRR) [73], thresholding-based subspace clustering (TSC) [74], subspace clustering using nearest neighborhood search (NSN+Spec) [75], and tensor trace maximization (TTM) [8]. Note that SCC [36], SGC [76], and Tetris [8]) are not applicable to the sketching scenario due to their iterative natures.



(a)

(b)

(c)

Fig. 1. (a) Distribution of edge weights for the Hopkins 155 data set. Notice that homogeneous edges have larger weights on average. This implies that the hypergraph constructed from tensor-based approaches respects our model. (b) Fractional error of various algorithms. We report the fractional error of each algorithm for varying n/k and σ (a lighter color implies a lower error fraction). Note that we include iterative algorithms (SCC, SGC, Tetris) although they cannot be utilized in the sketching scenario. We can see that our approach has a comparable performance to the state of the arts. (c) Average run time comparison with prior subspace clustering algorithms. We can observe that our proposed algorithm scales nearly linearly in n while others do not.

and co-authorship networks [78]. We remark that while our model assumes equal-sized hyperedges, the HSCLR algorithm is applicable even when the size of hyperedges vary, which is the case for some of these applications. However, the success of the refinement step is contingent upon whether or not the average weight of homogeneous edges is larger than that of heterogeneous edges. While this assumption is shown to hold for the subspace clustering problem, whether or not this assumption holds for the other applications is an interesting future direction.

VII. Conclusion

In this paper, we develop two hypergraph clustering algorithms: HSC and HSCLR. Our main contribution lies in performance analysis of them under a new hypergraph model, which we call the weighted SBM. Our results improve upon the state of the arts, and firstly settle the order-optimal results. Further, we show that HSCLR achieves the information-theoretic limits of a certain hypergraph model. We also develop a sketching algorithm for subspace clustering based on HSCLR, and empirically show that the new algorithm outperforms the existing ones.

We conclude our paper with future research directions.

- *Detection threshold:* In [31], a sharp threshold for detection is conjectured. Further, the non-backtracking method is conjectured to be optimal. Proving these conjectures still remains open. The optimality of HSC is also open.
- *Consistency threshold:* The fundamental limits for weak/strong consistency under the general weighted SBM are unknown. An important open problem is to characterize the general limits in terms of the model parameters $(n, d, k, p, q, \alpha_n)$.

Appendix A

We first note that the overall structure of the proof resembles the ones in [44], [63], except that the entries of A are not independent. This is because each hyperedge's weight is added to more than one entries of $\mathbf{A}^0$ in our case, resulting in dependency structure between all elements of the matrix. See Remark 7 for more details.

We begin with some preliminaries: Let $\nu := \binom{n-2}{d-2} p\alpha_n \geq \max_{i,j} \mathbb{E}[\mathbf{A}_{i,j}]$; let $B := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 \leq 1\}$; let $D_\delta := \{\mathbf{x} = (x_1, x_2, \ldots, x_n) \in B : \frac{\sqrt{n}x_i}{\delta} \in \mathbb{Z}\}$; for a matrix $\mathbf{C}$, $\text{den}_{\mathbf{C}}(\mathcal{A}, \mathcal{B}) := \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{B}} \mathbf{C}_{i,j}$; and for a matrix $\mathbf{C}$ and a subset I, let $\mathbf{C}^I$ be the matrix obtained from $\mathbf{C}$ by zeroing out all rows and columns in subset I. The following large deviation results will be frequently used throughout the proof:

Lemma 5: Let $S = \sum_{i=1}^n X_i$, where $0 \leq X_i \leq b$ for each i for $b > 0$. There exist constants $c_7 > 0$ depending only on b such that the following holds for any $a \geq \mathbb{E}[S]$ and $k \geq c_7$:

$$\Pr(S > k \cdot a) \leq \exp\left(-\frac{1}{2b} k \log k \cdot a\right).$$

■

Proof: See Appendix D-A.

We consider the most challenging case where $\binom{n}{d}\alpha_n = \Theta(n)$, i.e., $\nu = \Theta(1/n)$. First, note that $\mathbf{P} - \mathbb{E}[\mathbf{A}]$ is a diagonal matrix whose entries are $O(\nu)$. Hence, $\|\mathbf{P} - \mathbb{E}[\mathbf{A}]\| = O(\sqrt{n\nu})$. Thus, it suffices to show that

$$\left\|\mathbf{A}^0 - \mathbb{E}[\mathbf{A}]\right\| = O(\sqrt{n\nu}). \tag{11}$$

Lemma 6: Let $\mathbf{C}$ be a $n \times n$ matrix. For $0 < \delta < 1$,

$$\|\mathbf{C}\| \leq (1 - 3\delta)^{-1} \max_{\mathbf{x} \in D_\delta} \left|\mathbf{x}^T \mathbf{C} \mathbf{x}\right|.$$

Proof: See Appendix D-C. ■

Due to Lemma 6, one can replace (11) with a more tractable statement at the cost of the constant: $\sup_{\mathbf{x} \in D_\delta} |\mathbf{x}^T(\mathbf{A}^0 - \mathbb{E}[\mathbf{A}])\mathbf{x}| = O(\sqrt{n\nu})$. For a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in D_\delta$, define $S_\delta(\mathbf{x}) := \left\{(i,j) : |x_i x_j| < \delta^2 \sqrt{\frac{\nu}{n}}\right\}$ for $0 < \delta < 1$. Then, one has:

$$\sup_{\mathbf{x} \in B} \left|\mathbf{x}^T(\mathbf{A}^0 - \mathbb{E}[\mathbf{A}])\mathbf{x}\right| = \sup_{\mathbf{x} \in B} \left|\sum_{(i,j)} \left[\mathbf{A}^0_{i,j} x_i x_j\right] - \mathbf{x}^T \mathbb{E}[\mathbf{A}]\mathbf{x}\right|.$$

Let $(T1) = \sup_{\mathbf{x} \in B} |\sum_{(i,j) \in S_\delta(\mathbf{x})} [\mathbf{A}^0_{i,j} x_i x_j] - \mathbf{x}^T \mathbb{E}[\mathbf{A}]\mathbf{x}|$ and $(T2) = \sup_{\mathbf{x} \in B} |\sum_{(i,j) \in S_\delta(\mathbf{x})^c} [\mathbf{A}^0_{i,j} x_i x_j]|$. Then, the above quantity is bounded above by $(T_1) + (T_2)$. We now show that each of (T_1) and (T_2) is $O(\sqrt{n\nu})$.

A. Proof of $(T1)$

We denote by J the random subset of $[n]$ that corresponds to the removed rows and columns during the processing step (see step 2 of Algorithm 1). For a sufficiently large constant $c_{12} > 0$ (to be chosen later) and $I \subset [n]$, define the event

$$E_I = \left\{ \sup_{\mathbf{x}\in D_\delta} \left| \sum_{(i,j)\in S_\delta(\mathbf{x})} \left[\mathbf{A}^I_{i,j} x_i x_j\right] - \mathbf{x}^T \mathbb{E}[\mathbf{A}]\mathbf{x} \right| > c_{12}\cdot\sqrt{n\nu} \right\}.$$

Then, it is sufficient to show that $\Pr(E_J) \to 0$. Note that the following upper bound holds for:

$$\begin{aligned}
\Pr(E_J) &= \sum_{I\subset[n]} \left[\Pr(E_J, J = I)\right] \\
&\le \sum_{|I|\le(n\nu)^{-3}n} \left[\Pr(E_I, J = I)\right] \\
&\quad + \sum_{|I|\ge(n\nu)^{-3}n} \left[\Pr(E_I, J = I)\right] \\
&\le \sum_{|I|\le(n\nu)^{-3}n} \left[\Pr(E_I)\right] + \sum_{|I|\ge(n\nu)^{-3}n} \Pr(J = I) \\
&= \sum_{|I|\le(n\nu)^{-3}n} \left[\Pr(E_I)\right] + \Pr\left(|J| \ge (n\nu)^{-3}n\right).
\end{aligned}$$

The following lemma bounds the number of removed rows (and columns).

Lemma 7: For some $c_{thr} > 0$ (depending only on d), there exists a constant $c_8 > 0$ such that if $n\nu \ge c_8$, then w.p. $1 - \exp(-\Omega(n))$, $|\{i : \mathrm{den}_{\mathbf{A}}(i,[n]) \ge c_{thr}\cdot n\nu\}| \le (n\nu)^{-3}n$.

Proof: See Appendix D-D. ■

By Lemma 7, for $n\nu \ge c_8$, $\Pr\left(|J| \ge (n\nu)^{-3}n\right) \le e^{-\Omega(n)}$.

