Information Recovery From Pairwise Measurements

Yuxin Chen, Changho Suh, and Andrea J. Goldsmith

***Abstract*— This paper is concerned with jointly recovering n node variables $\{x_i\}_{1\le i\le n}$ from a collection of pairwise difference measurements. Imagine we acquire a few observations taking the form of $x_i - x_j$; the observation pattern is represented by a measurement graph $\mathcal{G}$ with an edge set $\mathcal{E}$, such that $x_i - x_j$ is observed if and only if $(i, j) \in \mathcal{E}$. To account for noisy measurements in a general manner, we model the data acquisition process by a set of channels with given input/output transition measures. Employing information-theoretic tools applied to channel decoding problems, we develop a *unified* framework to characterize the fundamental recovery criterion, which accommodates general graph structures, alphabet sizes, and channel transition measures. In particular, our results isolate a family of *minimum channel divergence measures* to characterize the degree of measurement corruption, which together with the size of the minimum cut of $\mathcal{G}$ dictates the feasibility of exact information recovery. For various homogeneous graphs, the recovery condition depends almost only on the edge sparsity of the measurement graph irrespective of other graphical metrics; alternatively, the minimum sample complexity required for these graphs scales like $(n \log n)/(\mathsf{Hel}_{1/2}^{\min})$ for certain information metric $\mathsf{Hel}_{1/2}^{\min}$ defined in the main text, as long as the alphabet size is not super-polynomial in n. We apply our general theory to three concrete applications, including the stochastic block model, the random corruption model, and the haplotype assembly problem. Our theory leads to orderwise tight recovery conditions for all these scenarios.**

***Index Terms*— Pairwise difference, information divergence, random graphs, geometric graphs, homogeneous graphs.**

I. Introduction

IN VARIOUS data processing scenarios, one wishes to acquire information about a large collection of objects, but it is infeasible or difficult to directly measure each individual object in isolation. Instead, only certain pairwise relations over a few object pairs can be measured. Partial examples of pairwise relations include cluster agreements, relative rotation and translation, pairwise matches, and paired sequencing reads, as will be discussed in details later. Taken collectively, these pairwise observations often carry a substantial amount of information across all objects of interest. As a consequence, reliable joint information recovery becomes feasible as soon as a sufficiently large number of pairwise measurements are obtained.

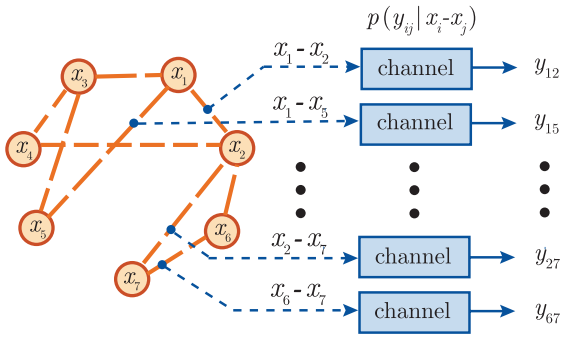

Fig. 1. Measurement graph and equivalent channel model. For each edge (i, j) in the measurement graph, $x_i - x_j$ is independently passed through a channel with output y_{ij} and transition probability $p(y_{ij} \mid x_i - x_j)$.

This paper explores a large family of pairwise measurements, which we term *pairwise difference measurements*. Consider n variables $x_1, \cdots, x_n$, and imagine we obtain independent measurements of the differences[1] $x_i - x_j$ over a few pairs (i, j). This pairwise difference functional is represented by a measurement graph $\mathcal{G}$ with an edge set $\mathcal{E}$ such that $x_i - x_j$ is observed if and only if $(i, j) \in \mathcal{E}$. To accommodate the noisy nature of data acquisition in a general manner, we model the observations $\{y_{ij}\}$ as the output of the following channel:

$$x_i - x_j \quad \xrightarrow{p(y_{ij} \mid x_i - x_j)} \quad y_{ij}, \quad \forall (i, j) \in \mathcal{E}, \tag{1}$$

as illustrated in Fig. 1. Here, the output distribution is specified solely by the associated channel input $x_i - x_j$, with $p(\cdot \mid \cdot)$ representing the channel transition probability. The goal is to recover $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ based on these channel outputs $\{y_{ij}\}$. Note that for any connected graph $\mathcal{G}$, the ground truth $\boldsymbol{x}$ is uniquely determined by the pairwise difference functional $\{x_i - x_j \mid (i, j) \in \mathcal{E}\}$, up to some global offset. Therefore, the problem can alternatively be posed as decoding the input of the channel (1) based on $\{y_{ij}\}$.

