Community Recovery in Hypergraphs

Kwangjun Ahn*, Kangwook Lee*, and Changho Suh

Abstract

Community recovery is a central problem that arises in a wide variety of applications such as network clustering, motion segmentation, face clustering and protein complex detection. The objective of the problem is to cluster data points into distinct communities based on a set of measurements, each of which is associated with the values of a certain number of data points. While most of the prior works focus on a setting in which the number of data points involved in a measurement is two, this work explores a generalized setting in which the number can be more than two. Motivated by applications particularly in machine learning and channel coding, we consider two types of measurements: (1) *homogeneity* measurement which indicates whether or not the associated data points belong to the same community; (2) *parity* measurement which denotes the modulo-2 sum of the values of the data points. Such measurements are possibly corrupted by Bernoulli noise. We characterize the fundamental limits on the number of measurements required to reconstruct the communities for the considered models.

I. Introduction

Clustering of data is one of the central problems, and it arises in many fields of science and engineering. Among many related problems, *community recovery in graphs* has received considerable attention with applications in numerous domains such as social networks [3]–[5], computational biology [6], and machine learning [7], [8]. The goal of the problem is to cluster data points into different communities based on *pairwise* information. Among a variety of models for the community recovery problem, the stochastic block model (SBM) [9] and the censored block model (CBM) [10] have received significant attention in recent years. In SBM, two data points in the same communities are more likely to be connected by an edge than the other edges. In the case of CBM, each measurement returns the modulo-2 sum of the values assigned to the two nodes, possibly corrupted by Bernoulli noise.

While these models reflect interactions between a pair of two nodes, there are numerous applications in which interactions occur across more than two nodes. One such application is a folksonomy, a social network in which users can annotate items with different tags [11]. In this application, the graph consists of nodes corresponding to different users, different items, and different tags. When user i annotate item j with tag k, one can view this as a hyperedge connecting node i, node j and node k. Therefore, in order to cluster nodes of such a graph based on such interactions, one needs a model that can capture such three-way interactions. Another application is molecular biology, in which multi-way interactions between distinct systems capture complex molecular interactions [12]. There are also a broad range of applications in other domains including computer vision [13], VLSI circuits [14], and categorical databases [15].

These applications naturally motivate us to investigate a *hypergraph* setting in which measurements are of *multi-way* information type. Specifically, we consider a simple yet practically-relevant model, which we name the generalized censored block model (GCBM). In the GCBM, the n data points are modeled as nodes in a *hypergraph*, and their interactions are encoded as hyperedges between the nodes. As an initial effort, we focus on a simple setting in which there are two communities: each node taking either 0 or 1 depending on its affiliation. More concretely, we consider a random d-uniform hypergraph in which each hyperedge connecting a set of d nodes exists with probability p and takes a function of the values assigned to the d nodes. In this work, inspired by applications in machine learning and channel coding, we study the following two types of measurements:

- *the homogeneity measurement* that reveals whether or not the d nodes are in the same community; and
- *the parity measurement* that reveals the modulo-2 sum of the affiliation of the d nodes.

Further, we study both the noiseless case and the noisy case.

A. Main contributions

Specialized to the $d = 2$ case, the above two measurement models reduce to the CBM, in which the information-theoretic limit on the expected number of edges required for exact recovery is characterized as $p\binom{n}{2} = \frac{1}{2} \cdot \frac{n \log n}{\left(\sqrt{1-\theta}-\sqrt{\theta}\right)^2}$ [16], [17]. On the other hand, the information-theoretic limits for the case of arbitrary d has not been settled. This precisely sets the goal of our paper: We seek to characterize the information-theoretic limits on the sample complexity for exact recovery under the two models. A summary of our findings is as follows. For a fixed constant d, the information-theoretic limits are:

Kwangjun Ahn is with the Department of Mathematical Sciences, KAIST (e-mail: kjahnkorea@kaist.ac.kr). Kangwook Lee and Changho Suh are with the School of Electrical Engineering, KAIST (e-mail: {kw1jjang, chsuh}@kaist.ac.kr).

This paper was presented in part at the 54th Annual Allerton Conference on Communication, Control, and Computing 2016 [1], and the IEEE International Symposium on Information Theory 2017 [2].

∗ Kwangjun Ahn and Kangwook Lee contributed equally to this work

TABLE I: **Summary of main results.** The information-theoretic limits on sample complexity ($p\binom{n}{d}$) are summarized. Here, n denotes the number of nodes, θ denotes the noise probability, and d denotes the size of hyperedges. "$d = f(n)$" means that d can scale with n, and "$\Theta_{\theta,d}$" implies that the constant involved depends on θ and d.

	$d = 2$	$d > 2$ (const.)	$d = f(n)$
Homogeneity	$\frac{1}{2} \cdot \frac{n \log n}{\left(\sqrt{1-\theta}-\sqrt{\theta}\right)^2}$	$\frac{2^{d-2}}{d} \cdot \frac{n \log n}{\left(\sqrt{1-\theta}-\sqrt{\theta}\right)^2}$	N/A
Parity	$\frac{1}{2} \cdot \frac{n \log n}{\left(\sqrt{1-\theta}-\sqrt{\theta}\right)^2}$	$\frac{1}{d} \cdot \frac{n \log n}{\left(\sqrt{1-\theta}-\sqrt{\theta}\right)^2}$	$\Theta_{\theta,d}\left(\max\left\{n,\ \frac{n \log n}{d}\right\}\right)$

- (the homogeneity measurement case) $p\binom{n}{d} = \frac{2^{d-2}}{d} \cdot \frac{n \log n}{\left(\sqrt{1-\theta}-\sqrt{\theta}\right)^2}$ if d is a fixed constant; and
- (the parity measurement case) $p\binom{n}{d} = \frac{1}{d} \cdot \frac{n \log n}{\left(\sqrt{1-\theta}-\sqrt{\theta}\right)^2}$ if d is a fixed constant.

For the parity measurement case, we also characterize the information-theoretic limits for a more general setting where d can arbitrarily scale with n.

- (the parity measurement case) $p\binom{n}{d} = \Theta\left(\frac{n \log n}{d}\right)$ if $d = o(\log n)$; and
- (the parity measurement case) $p\binom{n}{d} = \Theta(n)$ if $d = \Omega(\log n)$.

These results provide some interesting implications to relevant applications such as subspace clustering and channel coding. In particular, the results offer concrete guidelines as to how to choose d that minimizes sample complexity while ensuring successful clustering. See details in Sec. II-A and Sec. III.

B. Related work

1) The $d = 2$ case: The exact recovery problem in standard graphs ($d = 2$) has been studied in great generality. In SBM, both the fundamental limits and computationally efficient algorithms are investigated initially for the case of two communities [17]–[19], and recently for the case of an arbitrary number of communities [20]. In CBM, [16] characterizes the sample complexity limit, and [17] develops a computationally efficient algorithm that achieves the limit.

Another important recovery requirement is *detection*, which asks whether one can recover the clusters better than a random guess. The modern study of the detection problem in SBM is initiated by a paper by Decelle et al. [21], which conjectures phase transition phenomena for the detection problem[1]. This conjecture is initially tackled for the case of two communities. The impossibility of the detection below the conjectured threshold is established in [27], and it is proved in [28]–[30] that the conjectured threshold can be achieved efficiently. The conjecture for the arbitrary number of communities is recently settled by Abbe and Sandon [26]. For another line of researches, minimax-optimal rates are derived in [31], and algorithms that achieve the rates are developed in [32]. We refer to a recent survey by Abbe [33] for more exhaustive information.

2) The homogeneity measurement case: Recently, [34], [35] consider a general model that includes our model as a special case (to be detailed in Sec. II), and provide an upper bound on sample complexity for *almost exact* recovery, which allows a vanishing fraction of misclassified nodes. Applying their results to our model, their upper bound reduces to $p\binom{n}{d} = \Omega(n \log^2 n)$. Whether or not the sufficient condition is also necessary has been unknown. In this work, we show that it is not the case, demonstrating that the minimal sample complexity even for exact recovery is $\Theta(n \log n)$.

We note that the homogeneity measurement case is closely related to subspace clustering, one of the popular problems in computer vision [13], [36], [37]; See Sec. II-A1 for details.

3) The parity measurement case: The parity measurement case has been explored by [38] in the context of random constraint satisfaction problems. The case of $d = 3$ has been well-studied: it is shown that the maximum likelihood decoder succeeds if $p\binom{n}{3} \geq 2 \cdot \frac{n \log n}{(0.5-\theta)^2}$ [38]. Unlike the prior result which only considers the case of $d = 3$, we cover an arbitrary constant d, and characterize the sharp threshold on the sample complexity.

Abbe-Montanari [10] relate the parity measurement model to a channel coding problem in which random LDGM codes with a constant right-degree d are employed. By proving the concentration phenomenon of the mutual information between channel input and output, they demonstrate the existence of phase transition for an even d. Our results span *any* fixed d, and hence fully settle the phase transition (see Sec. III).

4) The stochastic block model for hypergraphs: There are several works which study the community recovery under SBM for hypergraphs. In [39], the authors explore the case of two equal-sized communities[2]. Specializing it to our model, one can readily show that detection is possible if $\binom{n}{d}p = \Omega(n)$. Moreover, [40] recently conjectures phase transition thresholds for detection. Lastly, [41] derives the minimax-optimal error rates, and generalizes the results in [31] to the hypergraph case.

[1] In the paper, it is also conjectured that an information-computation gap might exist for the case of more than 3 communites ($k \geq 4$). This conjecture is also extensively studied in [22]–[25], and is recently settled in [26].

[2] Actually, the main model in the paper is *the bipartite stochastic block model*, which is not a hypergraph model. However, the result for the hypergraph case follows as a corollary (see Theorem 5 therein).

5) Other relevant problems: Community recovery in hypergraphs bears similarities to other inference problems, in which the goal is to reconstruct data from multiple queries. Those problems include crowdsourced clustering [42], [43], group testing [44] and data exactration from histogram-type information [45], [46]. Here, one can make a connection to our problem by viewing each query as a hyperedge measurement. However, a distinction lies in the way that queries are collected. For instance, an adaptive measurement model is considered in the crowdsourced setting [42], [43] unlike our non-adaptive setting in which hyperedges are sampled uniformly at random. Histogram-type information acts as a query in [44]–[46].

C. *Paper organization*

Sec. II introduces the considered model; in Sec. III, our main results are presented along with some implications; in Sec. IV, V and VI, we provide the proofs of the main theorems; Sec. VII presents experimental results that corroborate our theoretical findings and discuss interesting aspects in view of applications; and in Sec. VIII, we conclude the paper with some future research directions.

D. *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 $\binom{A}{m} := \{B \subset A : |B| = m\}$. Let $[n]$ denote $\{1, \cdots, n\}$. 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. We use $\mathbb{I}\{\cdot\}$ to denote an indicator function. Let $\mathsf{D}_{\mathsf{KL}}(p\|q)$ be the Kullback-Leibler (KL) divergence between $\mathsf{Bern}(p)$ and $\mathsf{Bern}(q)$, i.e., $\mathsf{D}_{\mathsf{KL}}(p\|q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$. We shall use $\log(\cdot)$ to indicate the natural logarithm. We use $H(\cdot)$ to denote the binary entropy function.

II. GENERALIZED CENSORED BLOCK MODELS

Consider a collection of n nodes $\mathcal{V} = [n]$, each represented by a binary variable $X_i \in \{0, 1\}$, $1 \leq i \leq n$. Let $\mathbf{X} := \{X_i\}_{1\leq i\leq n}$ be the ground-truth vector. Let d denote the size of a hyperedge. Samples are obtained as per a *measurement hypergraph* $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{E} \subset \binom{[n]}{d}$. We assume that each element in $\binom{[n]}{d}$ belongs to $\mathcal{E}$ independently with probability $p \in [0, 1]$. *Sample complexity* is defined as the number of hyperedges in a random measurement hypergraph, which is concentrated around $p\binom{n}{d}$ in the limit of n. Each sampled edge $E \in \mathcal{E}$ is associated with a noisy binary measurement Y_E:

$$Y_E = f(X_{i_1}, X_{i_2}, \cdots, X_{i_d}) \oplus Z_E, \tag{1}$$

where $f : \{0, 1\}^d \to \{0, 1\}$ is some binary-valued function, $\oplus$ denotes modulo-2 sum, and $Z_E \overset{\text{i.i.d.}}{\sim} \mathsf{Bern}(\theta)$ is a random variable with noise rate $0 \leq \theta < \frac{1}{2}$. For the choice of f, we focus on the two cases:

- *the homogeneity measurement:*

$$f_h(X_{i_1}, X_{i_2}, \cdots, X_{i_d}) = \mathbb{I}\{X_{i_1} = X_{i_2} = \cdots = X_{i_d}\};$$

- *the parity measurement:*

$$f_p(X_{i_1}, X_{i_2}, \cdots, X_{i_d}) = X_{i_1} \oplus X_{i_2} \oplus \cdots \oplus X_{i_d}.$$

Let $\mathbf{Y} := \{Y_E\}_{E\in\mathcal{E}}$. We remark that when $d = 2$, this reduces to CBM [16].

The goal of this problem is to recover $\mathbf{X}$ from $\mathbf{Y}$. In this work, we will focus on the case of even d since the case of odd d readily follows from the even case [1]. When d is even, the conditional distribution of $\mathbf{Y}|\mathbf{X}$ is equal to that of $\mathbf{Y}|\mathbf{X} \oplus \mathbf{1}$. Hence, given a recovery scheme ψ, the probability of error is defined as

$$P_e(\psi) := \max_{\mathbf{X}\in\{0,1\}^n} \Pr\left(\psi(\mathbf{Y}) \notin \{\mathbf{X},\ \mathbf{X} \oplus \mathbf{1}\}\right).$$

We intend to characterize the minimum sample complexity, above which there exists a recovery algorithm ψ such that $P_e(\psi) \to 0$ as n tends to infinity, and under which $P_e(\psi) \nrightarrow 0$ for all algorithms.

A. *Relevant applications*

1) Subspace clustering and the homogeneity measurement: Subspace clustering is a popular problem of which the task is to cluster n data points that approximately lie in a union of lower-dimensional affine spaces. The problem arises in a variety of applications such as motion segmentation [47] and face clustering [48], where data points corresponding to the same class (tracked points on a moving object or faces of a person) lie on a single lower-dimensional subspace; for details, see [49] and references therein. A common procedure of the existing algorithms for subspace clustering [37], [50]–[52] begins construction of a d-th order affinity tensor ($d \geq 2$) whose entries represent *similarities* between every d data points. Since this construction incurs a complexity that scales like n^d, sampling-based approaches are proposed in [13], [36], [37].

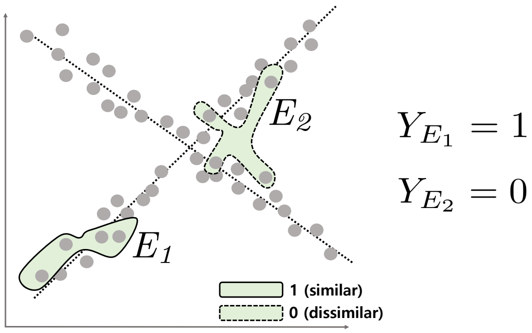


Fig. 1: **Connection to subspace clustering.** Subspace clustering is illustrated for a simple scenario in which the entire signal space is two-dimensional and data points are approximately lying on a union of two 1-dimensional affine spaces (lines). A common procedure in the existing algorithms includes construction of a d-th order affinity tensor ($d \geq 2$) each entry of which represents a quantity that captures a level of similarity across d data points, so taking either 0 or 1 depending on the similarity level. For instance, the four points involved in E_1 in the figure lie near the same affine space, so the similarity measure is decided as 1; on the other hand, the four points in E_2 span different affine spaces, so the similarity measure is decided as 0. Since each data point does not exactly lie in a subspace, an error can occur in the decision—the similarity measurement can be noisy. Hence one can view this problem as the GCBM under the homogeneity measurement model.

