Fifty-fourth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 27 - 30, 2016

Community Recovery in Hypergraphs

Kwangjun Ahn
Dept. of Mathematical Sciences
KAIST
kjahnkorea@kaist.ac.kr

Kangwook Lee
School of EE
KAIST
kw1jjang@kaist.ac.kr

Changho Suh
School of EE
KAIST
chsuh@kaist.ac.kr

***Abstract*—Data clustering is a core problem in many fields of science and engineering. *Community recovery in graphs* is one popular approach to data clustering, and it has received significant attention due to its wide applicability to social network applications, protein complex detection, shape matching, image segmentation, etc. While the community recovery in graphs has been extensively studied in the literature, the problem of community recovery in *hypergraphs* has not been studied much. In this paper, we study the generalized *Censored Block Model (CBM)*, where observations consist of randomly chosen hyperedges of size d, each of which is associated with the modulo-2 sum of the values of the nodes in the hyperedge, corrupted by Bernoulli noise. We characterize the information-theoretic limit of the community recovery in hypergraphs. Our results are for the general cases of arbitrarily scaling d.**

I. Introduction

Clustering of data is one of the central problems that arises in many fields of science and engineering. Among many related setups, *community recovery* [1], [2] has received significant attention due to its wide applicability to social network applications, protein complex detection [3], shape matching [4], image segmentation [5], etc. The goal of the community recovery problem is to cluster data points into different communities based on whether two data points belong to the same community or not.

There have been a proliferation of works on various models of the community recovery problem, and the *Censored Block Model* (CBM) is one of the most popular models in the literature [6]–[8]. We illustrate the generalized CBM for the problem of *community recovery in hypergraphs*. In this model, the n individuals, each of which belongs to either group 0 or group 1, are modeled as nodes of a hypergraph. The goal is to cluster the n nodes (or find the hidden communities) using noisy parity measurements obtained from a random d-uniform hypergraph [9]. More specifically, a random hypergraph (of hyperedge size d) with the n nodes is given as observation, in which each hyperedge exists with probability p. Further, each observed hyperedge is associated with the modulo-2 sum of the nodes of the hyperedge, i.e., the parity of the hyperedge. Further, such *parity* measurements are noisy in a way that each edge value is flipped with probability θ. For instance, when $d = 3$, if each of the three nodes of a hyperedge belongs to group-1,0,0, respectively, the value of the hyperedge is 1 with probability $1 - \theta$, and 0 with probability θ. See Fig. 1 for illustrations.

For the special case of $d = 2$, the information-theoretic limits as well as matching computation limits are characterized in [7], [8]. The prior works reveal that the minimum number of samples required for reliable community recovery is about $p\binom{n}{2} \simeq \frac{n \log n}{2\left(\sqrt{1-\theta}-\sqrt{\theta}\right)^2}$. However, for the case of general d, such characterization has been unknown in the literature, and this precisely sets the goal of our paper. More precisely, we seek to characterize the information-theoretic limit of the minimum number of samples (hyperedges) required for reliable community recovery.

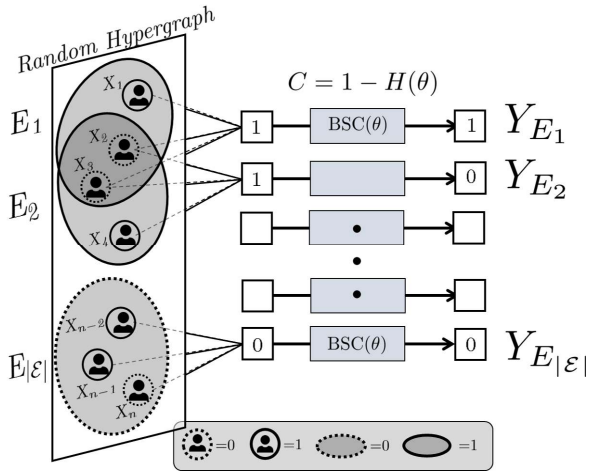

Fig. 1: **Community recovery in hypergraphs.** We illustrate the community recovery problem in hypergraphs under the generalized Censored Block Model (CBM). Shown on the left hand side is the group of n nodes belonging to group-0 (dotted circles) or group-1 (solid circles). The goal of the problem is to cluster nodes into two groups from an observed hypergraph. Under the generalized CBM, an observed hypergraph consists of randomly chosen hyperedges of size d. Each observed hyperedge is associated with the modulo-2 sum of the values of the nodes in the hyperedge, corrupted by Bernoulli noise. For instance, the second hyperedge E_2 connects two group-0 nodes (X_2, X_3) and one group-1 node (X_4), and hence the corresponding parity is 1, but it is corrupted by noise, making the actual observed value 0. Our goal is to characterize when the exact community recovery in hypergraphs is feasible in terms of p, θ, n, and d. Further, as illustrated in the middle of the figure, we note that the community recovery problem can be seen as a certain channel coding problem with a fixed encoding scheme.

A summary of important findings in this paper is as follows.

- if d grows asymptotically slower than $\log n$, the minimum number of samples required for reliable community recovery is $p\binom{n}{d} \simeq \beta \frac{n \log n}{d}$; and
- if d grows asymptotically faster than or equal to $\log n$, the minimum number of samples required for reliable community recovery is $p\binom{n}{d} \simeq \gamma n$,

where β and γ are some unknown constants, which can be bounded below and above. Thus, to make reliable community recovery with a linear number of samples possible, the size of hyperedges needs to scale at least as fast as $\log n$.