As there are at most $2^n = e^{n\log 2}$ many subsets of $[n]$, due to the union bound, the proof for $(T1)$ will be completed after showing that for a fixed $|I| \le (n\nu)^{-3}n$, $\Pr(E_I) \le O\left(e^{-2\log 2n}\right)$. Observe that

$$\begin{aligned}
&\left| \sum_{(i,j)\in S_\delta(\mathbf{x})} \left[\mathbf{A}^I_{i,j} x_i x_j\right] - \mathbf{x}^T \mathbb{E}[\mathbf{A}]\mathbf{x} \right| \\
&\le \underbrace{\left| \mathbf{x}^T \left(\mathbb{E}[\mathbf{A}^I] - \mathbb{E}[\mathbf{A}]\right)\mathbf{x} \right|}_{(E1)} + \underbrace{\left| \textstyle\sum_{(i,j)\in S_\delta(\mathbf{x})^c} \left[\mathbb{E}[\mathbf{A}^I_{i,j}] x_i x_j\right] \right|}_{(E2)} \\
&\quad + \underbrace{\left| \textstyle\sum_{(i,j)\in S_\delta(\mathbf{x})} \left[\left(\mathbf{A}^I_{i,j} - \mathbb{E}[\mathbf{A}^I_{i,j}]\right) x_i x_j\right] \right|}_{(E3)},
\end{aligned}$$

and hence we will show that there exist $c_{13}, c_{14}, c_{15} > 0$ such that $\sup_{\mathbf{x}\in D_\delta}(E1) \le c_{13}\sqrt{n\nu}$, $\sup_{\mathbf{x}\in D_\delta}(E2) \le c_{14}\sqrt{n\nu}$, and $\sup_{\mathbf{x}\in D_\delta}(E3) \le c_{15}\sqrt{n\nu}$ with probability $1 - O(e^{-2n\log 2})$, respectively. Having shown these, the proof for $(T1)$ is completed by taking $c_{12} := c_{13} + c_{14} + c_{15}$.

i) $(E1)$: As $|I| \le (n\nu)^{-3}n$,

$$\begin{aligned}
&\left|\mathbf{x}^T \left(\mathbb{E}[\mathbf{A}^I] - \mathbb{E}[\mathbf{A}]\right)\mathbf{x}\right| \le \left\|\mathbb{E}[\mathbf{A}^I] - \mathbb{E}[\mathbf{A}]\right\| \\
&\le \left\|\mathbb{E}[\mathbf{A}^I] - \mathbb{E}[\mathbf{A}]\right\|_F \le \sqrt{2(n\nu)^{-3}n^2\cdot\nu^2} = \sqrt{2}(n\nu)^{-1/2}.
\end{aligned}$$

Hence, by taking $c_{13} = \sqrt{2}$, $\sup_{\mathbf{x}\in D_\delta}(E1) \le c_{13}\sqrt{n\nu}$ holds with probability 1 for $n\nu \ge 1$.

ii) $(E2)$: As $\nu \ge \max_{i,j}\mathbb{E}[\mathbf{A}_{i,j}]$,

$$\begin{aligned}
&\left| \sum_{(i,j)\in S_\delta^c} \left[\mathbb{E}[\mathbf{A}^I_{i,j}] x_i x_j\right] \right| \le \nu \sum_{\substack{(i,j)\in S_\delta(\mathbf{x})^c \\ i\neq j}} |x_i x_j| \\
&= \nu \sum_{\substack{(i,j)\in S_\delta(\mathbf{x})^c \\ i\neq j}} \frac{x_i^2 x_j^2}{|x_i x_j|} \overset{(a)}{\le} \frac{1}{\delta^2}\sqrt{n\nu} \\
&\sum_{\substack{(i,j)\in S_\delta(\mathbf{x})^c \\ i\neq j}} x_i^2 x_j^2 \overset{(b)}{\le} \frac{1}{\delta^2}\sqrt{n\nu},
\end{aligned}$$

where (a) is due to the definition of $S_\delta(\mathbf{x})$, and (b) follows since $\|\mathbf{x}\| \le 1$. Hence, by taking $c_{14} = \frac{1}{\delta^2}$, $\sup_{\mathbf{x}\in D_\delta}(E2) \le c_{14}\sqrt{n\nu}$ holds with probability 1.

iii) $(E3)$: Let $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in D_\delta$ be fixed. We have

$$\begin{aligned}
&\sum_{(i,j)\in S_\delta(\mathbf{x})} \left[\left(\mathbf{A}^I_{i,j} - \mathbb{E}[A^I_{i,j}]\right) x_i x_j\right] \\
&= \sum_{\substack{(i,j)\in S_\delta(\mathbf{x}) \\ i\neq j}} \left[x_i x_j \mathbb{I}\{i\notin I, j\notin I\} \sum_{\substack{e\in\mathcal{E} \\ \{i,j\}\subset e}} \left[W_e - \mathbb{E}[W_e]\right] \right] \\
&= \sum_{e\in\mathcal{E}} \underbrace{\left[\left(W_e - \mathbb{E}[W_e]\right) \textstyle\sum_{\substack{(i,j)\in S_\delta(\mathbf{x}) \\ i\neq j,\, \{i,j\}\subset e}} \left[x_i x_j \mathbb{I}\{i\notin I, j\notin I\}\right] \right]}_{=:Y_e}.
\end{aligned}$$

Note that $\{Y_e\}_{e\in\mathcal{E}}$ is a collection of independent random variables. To apply Bernstein inequality to $\sum_{e\in\mathcal{E}} Y_e$, we do some preliminary calculations. First, it easily follows from the definition of S_δ that

$$\begin{aligned}
|Y_e| &\le \left| \left(W_e - \mathbb{E}[W_e]\right) \sum_{\substack{(i,j)\in S_\delta(\mathbf{x}) \\ i\neq j,\, \{i,j\}\subset e}} \left[x_i x_j \mathbb{I}\{i\notin I, j\notin I\}\right] \right| \\
&\le \sum_{\substack{(i,j)\in S_\delta(\mathbf{x}) \\ i\neq j,\, \{i,j\}\subset e}} |x_i x_j| \le \delta^2 \sqrt{\frac{\nu}{n}} \cdot 2\binom{d}{2} \le d^2\delta^2\sqrt{\frac{\nu}{n}}.
\end{aligned}$$

Next, we compute a bound on the sum of variances:

$$\begin{aligned}
&\sum_{e\in\mathcal{E}} \mathbb{E}[Y_e^2] \\
&\overset{(a)}{\le} \sum_{e\in\mathcal{E}} \left[d^2\mathbb{E}[W_e^2] \sum_{\substack{(i,j)\in S_\delta(\mathbf{x}) \\ i\neq j,\, \{i,j\}\subset e}} \left[x_i^2 x_j^2 \mathbb{I}\{i\notin I, j\notin I\}\right] \right] \\
&\overset{(b)}{\le} d^2 \sum_{e\in\mathcal{E}} \left[\mathbb{E}[W_e] \sum_{\substack{(i,j) \\ i\neq j, \{i,j\}\subset e}} \left[x_i^2 x_j^2\right] \right] \\
&= d^2 \sum_{\substack{(i,j) \\ i\neq j}} \left[x_i^2 x_j^2 \mathbb{E}[\mathbf{A}_{i,j}]\right] \le d^2\nu \sum_{\substack{(i,j) \\ i\neq j}} x_i^2 x_j^2 \overset{(c)}{\le} d^2\nu,
\end{aligned}$$

where (a) is due to $(\sum_{i=1}^k a_i)^2 \le k\sum_{i=1}^k a_i^2$; (b) follows since $W_e \in [0,1]$; (c) follows since $\|\mathbf{x}\| \le 1$.

Thus, Bernstein inequality yields: $\Pr\left(\left|\sum_{e\in\mathcal{E}} Y_e\right| \geq t\right) \leq 2\exp\left(-\frac{t^2/2}{d^2\nu+\frac{1}{3}d^2\delta^2\sqrt{\frac{\nu}{n}}t}\right)$, we have

$$\Pr\left(\left|\sum_{e\in E} Y_e\right| \geq c_{15}\sqrt{n\nu}\right) \leq 2\exp\left(-\frac{c_{15}^2}{2d^2+\frac{2}{3}d^2\delta^2 c_{15}}n\right).$$

As $|D_\delta| = e^{\Theta(n)}$, the union bound yields

$$\Pr\left(\sup_{\mathbf{x}\in D_\delta}\left|\sum_{e\in E} Y_e\right| \geq c_{15}\sqrt{n\nu}\right) \leq e^{\Theta(n)}\Pr\left(\left|\sum_{e\in E} Y_e\right| \geq c_{15}\sqrt{n\nu}\right) \leq 2e^{\Theta(n)}\exp\left(-\frac{c_{15}^2}{2d^2+\frac{2}{3}d^2\delta^2 c_{15}}n\right),$$

and hence by choosing c_{15} sufficiently large, one can ensure that $\sup_{\mathbf{x}\in D_\delta}(E3) \leq c_{15}\sqrt{n\nu}$ w.p. $1 - O(e^{-2n\log 2})$.

Since $n\nu \geq c_8$ and $n\nu \geq 1$, $n\nu$ should be greater than or equal to $\max\{c_8, 1\}$, so one can take $c_5 = \max\{c_8, 1\}$.