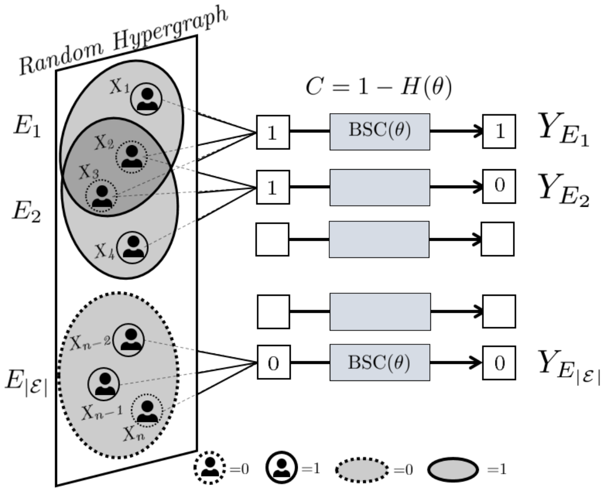


Fig. 2: **Connection to channel coding.** GCBM with the parity information can be seen as a channel coding problem which employs random LDGM codes with a constant right-degree d. To see this, we first draw a random d-uniform hypergraph with n nodes, where each edge of size d appears with probability p. Given the input sequence of n information bits, the parity bits corresponding to all the sampled hyperedges are concatenated, forming a codeword. The noisy measurement can be mapped to the output of a binary symmetric channel (BSC) with crossover probability θ, when fed by the codeword. A recovery algorithm ψ corresponds to the decoder which wishes to infer the n information bits from the received signals. One can then see that recovering communities in hypergraphs is equivalent to the above channel coding problem.

A similarity between d data points in prior works [13], [36], [37] is defined such that it tends to 1 if all of the d points are on the same subspace and 0 otherwise. Hence, restricted to the two-subspace case, one can view a similarity over a d-tuple E as a homogeneity measurement [3]. By setting the probability of each entry being sampled as p, one can relate this to our homogeneity measurement model; see Fig. 1 for visual illustration.

2) Channel coding and the parity measurement: The community recovery problem has an inherent connection with channel coding problems [16], [18]. To see this, consider a communication setting which employs random LDGM codes with a constant right-degree d. To make a connection, we begin by constructing a random d-uniform hypergraph with n nodes, where each edge of size d appears with probability p. Given the input sequence of n information bits, we then concatenate the parity bits with respect to the sampled hyperedges to form a codeword of average length $p\binom{n}{d}$. Note that the expected code rate is $\frac{n}{p\binom{n}{d}}$. The noisy measurement can be mapped to the output of a binary symmetric channel (BSC) with crossover probability θ, when fed by the codeword. A recovery algorithm ψ corresponds to the decoder which wishes to infer the n information bits from the received signals. One can then see that recovering communities in hypergraphs is equivalent to the above channel coding problem; see Fig. 2 for visual illustration.

[3] In subspace clustering, similarities can be sometimes noisy in that even though the d data points are from the same (different) subspace, similarity can be 0 (1). Note that Z_E in (1) precisely captures this noise.

III. Main results

A. The homogeneity measurement case

Theorem 1. *Fix $d \geq 2$ and $\epsilon > 0$. Under the homogeneity measurement case ($f = f_h$),*

$$\begin{cases} \inf_\psi P_e(\psi) \to 0 & \textit{if } \binom{n}{d}p \geq (1+\epsilon)\frac{2^{d-2}}{d}\frac{n \log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2}; \\ \inf_\psi P_e(\psi) \not\to 0 & \textit{if } \binom{n}{d}p \leq (1-\epsilon)\frac{2^{d-2}}{d}\frac{n \log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2}. \end{cases}$$

Proof: See Sec. IV. ■

We first make a comparison to the result in [34]. While [34] models a fairly general similarity measurement, it considers a more relaxed performance metric, so called almost exact recovery, which allows a vanishing fraction of misclassified nodes; and provides a sufficient condition on sample complexity under the setting [53]. On the other hand, we identify the sufficient and necessary condition for *exact* recovery, thereby characterizing the fundamental limit. Specializing their result to the model of our interest, the sufficient condition in [34] reads $\Omega(n \log^2 n)$, which comes with an extra $\log n$ factor gap to the optimality.

One interesting observation in Theorem 1 is that the sample complexity limit is proportional to $\frac{2^{d-2}}{d}$. This suggests that the amount of information that one hyperedge reveals on average is approximately $\frac{d}{2^{d-2}}$ bits. To understand why this is the case, consider a setting in which $\theta = 0$ and an hyperedge $E = \{i_1, i_2, \cdots, i_d\}$ is observed. The case of $Y_E = 1$ implies $X_{i_1} = X_{i_2} = \cdots = X_{i_d}$, in which there are only two uncertain cases (all zeros and all ones), i.e., the $d-1$ bits of information are revealed. On the other hand, the case of $Y_E = 0$ provides much less information as it rules out only two possible cases ($X_{i_1} = X_{i_2} = \cdots = X_{i_d} = 0$ and $X_{i_1} = X_{i_2} = \cdots = X_{i_d} = 1$) out of 2^d possible candidates. This amounts to roughly $d \cdot \frac{2}{2^d}$ bits. Since $Y_E = 1$ occurs with probability $\frac{1}{2^{d-1}}$, the amount of information that one hyperedge can carry on average should read about $\frac{1}{2^{d-1}}(d-1) + \left(1 - \frac{1}{2^{d-1}}\right)\frac{d}{2^{d-1}} \approx \frac{d}{2^{d-2}}$.

Relying on the connection to subspace clustering elaborated in Sec. II-A, one can make an interesting implication from Theorem 1. The result offers a detailed guideline as to how to choose d for sample-efficient subspace clustering. In the case where the measurement quality reflected in θ is irrelevant of the number d of data points involved in a measurement, the limit increases in d. In practical applications, however, θ may depend on d. Actually, the quality of similarity measure can improve as more data points get involved, making θ decrease as d increases. In this case, choosing d as small as possible minimizes $\frac{2^{d-2}}{d}$ but may make θ too large. Hence, there might be a *sweet spot* on d that minimizes the sample complexity. It turns out this is indeed the case in practice. Actually we identify such optimal d^* for motion segmentation application; see Sec. VII-A for details.

B. The parity measurement case

Theorem 2. *Fix $d \geq 2$ and $\epsilon > 0$. Under the parity measurement case ($f = f_p$),*

$$\begin{cases} \inf_\psi P_e(\psi) \to 0 & \textit{if } \binom{n}{d}p \geq (1+\epsilon)\frac{1}{d}\frac{n \log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2}\ ; \\ \inf_\psi P_e(\psi) \not\to 0 & \textit{if } \binom{n}{d}p \leq (1+\epsilon)\frac{1}{d}\frac{n \log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2}\ . \end{cases}$$

Proof: See Sec.V. ■

Notice that for a fixed θ and n, the minimum sample complexity is proportional to $\frac{1}{d}$, hence decreases in d unlike the homogeneity measurement case.

In view of the connection made in Sec. II-A, a natural question that arises in the context of channel coding is to ask how far the rate of the random LDGM code is from the capacity of the BSC channel. The connection can help immediately answer the question. We see from Theorem 2 that the rate of the LDGM code is

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

This suggests that the code rate increases in d. Note that as long as d is constant, the rate vanishes, being far from the capacity of BSC channel $1 - H(\theta)$. On the other hand, it is not clear as to whether or not the random LDGM code can achieve a non-vanishing code rate possibly by increasing the value of d. To check this, we explore the case where d can scale with n. By symmetry, it suffices to consider the case $2 \leq d \leq n/2$. Moreover, to avoid pathological cases where d fluctuates as n increases, we assume that d is a monotone function.

Theorem 3. *Fix d, a monotone function of n such that $2 \leq d \leq n/2$, and $\epsilon > 0$. Under the parity measurement case ($f = f_p$),*

- *(upper bound)* $\inf_\psi P_e(\psi) \to 0$ *if*

$$\binom{n}{d}p \geq (1+\epsilon)\frac{5/2}{d}\frac{n \log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \textit{ and} \tag{2}$$

$$\binom{n}{d}p \geq (1+\epsilon)5\log 2\frac{n}{(\sqrt{1-\theta}-\sqrt{\theta})^2}\ ; \tag{3}$$

- *(lower bound)* $\inf_\psi P_e(\psi) \not\to 0$ *if*

$$\binom{n}{d} p \le (1-\epsilon)\frac{1}{d}\frac{n\log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \text{ or} \tag{4}$$

$$\binom{n}{d} p \le \frac{n}{1-H(\theta)} \,. \tag{5}$$

Proof: See Sec. VI. ■

To see what these results mean, consider the two cases: $d=\Omega(\log n)$ and $d=o(\log n)$. In the case $d=\Omega(\log n)$, the theorem says that for a fixed θ,

$$\inf_\psi P_e(\psi) \to 0 \text{ if } \binom{n}{d}p > \beta_1 n \text{ and}$$
$$\inf_\psi P_e(\psi) \not\to 0 \text{ if } \binom{n}{d}p < \beta_2 n\,,$$

where $\beta_1 = \max\left\{\frac{5/2\log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2 d}, \frac{5\log 2}{(\sqrt{1-\theta}-\sqrt{\theta})^2}\right\} \asymp 1$ and $\beta_2 = \max\left\{\frac{\log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2 d}, \frac{1}{1-H(\theta)}\right\} \asymp 1$. This suggests that as long as d grows asymptotically larger than $\log n$, we can achieve an order-wise tight sample complexity that is linear in n. On the other hand, in the case $d=o(\log n)$, the theorem asserts that

$$\inf_\psi P_e(\psi) \to 0 \text{ if } \binom{n}{d}p > \frac{5/2}{d}\frac{n\log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2} \text{ and}$$
$$\inf_\psi P_e(\psi) \not\to 0 \text{ if } \binom{n}{d}p < \frac{1}{d}\frac{n\log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2}\,.$$

This implies that one cannot achieve the linear-order sample complexity if d grows slower than $\log n$. The implication of the above two can be formally stated as follows.

Corollary 1. *For $d=o(\log n)$, reliable recovery is impossible with linear-order sample complexity, while it is possible for $d=\Omega(\log n)$.*

From this, we see that the random LDGM code can achieve a constant rate as soon as $d=\Omega(\log n)$.

IV. Proof of Theorem 1

The achievability and converse proofs are streamlined with the help of Lemmas 1 and 2, of which the proofs are left in Appendix A. For illustrative purpose, we focus on the noisy case ($\theta > 0$) and assume that n is even. For a vector $\mathbf{V} := \{V_i\}_{1\le i\le n} \in \{0,1\}^n$, we define

$$\begin{cases} f_{\{i_1,i_2,\cdots,i_d\}}(\mathbf{V}) & := f(V_{i_1},V_{i_2},\cdots,V_{i_d}); \\ \mathbf{F}(\mathbf{V}) & := \{f_E(\mathbf{V})\}_{E\in\mathcal{E}}; \\ \mathsf{d_H}(\mathbf{V}) & := \|\mathbf{Y}-\mathbf{F}(\mathbf{V})\|_1\,. \end{cases} \tag{6}$$

Let ψ_{ML} be the maximum likelihood (ML) decoder. One can easily verify that

$$\psi_{\text{ML}}(\mathbf{Y}) = \arg\min_{\mathbf{V}\in\{0,1\}^n} \mathsf{d_H}(\mathbf{V}),$$

where ties are randomly broken.

A. Achievability proof

We intend to prove that

$$\max_{\mathbf{X}\in\{0,1\}^n} \Pr(\psi_{\text{ML}}(\mathbf{Y}) \notin \{\mathbf{X},\mathbf{X}\oplus\mathbf{1}\}) \to 0$$

under the claimed condition. Let $\mathbf{A}\in\{0,1\}^n$ be the ground-truth vector. Without loss of generality, assume that the first k coordinates are 0's and the next $n-k$ coordinates are 1's, where $0\le k\le n/2$.

Let $\mathcal{A}_{i,j}$ denote the collection of all vectors whose coordinates are different from that of $\mathbf{A}$ in i many positions among the first k coordinates and in j many positions among the next $n-k$ coordinates. Note that $\mathcal{A}_{0,0}=\{\mathbf{A}\}$ and $\mathcal{A}_{k,n-k}=\{\mathbf{A}\oplus\mathbf{1}\}$. Thus, a decoding algorithm ψ is successful if and only if the output $\psi(\mathbf{Y})\in\mathcal{A}_{0,0}\cup\mathcal{A}_{k,n-k}$. Let $\mathcal{I} := \{(i,j) \;:\; (i,j)\notin\{(0,0),(k,n-k)\},\; 0\le i\le k,\text{ and } 0\le j\le n-k\}$. We also define

$$\mathbf{V}_{i,j} := (\underbrace{\underbrace{1,\cdots,1}_{i},0,\cdots,0}_{k},\underbrace{\underbrace{0,\cdots,0}_{j},1,\cdots,1}_{n-k})\,,$$

which is a representative vector of $\mathcal{A}_{i,j}$.

Using these notations and the union bound, we get:

$$\begin{aligned}
&\Pr(\psi_{\text{ML}}(\mathbf{Y}) \notin \{\mathbf{X}, \mathbf{X} \oplus \mathbf{1}\} \mid \mathbf{X} = \mathbf{A}) \\
&\overset{(a)}{\leq} \Pr\left(\bigcup_{(i,j)\in\mathcal{I}} \bigcup_{\mathbf{V}\in\mathcal{A}_{i,j}} [\mathsf{d}_{\mathsf{H}}(\mathbf{V}) \leq \mathsf{d}_{\mathsf{H}}(\mathbf{A})] \right) \\
&\leq \sum_{(i,j)\in\mathcal{I}} \sum_{\mathbf{V}\in\mathcal{A}_{i,j}} \Pr\left(\mathsf{d}_{\mathsf{H}}(\mathbf{V}) \leq \mathsf{d}_{\mathsf{H}}(\mathbf{A})\right) \\
&= \sum_{(i,j)\in\mathcal{I}} \binom{k}{i}\binom{n-k}{j} \Pr\left(\mathsf{d}_{\mathsf{H}}(\mathbf{V}_{i,j}) \leq \mathsf{d}_{\mathsf{H}}(\mathbf{A})\right), \qquad (7)
\end{aligned}$$

where the step (a) follows from the fact that the ML decoder outputs $\mathbf{V} \notin \{\mathbf{A}, \mathbf{A} \oplus \mathbf{1}\}$ if $\mathsf{d}_{\mathsf{H}}(\mathbf{V}) \leq \mathsf{d}_{\mathsf{H}}(\mathbf{A})$.