978-1-5090-4550-1/16/$31.00 ©2016 IEEE

The rest of the paper is organized as follows. In the rest of this section, we relate our problem to a *d-right-degree linear code*, and then discuss related works; Sec. II introduces the problem formulation; in Sec. III, our main theorem is presented, along with some implications and remarks; in Sec. IV, we prove the achievability statement as well as the converse statement; Sec. V presents numerical simulation results that corroborate our theoretical findings; and in Sec. VI, we conclude the paper. The proofs of technical lemmas are presented in the full version of this paper [10]

A. Connection with channel coding

The community recovery problem has an inherent connection with channel coding problems [7]. In order to see such connection, consider the following point-to-point communication setup. The encoder employs the following random linear code. It first draws a random d-uniform hypergraph with n nodes, where each hyperedge exists with probability p. Given the input sequence of n bits, the parity bits corresponding to all the existing hyperedges are concatenated, forming a codeword. Note that the expected rate of this random code is $\frac{n}{p\binom{n}{d}}$. A codeword chosen from this code is transmitted through a Binary Symmetric Channel (BSC) with error probability θ, whose capacity is $1-H(\theta)$ [11]. Given the received symbols, the decoder wishes to infer the n input bits. By associating $0,1$ symbols with labels of the observed hyperedges, one can see that recovering communities in hypergraphs is equivalent to the above channel coding problem. See Fig. 1.

A natural channel coding question arises: "How far is the rate of the random code from the capacity of the BSC channel". Due to the equivalence, the information-theoretic limits of community recovery in hypergraphs can help immediately answer the above question. For instance, when $d = 2$, as mentioned earlier in this section, it is shown in [7], [8] that the exact community recovery is possible if $p\binom{n}{2} \gtrsim \frac{n\log n}{2\left(\sqrt{1-\theta}-\sqrt{\theta}\right)^2}$, implying that the following expected code rate can be achieved:

$$\frac{n}{p\binom{n}{2}} \lesssim \frac{2(\sqrt{1-\theta}-\sqrt{\theta})^2}{\log n}.$$

That is, while the capacity of a BSC channel is a fixed constant $1-H(\theta)$, the rate of the hypergraph-based random code vanishes as $n\to\infty$.

One natural question is whether one can achieve a non-vanishing code rate by increasing d. Our results on the information-theoretic limits of community recovery in hypergraphs answer this coding-theoretic question: when d scales as fast as $\log n$, the random-hypergraph-based linear code can achieve a constant rate.

B. Related Work

Community recovery in standard graphs has been extensively studied in the literature. Under the Stochastic Block Model (SBM) [12], [13], the probability of an edge appearing in the observed graph is assumed to depend on whether the edge is connecting the nodes in the same group or not. For instance, the SBM can capture the case where random graphs have statistically more edges between nodes within the same community than between nodes across two different communities. The information-theoretic limits as well as matching computation limits are characterized for the SBM [14]–[17]. Under the Censored Block Model (CBM) [6], [7], each edge is associated with a random label, whose distribution depends on whether the edge is connecting the nodes in the same group or not. The information-theoretic limits as well as matching computation limits are characterized in [7], [8]. Abbe et al. [7] show that the exact community recovery is impossible if $p\binom{n}{2} < \frac{(1-\epsilon)n\log n}{2\left(\sqrt{1-\theta}-\sqrt{\theta}\right)^2}$ for $\epsilon>0$. Hajek et al. show that the exact community recovery is possible via an efficient algorithm is possible if $p\binom{n}{2} > \frac{(1+\epsilon)n\log n}{2\left(\sqrt{1-\theta}-\sqrt{\theta}\right)^2}$ in [8]. In [18], the labeled stochastic block model is proposed as a general observation model that includes both SBM and CBM as special cases. We focus on the generalized CBM for community recovery in hypergraphs, and to the best of our knowledge, we are the first to characterize the information-theoretic limits for one of such generalized models.

The community recovery problem in graphs or hypergraphs is closely related to *MLS-dLIN* problems [19], [20], of which the goal is to find a binary vector x that is maximally consistent with a given set of parities of d variables. Under this context, the case of $d=3$ has been well-studied. For $d=3$, it is shown that the maximum likelihood decoder can succeed if $p \geq \frac{12\log n}{(0.5-\theta)^2 n^2}$ [20]. Unlike the prior result, our upper bounds on p, to be formally stated in Corollary 1, are for arbitrary constant d. Further, we provide the matching lower bounds as well. Among a few efficient algorithms for the MLS-3LIN problem, one proposed in [21], based on an efficient low-rank tensor factorization algorithm, is shown to find the optimal solution if $p=\Omega(\frac{\log^4 n}{n^{1.5}})$.

C. Notations

For any two sequences $f(n)$ and $g(n)$: $f(n)=\Omega(g(n))$ if there exists a positive constant c such that $f(n)\geq cg(n)$; $f(n)=O(g(n))$ if there exists a positive constant c such that $f(n)\leq cg(n)$; $f(n)=\omega(g(n))$ if $\lim_{n\to\infty}\frac{f(n)}{g(n)}=\infty$; $f(n)=o(g(n))$ if $\lim_{n\to\infty}\frac{f(n)}{g(n)}=0$; and $f(n)\asymp g(n)$ or $f(n)=\Theta(g(n))$ if there exist positive constants c_1 and c_2 such that $c_1 g(n)\leq f(n)\leq c_2 g(n)$. For a set A and an integer $m\leq |A|$, we denote $\binom{A}{m} \stackrel{\text{def}}{=} \{B\subset A\,|\,|B|=m\}$. Let $[n]$ denotes $\{1,\cdots,n\}$. Let $\mathbf{e}_i$ be the i^{th} standard unit vector. Let $\mathbf{0}$ be the all-zero-vector and $\mathbf{1}$ be the all-one-vector. Let $D(0.5\|\theta)$ be the Kullback-Liebler (KL) divergence between Bernoulli(0.5) and Bernoulli(θ), i.e., $D(0.5\|\theta)\stackrel{\text{def}}{=} 0.5\log(\frac{0.5}{\theta})+0.5\log(\frac{0.5}{1-\theta})$. We shall use $\log(\cdot)$ to indicate the natrual logarithm. We use $H(\cdot)$ to denote the binary entropy function.