Problems of this kind have received considerable attention across various fields like social networks, computer science, and computational biology. A small sample of them are listed as follows.

[1]Here, "$-$" represents some algebraic subtraction operation (broadly defined), as we detail in Section II.

Manuscript received July 15, 2015; revised May 5, 2016; accepted July 28, 2016. Date of publication August 16, 2016; date of current version September 13, 2016. Y. Chen and A. J. Goldsmith were supported in part by the NSF Center for Science of Information and in part by AFOSR through MURI under Grant FA9550-12-1-0215. C. Suh was supported by the Korea Space Launch Vehicle (KSLVII) Program through the Ministry of Science, ICT and Future Planning, Korean Government. This paper was presented at the International Symposium on Information Theory [1], [2].

Y. Chen is with the Department of Statistics, Stanford University, Stanford, CA 94305 USA (e-mail: yxchen@stanford.edu).

C. Suh is with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, South Korea (e-mail: chsuh@kaist.ac.kr).

A. J. Goldsmith is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: andrea@wsl.stanford.edu).

Communicated by O. Simeone, Associate Editor for Communications.

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

Digital Object Identifier 10.1109/TIT.2016.2600566

- *Community detection and graph partitioning*. Various real-world networks exhibit community structures [3], and the nodes are grouped into a few clusters based on shared features. The aim is to uncover the hidden community structure by observing the similarities between members. For instance, in the simplest two-community model, the vertex-variables represent the community assignment, and the edge variables encode whether two vertices belong to the same community. This two-community recovery problem, sometimes referred to as graph partitioning (e.g. [4], [5]), is a special instance of the aforementioned pairwise difference model.
- *Alignment, registration and synchronization*. Consider n views of a single scene from different angles and positions[2] [6]. One is allowed to estimate the *relative* translation / rotation across several pairs of views. The problem aims at simultaneously aligning all views based on these noisy pairwise estimates. This arises in many applications including structure from motion in computer vision [7], [8], spectroscopy imaging and structural biology [9], [10], and multi-reference alignment [11].
- *Joint matching*. Given n images / shapes representing the same physical object, one wishes to identify common features across them. The input to a cutting-edge joint matching paradigm is typically a set of noisy pairwise matches computed between several pairs of images in isolation [8], [12]–[15], which falls under the category of pairwise difference measurements. The goal is to recover globally consistent maps across the features of all images, by refining these noisy pairwise inputs. This problem arises in numerous applications in computer vision and graphics, solving jigsaw puzzles, etc.
- *Genome assembly*. The genomes of two unrelated people mostly differ at specific nucleotide positions called single nucleotide polymorphisms (SNPs). A haplotype is a collection of associated SNPs on a chromatid, which is important in understanding genetic causes of various diseases and developing personalized medicine. Among various sequencing methods, haplotype assembly is particularly effective from paired sequencing reads [16]–[18], which amounts to reconstructing the haplotype based on disagreement between pairs of single reads [19]–[21] — a special instance of pairwise difference measurement with binary alphabet.

Many of these practical applications have witnessed a flurry of recent activity in algorithm development, which are primarily motivated by computational considerations. For instance, inspired by recent success in spectral methods [22], [23] and convex relaxation [24]–[26] (particularly those developed for low-rank matrix recovery problems), many provably efficient algorithms have been proposed for graph clustering [4], joint matching [12], [14], synchronization [10], and so on. While these algorithms have been shown to enjoy intriguing recovery guarantees under simple randomized models, the choices of performance metrics have mainly been studied in a model-specific manner. On the fundamental-limit side, there have been several results in place for a few applications, e.g. stochastic block models [27], [28], synchronization [1], and haplotype assembly [19]–[21]. Despite their intrinsic connections, these results were developed primarily on a case-by-case basis instead of accounting for the most general observation models.

In the present paper, we emphasize the similarities and connections among all these motivating applications, by viewing them as a graph-based functional fed into a collection of general channels. We wish to explore the following questions from an information-theoretic perspective:

1) Are there any distance metrics of the channel transition measures and graphical properties that dictate the success of exact information recovery from pairwise difference measurements?
2) If so, can we characterize the interplay between these channel separation metrics and graphical constraints and provide insights into the feasibility of simultaneous recovery?