To compare $\mathsf{d}_{\mathsf{H}}(\mathbf{V}_{i,j})$ with $\mathsf{d}_{\mathsf{H}}(\mathbf{A})$, we define the set of *distinctive* hyperedges, i.e., the set of hyperedges such that $f_E(\mathbf{A}) \neq f_E(\mathbf{V}_{i,j})$:

$$\mathcal{F}_{i,j} := \left\{ E \in \binom{[n]}{d} \;:\; f_E(\mathbf{A}) \neq f_E(\mathbf{V}_{i,j}) \right\} \qquad (8)$$

and $\mathcal{E}_{i,j} := \mathcal{E} \cap \mathcal{F}_{i,j}$. By definition, for $E \in \mathcal{E}_{i,j}$, $Y_E = f_E(\mathbf{A})$ if $Z_E = 0$; $Y_E = f_E(\mathbf{V}_{i,j})$ otherwise. Hence, $\mathsf{d}_{\mathsf{H}}(\mathbf{V}_{i,j}) \leq \mathsf{d}_{\mathsf{H}}(\mathbf{A})$ if and only if $\sum_{E\in\mathcal{E}_{i,j}} Z_E \geq \frac{|\mathcal{E}_{i,j}|}{2}$. This leads to:

$$\begin{aligned}
&\Pr\left(\mathsf{d}_{\mathsf{H}}(\mathbf{V}_{i,j}) \leq \mathsf{d}_{\mathsf{H}}(\mathbf{A})\right) \\
&= \sum_{\ell=1}^{|\mathcal{F}_{i,j}|} \Pr\left(\mathsf{d}_{\mathsf{H}}(\mathbf{V}_{i,j}) \leq \mathsf{d}_{\mathsf{H}}(\mathbf{A}) \mid |\mathcal{E}_{i,j}| = \ell\right) \Pr(|\mathcal{E}_{i,j}| = \ell) \qquad (9) \\
&= \sum_{\ell=1}^{|\mathcal{F}_{i,j}|} \Pr\left(\sum_{E\in\mathcal{E}_{i,j}} Z_E \geq \frac{\ell}{2} \;\middle|\; |\mathcal{E}_{i,j}| = \ell \right) \cdot \binom{|\mathcal{F}_{i,j}|}{\ell} p^{\ell}(1-p)^{|\mathcal{F}_{i,j}|-\ell} \\
&\overset{(a)}{\leq} \sum_{\ell=1}^{|\mathcal{F}_{i,j}|} e^{-\ell D(0.5\|\theta)} \binom{|\mathcal{F}_{i,j}|}{\ell} p^{\ell}(1-p)^{|\mathcal{F}_{i,j}|-\ell} \\
&= (1 - (1 - e^{-D(0.5\|\theta)})p)^{|\mathcal{F}_{i,j}|}, \qquad (10)
\end{aligned}$$

where (a) is due to Chernoff-Hoeffding [54]. By letting $p' := (1 - e^{-D(0.5\|\theta)})p$ and applying this to (7), we get:

$$\begin{aligned}
&\Pr(\psi_{\text{ML}}(\mathbf{Y}) \notin \{\mathbf{X}, \mathbf{X} \oplus \mathbf{1}\} \mid \mathbf{X} = \mathbf{A}) \\
&\leq \sum_{(i,j)\in\mathcal{I}} \binom{k}{i}\binom{n-k}{j} (1-p')^{|\mathcal{F}_{i,j}|}. \qquad (11)
\end{aligned}$$

To give a tight upper bound on (11), one needs a tight lower bound on the size of the set of distinctive hyperedges, i.e., $|\mathcal{F}_{i,j}|$. It turns out that bounding $|\mathcal{F}_{i,j}|$ when $d > 2$ requires non-trivial combinatorial counting. Note that this was not the case when $d = 2$ since $|\mathcal{F}_{i,j}|$ can be exactly computed via simple counting. Indeed, one of our main technical contributions lies in the derivation of tight bounds on $|\mathcal{F}_{i,j}|$, which we detail below.

Fact 1. *The number of distinctive hyperedges can be calculated as follows:*

$$|\mathcal{F}_{i,j}| = \sum_{\ell=1}^{d-1} \binom{i}{\ell}\binom{k-i}{d-\ell} + \sum_{\ell=1}^{d-1} \binom{j}{\ell}\binom{n-k-j}{d-\ell} + \sum_{\ell=1}^{d-1} \binom{i}{\ell}\binom{n-k-j}{d-\ell} + \sum_{\ell=1}^{d-1} \binom{k-i}{\ell}\binom{j}{d-\ell}. \qquad (12)$$

Proof: Consider a hyperedge $E = \{i_1, i_2, \cdots, i_d\}$ such that $f_E(\mathbf{A}) = 1$. That is, the hyperedge is connected only to a subset of the first k nodes or only to a subset of the last $n-k$ nodes. That is, $\{i_1, i_2, \cdots, i_d\} \subset \{1, 2, \cdots, k\}$ or $\{i_1, i_2, \cdots, i_d\} \subset \{k+1, k+2, \cdots, n\}$. Consider the first case, i.e., $\{i_1, i_2, \cdots, i_d\} \subset \{1, 2, \cdots, k\}$. In order for this hyperedge to be distinctive, i.e., $f_E(\mathbf{V}_{i,j}) = 0$, at least one element of E must be in $\{1, 2, \cdots, i\}$, and at least one element of E must be in $\{i+1, \cdots, k\}$. Thus, the total number of such distinctive hyperedges is $\sum_{\ell=1}^{d-1} \binom{i}{\ell}\binom{k-i}{d-\ell}$. Similarly, one can count the number of distinctive hyperedges for the case $\{i_1, i_2, \cdots, i_d\} \subset \{k+1, k+2, \cdots, n\}$: $\sum_{\ell=1}^{d-1} \binom{j}{\ell}\binom{n-k-j}{d-\ell}$. By considering the opposite case where $f_E(\mathbf{A}) = 0$ and $f_E(\mathbf{V}_{i,j}) = 1$, one can also obtain the remaining two terms, proving the statement. ■

By symmetry, we see that $|\mathcal{F}_{i,j}| = |\mathcal{F}_{k-i,n-k-j}|$. Hence,

$$\sum_{(i,j)\in\mathcal{I}} \binom{k}{i}\binom{n-k}{j}(1-p')^{|\mathcal{F}_{i,j}|} \tag{13}$$

$$\leq \sum_{(i,j)\in\mathcal{I},\ j\leq\lfloor\frac{n-k}{2}\rfloor} \binom{k}{i}\binom{n-k}{j}(1-p')^{|\mathcal{F}_{i,j}|} + \sum_{(i,j)\in\mathcal{I},\ j\geq\lceil\frac{n-k}{2}\rceil} \binom{k}{i}\binom{n-k}{j}(1-p')^{|\mathcal{F}_{i,j}|} \tag{14}$$

$$= \sum_{(i,j)\in\mathcal{I},\ j\leq\lfloor\frac{n-k}{2}\rfloor} \binom{k}{i}\binom{n-k}{j}(1-p')^{|\mathcal{F}_{i,j}|} + \sum_{(i,j)\in\mathcal{I},\ j\leq\lfloor\frac{n-k}{2}\rfloor} \binom{k}{k-i}\binom{n-k}{n-k-j}(1-p')^{|\mathcal{F}_{k-i,n-k-j}|} \tag{15}$$

$$= 2\sum_{(i,j)\in\mathcal{I},\ j\leq\lfloor\frac{n-k}{2}\rfloor} \binom{k}{i}\binom{n-k}{j}(1-p')^{|\mathcal{F}_{i,j}|} =: 2V. \tag{16}$$

In order to bound V, for a fixed constant $\delta > 0$, we define the following index sets: $\mathcal{I}_{\text{big}} := \{(i,j)\in\mathcal{I} : \left[j \leq \frac{n-k}{2}\right] \cap ([i \geq \delta n] \cup [j \geq \delta n])\}$ and $\mathcal{I}_{\text{small}} := \{(i,j)\in\mathcal{I} : \left[j \leq \frac{n-k}{2}\right] \cap ([i < \delta n] \cap [j < \delta n])\}$. Then,

$$V = \sum_{(i,j)\in\mathcal{I}_{\text{big}}\cup\mathcal{I}_{\text{big}}} \binom{k}{i}\binom{n-k}{j}(1-p')^{|\mathcal{F}_{i,j}|} \tag{17}$$

$$= \sum_{(i,j)\in\mathcal{I}_{\text{big}}} \binom{k}{i}\binom{n-k}{j}(1-p')^{|\mathcal{F}_{i,j}|} \tag{18}$$

$$+ \sum_{(i,j)\in\mathcal{I}_{\text{small}}} \binom{k}{i}\binom{n-k}{j}(1-p')^{|\mathcal{F}_{i,j}|}. \tag{19}$$

Let us first consider (18). Without loss of generality, assume $i \geq \delta n$. Then it follows from (12) that

$$\begin{aligned}|\mathcal{F}_{i,j}| &\geq \sum_{\ell=1}^{d-1}\binom{i}{\ell}\binom{n-k-j}{d-\ell} \overset{(a)}{\geq} \sum_{\ell=1}^{d-1}\binom{i}{\ell}\binom{n/4}{d-\ell} \\ &\geq \binom{i}{1}\binom{n/4}{d-1} \geq \delta n\binom{n/4}{d-1} = \Omega(n^d),\end{aligned}$$

where (a) follows from the hypothesis that $j \leq \frac{n-k}{2}$ and $k \leq \frac{n}{2}$. Then it is easy to show that (18)$\to 0$:

$$\begin{aligned}(18) &\leq \sum_{(i,j)\in\mathcal{I}} \binom{k}{i}\binom{n-k}{j} e^{-p'\Omega(n^d)} \\ &\overset{(a)}{=} e^{-\Omega(n\log n)} \sum_{(i,j)\in\mathcal{I}} \binom{k}{i}\binom{n-k}{j} \leq e^{-\Omega(n\log n)}2^n \to 0,\end{aligned}$$

where (a) follows from the fact that $p'\Omega(n^d) \asymp p\binom{n}{d} = \Omega(n\log n)$.

Now we consider (19). The following lemma gives a tight lower bound on $|\mathcal{F}_{i,j}|$ for this case:

Lemma 1. *For $i < \delta n$ and $j < \delta n$,*

$$|\mathcal{F}_{i,j}| \geq (i+j)\cdot\frac{(1-2\delta)^{d-1}}{2^{d-2}}\binom{n-1}{d-1}.$$

Proof: See Sec. A-A. ■

Applying Lemma 1 to (19), we get:

$$\begin{aligned}(19) &= \sum_{(i,j)\in\mathcal{I}_{\text{small}}} \binom{k}{i}\binom{n-k}{j}(1-p')^{|\mathcal{F}_{i,j}|} \\ &\overset{(a)}{\leq} \sum_{(i,j)\in\mathcal{I}_{\text{small}}} n^i n^j e^{-p'(i+j)\cdot\frac{(1-2\delta)^{d-1}}{2^{d-2}}\binom{n-1}{d-1}} \\ &= \sum_{(i,j)\in\mathcal{I}_{\text{small}}} \exp\left((i+j)\left\{\log n - \frac{p'(1-2\delta)^{d-1}\binom{n-1}{d-1}}{2^{d-2}}\right\}\right),\end{aligned} \tag{20}$$

where (a) follows due to $\binom{k}{i} \le n^i$, $\binom{n-k}{j} \le n^j$ and Lemma 1. A straightforward computation yields $(1 - e^{-\mathsf{D}_{\mathsf{KL}}(0.5\|\theta)}) = (\sqrt{1-\theta} - \sqrt{\theta})^2$, so the claimed condition

$$\binom{n}{d} p \ge (1+\epsilon)\frac{2^{d-2}}{d}\frac{n\log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2}$$

becomes

$$\binom{n}{d} p' \ge (1+\epsilon)\frac{2^{d-2}}{d} n\log n \,. \tag{21}$$

Under the claimed condition, we get:

$$\begin{aligned}\frac{p'(1-2\delta)^{d-1}\binom{n-1}{d-1}}{2^{d-2}} &= \frac{p'(1-2\delta)^{d-1}\binom{n}{d}\frac{d}{n}}{2^{d-2}}\\ &\overset{(a)}{\ge} (1+\epsilon)(1-2\delta)^{d-1}\log n\\ &\overset{(b)}{\ge} (1+\epsilon/2)\log n,\end{aligned}$$

where (a) follows from (21); (b) follows by choosing δ sufficiently small ($(1-2\delta)^{d-1} \to 0$ as $\delta \to 0$). Thus, (20) converges to 0 as n tends to infinity. This completes the proof.

B. *Converse proof*

Let $\mathcal{V}_{1/2}$ be the collection of n-dimensional vectors, each consisting of $n/2$ number of 0's and $n/2$ number of 1's. Moreover, let $\mathbf{X}_{1/2}$ be the random vector sampled uniformly at random over $\mathcal{V}_{1/2}$. For any scheme ψ, by definition of $P_e(\psi)$, we see that

$$\Pr\left(\psi(\mathbf{Y}) \notin \{\mathbf{X},\ \mathbf{X}\oplus\mathbf{1}\} \mid \mathbf{X} = \mathbf{X}_{1/2}\right) \le P_e(\psi)$$

and hence

$$\inf_{\psi}\Pr\left(\psi(\mathbf{Y}) \notin \{\mathbf{X},\ \mathbf{X}\oplus\mathbf{1}\} \mid \mathbf{X} = \mathbf{X}_{1/2}\right) \le \inf_{\psi} P_e(\psi).$$

Relying on this inequality, our proof strategy is to show that the left hand side is strictly bounded away from 0. Note that the infimum in the left hand side is achieved by $\psi_{\text{ML},1/2}$:

$$\psi_{\text{ML},1/2}(\mathbf{Y}) = \arg\min_{\mathbf{V}\in\mathcal{V}_{1/2}} \mathsf{d}_{\mathsf{H}}(\mathbf{V})\,.$$

By letting $\mathbf{A} = (\underbrace{0,\cdots,0}_{n/2},\underbrace{1,\cdots,1}_{n/2})$, we obtain

$$\begin{aligned}&\Pr\left(\psi_{\text{ML},1/2}(\mathbf{Y}) \notin \{\mathbf{X},\ \mathbf{X}\oplus\mathbf{1}\} \mid \mathbf{X} = \mathbf{X}_{1/2}\right)\\ =&\Pr\left(\psi_{\text{ML},1/2}(\mathbf{Y}) \notin \{\mathbf{A},\ \mathbf{A}\oplus\mathbf{1}\} \mid \mathbf{X} = \mathbf{A}\right).\end{aligned}$$

Let S be the success event:

$$S := \bigcap_{\mathbf{V}\in\mathcal{V}_{1/2}\setminus\{\mathbf{A},\mathbf{A}\oplus\mathbf{1}\}} \left[\mathsf{d}_{\mathsf{H}}(\mathbf{V}) > \mathsf{d}_{\mathsf{H}}(\mathbf{A})\right]\,.$$

One can show that $\Pr\left(\psi_{\text{ML},1/2}(\mathbf{Y}) \notin \{\mathbf{A},\ \mathbf{A}\oplus\mathbf{1}\} \mid \mathbf{X} = \mathbf{A}\right) \ge \frac{1}{3}\Pr(S^c)$. This is due to the fact that given S^c, there are more than two candidates for $\arg\min_{\mathbf{V}\in\mathcal{V}_{1/2}} \mathsf{d}_{\mathsf{H}}(\mathbf{V})$, so

$$\Pr\left(\psi_{\text{ML},1/2}(\mathbf{Y}) \notin \{\mathbf{A},\ \mathbf{A}\oplus\mathbf{1}\} \mid \mathbf{X} = \mathbf{A},\ S^c\right) \ge \frac{1}{3}.$$

Hence, it suffices to show $\Pr(S) \to 0$. To give a tight upper bound on $\Pr(S)$, we construct a subset of nodes such that any two nodes in the subset do not share the same hyperedge. To this end, we use the deletion technique (alteration technique) [55]. We first choose a big subset

$$\mathcal{R}_{\text{big}} = \{1,2,\cdots,r\}\bigcup\left\{\frac{n}{2}+1,\frac{n}{2}+2,\cdots,\frac{n}{2}+r\right\},$$

where $r = \lceil\frac{n}{\log^7 n}\rceil$; then erase every node in $\mathcal{R}_{\text{big}}$ which shares hyperedges with other nodes in $\mathcal{R}_{\text{big}}$ to obtain $\mathcal{R}_{\text{res}}$. The following lemma guarantees that $\mathcal{R}_{\text{res}}$ has a comparable size as that of $\mathcal{R}_{\text{big}}$ with high probability. For the later usage, we allow d to scale with n.