II. Problem Formulation

A. Sampling Model: the Generalized CBM

Consider a collection of n vertices $\mathcal{V}=[n]$, each represented by a binary variable $X_i\in\{0,1\}$, $1\leq i\leq n$. Let $\mathbf{X}\stackrel{\text{def}}{=}[X_i]_{1\leq i\leq n}$ be *the ground truth vector*. Suppose that d is given as a monotone function of n satisfying $2\leq d\leq n/2$; for instance, we may choose d to be constants or be functions

that scale with n such as $d = \lceil \sqrt{n+1} \rceil$. Samples are obtained according to a *measurement hypergraph* $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{E} \subset \binom{[n]}{d}$. We assume $\mathcal{H} \sim \mathcal{H}_{n,p}^d$, i.e., each element in $\binom{[n]}{d}$ belongs to $\mathcal{E}$ with probability $p \in [0,1]$.

Fix a hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and a ground truth vector $\mathbf{X}$. For each edge $E \in \mathcal{E}$, a noisy binary observation is given by

$$Y_E = \left[\bigoplus_{i \in E} X_i \right] \oplus Z_E,$$

where $\oplus$ denotes modulo-2 addition, and $Z_E \sim \text{Bernoulli}(\theta)$ for $0 \le \theta < \frac{1}{2}$. We assume that $\{Z_E\}_{E \in \mathcal{E}}$ is a collection of mutually independent random variables. We define *the observation vector* $\mathbf{Y}$ as follows:

$$\mathbf{Y} \stackrel{\text{def}}{=} [Y_E]_{E \in \mathcal{E}}.$$

Note that when $d = 2$, this setup is reduced to the well-known community detection problem.

Note that when d is even and the measurement hypergraph is fixed, the conditional distribution of $\mathbf{Y}|\mathbf{X}$ is equal to that of $\mathbf{Y}|\mathbf{X} \oplus \mathbf{1}$, implying that decoding $\mathbf{X}$ is possible only up to a global shift. In the rest of the paper, we assume that d is odd for ease of presentation. The even case can be readily dealt with by allowing for the global shift error.

B. Our Goals

This paper concerns the exact recovery, that is, to reconstruct the ground truth $\mathbf{X}$ given observation vector $\mathbf{Y}$. More precisely, for any recovery procedure, or decoder, ψ the error probability is defined as follows.

Definition 1 (Error probability and admissibility)**.**

$$P_e(\psi) = \min_{\psi} \max_{\mathbf{X} \in \{0,1\}^n} \Pr[\psi(\mathbf{Y}) \neq \mathbf{X}].$$

Moreover, we say that p is *admissible* if $\lim_{n \to \infty} P_e = 0$.

The goal is to characterize necessary or sufficient conditions on (n, p, d, θ) for reliable recovery. In particular, we will often rewrite the conditions using *sample complexity* $p\binom{n}{d}$, which represents the expected number of hyperedges in the measurement random hypergraph.

III. Main Results

We begin with the main theorem of the paper, which characterizes sufficient and necessary conditions for reliable recovery.

Theorem 1. *Suppose that $\mathcal{H} \sim \mathcal{H}_{n,p}^d$ for $2 \le d \le \frac{n}{2}$. For a fixed $\epsilon > 0$,*

$$\begin{cases} P_e = O(n^{-\epsilon}) \text{ if } p \ge \frac{5/2(1+\epsilon)}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \frac{n \log n}{d\binom{n}{d}} \cdot \max\left\{1, \frac{2d \log 2}{\log n}\right\}, \\ P_e \not\to 0 \qquad \text{if } p \le (1-\epsilon) \max\left\{ \frac{1}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \frac{n \log n}{d\binom{n}{d}}, \right. \\ \qquad\qquad\qquad\qquad \left. \frac{1}{(1-H(\theta))} \frac{n}{\binom{n}{d}} \right\}. \end{cases}$$

We refer the readers to Sec. IV for the proof.

Let us interpret our main theorem for the following two cases: $d = \Omega(\log n)$ and $d = o(\log n)$. Note that when $d = \Omega(\log n)$, $P_e \to 0$ if $p > \beta_1 \frac{n}{\binom{n}{d}}$ and $P_e \not\to 0$ if $p < \beta_2 \frac{n}{\binom{n}{d}}$, where $\beta_1 = \max\left\{ \frac{5/2 \log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2 d}, \frac{5 \log 2}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \right\} \asymp 1$, $\beta_2 = \max\left\{ \frac{\log n}{d(\sqrt{1-\theta}-\sqrt{\theta})^2}, \frac{1}{(1-H(\theta))} \right\} \asymp 1$, for a fixed θ. Hence, Theorem 1 provides an order-wise tight characterization of the admissible region. Next, when $d = o(\log n)$, $P_e \to 0$ if $p > \frac{5/2}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \frac{n \log n}{d\binom{n}{d}}$, and $P_e \not\to 0$ if $p < \frac{1}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \frac{n \log n}{d\binom{n}{d}}$. Thus, the theorem offers a tighter characterization relative to the case $d = \Omega(\log n)$. Notice that the multiplicative gap between the lower and upper bounds is a small constant $5/2$ regardless of θ. Especially, when is d is a constant order, the result is even more enhanced, formally stated in the following corollary.