B. Proof of $(T2)$

This case immediately follows from a celebrated combinatorial technique proposed in [63]. We summarize their results.

Definition 8: We say *the bounded density property* holds with constants $\alpha, \beta, \gamma > 0$ if the following two hold:

1) For each node u, $\mathrm{den}_{\mathbf{A}^0}(u, [n]) \leq \alpha \cdot n\nu$.
2) For any two subsets A, B, either $\mathrm{den}_{\mathbf{A}^0}(\mathcal{A}, \mathcal{B}) \leq \beta \cdot \nu |\mathcal{A}||\mathcal{B}|$ or $\mathrm{den}_{\mathbf{A}^0}(\mathcal{A}, \mathcal{B}) \log\frac{\mathrm{den}_{\mathbf{A}_0}(\mathcal{A},\mathcal{B})}{\nu|\mathcal{A}||\mathcal{B}|} \leq \gamma \cdot \max\{|\mathcal{A}|, |\mathcal{B}|\}\log\frac{n}{\max\{|\mathcal{A}|,|\mathcal{B}|\}}$.

Proposition 4 ([44], [63]): If the bounded density property holds with some constant α, β, γ, then $(T2) = O(\sqrt{n\nu})$.

Therefore, one only needs to show that the bounded density property holds with high probability to finish the proof.

Lemma 8: With probability $1 - O(n^{-1})$, the bounded density property holds with some constants c_9, c_{10}, c_{11}.

Proof: See Appendix D-E. ■

APPENDIX B

For notational simplicity, as $k = 2$, we represent partition functions Φ_{HSC}, Ψ by binary vectors $\mathbf{X}, \mathbf{Z} \in \{0, 1\}^n$. We define some notations: Let $\mathbf{W} = [W_e]_{e\in\mathcal{E}}$; for a vector $\mathbf{V} = [\mathbf{V}_i]_{1\leq i\leq n} \in \{0, 1\}^n$ and $e = \{i_1, i_2, \ldots, i_d\} \in \binom{[n]}{d}$, let $f_e(\mathbf{V}) = \mathbb{I}\{V_{i_1} = V_{i_2} = \ldots V_{i_d}\}$; let $\mathbf{F}(\mathbf{V}) = [f_e(\mathbf{V})]_{e\in\mathcal{E}}$.

A straightforward calculation yields for any two binary vectors $\mathbf{X}$ and $\mathbf{Y}$, the likelihood of $\mathbf{X}$ is greater than that of $\mathbf{Y}$ if and only if $\mathsf{d}(\mathbf{W}, \mathbf{F}(\mathbf{X})) < \mathsf{d}(\mathbf{W}, \mathbf{F}(\mathbf{Y}))$, where $\mathsf{d}(\mathbf{X}, \mathbf{Y}) := |\{i \in [n] : \mathbf{X}_i \neq \mathbf{Y}_i\}|$ for any $\mathbf{X}$ and $\mathbf{Y}$.

To make HSCLR better adapted to the model, we modify the algorithm as follows:

1) We apply HSC to $([n], \mathbf{W}')$, where $\mathbf{W}'$ is obtained from $\mathbf{W}$ by replacing the erasure weights x's with 0's.
2) We then employ a likelihood-based refinement rule:
$$\mathbf{X}_i \leftarrow \begin{cases}\mathbf{X}_i & \text{if } \mathsf{d}(\mathbf{W}, \mathbf{F}(\mathbf{X})) < \mathsf{d}(\mathbf{W}, \mathbf{F}(\mathbf{X} \oplus \mathbf{e}_i));\\ \mathbf{X}_i \oplus 1 & \text{otherwise.}\end{cases}$$

Remark 9: Notice that one can employ such a likelihood-based estimator only when edge distributions are fully specified.

We now begin the main proof. We consider the most challenging regime where $\binom{n}{d}p = \Theta(n\log n)$, and suppose

$$\binom{n}{d}\alpha_n \geq (1+\epsilon)\frac{2^{d-2}}{d}\frac{n\log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \tag{12}$$

for a fixed $\epsilon > 0$. For simplicity, we assume that n is even, and fix the ground truth to be $\mathbf{A} = (\underbrace{1, \ldots, 1}_{n/2}, \underbrace{0, \ldots, 0}_{n/2})$; for other cases, the proof follows similarly.

Let $\mathbf{X}$ be the output of the first stage. By Theorem 1, one can see that $\mathbf{X}$ is weakly consistent. Without loss of generality, we assume for an arbitrarily small $\eta > 0$ that

$$\mathbf{X} = (\underbrace{0, \ldots, 0}_{\eta n}, \underbrace{1, 1, 1, \ldots, 1, 1, 1}_{n/2-\eta n}, \underbrace{1, \ldots, 1}_{\eta n}, \underbrace{0, 0, 0, \ldots, 0, 0, 0}_{n/2-\eta n}).$$

Indeed, $\mathbf{X}$ needs not have the same number of 0's and 1's but the other cases can be handled similarly using the same arguments.

As in the proof of Theorem 2 (see Section IV-B), due to the union bound, it is enough to show that the probability of having node 1's affiliation incorrect after refinement is $o(n^{-1})$, i.e.,

$$\Pr(\text{node 1 is incorrect after refinement}) = o(n^{-1}).$$

By the new refinement rule,

$$\Pr(\text{node 1 is incorrect after refinement}) = \Pr\Big(0 < \mathsf{d}(\mathbf{W}, \mathbf{F}(\mathbf{X} \oplus \mathbf{e}_1)) - \mathsf{d}(\mathbf{W}, \mathbf{F}(\mathbf{X}))\Big).$$

The following lemma states that the difference of hamming distances can be viewed as the sum of random variables.

Lemma 9: $P_i, P_i' \overset{\text{i.i.d.}}{\sim} \mathsf{Bern}(\alpha_n)$ and $\Theta_i, \Theta_i' \overset{\text{i.i.d.}}{\sim} \mathsf{Bern}(\theta)$. Then,

$$\mathsf{d}(\mathbf{W}, \mathbf{F}(\mathbf{X} \oplus \mathbf{e}_1)) - \mathsf{d}(\mathbf{W}, \mathbf{F}(\mathbf{X})) = \sum_{i=1}^{2\binom{n/2-\eta n}{d-1}} P_i(2\Theta_i - 1)) + \sum_{i=1}^{2\binom{n/2}{d-1}-2\binom{n/2-\eta n}{d-1}} P_i'(1 - 2\Theta_i').$$

Proof: See Appendix D-F. ■

Let $V_1 = \binom{n/2}{d-1}$ and $V_2 = \binom{n/2-\eta n}{d-1}$. By Lemma 9,

$$\Pr\Big(0 < \mathsf{d}(\mathbf{W}, \mathbf{F}(\mathbf{X} \oplus \mathbf{e}_1)) - \mathsf{d}(\mathbf{W}, \mathbf{F}(\mathbf{X}))\Big) = \Pr\left(-\sum_{i=1}^{2V_1-2V_2} P_i'(1-2\Theta_i') < \sum_{i=1}^{2V_2} P_i(2\Theta_i - 1))\right) \overset{(a)}{\leq} \Pr\left(-\sum_{i=1}^{2V_1-2V_2} P_i' < \sum_{i=1}^{2V_2} P_i(2\Theta_i - 1))\right) \overset{(b)}{=} \Pr\left(-\sum_{i=1}^{O(\eta)V_1} P_i' < \sum_{i=1}^{2V_2} P_i(2\Theta_i - 1))\right). \tag{13}$$

where (a) is due to $|1 - 2\Theta'_i| \le 1$; (b) is due to $2V_1 - 2V_2 = O(\eta)V_1$.