Lemma 2. *Suppose $\binom{n}{d}p = O(n\log n)$ and $d = O(\log n)$. Let $\mathcal{R}_{big}$ be a subset of $[n]$ and $\mathcal{R}_{res}$ be a subset obtained from $\mathcal{R}_{big}$ by deleting every node which shares hyperedges with other nodes in $\mathcal{R}_{big}$. If $|\mathcal{R}_{big}| = O(n/\log^7 n)$, then with probability approaching* 1,

$$|\mathcal{R}_{res}| = (1-o(1))|\mathcal{R}_{big}|\,.$$

Proof: See Sec. A-B. ∎

Let Δ be the event that $|\mathcal{R}_{\text{res}}| \ge (1-o(1))|\mathcal{R}_{\text{big}}|$. Given the event Δ, both $\{1,2,\cdots,n/2\}\cap\mathcal{R}_{\text{res}}$ and $\left\{\frac{n}{2}+1,\frac{n}{2}+2,\cdots,n\right\}\cap\mathcal{R}_{\text{res}}$ contain more than $r/2$ elements. We collect $r/2$ elements from each of these sets and denote by $\{b_1,b_2,\cdots,b_{r/2}\}$ and $\{c_1,c_2,\cdots,c_{r/2}\}$, respectively. Suppose that there exist (k,ℓ) such that $\mathsf{d_H}(\mathbf{A}\oplus\mathbf{e}_{b_k}) \le \mathsf{d_H}(\mathbf{A})$ and $\mathsf{d_H}(\mathbf{A}\oplus\mathbf{e}_{c_\ell}) \le \mathsf{d_H}(\mathbf{A})$. Conditioning on Δ, there are no hyperedges that contain both b_k and c_ℓ, so $\mathsf{d_H}(\mathbf{A}\oplus\mathbf{e}_{b_k}\oplus\mathbf{e}_{c_\ell}) \le \mathsf{d_H}(\mathbf{A})$. Hence conditioning on Δ,

$$\begin{aligned} S &\subset \bigcap_{k=1}^{r/2}\left[\mathsf{d_H}(\mathbf{A}\oplus\mathbf{e}_{b_k}) > \mathsf{d_H}(\mathbf{A})\right]\bigcup\bigcap_{k=1}^{r/2}\left[\mathsf{d_H}(\mathbf{A}\oplus\mathbf{e}_{c_k}) > \mathsf{d_H}(\mathbf{A})\right] \\ &=: S'. \end{aligned}$$

Since the event Δ occurs with probability approaching 1 and $S\subset S'$, $\Pr(S)\simeq\Pr(S\mid\Delta)\le\Pr(S'\mid\Delta)$. Hence,

$$\begin{aligned} \Pr(S) &\lesssim \Pr\left(S' \mid \Delta\right) \\ &\le 2\Pr\left(\bigcap_{k=1}^{r/2}\left[\mathsf{d_H}(\mathbf{A}\oplus\mathbf{e}_{b_k}) > \mathsf{d_H}(\mathbf{A})\right] \;\middle|\; \Delta\right) \\ &\overset{(a)}{=} 2\Pr\left(\mathsf{d_H}(\mathbf{A}\oplus\mathbf{e}_{b_1}) > \mathsf{d_H}(\mathbf{A}) \;\middle|\; \Delta\right)^{r/2}, \end{aligned}$$

where (a) follows from the fact that the events $\{[\mathsf{d_H}(\mathbf{A}\oplus\mathbf{e}_{b_k}) > \mathsf{d_H}(\mathbf{A})]\}_{1\le k\le r/2}$ are mutually independent conditioned on Δ. Let $p' = (1-e^{-\mathsf{D_{KL}}(0.5\|\theta)})p$ as in the achievability proof. We intend to give an upper bound on $\Pr\left(\mathsf{d_H}(\mathbf{A}\oplus\mathbf{e}_{b_1}) > \mathsf{d_H}(\mathbf{A}) \;\middle|\; \Delta\right)$, i.e., a lower bound on $\Pr\left(\mathsf{d_H}(\mathbf{A}\oplus\mathbf{e}_{b_1}) \le \mathsf{d_H}(\mathbf{A}) \;\middle|\; \Delta\right)$. Recall from the proof of achievability (see (10)) that

$$\Pr\left(\mathsf{d_H}(\mathbf{V}_{i,j}) \le \mathsf{d_H}(\mathbf{A})\right) \le \left(1-(1-e^{-\mathsf{D_{KL}}(0.5\|\theta)})p\right)^{|\mathcal{F}_{i,j}|}.$$

For the case of $\mathbf{V}_{i,j} = \mathbf{A}\oplus\mathbf{e}_{b_1}$, $|\mathcal{F}_{i,j}| = \binom{n/2-1}{d-1}+\binom{n/2}{d-1}$ (note that $k=n/2, i=1, j=0$). So we get:

$$\Pr\left(\mathsf{d_H}(\mathbf{A}\oplus\mathbf{e}_{b_1}) \le \mathsf{d_H}(\mathbf{A})\right) \le e^{-p'\left(\binom{n/2-1}{d-1}+\binom{n/2}{d-1}\right)}. \tag{22}$$

On the other hand, what we need for the converse proof is a lower bound. In what follows, we will show that (22) is tight enough, more precisely,

$$\Pr\left(\mathsf{d_H}(\mathbf{A}\oplus\mathbf{e}_{b_1}) \le \mathsf{d_H}(\mathbf{A}) \;\middle|\; \Delta\right) \ge (1-o(1))e^{-2p'\binom{n/2-1}{d-1}}. \tag{23}$$

What this means at a high level is that Chernoff-Hoeffding is tight enough. Let us condition on the event Δ for the time being. As in (8), we define the following sets:

$$\mathcal{F}_{b_1} := \left\{E\in\binom{[n]}{d} \;:\; f_E(\mathbf{A}) \ne f_E(\mathbf{A}\oplus\mathbf{e}_{b_1})\right\}$$

and $\mathcal{E}_{b_1} := \mathcal{E}\cap\mathcal{F}_{b_1}$. By definition, for $E\in\mathcal{E}_{b_1}$, $Y_E = f_E(\mathbf{A})$ if $Z_E = 0$; $Y_E = f_E(\mathbf{A}\oplus\mathbf{e}_{b_1})$ otherwise. We see that

$$\mathsf{d_H}(\mathbf{A}\oplus\mathbf{e}_{b_1}) \le \mathsf{d_H}(\mathbf{A}) \Leftrightarrow \sum_{E\in\mathcal{E}_{b_1}} Z_E \ge \frac{|\mathcal{E}_{b_1}|}{2}.$$

Now we want to manipulate $\Pr\left(\mathsf{d_H}(\mathbf{A}\oplus\mathbf{e}_{b_1}) \le \mathsf{d_H}(\mathbf{A}) \;\middle|\; \Delta\right)$ as we did in (9). However, here we need to give a careful attention to the range of summation as $\mathcal{E}_{b_1}$ cannot be equal to $\mathcal{F}_{b_1}$ due to the following reason. Since we conditioned on Δ, no hyperedge in $\mathcal{E}_{b_1}$ intersects $\mathcal{R}_{\text{big}}$ at more than one node (indeed, b_1 is the only node where they intersect); in other words, $\mathcal{E}_{b_1}$ is always contained in a proper subset of $\mathcal{F}_{b_1}$:

$$\mathcal{E}_{b_1} \subset \mathcal{F}_{b_1}\setminus\left\{E\in\binom{[n]}{d} \;:\; |E\cap\mathcal{R}_{\text{big}}| \ge 2\right\} =: \mathcal{G}_{b_1}. \tag{24}$$

Now a manipulation similar to (9) yields:

$$\begin{aligned} &\Pr\left(\mathsf{d_H}(\mathbf{A}\oplus\mathbf{e}_{b_1}) \le \mathsf{d_H}(\mathbf{A}) \mid \Delta\right) \\ &= \sum_{\ell=1}^{|\mathcal{G}_{b_1}|}\Pr\left(\mathsf{d_H}(\mathbf{A}\oplus\mathbf{e}_{b_1}) \le \mathsf{d_H}(\mathbf{A}) \;\middle|\; |\mathcal{E}_{b_1}| = \ell,\ \Delta\right)\Pr(|\mathcal{E}_{b_1}| = \ell|\Delta). \end{aligned}$$

Since the event Δ is related to the occurrence of edges in

$$\left\{E \in \binom{[n]}{d} \ : \ |E \cap \mathcal{R}_{\text{big}}| \geq 2\right\}$$

and $\mathcal{E}_{b_1}$ is subject to (24), Δ and $[|\mathcal{E}_{b_1}| = \ell]$ are independent. Thus, we get:

$$\begin{aligned}
&\Pr\left(\mathsf{d_H}(\mathbf{A} \oplus \mathbf{e}_{b_1}) \leq \mathsf{d_H}(\mathbf{A}) \mid \Delta\right) \\
&= \sum_{\ell=1}^{|\mathcal{G}_{b_1}|} \Pr\left(\mathsf{d_H}(\mathbf{A} \oplus \mathbf{e}_{b_1}) \leq \mathsf{d_H}(\mathbf{A}) \;\middle|\; |\mathcal{E}_{b_1}| = \ell,\ \Delta\right) \Pr(|\mathcal{E}_{b_1}| = \ell) \\
&= \sum_{\ell=1}^{|\mathcal{G}_{b_1}|} \Pr\left(\sum_{E \in \mathcal{E}_{b_1}} Z_E \geq \frac{\ell}{2} \middle| |\mathcal{E}_{b_1}| = \ell\right) \binom{|\mathcal{G}_{b_1}|}{\ell} \frac{p^\ell}{(1-p)^{\ell - |\mathcal{G}_{b_1}|}}. \qquad (25)
\end{aligned}$$

By the reverse Chernoff-Hoeffding bound [54], for a fixed $\delta > 0$, there exists $n_\delta > 0$ such that

$$\Pr\left(\sum_{E \in \mathcal{E}_{b_1}} Z_E \geq \frac{\ell}{2} \middle| |\mathcal{E}_{b_1}| = \ell\right) \geq e^{-(1+\delta)\ell \mathsf{D_{KL}}(0.5\|\theta)}$$

for all $\ell \geq n_\delta$. Let g_n be a sequence (to be determined) such that $g_n \to \infty$ as $n \to \infty$. For sufficiently large n,

$$\begin{aligned}
(25) \geq & \sum_{\ell=1}^{|\mathcal{G}_{b_1}|} \binom{|\mathcal{G}_{b_1}|}{\ell} \frac{(e^{-(1+\delta)\mathsf{D_{KL}}(0.5\|\theta)} p)^\ell}{(1-p)^{\ell - |\mathcal{G}_{b_1}|}} \qquad (26) \\
& - \sum_{\ell=1}^{g_n - 1} \binom{|\mathcal{G}_{b_1}|}{\ell} \frac{(e^{-(1+\delta)\mathsf{D_{KL}}(0.5\|\theta)} p)^\ell}{(1-p)^{\ell - |\mathcal{G}_{b_1}|}}\,. \qquad (27)
\end{aligned}$$

Actually one can choose g_n so that (27) is negligible compared to (26). To see this, we consider:

$$\begin{aligned}
\frac{(27)}{(26)} &\leq \frac{(1-p)^{|\mathcal{G}_{b_1}|} \sum_{\ell=1}^{g_n-1} \left(|\mathcal{G}_{b_1}| \frac{pe^{-(1+\delta)\mathsf{D_{KL}}(0.5\|\theta)}}{1-p}\right)^\ell}{(1-p)^{|\mathcal{G}_{b_1}|} \sum_{\ell=1}^{|\mathcal{G}_{b_1}|} \binom{|\mathcal{G}_{b_1}|}{\ell} \left(\frac{pe^{-(1+\delta)\mathsf{D_{KL}}(0.5\|\theta)}}{1-p}\right)^\ell} \\
&= \frac{\sum_{\ell=1}^{g_n-1} \left(|\mathcal{G}_{b_1}| \frac{pe^{-(1+\delta)\mathsf{D_{KL}}(0.5\|\theta)}}{1-p}\right)^\ell}{\left(1 + \frac{pe^{-(1+\delta)\mathsf{D_{KL}}(0.5\|\theta)}}{1-p}\right)^{|\mathcal{G}_{b_1}|}} \\
&\overset{(a)}{=} \frac{\sum_{\ell=1}^{g_n-1} \left(|\mathcal{G}_{b_1}| \frac{pe^{-(1+\delta)\mathsf{D_{KL}}(0.5\|\theta)}}{1-p}\right)^\ell}{(1+o(1)) \exp\left(|\mathcal{G}_{b_1}| \frac{pe^{-(1+\delta)\mathsf{D_{KL}}(0.5\|\theta)}}{1-p}\right)} \\
&=: \frac{\sum_{\ell=1}^{g_n-1} q^\ell}{(1+o(1)) e^q}, \qquad (28)
\end{aligned}$$

where (a) follows from the fact that $\lim_{x \to 0+} \frac{1+x}{e^x} = 1$, and the last equation is due to the following definition: $q := |\mathcal{G}_{b_1}| \frac{pe^{-(1+\delta)\mathsf{D_{KL}}(0.5\|\theta)}}{1-p}$. One can easily verify that $|\mathcal{F}_{b_1}| = \binom{n/2-1}{d-1} + \binom{n/2}{d-1}$ and $|\mathcal{G}_{b_1}| = \binom{n/2-1-r}{d-1} + \binom{n/2-r}{d-1}$. Since $r = o(n)$, $\lim_{n \to \infty} |\mathcal{G}_{b_1}|/|\mathcal{F}_{b_1}| \to 1$. Thus,

$$\begin{aligned}
q &= |\mathcal{G}_{b_1}| \frac{pe^{-(1+\delta)\mathsf{D_{KL}}(0.5\|\theta)}}{1-p} \qquad (29) \\
&\asymp |\mathcal{F}_{b_1}| \frac{pe^{-(1+\delta)\mathsf{D_{KL}}(0.5\|\theta)}}{1-p} \asymp n^{d-1} p = \Omega(\log n)\,. \qquad (30)
\end{aligned}$$

Therefore, if one chooses $g_n = \lfloor \log q \rfloor$,

$$\frac{(27)}{(26)} = \frac{\sum_{\ell=1}^{g_n-1} q^\ell}{e^q} \leq \frac{g_n q^{g_n}}{e^q} \leq \frac{\log q \cdot q^{\log q}}{e^q} = \frac{\log q \cdot e^{(\log q)^2}}{e^q} \to 0,$$

and thus $(27) = o(1) \cdot (26)$.