All in all, the aim of this work is to gain a *unified* understanding about the performance limits that underlie various applications falling in the realm of pairwise-measurement based recovery. In turn, these fundamental criteria will provide a general benchmark for algorithm evaluation and comparison.

A. Main Contributions

The main contribution of this paper is towards a unified characterization of the fundamental information recovery criterion, using both information-theoretic and graph-theoretic tools. In particular, we single out and emphasize a family of minimum channel separation measures (i.e. the minimum Kullback–Leibler (KL), Hellinger, and Rényi divergence), as well as two graphical metrics (i.e. the minimum cut size and the cut-homogeneity exponent defined in Section IV-A), that play central roles in determining the feasibility of exact recovery. Equipped with these metrics, we develop a sufficient and a necessary condition for information recovery, which apply to general graphs, any type of input alphabets, and general channel transition measures. Encouragingly, as long as the alphabet size is not super-polynomial in n, these two conditions coincide (modulo some explicit universal constant) for the broad class of homogeneous graphs, subsuming as special cases Erdős–Rényi models, homogeneous geometric graphs (e.g. generalized rings and grids), and many other expander graphs.

In a nutshell, the fundamental recovery criterion is specified by the product of the minimum channel divergence measures and the size of the minimum cut. Intuitively, this product characterizes the amount of information one has available to differentiate two minimally separated input hypotheses. Somewhat surprisingly, for a variety of homogeneous graphs, the recovery criterion relies only on the edge sparsity of the measurement graph. Equivalently, the minimum sample complexity required for exact recovery in these homogeneous

[2]In a variety of applications including structure from motion and cryo-EM, these views (e.g. photos of some architectures, or projected images of 3D molecules) are given to us without revealing their absolute camera poses / angles with respect to the 3D structure of interest.

graphs scales as

$$\text{minimum sample complexity} \asymp \frac{n \log n}{\mathsf{Hel}_{1/2}^{\min}}$$

for some information metric $\mathsf{Hel}_{1/2}^{\min}$ to be specified later, provided that the alphabet size is polynomial in n. This result holds irrespective of other second-order graphical metrics like the spectral gap.

The unified framework we develop is *non-asymptotic*, in the sense that it accommodates the most general settings without fixing either the alphabet size or channel transition probabilities. This allows full characterization of the high-dimensional regime where all parameters are allowed to scale (possibly with different rates) — a setting that has received increasing attention compared to the classical asymptotics where only n is tending to infinity.

Finally, to illustrate the effectiveness of our general theory, we develop concrete consequences for three canonical applications that have been investigated in prior literature, including the stochastic block model, the random corruption model, and the haplotype assembly problem. In each case, our theory recovers order-wise correct recovery guarantees, and even strengthens existing results in certain regimes.

B. Related Work

On the information-theoretic side, most prior works focused on *binary input and output alphabets*. Among them, Abbe *et al.* [29] characterized the orderwise information-theoretic limits under the Erdős–Rényi model, uncovering the intriguing observation that a decoding method based on convex relaxation achieves nearly-optimal recover guarantees under sparsely connected graphs. In addition, Si *et al.* [20] and Kamath *et al.* [21] determined the information-theoretic limits for a similar setup motivated from genome sequencing, which correspond to random graphs and (generalized) ring graphs, respectively. A sufficient recovery condition for general graphs has also been derived in [29], although it was not guaranteed to be order optimal. Our preliminary work [1] explored the fundamental recovery limits under general alphabets and graph structures, but was restricted to the simplistic random corruption model (or random corruption model) rather than general channel distributions. In contrast, the framework developed in the current work allows orderwise tight characterization of the recovery criterion for general alphabets and channel characteristics.

The pairwise measurement models considered in this paper and the aforementioned works [1], [20], [21], [29] can all be treated as a special type of "graphical channel" as coined by Abbe and Montanari [30], [31], which refers to a general family of channels whose transition probabilities factorize over a set of hyper-edges. This previous work on graphical channels centered on the metric of conditional entropy that quantifies the residual input uncertainty given the channel output, and uncovered the stability and concentration of this metric under random sparse graphs. In comparison, the present paper primarily aims to investigate how the channel transition measures affect the recovery limits in the absence of channel coding (or equivalently, the limits under a specific code that is often suboptimal), which was previously out of reach. Specifically, the information limit under optimal channel coding is determined by the mutual information metric; in contrast, the information limit without channel coding is often dictated by certain minimum divergence metrics, which could sometimes be much smaller than the mutual information. This arises because optimal encoding enables us to code against the channel variation by maximizing the output separation between distinct input hypotheses, while in the non-coding applications one has to deal with the minimally separated input hypotheses determined by the practical applications. In addition, we focus on full recovery in this work, but in some applications this might be too stringent. Recent interesting work [32], [33] explored the notion of *partial recovery* under binary alphabets, which highlighted the two-dimensional grids and supplied a two-step polynomial-time recovery algorithm. A more general theory regarding partial or approximate recovery is left for future work.