In view of Lemma 4, one can similarly show that

$$\Pr\left(\sum_{i=1}^{O(\eta)V_1} P'_i > \frac{V_1 \alpha_n}{\sqrt{\log(1/\eta)}}\right) = o(n^{-1}), \tag{14}$$

provided that η is sufficiently small. Thus,

$$(13) \le \Pr\left(-\frac{V_1\alpha_n}{\sqrt{\log(1/\eta)}} < \sum_{i=1}^{2V_2} P_i(2\Theta_i - 1))\right) + o(n^{-1})$$

Lemma 10: For an integer $K > 0$, let $\{P_i\}_{i=1}^K \overset{\text{i.i.d.}}{\sim} \mathsf{Bern}(\alpha_n)$ and $\{\Theta_i\}_{i=1}^K \overset{\text{i.i.d.}}{\sim} \mathsf{Bern}(\theta)$. Then, for any $\ell > 0$

$$\Pr\left(\log\left(\frac{1-\theta}{\theta}\right)\sum_{i=1}^K P_i(2\Theta_i - 1) \ge -\ell\right)$$
$$\le e^{\frac{1}{2}\ell - K\left(\alpha_n(\sqrt{1-\theta}-\sqrt{\theta})^2 + O(\alpha_n^2)\right)}.$$

Proof: See Appendix D-G. ■

By Lemma 10,

$$\Pr\left(-\frac{V_1\alpha_n}{\sqrt{\log(1/\eta)}} < \sum_{i=1}^{2V_2} P_i(2\Theta_i - 1))\right)$$
$$\le e^{\frac{1}{2}\frac{\log\left(\frac{1-\theta}{\theta}\right)}{\sqrt{\log(1/\eta)}}V_1\alpha_n - 2V_2\left(\alpha_n(\sqrt{1-\theta}-\sqrt{\theta})^2 + O(\alpha_n^2)\right)}. \tag{15}$$

Note that as $\alpha_n = o(1)$,

$$\frac{1}{2}\frac{\log\left(\frac{1-\theta}{\theta}\right)}{\sqrt{\log(1/\eta)}}V_1\alpha_n - 2V_2\left(\alpha_n(\sqrt{1-\theta}-\sqrt{\theta})^2 + O(\alpha_n^2)\right)$$
$$= (1+o(1))\left[\frac{\log\left(\frac{1-\theta}{\theta}\right)}{\sqrt{\log(1/\eta)}2^d}\frac{d}{n}\binom{n}{d}\alpha_n - 2\left(\frac{1}{2}-\eta\right)^{d-1}\right.$$
$$\left.\cdot \frac{d}{n}\binom{n}{d}\left(\alpha_n(\sqrt{1-\theta}-\sqrt{\theta})^2 + O(\alpha_n^2)\right)\right]$$
$$= (1+o(1))\left[\frac{\log\left(\frac{1-\theta}{\theta}\right)}{\sqrt{\log(1/\eta)}2^d}\frac{d}{n}\binom{n}{d}\alpha_n - 2\left(\frac{1}{2}-\eta\right)^{d-1}\right.$$
$$\left.\cdot \frac{d}{n}\binom{n}{d}\alpha_n(\sqrt{1-\theta}-\sqrt{\theta})^2\right]$$
$$\to -\frac{1}{2^{d-2}}\frac{d}{n}\binom{n}{d}\alpha_n(\sqrt{1-\theta}-\sqrt{\theta})^2$$

as $\eta \to 0^+$ and $n \to \infty$. Thus, $(15) \le e^{-(1+\epsilon/2)\log n} = o(n^{-1})$ for sufficiently large n and small η.

Appendix C

To extend the analysis to the planted clique model, we need another type of error fraction, which is defined as follows:

$$\mathrm{err}'(\Phi) := \min_{\Pi \in \mathcal{P}} \max_{1 \le j \le k} \frac{1}{n_j}\left|\{i \in \Psi^{-1}(j) : \Pi(\Phi(i)) \ne j\}\right|.$$

Note that err′ characterizes the maximum value of within-cluster error fraction over all clusters. Let us denote the smallest singular value of $\mathbf{B}$ (defined in Section III-A) by σ and the size of smallest cluster by $n_{\min}$. Then, following [5], one can prove the following result by tweaking the proof of Theorem 1:

Theorem 4: For some c_{14}, c_{15}, the following holds: if $\binom{n}{d}\alpha_n \ge c_{14}n$ and $c_{15}(1+\epsilon)\frac{kn\binom{n-2}{d-2}}{\alpha_n n_{\min}^2\sigma^2} < 1$, then w.p. exceeding $1 - O(n^{-1})$,

$$\mathrm{err}'(\Phi_{\text{HSC}}) \le c_{15}(1+\epsilon)\frac{kn\binom{n-2}{d-2}}{\alpha_n n_{\min}^2\sigma^2}. \tag{16}$$

We now demonstrate how Theorem 4 guarantees the detection of planted clique when $s \ge c^* \cdot \sqrt{n}$ for some constant c^*. To apply Theorem 4, we need to first compute σ of

$$\mathbf{B} = \begin{pmatrix} \frac{1}{2}\binom{s-2}{d-2} + \frac{1}{2}\binom{n-2}{d-2} & \frac{1}{2}\binom{n-2}{d-2} \\ \frac{1}{2}\binom{n-2}{d-2} & \frac{1}{2}\binom{n-2}{d-2} \end{pmatrix}.$$

Using the fact that the minimum singular value of $\left(\begin{smallmatrix} a+b & a \\ a & a \end{smallmatrix}\right)$ is $\frac{2b}{\sqrt{4+(\frac{b}{a})^2}+2+\frac{b}{a}}$, we have

$$\sigma = \frac{2\binom{s-2}{d-2}}{\sqrt{4+\left(\frac{\binom{s-2}{d-2}}{\binom{n-2}{d-2}}\right)^2}+2+\frac{\binom{s-2}{d-2}}{\binom{n-2}{d-2}}} = \Theta\left(s^{d-2}\right).$$

Hence, by Theorem 4 (as $\alpha_n = 1$), Thus, whenever $s = \Omega(\sqrt{n})$,

$$\mathrm{err}'(\Phi_{\text{HSC}}) \le c_{15}(1+\epsilon)\frac{2n\binom{n-2}{d-2}}{s^2\sigma^2} = O\left(\frac{n^{d-1}}{s^{2(d-1)}}\right) = O(1).$$

Appendix D

A. Proofs of Lemma 3 and Lemma 5

Without loss of generality, we will prove the lemmas assuming that $\mathbb{E}[X_i] > 0$ for all i. We first obtain a useful bound on the moment generating function (mgf) of S. For an arbitrary $\lambda > 0$,

$$\mathbb{E}[\exp\{\lambda S\}] = \mathbb{E}\left[\exp\left\{\lambda\left(\sum_{i=1}^n X_i\right)\right\}\right]$$
$$\le \prod_{i=1}^n\left(1 + \frac{(e^{\lambda b}-1)}{b}\mathbb{E}[X_i]\right)$$
$$\le \left(1 + \frac{(e^{\lambda b}-1)}{b}\frac{\sum_{i=1}^n \mathbb{E}[X_i]}{n}\right)^n, \tag{17}$$

where the first inequality holds since $\frac{e^{\lambda x}-1}{x} \le \frac{e^{\lambda b}-1}{b}$ holds for all $0 < x \le b$, and the second inequality holds due to the AM-GM inequality. We now prove the lemmas using this bound.

1) Proof of Lemma 5: Using Markov's inequality and (17),

$$\Pr(S > x) = \Pr(e^{\lambda S} > e^{\lambda x}) \le \exp\{-\lambda x\}\mathbb{E}[\exp\{\lambda S\}]$$
$$\le \exp\{-\lambda x\}\left(1 + \frac{(e^{\lambda b}-1)}{b}\frac{\mathbb{E}[S]}{n}\right)^n.$$

By choosing $\lambda = \frac{1}{b}\log(1+\frac{bx}{\mathbb{E}[S]})$, i.e., $x = \frac{(e^{b\lambda}-1)}{b}\mathbb{E}[S]$, we have

$$\begin{aligned}\Pr(S > x) &\le \exp\left\{-\frac{x}{b}\log\left(1+\frac{bx}{\mathbb{E}[S]}\right)\right\}\left(1+\frac{x}{n}\right)^n \\ &\le \exp\left\{-\frac{x}{b}\log\left(1+\frac{bx}{\mathbb{E}[S]}\right)\right\}\exp(x) \\ &= \exp\left[-\frac{x}{b}\cdot\left\{\log\left(1+\frac{bx}{\mathbb{E}[S]}\right)-b\right\}\right].\end{aligned}$$

By setting $x = ka$, we have

$$\begin{aligned}\Pr(S > k\cdot a) &= \exp\left[-\frac{ka}{b}\cdot\left\{\log\left(1+\frac{bka}{\mathbb{E}[S]}\right)-b\right\}\right] \\ &\le \exp\left[-k\frac{\log(1+bk)-b}{b}\cdot a\right],\end{aligned}$$

where the inequality holds since $a \ge \mathbb{E}[S]$. Since $[\log(1+bk)-b] \sim \log(k)$, $\log(1+bk)-b \ge \frac{1}{2}\log(k)$ holds for all $k \ge c_7$, where c_7 is some positive constant depending only on b. Applying this inequality to the above bound completes the proof.

2) Proof of Lemma 3: Since the proof bears great similarity to the conventional case [67], we only show the upper bound. Using Markov's inequality and (17) with $b = 1$,

$$\begin{aligned}\Pr\left(S > (1+\delta)\mathbb{E}[S]\right) &= \Pr\left(e^{\lambda S} > e^{\lambda(1+\delta)\mathbb{E}[S]}\right) \\ &\le e^{-\lambda(1+\delta)\mathbb{E}[S]}\left(1+(e^{\lambda}-1)\frac{\mathbb{E}[S]}{n}\right)^n\end{aligned}$$

By taking $\lambda = \log(1+\delta)$, we obtain

$$\begin{aligned}\Pr\left(S > (1+\delta)\mathbb{E}[S]\right) &\le e^{-(1+\delta)\log(1+\delta)\mathbb{E}[S]}\left(1+\frac{\delta\mathbb{E}[S]}{n}\right)^n \\ &\le e^{-(1+\delta)\log(1+\delta)\mathbb{E}[S]+\delta\mathbb{E}[S]} \le e^{-\frac{\delta^2}{2+\delta}\mathbb{E}[S]},\end{aligned} \tag{18}$$

where the last equality holds since $\log(1+\delta) \ge \frac{\delta}{1+\delta/2}$. This completes the proof of the upper bound.