Hence, we get:

$$\begin{aligned}(25) &= (26) - (27)\\ &\geq (1-o(1))\sum_{\ell=1}^{|\mathcal{G}_{b_1}|}\binom{|\mathcal{G}_{b_1}|}{\ell}\frac{(e^{-(1+\delta)\mathsf{D}_{\mathsf{KL}}(0.5\|\theta)}p)^{\ell}}{(1-p)^{\ell-|\mathcal{G}_{b_1}|}}\\ &= (1-o(1))\left(1-(1-e^{-(1+\delta)\mathsf{D}_{\mathsf{KL}}(0.5\|\theta)})p\right)^{|\mathcal{G}_{b_1}|}\\ &\overset{(a)}{\geq} (1-o(1))\left(1-(1-e^{-(1+\delta)\mathsf{D}_{\mathsf{KL}}(0.5\|\theta)})p\right)^{2\binom{n/2}{d-1}}\\ &\overset{(b)}{=} (1-o(1))\exp\left(-2\binom{n/2}{d-1}(1-e^{-(1+\delta)\mathsf{D}_{\mathsf{KL}}(0.5\|\theta)})p\right),\end{aligned}$$

where (a) follows since $|\mathcal{G}_{b_1}| \leq |\mathcal{F}_{b_1}| \leq 2\binom{n/2}{d-1}$; (b) follows from the fact that $\lim_{x\to 0+}\frac{1+x}{e^x} = 1$. As $\delta > 0$ can be chosen arbitrarily small, the term $e^{-(1+\delta)\mathsf{D}_{\mathsf{KL}}(0.5\|\theta)}$ can be made arbitrarily close to $e^{-\mathsf{D}_{\mathsf{KL}}(0.5\|\theta)}$, which in turn ensures that the last term is essentially equal to

$$(1-o(1))e^{-2p'\binom{n/2}{d-1}}.$$

Applying this to the previous upper bound on $\Pr(S)$, we get:

$$\begin{aligned}\Pr(S) &\leq \Pr\left(\mathsf{d}_{\mathsf{H}}(\mathbf{A}\oplus\mathbf{e}_{b_1}) > \mathsf{d}_{\mathsf{H}}(\mathbf{A}) \;\middle|\; \Delta\right)^{r/2}\\ &\leq \left(1-(1-o(1))e^{-2p'\binom{n/2}{d-1}}\right)^{r/2}\\ &\leq \exp\left(-(1-o(1))\frac{r}{2}e^{-2p'\binom{n/2}{d-1}}\right)\\ &= \exp\left(-(1-o(1))\frac{n}{2\log^7 n}e^{-(1+o(1))\cdot\frac{p'd\binom{n}{d}}{2^{d-2}n}}\right),\end{aligned}$$

where the last equality follows from the fact that

$$\lim_{n\to\infty}\frac{2p'\binom{n/2}{d-1}}{p'd\binom{n}{d}/2^{d-2}n} \to 1 \text{ and } r = \left\lceil\frac{n}{\log^7 n}\right\rceil.$$

The last term converges to 0 as $p' \leq (1-\epsilon)\frac{2^{d-2}}{d}\frac{n\log n}{\binom{n}{d}}$.

V. Proof of Theorem 2

In this section, we prove a similar statement for the parity measurement case.

A. Achievability proof

Note that the parity measurement is *symmetric* in a sense that for any two vector $\mathbf{A}$ and $\mathbf{B}$, $\Pr\left(\psi_{\text{ML}}(\mathbf{Y}) \notin \{\mathbf{X},\ \mathbf{X}\oplus\mathbf{1}\} \mid \mathbf{X}=\mathbf{A}\right) = \Pr\left(\psi_{\text{ML}}(\mathbf{Y}) \notin \{\mathbf{X},\ \mathbf{X}\oplus\mathbf{1}\} \mid \mathbf{X}=\mathbf{B}\right)$. Hence, we will prove that

$$\Pr\left(\psi_{\text{ML}}(\mathbf{Y}) \notin \{\mathbf{X},\ \mathbf{X}\oplus\mathbf{1}\} \mid \mathbf{X}=\mathbf{0}\right) \to 0$$

under the claimed condition. Conditioning on $\mathbf{X} = \mathbf{0}$,

$$\begin{aligned}
&\Pr\left(\psi_{\mathsf{ML}}(\mathbf{Y}) \notin \{\mathbf{0}, \mathbf{1}\}\right) \\
&\leq \Pr\left(\bigcup_{\mathbf{A} \neq \mathbf{0}, \mathbf{1}} [\mathsf{d_H}(\mathbf{A}) \leq \mathsf{d_H}(\mathbf{0})]\right) \\
&= \Pr\left(\bigcup_{k=1}^{n-1} \bigcup_{\|\mathbf{A}\|_1 = k} [\mathsf{d_H}(\mathbf{A}) \leq \mathsf{d_H}(\mathbf{0})]\right) \\
&\leq \sum_{k=1}^{n-1} \sum_{\|\mathbf{A}\|_1 = k} \Pr\left(\mathsf{d_H}(\mathbf{A}) \leq \mathsf{d_H}(\mathbf{0})\right) \\
&\overset{(a)}{=} 2 \cdot \sum_{k=1}^{n/2} \sum_{\|\mathbf{A}\|_1 = k} \Pr\left(\mathsf{d_H}(\mathbf{A}) \leq \mathsf{d_H}(\mathbf{0})\right) \\
&\overset{(b)}{=} 2 \cdot \sum_{k=1}^{n/2} \binom{n}{k} \Pr\left(\mathsf{d_H}\left(\sum_{i=1}^{k} \mathbf{e}_i\right) \leq \mathsf{d_H}(\mathbf{0})\right), \qquad (31)
\end{aligned}$$

where (a) follows form the fact that $\Pr\left(\mathsf{d_H}(\mathbf{A}) \leq \mathsf{d_H}(\mathbf{0})\right) = \Pr\left(\mathsf{d_H}(\mathbf{A} \oplus \mathbf{1}) \leq \mathsf{d_H}(\mathbf{0})\right)$; (b) follows due to symmetry. To compare $\mathsf{d_H}\left(\sum_{i=1}^{k} \mathbf{e}_i\right)$ and $\mathsf{d_H}(\mathbf{0})$, we define

$$\mathcal{F}_k := \left\{ E \in \binom{[n]}{d} \;:\; f_E(\mathbf{0}) \neq f_E\left(\sum_{i=1}^{k} \mathbf{e}_i\right) \right\}$$

and $\mathcal{E}_k := \mathcal{E} \cap \mathcal{F}_k$. As in (10), we obtain

$$\begin{aligned}
\Pr\left(\mathsf{d_H}\left(\sum_{i=1}^{k} \mathbf{e}_i\right) \leq \mathsf{d_H}(\mathbf{0})\right) &\leq (1 - (1 - e^{-\mathsf{D_{KL}}(0.5\|\theta)})p)^{|\mathcal{F}_k|} \\
&= (1 - p')^{|\mathcal{F}_k|},
\end{aligned}$$

yielding

$$\frac{1}{2} \cdot (31) \leq \sum_{k=1}^{n/2} \binom{n}{k} (1 - p')^{|\mathcal{F}_k|}. \qquad (32)$$

We again count $|\mathcal{F}_k|$ in an effort to obtain a tight upper bound on (32). Notice that $E \in \mathcal{F}_k$ if $|E \cap [k]|$ is odd, and hence

$$|\mathcal{F}_k| = \sum_{\substack{i \leq d \\ i \text{ is odd}}} \binom{k}{i} \cdot \binom{n-k}{d-i}. \qquad (33)$$

Let $\delta > 0$ be a small constant that will be determined later. For the case $k \geq \delta n$, it follows that

$$|\mathcal{F}_k| \geq \binom{k}{1}\binom{n-k}{d-1} \geq \delta n \binom{n/2}{d-1} = \Omega(n^d).$$

Then it is easy to show (32)→ 0 for this case:

$$\begin{aligned}
\sum_{k=\delta n}^{n/2} \binom{n}{k} (1 - p')^{|\mathcal{F}_k|} &\leq \sum_{k=\delta n}^{n/2} \binom{n}{k} e^{-p'\Omega(n^d)} \\
&\overset{(a)}{=} e^{-\Omega(n \log n)} \sum_{k=\delta n}^{n/2} \binom{n}{k} \leq e^{-\Omega(n \log n)} 2^n \to 0,
\end{aligned}$$

where (a) follows from the fact that $p'\Omega(n^d) \asymp p\binom{n}{d} = \Omega(n \log n)$. For the case $k < \delta n$, we see that

$$\begin{aligned}
|\mathcal{F}_k| &\geq \binom{k}{1}\binom{n-k}{d-1} \geq k\binom{(1-\delta)n}{d-1} \\
&\underset{n \to \infty}{\overset{(a)}{=}} (1 + o(1))k(1-\delta)^{d-1}\binom{n-1}{d-1}, \qquad (34)
\end{aligned}$$

where (a) follows since

$$\lim_{n\to\infty}\frac{\alpha^{d-1}\binom{n-1}{d-1}}{\binom{\alpha n}{d-1}}=1 \tag{35}$$

holds for a fixed d and $\alpha\in(0,1)$. Hence, we get

$$\begin{aligned}\sum_{k=1}^{\delta n}\binom{n}{k}(1-p')^{|\mathcal{F}_k|} &\le \sum_{k=1}^{\delta n} n^k e^{-(1+o(1))p'k(1-\delta)^{d-1}\binom{n}{d-1}}\\ &=\sum_{k=1}^{\delta n} e^{k\cdot\left\{\log n-(1+o(1))p'(1-\delta)^{d-1}\binom{n}{d-1}\right\}}\,. \end{aligned}\tag{36}$$

By choosing δ arbitrarily small, under the claimed condition, one can make

$$\begin{aligned}p'(1-\delta)^{d-1}\binom{n}{d-1} &= (1+o(1))(1-\delta)^{d-1}\binom{n}{d}p'\frac{d}{n}\\ &\ge (1+\epsilon/2)\log n\,,\end{aligned}$$

which implies that (36) converges to 0 as n tends to infinity.

B. Converse proof

As the parity measurement is symmetric,

$$\inf_{\psi} P_e(\psi)=\Pr\left(\psi_{\text{ML}}(\mathbf{Y})\notin\{\mathbf{X},\ \mathbf{X}\oplus\mathbf{1}\}\mid\mathbf{X}=\mathbf{0}\right)\,.$$

As before, we define the success event as:

$$S:=\bigcap_{\mathbf{V}\neq\mathbf{0},\mathbf{1}}\left[\mathsf{d}_{\mathsf{H}}(\mathbf{V})>\mathsf{d}_{\mathsf{H}}(\mathbf{0})\right]\,. \tag{37}$$

Again, it suffices to show that $\Pr(S)\to 0$, and to this end, we construct a subset of nodes such that any two nodes in the subset do not share the same hyperedge. Unlike the previous case, the subset is now defined as:

$$\mathcal{R}_{\text{big}}:=\{1,2,\cdots,r\} \tag{38}$$

where $r=\lceil\frac{n}{\log^7 n}\rceil$, and we erase every node in $\mathcal{R}_{\text{big}}$ which shares hyperedges with other nodes in $\mathcal{R}_{\text{big}}$ to obtain $\mathcal{R}_{\text{res}}$. In view of Lemma 2, we have $|\mathcal{R}_{\text{res}}|\ge(1-o(1))r$ almost surely; let Δ be such event. Conditioning on Δ, we enumerate $r/2$ many elements of $\mathcal{R}_{\text{res}}$ by $b_1,\cdots,b_{r/2}$. As there are no hyperedges that connect two nodes in $\mathcal{R}_{\text{res}}$, the events $\{[\mathsf{d}_{\mathsf{H}}(\mathbf{e}_{b_k})>\mathsf{d}_{\mathsf{H}}(\mathbf{0})]\}_{1\le k\le r/2}$ are mutually independent conditioned on Δ. Hence, we get:

$$\begin{aligned}\Pr(S)&\lesssim\Pr\left(S\mid\Delta\right)\\ &\le\Pr\left(\bigcap_{k=1}^{r/2}\left[\mathsf{d}_{\mathsf{H}}(\mathbf{e}_{b_k})>\mathsf{d}_{\mathsf{H}}(\mathbf{0})\right]\,\middle|\,\Delta\right)\\ &=\Pr\left(\mathsf{d}_{\mathsf{H}}(\mathbf{e}_{b_1})>\mathsf{d}_{\mathsf{H}}(\mathbf{0})\,\middle|\,\Delta\right)^{r/2}\,.\end{aligned}\tag{39}$$

Let $p'=(1-e^{-\mathsf{D}_{\mathsf{KL}}(0.5\|\theta)})p$ as before. Using similar arguments used in the previous section, we have

$$\Pr\left(\mathsf{d}_{\mathsf{H}}(\mathbf{e}_{b_1})\le\mathsf{d}_{\mathsf{H}}(\mathbf{0})\,\middle|\,\Delta\right)\ge(1-o(1))e^{-p'\binom{n-1}{d-1}}\,. \tag{40}$$

This gives:

$$\begin{aligned}&\Pr\left(\mathsf{d}_{\mathsf{H}}(\mathbf{e}_{b_1})>\mathsf{d}_{\mathsf{H}}(\mathbf{0})\,\middle|\,\Delta\right)^{r/2}\\ &\le\left(1-(1-o(1))e^{-p'\binom{n-1}{d-1}}\right)^{r/2}\\ &\le\exp\left(-(1-o(1))\frac{r}{2}\exp\left\{-p'\binom{n-1}{d-1}\right\}\right)\\ &\le\exp\left(-(1-o(1))\frac{n}{2\log^7 n}\exp\left\{-(1+o(1))\cdot\frac{p'\binom{n}{d}d}{n}\right\}\right)\,.\end{aligned}$$

Notice that the last term converges to 0 as $\binom{n}{d}p'\le(1-\epsilon)\frac{n\log n}{d}$, which completes the proof.

VI. Proof of Theorem 3

When d scales with n, a technical challenge arises, and we will focus on such technical difficulties, skipping most of the redundant parts.

A. Proof of the upper bound

From (32) and (33), we get

$$P_e(\psi_{\mathrm{ML}}) \le \sum_{k=1}^{n/2} \binom{n}{k} (1-p')^{N_k}\,, \tag{41}$$

where

$$N_k := \sum_{\substack{1 \le i \le d \\ i \text{ is odd}}} \binom{k}{i} \cdot \binom{n-k}{d-i} \tag{42}$$

and $p' := (\sqrt{1-\theta} - \sqrt{\theta})^2 p$. Let us focus on counting N_k. When $d \asymp 1$, $\binom{n}{d} \approx \frac{n^d}{d!}$ suffices to obtain a proper bound on N_k. However, in the general case where d scales with n, one needs a more delicate bounding technique to obtain sharp results. The following lemma presents our new bound.

Lemma 3. *Let $\beta := \lceil \frac{n-d+1}{2d+1} \rceil < n/2$ and $\alpha := \frac{n-d+1}{d}$. Then*

$$\sum_{\substack{1 \le i \le d \\ i \text{ is odd}}} \binom{k}{i}\binom{n-k}{d-i} \ge \begin{cases} \frac{2k}{5\alpha}\binom{n}{d}, & k < \beta; \\ \frac{1}{5}\binom{n}{d}, & \beta \le k \le n/2\,. \end{cases}$$

Proof: See Sec. VI-C. The proof requires an involved combinatorial counting, which is one of our main technical contributions. ■

Employing Lemma 3, we get:

$$\begin{aligned}
(41) &\le \sum_{k=1}^{\beta-1} \binom{n}{k} (1-p')^{N_k} + \sum_{k=\beta}^{n/2} \binom{n}{k} (1-p')^{N_k} \\
&\le \sum_{k=1}^{\beta-1} \binom{n}{k} (1-p')^{\frac{2k}{5\alpha}\binom{n}{d}} + \sum_{k=\beta}^{n/2} \binom{n}{k} (1-p')^{\frac{1}{5}\binom{n}{d}} \\
&\le \sum_{k=1}^{\beta-1} n^k e^{-p'\frac{2k}{5\alpha}\binom{n}{d}} + 2^n e^{-\frac{1}{5}p'\binom{n}{d}} \\
&\le \sum_{k=1}^{\beta-1} \exp\left\{ k \left(\log n - \frac{2p'\binom{n}{d}}{5\alpha} \right) \right\} \qquad (43) \\
&\quad + \exp\left\{ n \log 2 - \frac{1}{5} p' \binom{n}{d} \right\}. \qquad (44)
\end{aligned}$$

Note that (44) vanishes due to (3). In order to show that (43) vanishes as well, we consider two cases: $d = o(n)$ and $d \asymp n$. When $d = o(n)$,

$$\begin{aligned}
&\sum_{k=1}^{\beta-1} \exp\left\{ k \left(\log n - \frac{2p'\binom{n}{d}}{5\alpha} \right) \right\} \\
\le &\sum_{k=1}^{\beta-1} \exp\left\{ k \left(\log n - \frac{2dp'\binom{n}{d}}{5n} \right) \right\} \\
\le &\frac{\exp\left(\log n - \frac{2dp'\binom{n}{d}}{5n} \right)}{1 - \exp\left(\log n - \frac{2dp'\binom{n}{d}}{5n} \right)} \to 0,
\end{aligned}$$

since $\log n - \frac{2dp'\binom{n}{d}}{5n} \to -\infty$.