Finally, the input variables $\{x_i\}$ can be viewed as discrete signals on the graph $\mathcal{G}$. Recent years have seen much activity regarding discrete signal processing on graphs [34], [35]. For instance, it has been studied in [36] how to optimally subsample band-limited graphs signals, subject to a sampling rate constraint, while enabling perfect signal recovery. Our model differs from this line of work in that the samples we take are highly constrained—that is, we only allow pairwise difference samples taken over the edges—and hence the resulting sample complexity significantly exceeds the sampling rate limit.

C. Terminology and Notation

1) Graph Terminology: Let $\deg(v)$ represent the degree of a vertex v. For any two vertex sets $\mathcal{S}_1$ and $\mathcal{S}_2$, denote by $\mathcal{E}(\mathcal{S}_1, \mathcal{S}_2)$ (resp. $e(\mathcal{S}_1, \mathcal{S}_2)$) the set (resp. the number) of edges with exactly one endpoint in $\mathcal{S}_1$ and another in $\mathcal{S}_2$. A complete graph of n vertices, denoted by K_n, is a graph in which every pair of vertices is connected by an edge. Below we introduce several widely used (random) graph models; see [37], [38] and the references therein for in-depth discussion.

1) *Erdős–Rényi graph*. An Erdős–Rényi graph of n vertices, denoted by $\mathcal{G}_{n,p}$, is constructed in such a way that each pair of vertices is independently connected by an edge with probability p.
2) *Random geometric graph*. A random geometric graph, denoted by $\mathcal{G}_{n,r}$, is generated via a 2-step procedure: (i) place n vertices uniformly and independently on the surface of a unit sphere[3]; (ii) connect two vertices by an edge if the Euclidean distance between them is at most r.
3) *Expander graph*. A graph $\mathcal{G}$ is said to be an expander graph with edge expansion $h_{\mathcal{G}}$ if $e(\mathcal{S}, \mathcal{S}^{\mathrm{c}}) \geq h_{\mathcal{G}} |\mathcal{S}|$ for all vertex set $\mathcal{S}$ satisfying $|\mathcal{S}| \leq n/2$.

2) Divergence Measures: Our results are established upon a family of divergence measures. Formally, for any two probability measures P and Q, if P is absolutely continuous

[3]We consider $\mathcal{G}_{n,r,}$ on a unit sphere instead of $[0, 1]^2$ to eliminate edge effects.

with respect to Q, then the KL divergence of Q from P is defined as

$$\mathsf{KL}\left(P \,\|\, Q\right) := \int \mathrm{d}P \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right), \tag{2}$$

whereas the Hellinger divergence of order $\alpha \in (0,1)$ of Q from P is defined to be [39], [40]

$$\mathsf{Hel}_{\alpha}\left(P \,\|\, Q\right) := \frac{1}{1-\alpha}\left[1 - \int (\mathrm{d}P)^{\alpha} (\mathrm{d}Q)^{1-\alpha}\right]. \tag{3}$$

When $\alpha = 1/2$, this reduces to the so-called *squared Hellinger distance*[4]

$$\mathsf{Hel}_{\frac{1}{2}}\left(P \,\|\, Q\right) = 2 - 2\int \sqrt{\mathrm{d}P}\sqrt{\mathrm{d}Q} = \int \left(\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q}\right)^2. \tag{4}$$

The χ^2 divergence is defined as

$$\chi^2\left(P \,\|\, Q\right) = \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2 \mathrm{d}Q. \tag{5}$$