B. Proof of Lemma 4

Assume that $|\mathcal{E}^{(i)}(j)\cap\mathcal{E}_h| = c\cdot\gamma|\mathcal{E}^{(i)}(j)|$ for some constant $c > 0$. For simplicity, let us write $\sum_{e\in\mathcal{E}^{(i)}(j)\cap\mathcal{E}_h} W_e$ as $\sum_{i=1}^{c\cdot\gamma|\mathcal{E}^{(i)}(j)|} W_i$. Then we get:

$$\begin{aligned}&\Pr\left(\sum_{e\in\mathcal{E}^{(i)}(j)\cap\mathcal{E}_h} W_e > \frac{p\alpha_n|\mathcal{E}^{(i)}(j)|}{\sqrt{\log(1/\gamma)}}\right) \\ &= \Pr\left(\sum_{i=1}^{c\cdot\gamma|\mathcal{E}^{(i)}(j)|} W_i > \frac{1}{c\gamma\sqrt{\log(1/\gamma)}}\cdot c\gamma p\alpha_n|\mathcal{E}^{(i)}(j)|\right).\end{aligned} \tag{19}$$

From the proof of Lemma 3 (see (18)), one can deduce the following:

Corollary 4: Let S be the sum of m mutually independent random variables taking values in $[0, 1]$. For any $\delta > 0$, we have

$$\Pr\left(S > (1+\delta)\mathbb{E}[S]\right) \le e^{-\{(1+\delta)\log(1+\delta)-\delta\}\mathbb{E}[S]}. \tag{20}$$

We will apply Corollary 4 with $(1+\delta) = (c\gamma\sqrt{\log(1/\gamma)})^{-1}$.

As $(c\gamma\sqrt{\log(1/\gamma)})^{-1} \to \infty$ as $\gamma \to 0^+$, we may regard δ to be an arbitrarily large constant. Because $(1+\delta)\log(1+\delta)-\delta = (1+o(1))(1+\delta)\log(1+\delta)$ as $\delta \to \infty$, in what follows, we will replace the upper bound (20) with $e^{-(1+\delta)\log(1+\delta)\mathbb{E}[S]}$:

$$\begin{aligned}(19) &\le \left(\frac{1}{ec\gamma\sqrt{\log(1/\gamma)}}\right)^{-\frac{1}{c\gamma\sqrt{\log(1/\gamma)}}c\gamma p\alpha_n|\mathcal{E}^{(i)}(j)|} \\ &\le \left(\frac{1}{ec\gamma\sqrt{\log(1/\gamma)}}\right)^{-\frac{1}{c\gamma\sqrt{\log(1/\gamma)}}c\gamma p\alpha_n|\mathcal{E}^{(i)}(j)|}.\end{aligned} \tag{21}$$

Since we consider the regime $\binom{n}{d}\alpha_n = \Theta(n\log n)$, $p\alpha_n|\mathcal{E}^{(i)}(j)| = c'\cdot\log n$ for some constant $c' > 0$. Hence, the last term is equal to

$$\begin{aligned}&\left(\frac{1}{ec\gamma\sqrt{\log(1/\gamma)}}\right)^{-\frac{1}{\sqrt{\log(1/\gamma)}}c'\cdot\log n} \\ &= \exp\left(-\log n\cdot\left\{\sqrt{\log(1/\gamma)}-\frac{1+\log c+\frac{1}{2}\log(\log(1/\gamma))}{\sqrt{\log(1/\gamma)}}\right\}\right).\end{aligned}$$

Since the exponent diverges as $\gamma \to 0^+$, we prove the lemma.

C. Proof of Lemma 6

WLOG, assume that $\|C\| = \sup_{\mathbf{x}\in B}\mathbf{x}^T\mathbf{C}\mathbf{x}$; the case $\|C\| = -\inf_{\mathbf{x}\in B}\mathbf{x}^T\mathbf{C}\mathbf{x}$ follows similarly. Observe that the diameter of each cell resulting from discretization is δ. For a vector $\mathbf{d}$ such that $\|\mathbf{d}\|_2 \le \delta$ and $\mathbf{x}\in B$, $(\mathbf{x}+\mathbf{d})^T\mathbf{C}(\mathbf{x}+\mathbf{d}) - \mathbf{x}^T\mathbf{C}\mathbf{x} = 2\mathbf{d}^T\mathbf{C}\mathbf{x}+\mathbf{d}^T\mathbf{C}\mathbf{d}$. Thus, we get:

$$\begin{aligned}&\left|(\mathbf{x}+\mathbf{d})^T\mathbf{C}(\mathbf{x}+\mathbf{d})\right| - \left|\mathbf{x}^T\mathbf{C}\mathbf{x}\right| \\ &\le \left|(\mathbf{x}+\mathbf{d})^T\mathbf{C}(\mathbf{x}+\mathbf{d}) - \mathbf{x}^T\mathbf{C}\mathbf{x}\right| \le 2\left|\mathbf{d}^T\mathbf{C}\mathbf{x}\right| + \left|\mathbf{d}^T\mathbf{C}\mathbf{d}\right| \\ &\le 2\|\mathbf{d}\|\|\mathbf{C}\|\|\mathbf{x}\| + \|\mathbf{C}\|\|\mathbf{d}\|^2 \le 2\|\mathbf{d}\|\|\mathbf{C}\| + \|\mathbf{C}\|\|\mathbf{d}\|^2 \\ &\le 3\|\mathbf{d}\|\|\mathbf{C}\| \le 3\delta\|\mathbf{C}\|.\end{aligned}$$

Let $\mathbf{x}^* = \arg\sup_{\mathbf{x}\in B}\mathbf{x}^T\mathbf{C}\mathbf{x}$. Then, there exists $\mathbf{x}_0\in D_\delta$ such that $\|\mathbf{x}_0-\mathbf{x}^*\| \le \delta$, so

$$\begin{aligned}\|\mathbf{C}\| = (\mathbf{x}^*)^T\mathbf{C}\mathbf{x}^* &\le (\mathbf{x}_0)^T\mathbf{C}\mathbf{x}_0 + 3\delta\|\mathbf{C}\| \\ &\le \sup_{\mathbf{x}\in D_\delta}|\mathbf{x}^T\mathbf{C}\mathbf{x}| + 3\delta\|\mathbf{C}\|.\end{aligned}$$

By rearrangement, we get: $(1-3\delta)\|\mathbf{C}\| \le \sup_{\mathbf{x}\in D_\delta}|\mathbf{x}^T\mathbf{C}\mathbf{x}|$.

D. Proof of Lemma 7

Let us say node i is *bad* if $\text{den}_{\mathbf{A}}(i,[n]) \ge c_{\text{thr}}\cdot n\nu$ for some constant c_{thr} to be chosen later. Let $\delta := (n\nu)^{-3}$.

$$\begin{aligned}&\Pr(\text{there are more than } \delta n \text{ bad nodes}) \\ &\le \sum_{X\subset[n]:|X|=\delta n}\Pr(\text{every node in } X \text{ is bad}) \\ &\le \sum_{X\subset[n]:|X|=\delta n}\Pr\left(\text{den}_{\mathbf{A}}(X,[n]) \ge \delta n\cdot(c_{\text{thr}}\cdot n\nu)\right).\end{aligned}$$

Note that for any subsets $\mathcal{A}$ and $\mathcal{B}$,

$$\begin{aligned}\mathrm{den}_{\mathbf{A}}(\mathcal{A},\mathcal{B}) &= \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{B}}\mathbf{A}_{i,j} = \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{B}}\sum_{\substack{e\in\mathcal{E}\\ \{i,j\}\subset e}} W_e\\ &= \sum_{e\in\mathcal{E}}\left[W_e\left(\sum_{(i,j)\in\mathcal{A}\times\mathcal{B}:\{i,j\}\subset e} 1\right)\right], \qquad (22)\end{aligned}$$

i.e., $\mathrm{den}_{\mathbf{A}}(\mathcal{A},\mathcal{B})$ is a sum of independent random variables taking values in $[0, d^2]$. Hence, using the fact that $\mathbb{E}[\mathrm{den}_{\mathbf{A}}(X,[n])] \le \delta n^2\nu$ ($\because \nu \ge \max_{i,j}\mathbb{E}[\mathbf{A}_{i,j}]$) together with Lemma 5 (take $b = d^2$), there exists $c_7 > 0$ such that

$$\Pr\left(\mathrm{den}_{\mathbf{A}}(X,[n]) \ge k\delta n^2\nu\right) \le \exp\left(-\left(\frac{1}{2d^2}k\log k\right)\cdot\delta n^2\nu\right)$$

whenever $k \ge c_7$.