Corollary 1. *Suppose that $d \asymp 1$ and $\mathcal{H} \sim \mathcal{H}_{n,p}^d$. For a fixed $\epsilon > 0$,*

$$\begin{cases} P_e \to 0 & \text{if } p \ge \frac{1+\epsilon}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \frac{n \log n}{d\binom{n}{d}}, \\ P_e \not\to 0 & \text{if } p \le \frac{1-\epsilon}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \frac{n \log n}{d\binom{n}{d}}. \end{cases}$$

The proof can be readily deduced from the proof of Theorem 1, hence we omit the proof. Corollary 1 characterizes the sharp threshold on the admissible region for constant d. Note that this result recovers that in [8] as a special case.

Finally, we reinterpret our main theorem using sample complexity. When $d = o(\log n)$, reliable recovery is possible only if sample complexity is superlinear. However, when $d = \Omega(\log n)$, we see that reliable recovery can be done with linear sample complexity $\Theta(n)$. Hence, we can answer the question on how large d is needed to be to make reliable recovery with linear sample complexity possible.

Corollary 2. *For $d = o(\log n)$, reliable recovery is impossible with linear sample complexity, while it is possible for $d = \Omega(\log n)$: if $d = k \log n$ for some constant $k > 0$, there exists a constant $c_k > 0$ such that reliable recovery is feasible whenever sample complexity is larger than $c_k n$; If $d = \omega(\log n)$, there exists an absolute constant $c > 0$ such that reliable recovery is feasible whenever sample complexity is larger than cn.*

IV. Proof of the main theorem

We consider the noisy case ($\theta \neq 0$) only for conciseness. We remark that the proof of the noiseless case can be done analogously, and hence we omit the proof.

For both achievability and converse proofs, we use the optimal maximum likelihood (ML) decoder, where ties are randomly broken. One can easily verify that the ML decoder reduces to:

$$\hat{\mathbf{X}}_{\text{ML}} = \arg \min_{\mathbf{A} = [A_i]_{i \in [n]}} d(\mathbf{A}),$$

where $d(\mathbf{A}) \stackrel{\text{def}}{=} d_H\left(\mathbf{Y}, \left[\bigoplus_{i \in E} A_i\right]_{E \in \mathcal{E}}\right)$ and $d_H(\cdot, \cdot)$ is the Hamming distance between two binary vectors.

A. Achievability

We first give an upper bound on the error probability:

$$\begin{aligned}
P_e &= \max_{\mathbf{X}\in\{0,1\}^n} \Pr[\hat{\mathbf{X}}_{\mathrm{ML}} \neq \mathbf{X}] \\
&= \Pr(\hat{\mathbf{X}}_{\mathrm{ML}} \neq \mathbf{0} | \mathbf{X} = \mathbf{0}) \qquad (1)\\
&\leq \Pr\left(\bigcup_{\mathbf{A}\neq\mathbf{0}} [d(\mathbf{A}) \leq d(\mathbf{0})] \,\middle|\, \mathbf{X} = \mathbf{0} \right) \qquad (2)\\
&= \Pr\left(\bigcup_{k=1}^{n} \bigcup_{\|\mathbf{A}\|_1 = k} [d(\mathbf{A}) \leq d(\mathbf{0})] \,\middle|\, \mathbf{X} = \mathbf{0} \right) \\
&\leq \sum_{k=1}^{n} \sum_{\|\mathbf{A}\|_1 = k} \Pr\left(d(\mathbf{A}) \leq d(\mathbf{0}) \,\middle|\, \mathbf{X} = \mathbf{0} \right) \\
&\leq \sum_{k=1}^{n} \binom{n}{k} \Pr\left(d\left(\sum_{i=1}^{k} \mathbf{e}_i\right) \leq d(\mathbf{0}) \,\middle|\, \mathbf{X} = \mathbf{0} \right), \qquad (3)
\end{aligned}$$

where (1) follows by symmetry; (2) follows by the fact that the ML decoder fails only if there exists one or more than one non-zero vectors whose likelihood is greater than or equal to that of the zero vector; and (3) follows by symmetry.

We now find an upper bound to

$$\Pr\left(d\left(\sum_{i=1}^{k} \mathbf{e}_i\right) \leq d(\mathbf{0}) \,\middle|\, \mathbf{X} = \mathbf{0} \right),$$

for $1 \leq k \leq n$.

Let $\mathcal{O}_k \subset \mathcal{E}$ be the collection of hyperedges that contain an odd number of nodes among $[k]$. Observe that $E \notin \mathcal{O}_k$ contributes eqaully to $d\left(\sum_{i=1}^{k} \mathbf{e}_i\right)$ and $d(\mathbf{0})$. Hence, in order for $\sum_{i=1}^{k} \mathbf{e}_i$ to have higher likelihood than that of the zero vector, the number of bit flips among $\mathcal{O}_k$ has to be $\geq \frac{|\mathcal{O}_k|}{2}$. We define by N_k the size of the subset of $\binom{[n]}{d}$ such that each element contains an odd number of elements in $[k]$: $N_k = \sum_{\substack{i \leq d \\ i \text{ is odd}}} \binom{k}{i} \cdot \binom{n-k}{d-i}$. Using $\mathcal{O}_k$ and N_k notations, we can then get:

$$\begin{aligned}
&\Pr\left(d\left(\sum_{i=1}^{k} \mathbf{e}_i\right) \leq d(\mathbf{0}) \,\middle|\, \mathbf{X} = \mathbf{0} \right) \\
&\leq \sum_{m=1}^{N_k} \Pr(|\mathcal{O}_k| = m) \\
&\qquad \cdot \Pr\left(d\left(\sum_{i=1}^{k} \mathbf{e}_i\right) \leq d(\mathbf{0}) \;\middle|\; \mathbf{X} = \mathbf{0}, |\mathcal{O}_k| = m \right) \\
&\leq \sum_{m=0}^{N_k} \binom{N_k}{m} p^m (1-p)^{N_k - m} \\
&\qquad\qquad \cdot \Pr\left(\sum_{E\in\mathcal{O}_k} Z_E \geq \frac{|\mathcal{O}_k|}{2} \;\middle|\; |\mathcal{O}_k| = m \right) \\
&\leq \sum_{m=0}^{N_k} \binom{N_k}{m} p^m (1-p)^{N_k - m} e^{-mD(0.5\|\theta)} \\
&= \left(1 - p(1 - e^{-D(0.5\|\theta)})\right)^{N_k} \stackrel{\text{def}}{=} (1-p')^{N_k},
\end{aligned}$$

where the last inequality is due to the Chernoff bound [22].