If $d \asymp n$,

$$\sum_{k=1}^{\beta-1} \exp\left\{k\left(\log n - \frac{2p'\binom{n}{d}}{5\alpha}\right)\right\} \leq \beta \max_{1\leq k\leq \beta-1} \exp\left\{k\left(\log n - \frac{2p'\binom{n}{d}}{5\alpha}\right)\right\} = \beta \exp\left(\log n - \frac{2p'\binom{n}{d}}{5\alpha}\right),$$

where the last equality holds since $\log n - \frac{2p'\binom{n}{d}}{5\alpha} < 0$, and hence $k = 1$ achieves the maximum value. Note that this vanishes since β is asymptotically bounded by a constant. Therefore, (43) always vanishes, completing the proof.

B. Proof of the lower bound

The lower bound statement can be rewritten as follows: $\inf_\psi P_e(\psi) \not\to 0$ if $\binom{n}{d}p \leq \max\left((1-\epsilon)\frac{1}{d}\frac{n\log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2}, \frac{n}{1-H(\theta)}\right)$. Note that when $d = \omega(\log n)$, the condition reduces to $\binom{n}{d}p \leq \frac{n}{1-H(\theta)}$. Hence, it is sufficient to show the following two statements.

- If $d = O(\log n)$: $\inf_\psi P_e(\psi) \not\to 0$ if $\binom{n}{d}p \leq \max\left((1-\epsilon)\frac{1}{d}\frac{n\log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2}, \frac{n}{1-H(\theta)}\right)$.
- If $d = \omega(\log n)$: $\inf_\psi P_e(\psi) \not\to 0$ if $\binom{n}{d}p \leq \frac{n}{1-H(\theta)}$.

We first show that $\binom{n}{d}p \leq \frac{n}{1-H(\theta)}$ implies $\inf_\psi P_e(\psi) \not\to 0$ for all d. By rearranging terms, we have $\binom{n}{d}p \leq \frac{n}{1-H(\theta)} \Leftrightarrow \frac{n}{\binom{n}{d}p} \geq 1 - H(\theta)$. One can immediately observe that this implies $\inf_\psi P_e(\psi) \not\to 0$ since $\frac{n}{\binom{n}{d}p}$ (which can be viewed as the rate of a code) cannot exceed the Shannon capacity of the channel $1 - H(\theta)$.

We now prove that $\binom{n}{d}p \leq (1-\epsilon)\frac{1}{d}\frac{n\log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2}$ implies $\inf_\psi P_e(\psi) \not\to 0$ if $d = O(\log n)$. Further, we will focus on the case of $\binom{n}{d}p \asymp \frac{n\log n}{d}$ since this is the regime where the largest amount of information is available. Again, it is enough to show that $\Pr(S) \to 0$, where S is defined as (37). By defining $\mathcal{R}_{\text{big}}, \mathcal{R}_{\text{res}}, \Delta$ and $b_1, \cdots, b_{r/2}$ as before, we again obtain (39):

$$\Pr(S) \leq \Pr\left(\mathsf{d}_\mathsf{H}(\mathbf{e}_{b_1}) > \mathsf{d}_\mathsf{H}(\mathbf{0}) \;\middle|\; \Delta\right)^{r/2}. \tag{45}$$

We finish the proof by showing the following for the considered case:

$$\Pr\left(\mathsf{d}_\mathsf{H}(\mathbf{e}_{b_1}) \leq \mathsf{d}_\mathsf{H}(\mathbf{0}) \;\middle|\; \Delta\right) \geq (1-o(1))e^{-2p'\binom{n-1}{d-1}}.$$

While following the proof of (23), the key technical difficulty arises when checking $q = \Omega(\log n)$ (see (30)): a simple calculation yields $|\mathcal{F}_{b_1}| = \binom{n-1}{d-1}$ and $|\mathcal{G}_{b_1}| = \binom{n-|\mathcal{R}_{\text{big}}|}{d-1}$, but here it is not clear whether $\binom{n-|\mathcal{R}_{\text{big}}|}{d-1} \asymp \binom{n-1}{d-1}$ when d is not a constant. We resolve this using a careful estimation as follows. As $|\mathcal{R}_{\text{big}}| = \Theta(\frac{n}{\log^7 n})$ and $d = O(\log n)$, it is straightforward to verify

$$1 - \frac{1}{\log^2 n} \leq \frac{n - |\mathcal{R}_{\text{big}}| - j}{n-1-j}$$

for $0 \leq j \leq d-2$. This simple yet crucial inequality concludes:

$$\begin{aligned}\frac{\binom{n-|\mathcal{R}_{\text{big}}|}{d-1}}{\binom{n-1}{d-1}} &= \prod_{j=0}^{d-2} \frac{n-|\mathcal{R}_{\text{big}}|-j}{n-1-j} \\ &\geq \left(1 - \frac{1}{\log^2 n}\right)^{d-1} \approx \exp\left\{-\frac{d-1}{\log^2 n}\right\} \to 1.\end{aligned}$$

C. Proof of Lemma 3

Without loss of generality, we prove the lemma assuming that $k \geq d$. The proof for the other cases is similar.

We wish to obtain lower bounds on

$$N_k = \sum_{\substack{1\leq i\leq d\\ i \text{ is odd}}} \binom{k}{i}\binom{n-k}{d-i} = \underbrace{\binom{k}{1}\binom{n-k}{d-1}}_{\text{boundary odd term}} + \underbrace{\sum_{i=1,3,\cdots,d-3,d-1} \binom{k}{i}\binom{n-k}{d-i}}_{\text{intermediate odd terms}} + \underbrace{\binom{k}{d-1}\binom{n-k}{1}}_{\text{boundary odd term}} \tag{46}$$

in terms of $\binom{n}{d}$. First, observe that

$$\binom{n}{d} = \sum_{0\leq i\leq d} \binom{k}{i}\binom{n-k}{d-i} = \underbrace{\binom{k}{0}\binom{n-k}{d}}_{\text{boundary term}} + \underbrace{\sum_{i=1,2,\cdots,d-2,d-1} \binom{k}{i}\binom{n-k}{d-i}}_{\text{intermediate terms}} + \underbrace{\binom{k}{d}\binom{n-k}{0}}_{\text{boundary term}}. \tag{47}$$

Suppose we have the following bounds:

$$\underbrace{\binom{k}{0}\binom{n-k}{d}+\binom{k}{d}\binom{n-k}{0}}_{\text{sum of boundary terms}} \le A_1 \underbrace{\left[\binom{k}{1}\binom{n-k}{d-1}+\binom{k}{d-1}\binom{n-k}{1}\right]}_{\text{sum of boundary odd terms}}; \tag{48}$$

$$\underbrace{\sum_{i=1,2,\cdots,d-2,d-1}\binom{k}{i}\binom{n-k}{d-i}}_{\text{intermediate terms}} \le A_2 \cdot \underbrace{\sum_{i=1,3,\cdots,d-3,d-1}\binom{k}{i}\binom{n-k}{d-i}}_{\text{intermediate odd terms}} + A_3 N_k\,, \tag{49}$$

for some quantities $A_1, A_2, A_3 > 0$. Then, by summing up the two inequalities, one can obtain a lower bound on N_k:

$$\binom{n}{d} \le \left(\max(A_1, A_2) + A_3\right) N_k\,. \tag{50}$$

Thus, the proof is completed as long as one can find the quantities A_1, A_2 and A_3 that satisfy (48) and (49).

We begin with (49). The following lemma asserts that $A_2 = 2$ and $A_3 = 3$ satisfy (49).

Lemma 4. *For* $1 \le k \le n/2$*,*

$$\sum_{i=1,2,\cdots,d-2,d-1}\binom{k}{i}\binom{n-k}{d-i} \le 2 \cdot \sum_{i=1,3,\cdots,d-3,d-1}\binom{k}{i}\binom{n-k}{d-i} + 3N_k.$$

Proof: See Sec. A-C. ■

For (49), the following lemma characterizes A_1.

Lemma 5. *Let* $\beta := \left\lceil \frac{n-d+1}{2d+1} \right\rceil$*. For* $\beta \le k \le n/2$*,*

$$\binom{k}{0}\binom{n-k}{d}+\binom{k}{d}\binom{n-k}{0} \le 2\left[\binom{k}{1}\binom{n-k}{d-1}+\binom{k}{d-1}\binom{n-k}{1}\right]. \tag{51}$$

For $k < \beta$*,*

$$\binom{k}{0}\binom{n-k}{d}+\binom{k}{d}\binom{n-k}{0} \le \frac{\alpha}{k}\left[\binom{k}{1}\binom{n-k}{d-1}+\binom{k}{d-1}\binom{n-k}{1}\right] \tag{52}$$

and

$$\frac{\alpha}{k} \ge 2\,, \tag{53}$$

where $\alpha = \frac{n-d+1}{d}$*.*

Proof: See Sec. A-D. ■

That is, $A_1 = 2$ if $\beta \le k \le n/2$, and $A_1 = \frac{\alpha}{k}$ if $k < \beta$.

We now are ready to prove Lemma 3 with the help of Lemma 4, Lemma, 5 and (50). When $\beta \le k < n/2$,

$$\binom{n}{d} \le 5N_k.$$

When $k < \beta$,

$$\binom{n}{d} \le \left(\max\left(2, \frac{\alpha}{k}\right) + 3\right) N_k \le \frac{5\alpha}{2k} N_k,$$

where the last inequality holds since $\frac{\alpha}{k} \ge 2$. This completes the proof.

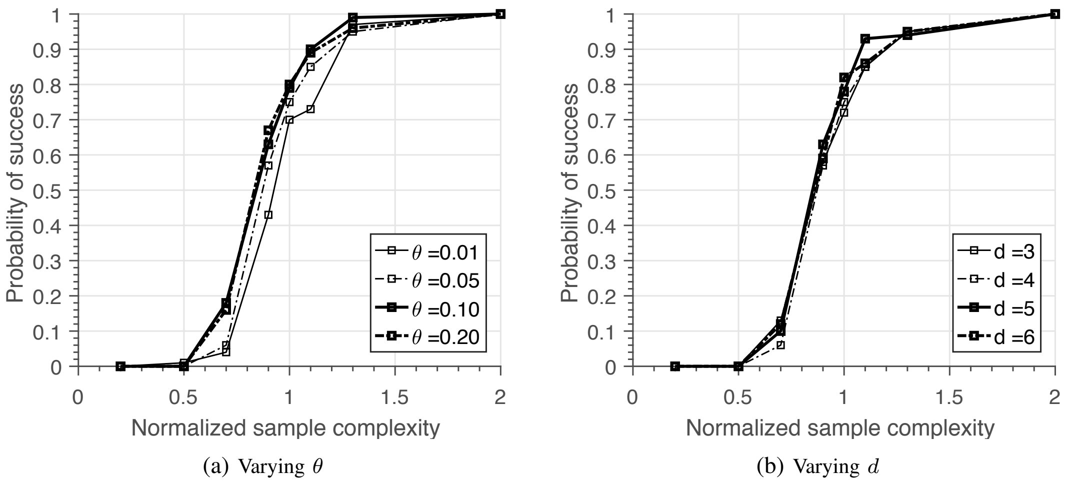
VII. Experimental results

A. The homogeneity measurement case

1) Efficient algorithms: We also develop a computationally-efficient algorithm that achieves the information-theoretic limit characterized in Theorem 1. Here we only present the algorithm while deferring a detailed analysis to our companion paper [56]. The algorithm operates in two stages, beginning with a decent initial estimate from Hypergraph Spectral Clustering [56] followed by iterative refinement. Detailed procedures are presented in Algorithm 1. Our algorithm is inspired by two-stage approaches that have been applied to a wide variety of problems including matrix completion [57], [58], phase retrieval [59], [60], robust PCA [61], community recovery [18], [20], [32], [62], [63], EM-algorithm [64], and rank aggregation [65].

Algorithm 1

```
For $E \in \binom{[n]}{d}$, define
    $W_E := \begin{cases} Y_E & \text{if } E \in \mathcal{E}; \\ 0, & \text{otherwise.} \end{cases}$
Apply Spectral Hypergraph Clustering [56] to a weighted hypergraph $([n], \{W_E\}_{E \in \binom{[n]}{d}})$ to obtain $\mathbf{X}^{(0)} = \{X_i^{(0)}\}_{1 \le i \le n} \in \{0,1\}^n$.
For $t = 0, 1, \cdots, T-1$ ($T = c \log n$ for some constant $c > 0$), update $\mathbf{X}^{(t)} = \{X_i^{(t)}\}_{1 \le i \le n}$ as per
    $X_i^{(t+1)} = \begin{cases} X_i^{(t)} & \text{if } \mathsf{d_H}(\mathbf{X}^{(t)}) < \mathsf{d_H}(\mathbf{X}^{(t)} \oplus \mathbf{e}_i); \\ X_i^{(t)} \oplus 1 & \text{if } \mathsf{d_H}(\mathbf{X}^{(t)}) \ge \mathsf{d_H}(\mathbf{X}^{(t)} \oplus \mathbf{e}_i), \end{cases}$
    for $i = 1, 2, \cdots, n$, where $\mathsf{d_H}(\cdot)$ is defined in (6).
Output $\mathbf{X}^{(T)} = \{X_i^{(T)}\}_{1 \le i \le n}$.
```



(a) Varying θ

(b) Varying d

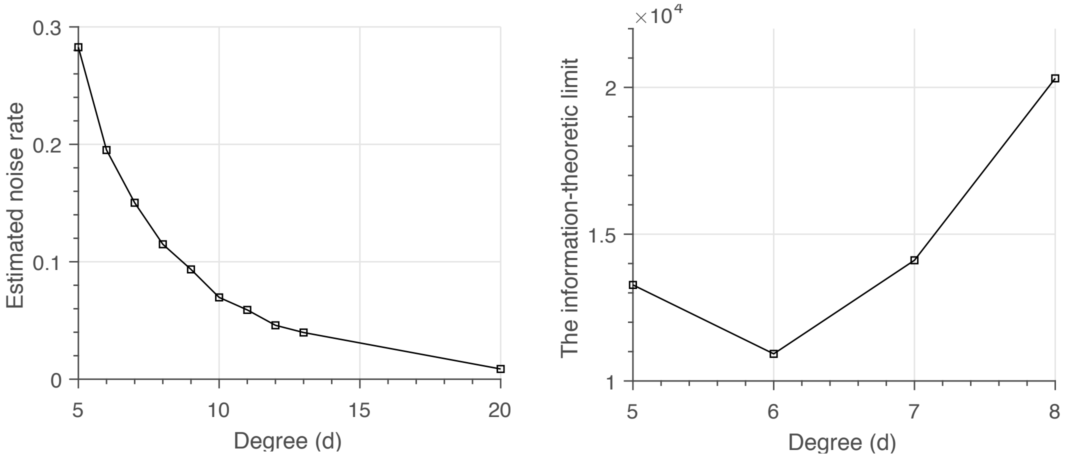
Fig. 3: **Algorithm 1 achieves the optimal sample complexity.** We run Monte Carlo simulations to estimate the probability of success when: (a) $n = 1000$, $d = 4$, and for various choices of θ; (b) $n = 1000$, $\theta = 0.05$, and for various choices of d. For each curve, we normalize the number of samples by the respective information theoretic limits, characterized in Theorem 1. Observe that the probability of success quickly approaches 1 as the normalized sample complexity crosses 1.