In particular, when $P = \mathsf{Bernoulli}\,(p)$ and $Q = \mathsf{Bernoulli}\,(q)$, we abuse the notation and let

$$\mathsf{KL}\left(p \,\|\, q\right) = \mathsf{KL}\left(P \,\|\, Q\right), \quad \mathsf{Hel}_{\alpha}\left(p \,\|\, q\right) = \mathsf{Hel}_{\alpha}\left(P \,\|\, Q\right)$$

and

$$\chi^2\left(p \,\|\, q\right) = \chi^2\left(P \,\|\, Q\right). \tag{6}$$

More generally, the f-divergence of Q from P is defined as

$$D_f\left(P \,\|\, Q\right) := \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q \tag{7}$$

for any convex function $f(\cdot)$ such that $f(1) = 0$ [39], [40]. Note that the Hellinger divergence of order α, the KL divergence, and the χ^2 divergence are special cases of f-divergence generated by $f(x) = \frac{1}{1-\alpha}(1 - x^{\alpha})$, $f(x) = x\log x$ (or $f(x) = x\log x - x + 1$), and $f(x) = (x-1)^2$, respectively. These divergence measures can often be efficiently estimated even under large alphabets; see, e.g., [42] and their subsequent work.

Finally, we introduce the Rényi divergence of positive order α, where $\alpha \neq 1$, of a distribution P from another distribution Q as [43], [44]

$$D_{\alpha}\left(P \,\|\, Q\right) := -\frac{1}{1-\alpha}\log\left(\int (\mathrm{d}P)^{\alpha}(\mathrm{d}Q)^{1-\alpha}\right) \tag{8}$$

$$= -\frac{1}{1-\alpha}\log\left(1 - (1-\alpha)\,\mathsf{Hel}_{\alpha}\right). \tag{9}$$

It follows from the elementary inequality $1 - x \leq e^{-x}$ that $D_{\alpha}\left(P \,\|\, Q\right) \geq \mathsf{Hel}_{\alpha}\left(P \,\|\, Q\right)$. This together with the monotonicity of D_{α} [44, Th. 3] gives

$$\mathsf{Hel}_{\alpha}\left(P \,\|\, Q\right) \leq D_{\alpha}\left(P \,\|\, Q\right) \leq \mathsf{KL}\left(P \,\|\, Q\right), \quad 0 < \alpha < 1. \tag{10}$$

[4]Several other sources introduce a prefactor of $1/2$ in order to normalize the squared Hellinger distance, resulting in the definition $\int \frac{1}{2}(\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q})^2$. Here, we adopt the unnormalized version as given in [41, Sec. 2.4].

3) Other Notation: Let $\mathbf{1}$ and $\mathbf{0}$ be the all-one and all-zero vectors, respectively. We denote by $\mathrm{supp}\,(\boldsymbol{x})$ (resp. $\|\boldsymbol{x}\|_0$) the support (resp. the support size) of $\boldsymbol{x}$. The standard notation $f(n) = o\left(g(n)\right)$ means $\lim_{n\to\infty} f(n)/g(n) = 0$; $f(n) = \omega\left(g(n)\right)$ means $\lim_{n\to\infty} g(n)/f(n) = 0$; $f(n) = \Omega\left(g(n)\right)$ or $f(n) \gtrsim g(n)$ mean there exists a constant c such that $f(n) \geq cg(n)$; $f(n) = \mathcal{O}\left(g(n)\right)$ or $f(n) \lesssim g(n)$ mean there exists a constant c such that $f(n) \leq cg(n)$; $f(n) = \Theta\left(g(n)\right)$ or $f(n) \asymp g(n)$ mean there exist constants c_1 and c_2 such that $c_1 g(n) \leq f(n) \leq c_2 g(n)$. Throughout this paper, $\log(\cdot)$ represents the natural logarithm.

D. Organization

The remainder of the paper is organized as follows. In Section II, we describe the formal problem setup and introduce the key channel distance measures. We develop non-asymptotic sufficient and necessary recovery conditions for the special Erdős–Rényi model in Section III, along with some intuitive interpretation of the results. Section IV presents the recovery conditions in full generality, which accommodate general alphabets, graph structures, and channel characteristics, with particular emphasis on the family of homogeneous graphs. To illustrate the effectiveness of our framework, we apply our general theory to a few concrete examples in Section V. Section VI concludes the paper with a summary of our findings and a discussion of future directions. The proofs of the main results and auxiliary lemmas are deferred to the appendices.

II. Problem Formulation and Key Metrics

A. Models

Imagine a collection of n vertices $\mathcal{V} = \{1, \cdots, n\}$, each represented by a vertex-variable x_i over the input alphabet $\mathcal{X} := \{0, 1, \cdots, M-1\}$, where M represents the alphabet size.