This together with (3) gives

$$P_e \leq \sum_{k=1}^{n-1} \binom{n}{k} (1-p')^{N_k}.$$

Now, we are left with estimating N_k to complete the proof. For $d \asymp 1$, N_k can be easily estimated using the fact that $\binom{n}{d} \approx \frac{n^d}{d!}$. However, in the general case where d can scale with n, the estimation is no longer valid. On the other hand, the following lemma, which is one of the key technical contributions in this paper, provides a clever lower bound:

Lemma 1. *Let* $\beta = \lceil \max\{\frac{n-d+1}{2d+1}, \frac{d+1}{2(n-d)+1}\} \rceil$ *and* $\alpha = \max\{\frac{n-d+1}{d}, \frac{d+1}{n-d}\}$. *Then*

$$\sum_{\substack{i\leq d \\ i \text{ is odd}}} \binom{k}{i}\binom{n-k}{d-i} \geq \begin{cases} \frac{2k}{5\alpha}\binom{n}{d}, & k < \beta, \\ \frac{1}{5}\binom{n}{d}, & \beta \leq k \leq n-\beta, \\ \frac{2(n-k)}{5\alpha}\binom{n}{d}, & n-\beta < k. \end{cases}$$

We present the proof in the full version [10]. Employing Lemma 1, we get:

$$\begin{aligned}
P_e \leq& \sum_{k=1}^{\beta-1} \binom{n}{k}(1-p')^{N_k} + \sum_{k=\beta}^{n-\beta} \binom{n}{k}(1-p')^{N_k} \\
&+ \sum_{k=n-\beta+1}^{n-1} \binom{n}{k}(1-p')^{N_k} + (1-p')^{\binom{n}{d}} \\
\leq& 2\sum_{k=1}^{\beta-1} \binom{n}{k}(1-p')^{\frac{2k}{5\alpha}\binom{n}{d}} + \sum_{k=\beta}^{n-\beta} \binom{n}{k}(1-p')^{\frac{1}{5}\binom{n}{d}} \\
&+ (1-p')^{\binom{n}{d}} \\
\leq& 2\sum_{k=1}^{\beta-1} n^k e^{-p'\frac{2k}{5\alpha}\binom{n}{d}} + 2^n e^{-\frac{1}{5}p'\binom{n}{d}} + e^{-p'\binom{n}{d}} \\
\leq& 2\sum_{k=1}^{\beta-1} \exp\left\{ k\left(\log n - \frac{2p'\binom{n}{d}}{5\alpha} \right) \right\} \qquad (4)\\
&+ \exp\left\{ n\log 2 - \frac{1}{5}p'\binom{n}{d} \right\} + e^{-p'\binom{n}{d}}. \qquad (5)
\end{aligned}$$

Note that (5) vanishes since $p' = p(1 - e^{-D(0.5\|\theta)}) \geq (1+\epsilon)\frac{n\cdot 5\log 2}{\binom{n}{d}}$. We now show that (4) also vanishes by considering two cases: $d = o(n)$ and $d \asymp n$. First, consider the case of $d = o(n)$. Note that $\alpha = \max\left\{\frac{n-d+1}{d}, \frac{d+1}{n-d}\right\} \leq \frac{n}{d}$, $\beta = \left\lceil \max\left\{\frac{n-d+1}{2d+1}, \frac{d+1}{2(n-d)+1}\right\}\right\rceil = \left\lceil \frac{n-d+1}{2d+1} \right\rceil \to \infty$, and $\gamma \stackrel{\text{def}}{=} \log n - \frac{2dp'\binom{n}{d}}{5n} \to -\infty$. Thus,

$$(4) \leq 2\sum_{k=1}^{\beta-1} \exp\{k\gamma\} = 2\frac{\exp\{\gamma\} - \exp\{\beta\gamma\}}{1-\exp\{\gamma\}} \to 0.$$

When $d \asymp n$, $\beta = \lceil \max\{\frac{n-d+1}{2d+1}, \frac{d+1}{2(n-d)+1}\} \rceil \asymp 1$ and $\alpha = \max\{\frac{n-d+1}{d}, \frac{d+1}{n-d}\} \asymp 1$. Also, $\delta \stackrel{\text{def}}{=} \log n - \frac{2p'\binom{n}{d}}{5\alpha} \to -\infty$.

Thus,

$$(4) = 2\frac{\exp\{\delta\} - \exp\{\beta\delta\}}{1 - \exp\{\delta\}} \to 0.$$

Therefore, $P_e \to 0$. Carefully following the arguments above, we can see that $P_e = O(n^{-\epsilon})$.