By taking $c_{\text{thr}} = c_7$,

$$\begin{aligned}&\sum_{X\subset[n]:|X|=\delta n}\Pr\left(\mathrm{den}_{\mathbf{A}}(X,[n]) \ge c_{\text{thr}}\delta n^2\nu\right)\\ &\le \binom{n}{\delta n}\exp\left(-\left(\frac{1}{2d^2}c_{\text{thr}}\log c_{\text{thr}}\right)\cdot\delta n^2\nu\right)\\ &\overset{(a)}{\le} \exp\left\{\left(\delta\log\frac{1}{\delta} + \delta - \left(\frac{1}{2d^2}c_{\text{thr}}\log c_{\text{thr}}\right)\cdot\delta n\nu\right)\cdot n\right\}.\end{aligned}$$

where (a) is due to the fact that $\binom{n}{m} \le \left(\frac{ne}{m}\right)^m$. Plugging back in $\delta = (n\nu)^{-3}$, we obtain

$$\begin{aligned}&\delta\log\frac{1}{\delta} + \delta - \left(\frac{1}{2d^2}c_{\text{thr}}\log c_{\text{thr}}\right)\cdot\delta n\nu\\ &= -3\log(n\nu)(n\nu)^{-3} + (n\nu)^{-3} - \left(\frac{1}{2d^2}c_{\text{thr}}\log c_{\text{thr}}\right)\cdot(n\nu)^{-2}. \qquad (23)\end{aligned}$$

Since $(23)\cdot(n\nu)^2 \to -\frac{1}{2d^2}c_{\text{thr}}\log c_{\text{thr}} < 0$ as $n\nu \to \infty$, there exists a constant c_8 such that $n\nu \ge c_8$ implies $(23) < 0$. This completes the proof.

E. Proof of Lemma 8

By taking $c_9 = c_{\text{thr}}$, the first part of Definition 8 follows easily by the definition of $\mathbf{A}^0$.

We now turn to the second part of Definition 8. Without loss of generality, we assume that $\mathcal{A}\cap\mathcal{B} = \emptyset$ and $|\mathcal{A}| \le |\mathcal{B}|$.

1) The case where $|\mathcal{B}| \ge \frac{n}{e}$:
It follows that $\nu|\mathcal{A}||\mathcal{B}| \ge \frac{\nu|\mathcal{A}|n}{e}$, and since we verified the first part of Definition 8, we obtain $\mathrm{den}_{\mathbf{A}^0}(\mathcal{A},\mathcal{B}) \le |\mathcal{A}|\cdot c_9 n\nu$. Hence, $\mathrm{den}_{\mathbf{A}^0}(A,B) \le c_9 e\nu|A||B|$.

2) The case where $|\mathcal{B}| < \frac{n}{e}$:
It suffices to show the property for the case where $\mathbf{A}^0$ is replaced by $\mathbf{A}$ due to the fact that $\mathrm{den}_{\mathbf{A}^0}(\mathcal{A},\mathcal{B}) \le \mathrm{den}_{\mathbf{A}}(\mathcal{A},\mathcal{B})$. Because of (22), $\mathrm{den}_{\mathbf{A}}(\mathcal{A},\mathcal{B})$ is a sum of independent random variables taking values in $[0, d^2]$. As $\mathbb{E}[\mathrm{den}_{\mathbf{A}}(\mathcal{A},\mathcal{B})] \le \nu|\mathcal{A}||\mathcal{B}|$ ($\because \nu \ge \max_{i,j}\mathbb{E}[\mathbf{A}_{i,j}]$), Lemma 5 ensures that there exist constants $c_7 > 0$ such that

$$\begin{aligned}&\Pr(\mathrm{den}_{\mathbf{A}}(A,B) > k\cdot\nu|A||B|)\\ &\qquad\le \exp\left(-\frac{1}{2d^2}k\log k\cdot\nu|A||B|\right)\end{aligned}$$

for any $k \ge c_7$ regardless of choices of $\mathcal{A}$ and $\mathcal{B}$.

Claim 2: Let

$$k_{a,b} := \max\left\{\min\left\{k \ge 1 \,:\, k\log k \ge \frac{14d^2}{\nu a}\log\frac{n}{b}\right\}, c_7\right\}.$$

Then, with probability $1 - O(n^{-1})$, the following holds: For any two subsets $\mathcal{A}$ and $\mathcal{B}$,

$$\mathrm{den}_{\mathbf{A}}(\mathcal{A},\mathcal{B}) \le k_{|\mathcal{A}|,|\mathcal{B}|}\nu|\mathcal{A}||\mathcal{B}|.$$

Proof: It is sufficient to prove the following: $\Pr(\bigcup_{\mathcal{A},\mathcal{B}}[\mathrm{den}_{\mathbf{A}}(\mathcal{A},\mathcal{B}) > k_{|\mathcal{A}|,|\mathcal{B}|}\nu|\mathcal{A}||\mathcal{B}|]) = O(n^{-1})$. Note that in the case of $|\mathcal{A}| = a$ and $|\mathcal{B}| = b$, $\Pr(\mathrm{den}_{\mathbf{A}}(|\mathcal{A}|,|\mathcal{B}|) > k_{\mathcal{A},\mathcal{B}}\cdot\nu|\mathcal{A}||\mathcal{B}|)$ is upper bounded by $\exp(-\frac{1}{2d^2}k_{a,b}\log k_{a,b}\cdot\nu ab)$ as $k_{a,b} \ge c_7$. Hence, the union bound yields:

$$\begin{aligned}&\Pr\left(\bigcup_{\mathcal{A},\mathcal{B}}[\mathrm{den}_{\mathbf{A}}(\mathcal{A},\mathcal{B}) > k_{\mathcal{A},\mathcal{B}}\cdot\nu|\mathcal{A}||\mathcal{B}|]\right)\\ &\le \sum_{a,b}\left[\binom{n}{a}\binom{n}{b}\exp\left(-\frac{1}{2d^2}k_{a,b}\log k_{a,b}\cdot\nu ab\right)\right].\end{aligned}$$

Since there are at most n^2 choices for (a,b), it is enough to show that $\binom{n}{a}\binom{n}{b}\exp\left(-\frac{1}{2d^2}k_{a,b}\log k_{a,b}\cdot\nu ab\right) \le \frac{1}{n^3}$ for any (a,b).

By the definition of "$k_{a,b}$", we have $k_{a,b}\log k_{a,b} \ge \frac{14d^2}{\nu a}\log\frac{n}{b}$. Hence,

$$\begin{aligned}\frac{1}{2d^2}k_{a,b}\log k_{a,b}\cdot\nu ab &\ge 7b\log\frac{n}{b} \overset{(a)}{\ge} a + b + 5b\log\frac{n}{b}\\ &\overset{(b)}{\ge} a + b + a\log na + b\log nb + 3\log n,\end{aligned}$$

where (a) follows since $a \le b \le \frac{n}{e}$; (b) follows since $x\log x$ is increasing on $[1, \frac{n}{e}]$. Thus, we have

$$\begin{aligned}&\exp\left(-\frac{1}{2d^2}k_{a,b}\log k_{a,b}\cdot\nu ab\right)\\ &\le \exp\left(-a\left(\log\frac{n}{a}+1\right) - b\left(\log\frac{n}{b}+1\right) - 3\log n\right).\end{aligned}$$

Further, since $\binom{n}{m} \le \left(\frac{ne}{m}\right)^m = \exp\left(m(\log n/m + 1)\right)$,

$$\binom{n}{a}\binom{n}{b} = \exp\left(a\left(\log n/a + 1\right) + b\left(\log n/b + 1\right)\right).$$

Thus, $\binom{n}{a}\binom{n}{b}e^{-\frac{1}{2d^2}k_{a,b}\log k_{a,b}\cdot\nu ab} \le e^{-3\log n}$. ■

By the above claim, we have that for any $\mathcal{A}$ and $\mathcal{B}$ such that $|\mathcal{A}| \le |\mathcal{B}| \le \frac{n}{e}$, either of the following holds:

$$\text{(i)}\ \mathrm{den}_{\mathbf{A}}(\mathcal{A},\mathcal{B}) \le c_7\nu|\mathcal{A}||\mathcal{B}| \text{ or;}$$

$$\text{(ii)}\ k_{|\mathcal{A}|,|\mathcal{B}|}\log k_{|\mathcal{A}|,|\mathcal{B}|} = \frac{14d^2}{\nu a}\log\frac{n}{b}.$$

For (ii), one can derive: $\mathrm{den}_{\mathbf{A}}(\mathcal{A},\mathcal{B}) \le k_{|\mathcal{A}|,|\mathcal{B}|}\cdot\nu|\mathcal{A}||\mathcal{B}| = (\frac{14d^2}{\nu|\mathcal{A}|\log k_{|\mathcal{A}|,|\mathcal{B}|}}\log\frac{n}{|\mathcal{B}|})\cdot\nu|\mathcal{A}||\mathcal{B}| \le (\frac{14d^2}{\nu|\mathcal{A}|\log\frac{\mathrm{den}_{\mathbf{A}}(\mathcal{A},\mathcal{B})}{\nu|\mathcal{A}||\mathcal{B}|}}\log\frac{n}{|\mathcal{B}|})\cdot\nu|\mathcal{A}||\mathcal{B}|$, and hence $\mathrm{den}_{\mathbf{A}}(\mathcal{A},\mathcal{B})\log\frac{\mathrm{den}_{\mathbf{A}}(\mathcal{A},\mathcal{B})}{\nu|\mathcal{A}||\mathcal{B}|} \le 14d^2|\mathcal{B}|\log\frac{n}{|\mathcal{B}|}$.