2) Performance of Algorithm 1: We demonstrate the performance of Algorithm 1 by running Monte Carlo simulations. Each point plotted in Fig. 3a and Fig. 3b indicates an empirical success rate. We take 100 Monte Carlo trials. Fig. 3a shows the probability of success when $n = 1000$, $d = 4$, and for various choices of θ. Shown in Fig. 3b is the performance of our algorithm with $n = 1000$, $\theta = 0.05$, and for various choices of d. For both figures, the x-axis denotes the number of samples normalized by the respective information-theoretic limits, characterized in Theorem 1. One can observe that the success probability due to Algorithm 1 quickly approaches 1 as the normalized sample complexity crosses 1, which corroborates our theoretical findings.

3) Optimal d for subspace clustering: We observe how the fundamental limit varies as a function of d. As we briefly discussed in Sec. III, if the noise rate θ is irrelevant to d, the optimal choice of d would be the minimum possible value of d. However, if the noise quality θ depends on d, there may be a sweet spot for d.

We demonstrate the existence of a sweet spot in one of subspace clustering applications: motion segmentation. We use the benchmark Hopkins 155 [66] dataset to compute an empirical noise rate θ as a function d as follows. For each sampled hyperedge $E = \{i_1, \cdots, i_d\}$, we adopt the method proposed in [37] to evaluate similarity between the corresponding d data points that we denote by D. Then, we set $Y_E = 1$ if and only if D is less than a fixed threshold, which is appropriately chosen so that $\Pr(Y_E = 0 \mid i_1, i_2, \cdots, i_d \text{ are from the same line}) \approx \Pr(Y_E = 1 \mid i_1, i_2, \cdots, i_d \text{ are not from the same line})$. We estimate the effective noise rate $\hat{\theta} := \Pr(Y_E = 0 \mid i_1, i_2, \cdots, i_d \text{ are from the same line})$ for various d, and observe that $\hat{\theta}$ quickly decreases as d increases; see Fig. 4a. We then plug these $\hat{\theta}$'s to the limit characterized in Theorem 1; see Fig. 4b. Note that $d = 5$ is not the optimal choice, but $d = 6$ is the sweet spot.

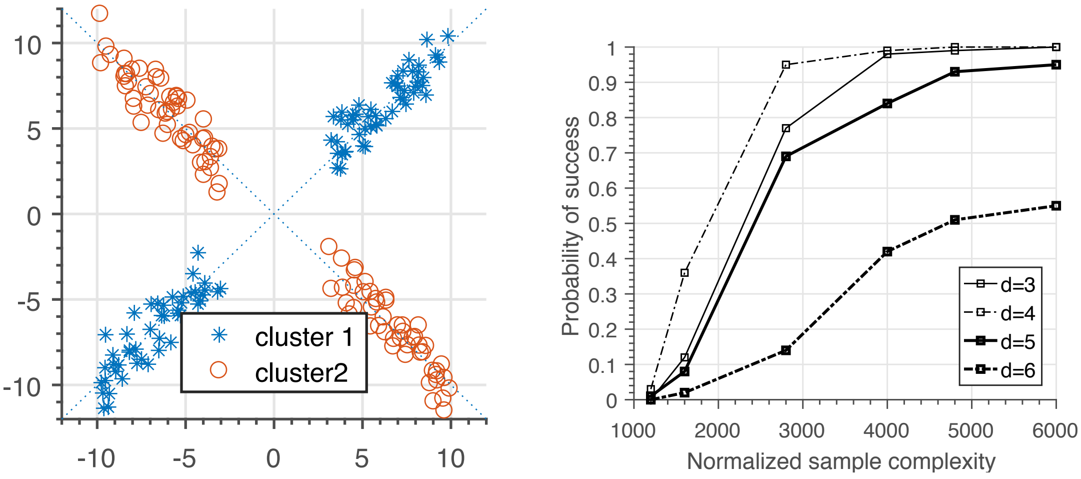
We also corroborate the existence of a sweet spot in a synthetic data set for subspace clustering, shown in Fig. 5a. Here the goal is to cluster n ($= 200$) 2-dimensional data points approximately lying on a union of two lines (1-dimensional subspaces).



(a) Estimated empirical noise rate $\hat{\theta}$

(b) d^*

Fig. 4: **Existence of d^* in motion segmentation.** (a) We estimate the empirical noise rate $\hat{\theta}$ as a function of d in motion segmentation. (b) We plug $\hat{\theta}$ to the limit characterized in Theorem 1 and verify that $d^* = 6$.



(a) Randomly generated data set

(b) Varying d

Fig. 5: **Optimal choice of d when θ decays with d.** We run Monte Carlo simulations to estimate the probability of success with the data set shown in (a). We observe that the effective noise rate decreases as d increases. For varying d from 3 to 6, the success probability of Algorithm 1 is shown in (b): the best performance of the algorithm is observed when $d = 4$.

We compute Y_E as above and evaluate the performance of Algorithm 1, shown in Fig. 5b. As a result, we observe that the optimal choice of d here is 4 rather than 3.

B. The parity measurement case

1) Efficient algorithms: For the parity measurement case, there are two efficient algorithms in the literature [38], [67]. In [38], it is shown that for $d = 3$, a variant of message passing algorithm successfully recovers the ground-truth vector provided that $\binom{n}{3}p = \Omega(n^2/\log n)$. Another efficient algorithm is based on a low-rank tensor factorization algorithm proposed in [67], and it is proved that reliable community recovery is feasible if $\binom{n}{3}p = \Omega(n^{1.5}\log^4 n)$. In either of the two cases, the sufficient condition comes with a polynomial term (n or $n^{1/2}$) to the fundamental limit characterized in Theorem 1. In fact, it is conjectured in [39] (see Conjecture 1 therein) that at least $n^{1.5}$ many samples are required for exact recovery.

On the other hand, focusing on the $\theta = 0$ case, recovering the ground-truth vector from the measurement vector $\mathbf{Y}$ is essentially the same as solving linear equations over the Galois field of two elements $\mathbb{F}_2$. Hence it immediately follows that efficient algorithms for solving linear equations such as Gaussian elimination can be employed in the noiseless case.

2) Information-theoretic limit: We first provide Monte Carlo simulation results which corroborate our theoretical findings in Theorem 2. Each point plotted in Fig. 6 and Fig. 7 is an empirical success rate. All results are obtained with 50 Monte Carlo trials. In Fig. 6, 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

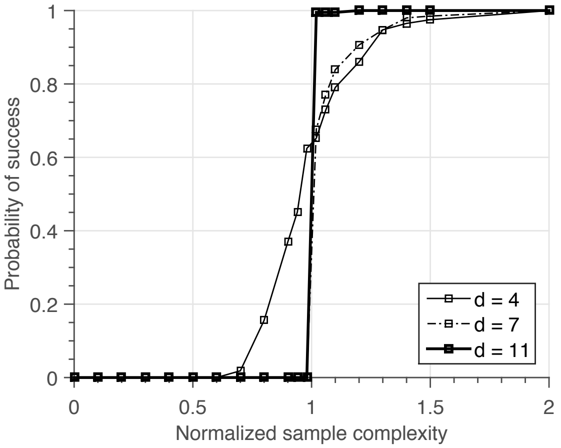


Fig. 6: We run the Monte Carlo simulations to estimate the probability of success for $n = 1000$, varying d, and $\theta = 0$. For each d, we normalize the number of samples by $\max(n, n \log n/d)$. Observe that the probability of success quickly approaches 1 as the normalized sample complexity crosses 1.

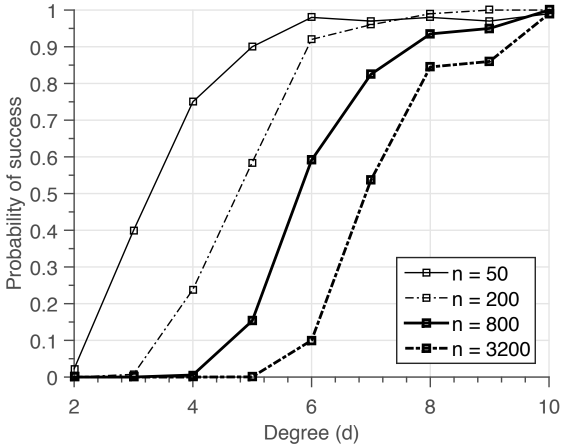


Fig. 7: We run the Monte Carlo simulations to estimate the probability of success for varying n, varying d, $\theta = 0$, and $p = 1.1n/\binom{n}{d}$. Note that when n increases by a multiplicative factor of 4, the curve shifts rightward about the same amount, supporting our result in Corollary 1

1 as the normalized sample complexity crosses 1.

3) Minimum d for linear sample complexity: Plotted in Fig. 7 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 1.

VIII. Conclusion

In this paper, we investigate the problem of community recovery in hypergraphs under the two generalized censored block models (GCBM), one based on the homogeneity measurement and the other based on the parity measurement. For these two models, we fully characterize the information-theoretic limits on sample complexity as a function of the number of nodes n, the size of edges d, the noise rate θ, and the edge observation probability p. We also corroborate our theoretical findings via experiments.

We conclude our paper by highlighting a few interesting open problems. One interesting question is whether or not one can sharpen Theorem 3 to characterize exact information-theoretic limits for the scaling d case. From the simulation results in Sec. VII-B, we make the following conjecture: Under the setting of Theorem 3, the information-theoretic limits is $\max\left\{\frac{n}{1-H(\theta)}, \frac{1}{d}\frac{n\log n}{(\sqrt{1-\theta}-\sqrt{\theta})^2}\right\}$. Another interesting open problem is about the computational gap for the parity measurement case: Investigating efficient algorithms for this case would shed some light on the study of information-computation gaps.

References

[1] K. Ahn, K. Lee, and C. Suh, "Community recovery in hypergraphs," *Allerton Conference on Communication, Control and Computing*, 2016.

[2] ——, "Information-theoretic limits of subspace clustering," in *IEEE International Symposium on Information Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 2473–2477.

[3] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

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

[5] 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.

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

[7] 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.

[8] 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.

[9] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social networks*, vol. 5, no. 2, pp. 109–137, 1983.

[10] E. Abbe and A. Montanari, "Conditional random fields, planted constraint satisfaction and entropy concentration," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*. Springer, 2013, pp. 332–346.

[11] G. Ghoshal, V. Zlatić, G. Caldarelli, and M. Newman, "Random hypergraphs and their applications," *Physical Review E*, vol. 79, no. 6, p. 066118, 2009.

[12] T. Michoel and B. Nachtergaele, "Alignment and integration of complex networks by hypergraph-based spectral clustering," *Physical Review E*, vol. 86, no. 5, p. 056111, 2012.

[13] S. Agarwal, K. Branson, and S. Belongie, "Higher order learning with graphs," in *ICML*. ACM, 2006, pp. 17–24.

[14] G. Karypis and V. Kumar, "Multilevel k-way hypergraph partitioning," *VLSI design*, vol. 11, no. 3, pp. 285–300, 2000.

[15] D. Gibson, J. Kleinberg, and P. Raghavan, "Clustering categorical data: An approach based on dynamical systems," *Databases*, vol. 1, p. 75, 1998.

[16] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer, "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.

[17] B. Hajek, Y. Wu, and J. Xu, "Achieving exact cluster recovery threshold via semidefinite programming: Extensions," *IEEE Transactions on Information Theory*, vol. 62, no. 10, pp. 5918–5937, Oct 2016.

[18] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 471–487, 2016.

[19] E. Mossel, J. Neeman, and A. Sly, "Consistency thresholds for binary symmetric block models," *arXiv preprint arXiv:1407.1591*, 2014.

[20] E. Abbe and C. Sandon, "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery," in *FOCS*. IEEE, 2015, pp. 670–688.

[21] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications," *Physical Review E*, vol. 84, no. 6, p. 066106, 2011.

[22] Y. Chen and J. Xu, "Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 882–938, 2016.

[23] J. Neeman and P. Netrapalli, "Non-reconstructability in the stochastic block model," *arXiv preprint arXiv:1404.6304*, 2014.

[24] A. Montanari, "Finding one community in a sparse graph," *Journal of Statistical Physics*, vol. 161, no. 2, pp. 273–299, 2015.

[25] J. Banks, C. Moore, J. Neeman, and P. Netrapalli, "Information-theoretic thresholds for community detection in sparse networks," in *Conference on Learning Theory (COLT)*, 2016, pp. 383–416.

[26] E. Abbe and C. Sandon, "Proof of the achievability conjectures in the general stochastic block model," *To Appear in Communications on Pure and Applied Mathematics*, 2017.

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

[28] ——, "A proof of the block model threshold conjecture," *arXiv preprint arXiv:1311.4115*, 2013.

[29] L. Massoulié, "Community detection thresholds and the weak ramanujan property," in *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. ACM, 2014, pp. 694–703.

[30] C. Bordenave, M. Lelarge, and L. Massoulie, "Non-backtracking spectrum of random graphs: Community detection and non-regular ramanujan graphs," in *FOCS*. IEEE, 2015, pp. 1347–1357.

[31] A. Y. Zhang, H. H. Zhou *et al.*, "Minimax rates of community detection in stochastic block models," *The Annals of Statistics*, vol. 44, no. 5, pp. 2252–2280, 2016.

[32] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou, "Achieving optimal misclassification proportion in stochastic block model," *Journal of Machine Learning Research*, 2017.

[33] E. Abbe, "Community detection and stochastic block models: Recent developments," *Journal of Machine Learning Research, Special Issue*, 2017.

[34] D. Ghoshdastidar and A. Dukkipati, "Uniform hypergraph partitioning: Provable tensor methods and sampling techniques," *Journal of Machine Learning Research*, vol. 18, no. 50, pp. 1–41, 2017. [Online]. Available: http://jmlr.org/papers/v18/16-100.html

[35] ——, "Consistency of spectral hypergraph partitioning under planted partition model," *The Annals of Statistics, 45(1), pp. 289-315*, 2017.

[36] V. M. Govindu, "A tensor decomposition for geometric grouping and segmentation," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2005, pp. 1150–1157.

[37] G. Chen and G. Lerman, "Spectral curvature clustering (scc)," *International Journal of Computer Vision*, vol. 81, no. 3, pp. 317–330, 2009.

[38] O. Watanabe, "Message passing algorithms for MLS-3LIN problem," *Algorithmica*, vol. 66, no. 4, pp. 848–868, 2013.

[39] L. Florescu and W. Perkins, "Spectral thresholds in the bipartite stochastic block model," *In Proceedings of the Conference on Learning Theory (COLT)*, pp. 943–959, 2016.

[40] M. Angelini, F. Caltagirone, F. Krzakala, and L. Zdeborova, "Spectral detection on sparse hypergraphs," in *Allerton Conference on Communication, Control, and Computing*, Sept 2015, pp. 66–73.

[41] C.-Y. Lin, C. I, and I.-H. Wang, "On the fundamental statistical limit of community detection in random hypergraphs," in *IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 2183–2187.

[42] N. Vesdapunt, K. Bellare, and N. Dalvi, "Crowdsourcing algorithms for entity resolution," *Proceedings of the VLDB Endowment*, vol. 7, no. 12, pp. 1071–1082, 2014.

[43] H. Ashtiani, S. Kushagra, and S. Ben-David, "Clustering with same-cluster queries," in *Advances in Neural Information Processing Systems*, 2016, pp. 3216–3224.

[44] R. Dorfman, "The detection of defective members of large populations," *The Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 436–440, 1943.

[45] I. H. Wang, S. L. Huang, and K. Y. Lee, "Extracting sparse data via histogram queries," in *Allerton Conference on Communication, Control, and Computing*, Sept 2016, pp. 39–45.

[46] I. H. Wang, S. L. Huang, K. Y. Lee, and K. C. Chen, "Data extraction via histogram and arithmetic mean queries: Fundamental limits and algorithms," in *IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 1386–1390.

[47] R. Vidal, R. Tron, and R. Hartley, "Multiframe motion segmentation with missing data using powerfactorization and GPCA," *International Journal of Computer Vision*, vol. 79, no. 1, pp. 85–105, 2008.

[48] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003.*, vol. 1. IEEE, 2003, pp. I–11.