B. Converse

First, it is obvious that $P_e \not\to 0$ for $p \le \frac{n}{(1-H(\theta))\binom{n}{d}}$. Hence, it suffices to consider the case $d = O(\log n)$ and $\binom{n}{d}p \asymp \frac{n\log n}{d}$. Let S be the event that the ground truth vector is the unique candidate for $\arg\min_{\mathbf{A}} d(\mathbf{A})$. As $\Pr(S) \to 0$ implies $\liminf_{n\to\infty} P_e \ge 1/2$, we will show that $\Pr(S) \to 0$ when $p \le (1-\epsilon)C_\theta \frac{n\log n}{d\binom{n}{d}}$. First, observe

$$\begin{aligned}\Pr(S) &= \Pr(S \,|\, \mathbf{X} = \mathbf{0}) \\ &\le \Pr\left(\bigcap_{\mathbf{A}\neq\mathbf{0}} [d(\mathbf{A}) > d(\mathbf{0})] \,\middle|\, \mathbf{X} = \mathbf{0}\right) \\ &\le \Pr\left(\bigcap_{i=1}^{n} [d(\mathbf{e}_i) > d(\mathbf{0})] \,\middle|\, \mathbf{X} = \mathbf{0}\right).\end{aligned}$$

We find an upper bound to the above quantity by finding a large enough subset of $[n]$ such that the corresponding collection of events of form $[d(\mathbf{e}_\cdot) > d(\mathbf{0})]$ are mutually independent. To this end, we first propose a crude construction of a subset of $[n]$ whose nodes do not share hyperedges. First, choose a big subset $\mathcal{R}_{\text{big}} = \left\{1, 2, \cdots, \lceil 2c\frac{n}{\log^6 n}\rceil\right\}$ for some absolute constant $c > 0$. Then erase every node in $\mathcal{R}_{\text{big}}$ which shares hyperedges with other nodes in $\mathcal{R}_{\text{big}}$ to obtain $\mathcal{R}_{\text{res}}$. Formally,

$$\mathcal{R}_{\text{res}} = \mathcal{R}_{\text{big}} - \bigcup_{k=2}^{d} \mathcal{R}_{\text{share}}^{(k)},$$

where

$$\mathcal{R}_{\text{share}}^{(k)} = \bigcup_{E\in\mathcal{F}_{\text{share}}^{(k)}} E \cap \mathcal{R}_{\text{big}},$$

and $\mathcal{F}_{\text{share}}^{(k)}$ is the collection of edges that meets $\mathcal{R}_{\text{big}}$ at exactly k nodes. The following lemma guarantees that $\mathcal{R}_{\text{res}}$ has comparable size with $\mathcal{R}_{\text{big}}$.

Lemma 2. *Suppose $d = O(\log n)$. Then with probability approaching 1,*

$$|\mathcal{R}_{res}| \ge c\frac{n}{\log^6 n}$$

for some absolute constant $c > 0$.

The proof can be found in the full version [10]. Let $\mathcal{E}_{\text{typ}} = [|\mathcal{R}_{\text{res}}| \ge c\frac{n}{\log^6 n}]$. Conditioned on $\mathcal{E}_{\text{typ}}$, the collection $[d(\mathbf{e}_i) > d(\mathbf{0})]_{i\in\mathcal{R}_{\text{res}}}$ is statistically independent. Hence,

$$\begin{aligned}\Pr(S) &\lesssim \Pr\left(\bigcap_{i\in\mathcal{R}_{\text{res}}} [d(\mathbf{e}_i) > d(\mathbf{0})] \,\middle|\, \mathbf{X} = \mathbf{0}, \mathcal{E}_{\text{typ}}\right) \\ &= \prod_{i\in\mathcal{R}_{\text{res}}} \Pr\left(d(\mathbf{e}_i) > d(\mathbf{0}) \,\middle|\, \mathbf{X} = \mathbf{0}, \mathcal{E}_{\text{typ}}\right),\end{aligned}$$

which leaves us to seek an upper bound of

$$\Pr\left(d(\mathbf{e}_i) > d(\mathbf{0}) \,\middle|\, \mathbf{X} = \mathbf{0}, \mathcal{E}_{\text{typ}}\right)$$

or a lower bound of

$$\Pr\left(d(\mathbf{e}_i) \le d(\mathbf{0}) \,\middle|\, \mathbf{X} = \mathbf{0}, \mathcal{E}_{\text{typ}}\right)$$

for $i \in \mathcal{R}_{\text{res}}$. We denote by $\mathcal{F}_i \subset \mathcal{E}$ the collection of hyperedges that contain i, for $i \in \mathcal{R}_{\text{res}}$. Due to the construction, edges in $\mathcal{F}_i$ must meet $\mathcal{R}_{\text{big}}$ only at i. Hence, $|\mathcal{F}_i| \le \binom{n-|\mathcal{R}_{\text{big}}|}{d-1} =: N$. Observe that $d(\mathbf{e}_i) \le d(\mathbf{0})$ when $\sum_{E\in\mathcal{F}_i} Z_E \ge |\mathcal{F}_i|/2$. Thus,

$$\begin{aligned}&\Pr\left(d(\mathbf{e}_i) \le d(\mathbf{0}) \,\middle|\, \mathbf{X} = \mathbf{0}, \mathcal{E}_{\text{typ}}\right) \\ &= \sum_{m=1}^{N} \Pr(|\mathcal{F}_i| = m) \Pr\left(d(\mathbf{e}_i) \le d(\mathbf{0}) \;\middle|\; \mathbf{X} = \mathbf{0}, |\mathcal{F}_i| = m\right) \\ &\ge \sum_{m=1}^{N} \binom{N}{m} p^m (1-p)^{N-m} \\ &\qquad \cdot \Pr\left(\sum_{E\in\mathcal{F}_i} Z_E \ge \frac{|\mathcal{F}_i|}{2} \;\middle|\; |\mathcal{F}_i| = m\right).\end{aligned}$$