Combining the above two cases 1) and 2), the proof is completed by taking $c_{10} = \max\{c_9 e,\ c_7\}$ and $c_{11} = 14d^2$.

F. Proof of Lemma 9

One can easily show that the LHS is equal to $\sum_{e\in\mathcal{E}:W_e\neq \mathsf{x}} [\mathbb{I}\{f_e(\mathbf{X}\oplus\mathbf{e}_1)\neq W_e\} - \mathbb{I}\{f_e(\mathbf{X})\neq W_e\}]$. Since the summand is nonzero only if $f_e(\mathbf{X}\oplus\mathbf{e}_1)\neq f_e(\mathbf{X})$, we count the number of such edges.

First, observe that if $1\notin e$, $f_e(\mathbf{X}\oplus\mathbf{e}_1) = f_e(\mathbf{X})$. Further, if two (or more) nodes other than node 1 are of different affiliations, then $f_e(\mathbf{X}\oplus\mathbf{e}_1) = f_e(\mathbf{X}) = 0$. Thus, e must include 1 and all the other nodes in e must be of the same affiliation: If all the nodes of e other than node 1 are affiliated with community 0, $f_e(\mathbf{X}\oplus\mathbf{e}_1) = 1$ and $f_e(\mathbf{X}) = 0$; and if all the nodes of e other than node 1 are affiliated with community 1, $f_e(\mathbf{X}\oplus\mathbf{e}_1) = 0$ and $f_e(\mathbf{X}) = 1$.

Define the set of edges corresponding to the former case as $\mathcal{E}_1$, and that corresponding to the latter case as $\mathcal{E}_2$, i.e., $\mathcal{E}_1 := \{e\in\mathcal{E} : 1\in e \text{ and } (e\setminus\{1\})\subset\{\eta n+1, \eta n+2, \ldots, \eta n + n/2\}\}$ and $\mathcal{E}_2 := \{e\in\mathcal{E} : 1\in e \text{ and } (e\setminus\{1\})\subset\{2,3,\ldots, \eta n, \eta n+n/2+1, \eta n+n/2+2,\ldots,n\}\}$. Consider all homogeneous edges in $\mathcal{E}_1$. The total contribution of the terms associated with these edges to the sum is $\sum_{e\in\mathcal{E}_1 :W_e\neq\mathsf{x}, e:\text{homogeneous}} [\mathbb{I}\{1\neq W_e\} - \mathbb{I}\{0\neq W_e\}]$. Each term is -1 if observation is not corrupted, and $+1$ if observation is corrupted. Thus, the total contribution is $\sum_{i=1}^{|\{e\in\mathcal{E}_1 \,:\, e \text{ is homogeneous}\}|} P_i(2\Theta_i - 1) = \sum_{i=1}^{\binom{n/2-\eta n}{d-1}} P_i(2\Theta_i-1)$, where $P_i \overset{\text{i.i.d.}}{\sim} \mathsf{Bern}(\alpha_n)$ and $\Theta_i \overset{\text{i.i.d.}}{\sim} \mathsf{Bern}(\theta)$. By rewriting other contributions in a similar way, we complete the proof.

G. Proof of Lemma 10

Let $Z := \log\left(\frac{1-\theta}{\theta}\right)\sum_{i=1}^{K} P_i(2\Theta_i - 1) + \ell$ and $\mathbb{M}(\lambda) := \mathbb{E}[e^{\lambda\log\left(\frac{1-\theta}{\theta}\right)P_1(2\Theta_1-1)}]$. Via simple calculation, we have

$$\begin{aligned}\Pr(Z>0) &= \Pr\left(e^{\frac{1}{2}Z}>1\right) \leq e^{\frac{1}{2}\ell}\left\{\mathbb{M}(1/2)\right\}^K \\ &= e^{\frac{1}{2}\ell + K\left\{-\alpha_n(\sqrt{1-\theta}-\sqrt{\theta})^2 + O(\alpha_n^2)\right\}}.\end{aligned}$$

REFERENCES

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

[2] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[3] F. McSherry, “Spectral partitioning of random graphs,” in *Proc. 42nd IEEE Symp. Found. Comput. Sci.*, 2001, pp. 529–537.

[4] K. Rohe, S. Chatterjee, and B. Yu, “Spectral clustering and the high-dimensional stochastic blockmodel,” *Ann. Statist.*, vol. 39, pp. 1878–1915, 2011.

[5] J. Lei and A. Rinaldo, “Consistency of spectral clustering in stochastic block models,” *Ann. Statist.*, vol. 43, no. 1, pp. 215–237, 2015.

[6] G. Ghoshal, V. Zlatić, G. Caldarelli, and M. Newman, “Random hypergraphs and their applications,” *Phys. Rev. E*, vol. 79, no. 6, 2009, Art. no. 066118.

[7] T. Michoel and B. Nachtergaele, “Alignment and integration of complex networks by hypergraph-based spectral clustering,” *Phys. Rev. E*, vol. 86, no. 5, 2012, Art. no. 056111.

[8] D. Ghoshdastidar and A. Dukkipati, “Uniform hypergraph partitioning: Provable tensor methods and sampling techniques,” *J. Mach. Learn. Res.*, vol. 18, no. 50, pp. 1–41, 2017.

[9] D. Ghoshdastidar and A. Dukkipati, “Consistency of spectral partitioning of uniform hypergraphs under planted partition model,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 397–405.

[10] 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.

[11] D. Ghoshdastidar and A. Dukkipati, “Consistency of spectral hypergraph partitioning under planted partition model,” *Ann. Statist.*, vol. 45, no. 1, pp. 289–315, 2017.

[12] L. Florescu and W. Perkins, “Spectral thresholds in the bipartite stochastic block model,” in *Proc. 29th Annu. Conf. Learn. Theory*, 2016, pp. 943–959.

[13] E. Abbe, “Community detection and stochastic block models: Recent developments,” *J. Mach. Learn. Res.*, arXiv: 1703.10146, 2017.

[14] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, “Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications,” *Phys. Rev. E*, vol. 84, 2011, Art. no. 066106.

[15] E. Mossel, J. Neeman, and A. Sly, “Reconstruction and estimation in the planted partition model,” *Probability Theory Related Fields*, vol. 162, no. 3/4, pp. 431–461, 2015.

[16] L. Massoulié, “Community detection thresholds and the weak Ramanujan property,” in *Proc. 46th Annu. ACM Symp. Theory Comput.*, 2014, pp. 694–703.

[17] C. Bordenave, M. Lelarge, and L. Massoulie, “Non-backtracking spectrum of random graphs: Community detection and non-regular Ramanujan graphs,” in *Proc. 56th Annu. Symp. Found. Computer Sci.*, 2015, pp. 1347–1357.

[18] Y. Chen and J. Xu, “Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices,” *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 882–938, 2016.

[19] J. Banks, C. Moore, J. Neeman, and P. Netrapalli, “Information-theoretic thresholds for community detection in sparse networks,” in *Proc. Conf. Learn. Theory*, 2016, pp. 383–416.

[20] A. Montanari, “Finding one community in a sparse graph,” *J. Statist. Phys.*, vol. 161, no. 2, pp. 273–299, 2015.

[21] J. Neeman and P. Netrapalli, “Non-reconstructability in the stochastic block model,” arXiv:1404.6304, 2014.

[22] E. Abbe and C. Sandon, “Proof of the achievability conjectures in the general stochastic block model,” *Commun. Pure Appl. Math.*, 2017, doi: 10.1002/cpa.21719.

[23] A. A. Amini and E. Levina, “On semidefinite relaxations for the block model,” *Ann. Statist.* vol. 46, pp. 149–179, 2014.

[24] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou, “Achieving optimal misclassification proportion in stochastic block model,” *J. Mach. Learn. Res.*, vol. 60, pp. 1–45, 2017.

[25] 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.