[49] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, March 2011.

[50] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013.

[51] E. L. Dyer, A. C. Sankaranarayanan, and R. G. Baraniuk, "Greedy feature selection for subspace clustering." *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2487–2517, 2013.

[52] R. Heckel and H. Bölcskei, "Robust subspace clustering via thresholding," *IEEE Transactions on Information Theory*, vol. 61, no. 11, pp. 6320–6342, 2015.

[53] B. Hajek, Y. Wu, and J. Xu, "Information limits for recovering a hidden community," in *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 1894–1898.

[54] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.

[55] N. Alon and J. H. Spencer, *The probabilistic method*. John Wiley & Sons, 2004.

[56] K. Ahn, K. Lee, and C. Suh, "Hypergraph spectral clustering in the weighted stochastic block model," 2017. [Online]. Available: https://sites.google.com/site/kw1jjang/HSC_WSBM.pdf

[57] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.

[58] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM, 2013, pp. 665–674.

[59] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Advances in Neural Information Processing Systems*, 2013, pp. 2796–2804.

[60] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[61] X. Yi, D. Park, Y. Chen, and C. Caramanis, "Fast algorithms for robust pca via gradient descent," in *Advances in Neural Information Processing Systems*, 2016, pp. 4152–4160.

[62] Y. Chen, G. Kamath, C. Suh, and D. Tse, "Community recovery in graphs with locality," in *ICML*, 2016.

[63] 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 *Proceedings of the Conference on Learning Theory (COLT)*, 2015, pp. 391–423.

[64] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the em algorithm: From population to sample-based analysis," *Ann. Statist.*, vol. 45, no. 1, pp. 77–120, 02 2017. [Online]. Available: http://dx.doi.org/10.1214/16-AOS1435

[65] Y. Chen and C. Suh, "Spectral MLE: Top-k rank aggregation from pairwise comparisons," in *ICML*, 2015, pp. 371–380.

[66] R. Tron and R. Vidal, "A benchmark for the comparison of 3-d motion segmentation algorithms," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[67] P. Jain and S. Oh, "Provable tensor factorization with missing data," in *Advances in Neural Information Processing Systems*, 2014, pp. 1431–1439.

Appendix A
Proofs of Lemmas

A. Proof of Lemma 1

Recall that

$$|\mathcal{F}_{i,j}| = \sum_{\ell=1}^{d-1}\binom{i}{\ell}\binom{k-i}{d-\ell} + \sum_{\ell=1}^{d-1}\binom{i}{\ell}\binom{n-k-j}{d-\ell} + \sum_{\ell=1}^{d-1}\binom{j}{\ell}\binom{n-k-j}{d-\ell} + \sum_{\ell=1}^{d-1}\binom{k-i}{\ell}\binom{j}{d-\ell}. \tag{54}$$

In order to prove the lemma, it is sufficient to prove the following inequalities:

$$Z := \sum_{\ell=1}^{d-1}\binom{i}{\ell}\binom{k-i}{d-\ell} + \sum_{\ell=1}^{d-1}\binom{i}{\ell}\binom{n-k-j}{d-\ell} \geq i \cdot \frac{(1-2\delta)^{d-1}}{2^{d-2}}\binom{n-1}{d-1} \tag{55}$$

and

$$\sum_{\ell=1}^{d-1}\binom{j}{\ell}\binom{n-k-j}{d-\ell} + \sum_{\ell=1}^{d-1}\binom{k-i}{\ell}\binom{j}{d-\ell} \geq j \cdot \frac{(1-2\delta)^{d-1}}{2^{d-2}}\binom{n-1}{d-1}. \tag{56}$$

Here, we will focus on proving (55). We remark that the proof of (56) is essentially identical.

Since $i < \delta n$ and $j < \delta n$,

$$Z \geq \sum_{\ell=1}^{d-1}\binom{i}{\ell}\binom{k-\delta n}{d-\ell} + \sum_{\ell=1}^{d-1}\binom{i}{\ell}\binom{n-k-\delta n}{d-\ell}. \tag{57}$$

We further bound Z by considering two cases separately: $k \geq \delta n$ and $k < \delta n$. When $k \geq \delta n$,

$$\begin{aligned}\sum_{\ell=1}^{d-1}\binom{i}{\ell}\binom{k-\delta n}{d-\ell} + \sum_{\ell=1}^{d-1}\binom{i}{\ell}\binom{n-k-\delta n}{d-\ell} &\geq i\binom{k-\delta n}{d-1} + i\binom{n-k-\delta n}{d-1} \\ &\approx i \cdot \left[\left(\frac{k}{n}-\delta\right)^{d-1} + \left(1-\frac{k}{n}-\delta\right)^{d-1}\right]\binom{n-1}{d-1}, \end{aligned} \tag{58}$$

where the last inequality holds since $\binom{an}{b} \approx a^b\binom{n}{b} \approx a^b\binom{n-1}{b}$ for constants a and b. We then apply Hölder's inequality: Given p, q such that $1/p + 1/q = 1$, we have $\sum_z |x_z y_z| \le \left(\sum_z |x_z|^p\right)^{1/p} \left(\sum_z |y_z|^q\right)^{1/q}$ for all sequences $\{x_z\}$ and $\{y_z\}$. By setting $(x_1, x_2) = (\alpha, \beta), (y_1, y_2) = (1, 1), p = d-1, q = \frac{d-1}{d-2}$, we have

$$\alpha + \beta \le (\alpha^{d-1} + \beta^{d-1})^{\frac{1}{d-1}} 2^{\frac{d-2}{d-1}}. \tag{59}$$

Applying (59) to (58), we have

$$i \cdot \left[\left(\frac{k}{n} - \delta\right)^{d-1} + \left(1 - \frac{k}{n} - \delta\right)^{d-1} \right] \binom{n-1}{d-1} \ge i \cdot \frac{(1-2\delta)^{d-1}}{2^{d-2}} \binom{n-1}{d-1}. \tag{60}$$

When $k < \delta n$, $\sum_{\ell=1}^{d-1} \binom{i}{\ell}\binom{n-k-j}{d-\ell}$ becomes the dominant term. Hence,

$$\sum_{\ell=1}^{d-1} \binom{j}{\ell}\binom{n-k-\delta n}{d-\ell} + \sum_{\ell=1}^{d-1} \binom{k-i}{\ell}\binom{\delta n}{d-\ell} > i\binom{n-k-\delta n}{d-1} > i\binom{n-2\delta n}{d-1} \tag{61}$$

$$\approx i \cdot (1-2\delta)^{d-1}\binom{n-1}{d-1} > i \cdot \frac{(1-2\delta)^{d-1}}{2^{d-2}}\binom{n-1}{d-1}. \tag{62}$$

This completes the proof.

B. Proof of Lemma 2

Denote by $\mathcal{R}_{\text{big}} \subset [n]$ the set of nodes of size $n/\log^7 n$. One can easily show that with high probability, some nodes of this set are connected by the same hyperedge(s). Denote by $\mathcal{R}_{\text{res}}$ the largest subset of $\mathcal{R}_{\text{big}}$, whose elements do not share the same hyperedges. The lemma states that with high probability, $|\mathcal{R}_{\text{res}}| \simeq |\mathcal{R}_{\text{big}}|$.

We now formally prove this statement. Note that for a hyperedge $E = (i_1, i_2, \cdots, i_d)$, $|E \cap \mathcal{R}_{\text{big}}|$ is the number of nodes in $\mathcal{R}_{\text{big}}$ that are connected by the hyperedge. Hence, if $2 \le |E \cap R_{\text{big}}| \le d$, this hyperedge connects more than one nodes in $\mathcal{R}_{\text{big}}$, and $E \cap \mathcal{R}_{\text{big}}$ is the set of the nodes that share the same hyperedge E.

Let us denote by $\mathcal{R}_{\text{share}}$ the subset of nodes that are connected by the same hyperedge(s). Then,

$$\mathcal{R}_{\text{share}} := \bigcup_{k=2}^{d} \mathcal{R}_{\text{share}}^{(k)} := \bigcup_{k=2}^{d} \bigcup_{\substack{E \in \mathcal{E}: \\ |E \cap \mathcal{R}_{\text{big}}| = k}} E \cap \mathcal{R}_{\text{big}}. \tag{63}$$

Our proof strategy is as follows. Since

$$\mathcal{R}_{\text{res}} = \mathcal{R}_{\text{big}} - \mathcal{R}_{\text{share}} = \mathcal{R}_{\text{big}} - \bigcup_{k=2}^{d} \mathcal{R}_{\text{share}}^{(k)}, \tag{64}$$

it is sufficient to show that

$$\left| \bigcup_{k=2}^{d} \mathcal{R}_{\text{share}}^{(k)} \right| = o(|\mathcal{R}_{\text{big}}|). \tag{65}$$

More specifically, we will show

$$\Pr\left(\exists k \in \{2, 3, \ldots, d\} \text{ s.t. } |\mathcal{R}_{\text{share}}^{(k)}| > \frac{n}{\log^9 n} \right) \to 0. \tag{66}$$

That is, with probability approaching 1, $|\mathcal{R}_{\text{share}}^{(k)}| = o(n/\log^8 n)$ for all k, $2 \le k \le d$. Note that this implies (65) since

$$\left| \bigcup_{k=2}^{d} \mathcal{R}_{\text{share}}^{(k)} \right| \le \sum_{k=2}^{d} |\mathcal{R}_{\text{share}}^{(k)}| = O(d) \times o\left(\frac{n}{\log^8 n}\right) = o\left(\frac{n}{\log^7 n}\right) = o(|\mathcal{R}_{\text{big}}|). \tag{67}$$

In order to bound (66), we first derive an upper bound on the expected value of $|\mathcal{R}_{\text{share}}^{(k)}|$. By definition,

$$|\mathcal{R}_{\text{share}}^{(k)}| \le \sum_{\substack{E \in \mathcal{E}: \\ |E \cap \mathcal{R}_{\text{big}}| = k}} |E \cap \mathcal{R}_{\text{big}}| \le \sum_{\substack{E \in \mathcal{E}: \\ |E \cap \mathcal{R}_{\text{big}}| = k}} k = |\{E \in \mathcal{E} \ : \ |E \cap \mathcal{R}_{\text{big}}| = k\}| \cdot k. \tag{68}$$

Observe that $|\{E \in \mathcal{E} \ : \ |E \cap \mathcal{R}_{\text{big}}|\}|$ is the sum of $\binom{|\mathcal{R}_{\text{big}}|}{k}\binom{n-|\mathcal{R}_{\text{big}}|}{d-k}$ i.i.d. Bernoulli random variables with probability p. Hence,

$$\mathbb{E}\left\{|\{E \in \mathcal{E} \ : \ |E \cap \mathcal{R}_{\text{big}}| = k\}| \cdot k\right\} = k\binom{|\mathcal{R}_{\text{big}}|}{k}\binom{n-|\mathcal{R}_{\text{big}}|}{d-k}p \tag{69}$$

$$= |\mathcal{R}_{\text{big}}|\binom{|\mathcal{R}_{\text{big}}|-1}{k-1}\binom{n-|\mathcal{R}_{\text{big}}|}{d-k}p\,. \tag{70}$$

As $|\mathcal{R}_{\text{big}}| = o(n)$, we have $\binom{|\mathcal{R}_{\text{big}}|-1}{k-1}\binom{n-|\mathcal{R}_{\text{big}}|}{d-k} \leq \binom{|\mathcal{R}_{\text{big}}|-1}{1}\binom{n-|\mathcal{R}_{\text{big}}|}{d-2}$, which in turn gives

$$\begin{aligned}
(70) &\leq |\mathcal{R}_{\text{big}}|\binom{|\mathcal{R}_{\text{big}}|-1}{1}\binom{n-|\mathcal{R}_{\text{big}}|}{d-2}p \\
&\leq 2|\mathcal{R}_{\text{big}}|^2\binom{n-2}{d-2}p \\
&= 2|\mathcal{R}_{\text{big}}|^2\binom{n-2}{d-2}\frac{n\log n}{\binom{n}{d}} \\
&= O\left(\frac{|\mathcal{R}_{\text{big}}|^2 d^2 \log n}{n}\right) = O\left(\frac{n}{\log^{11} n}\right),
\end{aligned}$$

where the last equality holds since $\binom{n-2}{d-2} \approx \binom{n}{d}\frac{d^2}{n^2}$. Note that this inequality holds for any $2 \leq k \leq d$. Using Markov's inequality,

$$\Pr\left(|\{E \in \mathcal{E} \ : \ |E \cap \mathcal{R}_{\text{big}}| = k\}| \cdot k > \frac{n}{\log^9 n}\right) \leq \frac{\log^9 n}{n} \cdot O\left(\frac{n}{\log^{11} n}\right) = O\left(\frac{1}{\log^2 n}\right). \tag{71}$$

Applying the union bound over all $2 \leq k \leq d$,

$$\Pr\left(\exists k \in \{2, 3, \ldots, d\} \text{ s.t. } |\{E \in \mathcal{E} \ : \ |E \cap \mathcal{R}_{\text{big}}| = k\}| \cdot k > \frac{n}{\log^9 n}\right) \leq d \cdot O\left(\frac{1}{\log^2 n}\right) = O\left(\frac{1}{\log n}\right). \tag{72}$$

This completes the proof.

C. Proof of Lemma 4

Since $1 \leq i \leq P-1$,

$$\begin{aligned}
&\frac{\binom{k}{i+1}\binom{n-k}{d-(i+1)} + \binom{k}{i-1}\binom{n-k}{d-(i-1)}}{\binom{k}{i}\binom{n-k}{d-i}} \\
&= \frac{(k-i)(d-i)}{(i+1)(n-k-d+i+1)} + \frac{i(n-k-d+i)}{(k-i+1)(d-i+1)} \\
&\geq 2\sqrt{\frac{(k-i)(d-i)}{(i+1)(n-k-d+i+1)} \cdot \frac{i(n-k-d+i)}{(k-i+1)(d-i+1)}} \\
&= 2\sqrt{\frac{(k-i)}{(k-i+1)} \cdot \frac{(d-i)}{(d-i+1)} \cdot \frac{i}{i+1} \cdot \frac{(n-k-d+i)}{(n-k-d+i+1)}} \\
&\geq 2\sqrt{\left(\frac{1}{2}\right)^4} = \frac{1}{2}\,.
\end{aligned}$$

D. Proof of Lemma 5

1) $\beta \leq k \leq n/2$

Since $d \leq n/2$ and $\beta < n/2$, one can verify the inequality using the following facts:

$$\binom{k}{0}\binom{n-k}{d} \leq 2\binom{k}{1}\binom{n-k}{d-1} \Leftrightarrow k \geq \frac{n-d+1}{2d+1};$$
$$\binom{k}{d}\binom{n-k}{0} \leq 2\binom{k}{d-1}\binom{n-k}{1} \Leftrightarrow k \leq n - \frac{n-d+1}{2d+1}\,.$$

2) $k < \beta$

We first show that $\alpha/k \geq 2$. Since $k \leq \lceil \frac{n-d+1}{2d+1} \rceil - 1 \leq \frac{n-d+1}{2d+1}$, $\frac{\alpha}{k} = \binom{n-d+1}{d}/k > \binom{n-d+1}{d} \Big/ \left(\frac{n-d+1}{2d+1} \right) = \frac{2d+1}{d} \geq 2$. Next, the inequality can be checked using the following facts:

$$\binom{k}{d}\binom{n-k}{0} \leq 2\binom{k}{d-1}\binom{n-k}{1} \Leftrightarrow k \leq n - \frac{n-d+1}{2d+1}$$

$$\frac{\binom{k}{0}\binom{n-k}{d}}{\binom{k}{1}\binom{n-k}{d-1}} = \frac{n-k-d+1}{kd} \leq \frac{n-d+1}{kd} = \frac{\alpha}{k}\,.$$