Applying the reverse Chernoff bound [22] with a fixed $\delta > 0$, there exists $n_\delta > 0$ such that

$$\Pr\left(\sum_{E\in\mathcal{F}_i} Z_E \ge \frac{|\mathcal{F}_i|}{2} \;\middle|\; |\mathcal{F}_i| = m\right) \ge e^{-(1+\delta)mD(0.5\|\theta)}$$

for all $m \ge n_\delta$. Let g_n be a sequence (to be determined) that diverges to ∞ as $n \to \infty$. Then for sufficiently large n,

$$\begin{aligned}&\sum_{m=1}^{N} \binom{N}{m} p^m (1-p)^{N-m} \Pr\left(\sum_{E\in\mathcal{F}_i} Z_E \ge \frac{|\mathcal{F}_i|}{2} \;\middle|\; |\mathcal{F}_i| = m\right) \\ &\ge \sum_{m=1}^{N} \binom{N}{m} p^m (1-p)^{N-m} e^{-(1+\delta)mD(0.5\|\theta)} \qquad (6) \\ &\quad - \sum_{m=1}^{g_n-1} \binom{N}{m} p^m (1-p)^{N-m} e^{-(1+\delta)mD(0.5\|\theta)}. \qquad (7)\end{aligned}$$

We shall show that (7) is negligible compared to (6). Note that

$$\begin{aligned}\frac{(7)}{(6)} &\le \frac{(1-p)^N \sum_{m=1}^{g_n-1} \left(N\frac{pe^{-(1+\delta)D(0.5\|\theta)}}{1-p}\right)^m}{(1-p)^N \sum_{m=1}^{N} \binom{N}{m} \left(\frac{p}{1-p}e^{-(1+\delta)D(0.5\|\theta)}\right)^m} \\ &\le \frac{\sum_{m=1}^{g_n-1} \left(N\frac{pe^{-(1+\delta)D(0.5\|\theta)}}{1-p}\right)^m}{\left(1 + \frac{p}{1-p}e^{-(1+\delta)D(0.5\|\theta)}\right)^N} \\ &\approx \frac{\sum_{m=1}^{g_n-1} \left(N\frac{pe^{-(1+\delta)D(0.5\|\theta)}}{1-p}\right)^m}{\exp\left\{N\frac{pe^{-(1+\delta)D(0.5\|\theta)}}{1-p}\right\}}. \qquad (8)\end{aligned}$$

As $|\mathcal{R}_{\text{big}}| = \Theta(\frac{n}{\log^6 n})$ and $d = O(\log n)$, simple algebra yields

$$(n-j)\left(1 - \frac{1}{\log^2 n}\right) \le (n - j - |\mathcal{R}_{\text{big}}|)$$

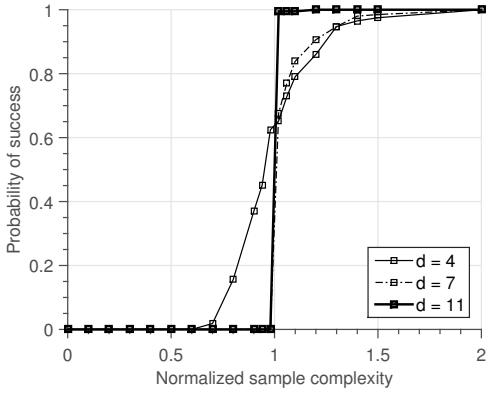

Fig. 2: We run the Monte Carlo simulations to estimate the probability of success for $n = 1000$, varying d, and $\theta = 0$. For each d, we normalize the number of samples by $\max(n, n \log n/d)$. Observe that the probability of success quickly approaches 1 as the normalized sample complexity crosses 1.

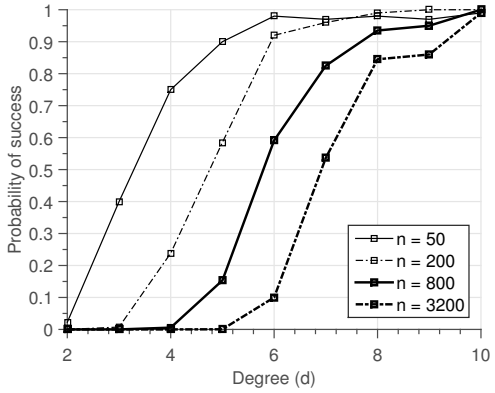

Fig. 3: We run the Monte Carlo simulations to estimate the probability of success for varying n, varying d, $\theta = 0$, and $p = 1.1n/\binom{n}{d}$. Note that when n increases by a multiplicative factor of 4, the curve shifts rightward about the same amount, supporting our result in Corollary 2

for $0 \le j \le d-2$, which in turn gives

$$\frac{\binom{n-|\mathcal{R}_{\text{big}}|}{d-1}}{\binom{n-1}{d-1}} \ge \left(1 - \frac{1}{\log^2 n}\right)^{d-1} \approx \exp\left\{-\frac{d-1}{\log^2 n}\right\} \to 1.$$

Thus, we obtain $N\frac{pe^{-(1+\delta)D(0.5\|\theta)}}{1-p} \asymp \binom{n-1}{d-1}p \asymp \log n$, and by taking $g_n = \left\lceil \log\left(N\frac{pe^{-(1+\delta)D(0.5\|\theta)}}{1-p}\right)\right\rceil$, we get (8)$\to 0$. Hence,

$$\begin{aligned}
&\Pr\left(d(\mathbf{e}_i) > d(\mathbf{0}) \,\middle|\, \mathbf{X} = \mathbf{0}, \mathcal{E}_{\text{typ}}\right) \\
&\lesssim 1 - \sum_{m=1}^{N} \binom{N}{m} p^m (1-p)^{N-m} e^{-(1+\delta)mD(0.5\|\theta)} \\
&= 1 - (1 - p(1 - e^{-(1+\delta)D(0.5\|\theta)}))^N \\
&\approx \exp\left[-\exp\left\{-Np(1 - e^{-(1+\delta)D(0.5\|\theta)})\right\}\right] \\
&\le \exp\left[-\exp\left\{-\binom{n-1}{d-1} p(1 - e^{-(1+\delta)D(0.5\|\theta)})\right\}\right] \\
&\le \exp\left[-\exp\left\{-(1-\epsilon)\frac{(1 - e^{-(1+\delta)D(0.5\|\theta)})}{(1 - e^{-D(0.5\|\theta)})} \log n\right\}\right] \\
&\le \exp\left[-\exp\left\{-\left(1 - \frac{\epsilon}{2}\right)\log n\right\}\right],
\end{aligned}$$