[26] S.-Y. Yun and A. Proutiere, “Community Detection via Random and Adaptive Sampling,” *COLT*, pp. 138–175, 2014.

[27] 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.*, 2015, pp. 670–688.

[28] 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.

[29] 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.

[30] A. Y. Zhang *et al.*, “Minimax rates of community detection in stochastic block models,” *Ann. Statist.*, vol. 44, pp. 2252–2280, 2016.

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

[32] K. Ahn, K. Lee, and C. Suh, “Community recovery in hypergraphs,” in *Proc. 54th Annu. Allerton Conf. Commun., Control, Comput.*, 2016, pp. 657–663.

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

[34] V. M. Govindu, “A tensor decomposition for geometric grouping and segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 1150–1157.

[35] S. Agarwal, K. Branson, and S. Belongie, “Higher order learning with graphs,” in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 17–24.

[36] G. Chen and G. Lerman, “Spectral curvature clustering (SCC),” *Int. J. Comput. Vis.*, vol. 81, no. 3, pp. 317–330, 2009.

[37] F. L. Hitchcock, “The expression of a tensor or a polyadic as a sum of products,” *Studies Appl. Math.*, vol. 6, pp. 164–189, 1927.

[38] B. Barak and A. Moitra, “Noisy tensor completion via the sum-of-squares hierarchy,” in *Proc. Conf. Learn. Theory*, 2016, pp. 417–445.

[39] C. Kim, A. S. Bandeira, and M. X. Goemans, “Community detection in hypergraphs, spiked tensor models, and sum-of-squares,” in *Proc. 12th Int. Conf. Sampling Theory Appl.*, 2017, pp. 124–128.

[40] V. Jog and P.-L. Loh, “Information-theoretic bounds for exact recovery in weighted stochastic block models using the renyi divergence,” in *Proc. 53rd Annu. Allerton Conf. Commun., Control, Comput.*, 2015, pp. 1308–1315.

[41] M. Xu, V. Jog, and P.-L. Loh, “Optimal rates for community estimation in the weighted stochastic block model,” arXiv:1706.01175, 2017.

[42] P. W. Holland, K. B. Laskey, and S. Leinhardt, “Stochastic blockmodels: First steps,” *Soc. Netw.*, vol. 5, no. 2, pp. 109–137, 1983.

[43] E. Mossel, J. Neeman, and A. Sly, “Consistency thresholds for binary symmetric block models,” arXiv:1407.1591, 2014.

[44] U. Feige and E. Ofek, “Spectral techniques applied to sparse random graphs,” *Random Struct. Algorithms*, vol. 27, pp. 251–275, 2005.

[45] A. Coja-Oghlan, “Graph partitioning via adaptive spectral techniques,” *Combinatorics, Probability Comput.*, vol. 19, pp. 227–284, 2010.

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

[47] O. Guédon and R. Vershynin, “Community detection in sparse networks via grothendiecks inequality,” *Probability Theory Related Fields*, vol. 165, no. 3/4, pp. 1025–1049, 2016.

[48] J. Matoušek, “On approximate geometric k-clustering,” *Discrete Comput. Geometry*, vol. 24, no. 1, pp. 61–84, 2000.

[49] 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. 28th Conf. Learn. Theory*, 2015, pp. 391–423.

[50] C. Boutsidis, P. Kambadur, and A. Gittens, “Spectral clustering via the power method-provably,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 40–48.

[51] R. H. Keshavan, A. Montanari, and S. Oh, “Matrix completion from a few entries,” *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.

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

[53] P. Netrapalli, P. Jain, and S. Sanghavi, “Phase retrieval using alternating minimization,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2796–2804.

[54] E. J. Candes, X. Li, and M. Soltanolkotabi, “Phase retrieval via wirtinger flow: Theory and algorithms,” *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.

[55] X. Yi, D. Park, Y. Chen, and C. Caramanis, “Fast algorithms for robust PCA via gradient descent,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 4152–4160.

[56] Y. Chen, G. Kamath, C. Suh, and D. Tse, “Community recovery in graphs with locality,” in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 689–698.

[57] S. Balakrishnan, M. J. Wainwright, and B. Yu, “Statistical guarantees for the EM algorithm: From population to sample-based analysis,” *Ann. Statist.*, vol. 45, pp. 77–120, 2017.

[58] Y. Chen and C. Suh, “Spectral MLE: Top-k rank aggregation from pairwise comparisons,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 371–380.

[59] F. Krzakala *et al.*, “Spectral redemption in clustering sparse networks,” *Proc. Nat. Acad. Sci. United States Amer.*, vol. 110, no. 52, pp. 20 935–20 940, 2013.

[60] I. Chien *et al.*, “On the minimax misclassification ratio of hypergraph community detection,” 2018, AISTATS, PMLR 84:871–879.

[61] A. Coja-Oghlan, C. Moore, and V. Sanwalani, “Counting connected graphs and hypergraphs via the probabilistic method,” *Random Struct. Algorithms*, vol. 31, no. 3, pp. 288–329, 2007.

[62] J. Nesetril, O. Serra, J. A. Telle, O. Cooley, M. Kang, and C. Koch, “Evolution of high-order connected components in random hypergraphs,” *Electron. Notes Discrete Math.*, vol. 49, pp. 569–575, 2015.

[63] J. Friedman, J. Kahn, and E. Szemeredi, “On the second eigenvalue of random regular graphs,” in *Proc. 21st Annu. ACM Symp. Theory Comput.*, 1989, pp. 587–598.

[64] T. Tao, *Topics in Random Matrix Theory*. Providence, RI, USA: Amer. Math. Soc.2012, vol. 132.

[65] J. A. Tropp, “User-friendly tail bounds for sums of random matrices,” *Found. Comput. Math.*, vol. 12, pp. 389–434, 2012.

[66] P. Jain and S. Oh, “Provable tensor factorization with missing data,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 1431–1439.

[67] N. Alon and J. H. Spencer, *The Probabilistic Method*. Hoboken, NJ, USA: Wiley, 2004.

[68] N. Alon, M. Krivelevich, and B. Sudakov, “Finding a large hidden clique in a random graph,” *Random Struct. Algorithms*, vol. 13, pp. 457–468, 1998.

[69] R. Tron and R. Vidal, “A benchmark for the comparison of 3-D motion segmentation algorithms,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[70] R. Heckel, M. Tschannen, and H. Bölcskei, “Dimensionality-reduced subspace clustering,” *Inf. Inference, A J. IMA*, vol. 6, no. 3, pp. 246–283, 2017.

[71] E. Elhamifar and R. Vidal, “Sparse subspace clustering: Algorithm, theory, and applications,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[72] E. L. Dyer, A. C. Sankaranarayanan, and R. G. Baraniuk, “Greedy feature selection for subspace clustering,” *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 2487–2517, 2013.

[73] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, “Robust recovery of subspace structures by low-rank representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[74] R. Heckel and H. Bölcskei, “Robust subspace clustering via thresholding,” *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 6320–6342, Nov. 2015.

[75] D. Park, C. Caramanis, and S. Sanghavi, “Greedy subspace clustering,” in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2753–2761.

[76] S. Jain and V. Madhav Govindu, “Efficient higher-order clustering on the Grassmann manifold,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3511–3518.

[77] O. Duchenne, F. Bach, I.-S. Kweon, and J. Ponce, “A tensor-based algorithm for high-order graph matching,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2383–2395, Dec. 2011.

[78] J. Yang and J. Leskovec, “Defining and evaluating network communities based on ground-truth,” *Knowl. Inf. Syst.*, vol. 42, no. 1, pp. 181–213, 2015.

Kwangjun Ahn received the B.S. degree from the Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2017.

He is currently a military police desk clerk in the US Army as a part of Korean Augmentation to the US Army. His research interests include applied mathematics.

Kangwook Lee received the Ph.D. degree in EECS from UC Berkeley, Berkeley, CA, USA, in 2016. He is a recipient of the KFAS Fellowship from 2010 to 2015. He is a Postdoctoral Researcher in the School of Electrical Engineering, Korea Advanced Institute of Science and Technology. His research interests include information theory and machine learning.

Changho Suh (S’10–M’12) received the B.S. and M.S. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST, Daejeon, South Korea, in 2000 and 2002, respectively, and the Ph.D. degree in electrical engineering and computer sciences from UC-Berkeley, Berkeley, CA, USA, in 2011. He is an Ewon Associate Professor in the School of Electrical Engineering, KAIST, since 2012. From 2011 to 2012, he was a Postdoctoral Associate with the Research Laboratory of Electronics, MIT. From 2002 to 2006, he had been with the Telecommunication R&D Center, Samsung Electronics. He received the 2015 Haedong Young Engineer Award from the Institute of Electronics and Information Engineers, the 2013 Stephen O. Rice Prize from the IEEE Communications Society, the David J. Sakrison Memorial Prize from the UC-Berkeley EECS Department in 2011, and the Best Student Paper Award of the IEEE International Symposium on Information Theory in 2009.