since $\delta > 0$ can be arbitrarily chosen. Thus, we conclude

$$\begin{aligned}
&\prod_{i \in \mathcal{R}_{\text{res}}} \Pr\left(d(\mathbf{e}_i) > d(\mathbf{0}) \,\middle|\, \mathbf{X} = \mathbf{0}, \mathcal{E}_{\text{typ}}\right) \\
&\le \exp\left[-\exp\left\{-\left(1 - \frac{\epsilon}{2}\right)\log n\right\}\right]^{|\mathcal{R}_{\text{res}}|} \\
&\le \exp\left[-c\frac{n}{\log^6 n}\exp\left\{-\left(1 - \frac{\epsilon}{2}\right)\log n\right\}\right] \to 0.
\end{aligned}$$

V. Experimental Results

In this section, we provide Monte Carlo simulation results which corroborate our theoretical findings. Each point plotted in Fig. 2 and Fig. 3 is an empirical success rate. All results are obtained with 50 Monte Carlo trials. In Fig. 2, we plot the probability of successful recovery for $n = 1000$, varying d, and $\theta = 0$. For each d, we normalize the number of samples by $\max(n, n \log n/d)$. One can observe that the probability of success quickly approaches 1 as the normalized sample complexity crosses 1.

Plotted in Fig. 3 are the simulation results for varying n, varying d, $\theta = 0$, and $p = 1.1n/\binom{n}{d}$. We note that when n increases by a multiplicative factor of 4, the curve shifts rightward about the same amount, supporting our result in Corollary 2.

VI. Conclusion

In this paper, we study the problem of community recovery in hypergraphs for the generalized Censored Block Model (CBM), and characterize the information-theoretic limits of the problem as a function of the number of nodes n, the size of hyperedges d, the noise probability θ, and the edge observation probability p. We also corroborate our theoretical results via Monte Carlo simulations. Our characterizations imply that the community recovery in hypergraphs with a linear number of measurements becomes possible when d is on the order of $\log n$.

References

[1] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3, pp. 75–174, 2010.

[2] M. A. Porter, J.-P. Onnela, and P. J. Mucha, "Communities in networks," *Notices of the AMS*, vol. 56, no. 9, pp. 1082–1097, 2009.

[3] J. Chen and B. Yuan, "Detecting functional modules in the yeast protein–protein interaction network," *Bioinformatics*, vol. 22, no. 18, pp. 2283–2290, 2006.

[4] Q.-X. Huang and L. Guibas, "Consistent shape maps via semidefinite programming," in *Computer Graphics Forum*, vol. 32, no. 5. Wiley Online Library, 2013, pp. 177–186.

[5] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[6] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer, "Linear inverse problems on erdős-rényi graphs: Information-theoretic limits and efficient recovery," in *2014 IEEE International Symposium on Information Theory*. IEEE, 2014, pp. 1251–1255.

[7] ——, "Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery," *IEEE Transactions on Network Science and Engineering*, vol. 1, no. 1, pp. 10–22, 2014.

[8] B. Hajek, Y. Wu, and J. Xu, "Exact recovery threshold in the binary censored block model," in *Information Theory Workshop-Fall (ITW), 2015 IEEE*. IEEE, 2015, pp. 99–103.

[9] R. Durrett, *Random graph dynamics*. Citeseer, 2007, vol. 200, no. 7.
[10] K. Ahn, K. Lee, and C. Suh, "Community recovery in hypergraphs," https://sites.google.com/site/kw1jjang/ALS16.pdf, 2016.
[11] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
[12] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social networks*, vol. 5, no. 2, pp. 109–137, 1983.
[13] A. Condon and R. M. Karp, "Algorithms for graph partitioning on the planted partition model," *Random Structures and Algorithms*, vol. 18, no. 2, pp. 116–140, 2001.
[14] A. Coja-Oghlan, "Graph partitioning via adaptive spectral techniques," *Combinatorics, Probability and Computing*, vol. 19, no. 02, pp. 227–284, 2010.
[15] K. Chaudhuri, F. C. Graham, and A. Tsiatas, "Spectral clustering of graphs with general degrees in the extended planted partition model." in *COLT*, vol. 23, 2012, pp. 35–1.
[16] E. Abbe and C. Sandon, "Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms," *arXiv preprint arXiv:1503.00609*, 2015.
[17] E. Mossel and J. Xu, "Density evolution in the degree-correlated stochastic block model," *arXiv preprint arXiv:1509.03281*, vol. 7, 2015.
[18] S. Heimlicher, M. Lelarge, and L. Massoulié, "Community detection in the labelled stochastic block model," *arXiv preprint arXiv:1209.2910*, 2012.
[19] O. Watanabe and M. Yamamoto, "Average-case analysis for the max-2sat problem," *Theoretical Computer Science*, vol. 411, no. 16, pp. 1685 – 1697, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304397510000022
[20] O. Watanabe, "Message passing algorithms for mls-3lin problem," *Algorithmica*, vol. 66, no. 4, pp. 848–868, 2013.
[21] P. Jain and S. Oh, "Provable tensor factorization with missing data," in *Advances in Neural Information Processing Systems*, 2014, pp. 1431–1439.
[22] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American statistical association*, vol. 58, no. 301, pp. 13–30, 1963.