- **Object representation and pairwise difference.** Consider an additive group formed over $\mathcal{X}$ together with an associative addition operation "+" (broadly defined). For any $x_i, x_j \in \mathcal{X}$, the pairwise difference operation is defined as

$$x_i - x_j := x_i + (-x_j), \tag{11}$$

where $-x$ stands for the unique additive inverse of x. We assume throughout that "+" satisfies the following bijective property:

$$\forall x_i \in \mathcal{X}: \quad \begin{cases} x_i + x_j \neq x_i + x_l, & \forall x_l \neq x_j; \\ x_i + x_j \neq x_l + x_j, & \forall x_l \neq x_i. \end{cases} \tag{12}$$

A partial list of examples includes:

1) *Modular arithmetic*: if we define "+" to be the modular addition over integers $\{0, 1, \cdots, M-1\}$, then $x_i - x_j$ $(\mathsf{mod}\ M)$ is a valid example of (11).
2) *Relative rotation*: set $x_i = \boldsymbol{R}_i$ for some rotation matrix $\boldsymbol{R}_i$ and let "+" denote matrix multiplication. Then $x_i - x_j$ stands for $\boldsymbol{R}_i \boldsymbol{R}_j^{-1}$, which represents the relative

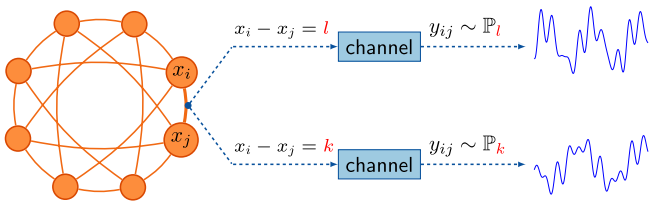

Fig. 2. The probability measure $\mathbb{P}_l(\cdot)$ is defined to be the distribution of y_{ij} given $x_i - x_j = l$.

rotation between i and j, and hence is a special case of (11).

3) *Pairwise map*: if we set x_i to be some permutation matrix $\mathbf{\Pi}_i$ and let "+" be matrix multiplication, then the pairwise map between two isomorphic sets—captured by $\mathbf{\Pi}_i \mathbf{\Pi}_j^\top$—also belongs to the pairwise difference model.

- **Measurement graph and channel model**. The measurement pattern is represented by a *measurement graph* $\mathcal{G}$ that comprises an undirected edge set $\mathcal{E}$, so that $x_i - x_j$ is measured if and only if $(i, j) \in \mathcal{E}$. As illustrated in Fig. 1, for each $(i, j) \in \mathcal{E}$ $(i > j)$, the pairwise difference $x_i - x_j$ is independently passed through a channel, whose output y_{ij} follows the conditional distribution

$$p\left(y_{ij} \middle| x_i - x_j = l\right) = \mathbb{P}_l\left(y_{ij}\right), \quad 0 \le l < M. \tag{13}$$

Here, $\mathbb{P}_l(\cdot)$ denotes the transition measure that maps a given input l to the output alphabet $\mathcal{Y}$; see Fig. 2 for an illustration. With a slight abuse of notation, we let $\mathbb{P}_i = \mathbb{P}_{i \bmod M}$ for any $i \notin \{0, 1, \cdots, M-1\}$. We assume throughout that the observations are symmetric[5] in the sense that there exists a one-to-one mapping between y_{ij} and y_{ji} for any $(i, j) \in \mathcal{E}$; that said, all information are contained in the upper triangular part $\{y_{ij}\}_{1 \le i < j \le n}$. The output alphabet $\mathcal{Y}$ can be either continuous or discrete, finite or infinite, which allows general modeling of distortion, corruption, etc. As opposed to conventional information theory settings, no coding is employed across channel uses.

This paper centers on exact information recovery, that is, to reconstruct all input variables $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ precisely, except for some *global offset*. This is all one can hope for since there is absolutely no basis to distinguish $\boldsymbol{x}$ from its shifted version $\boldsymbol{x} + l \cdot \mathbf{1} = \{x_1 + l, \cdots, x_n + l\}$ given only the output $\boldsymbol{y} := \{y_{ij} \mid (i, j) \in \mathcal{E}\}$. In light of this, we introduce the zero-one distance modulo a global offset factor as follows

$$\operatorname{dist}(\boldsymbol{w}, \boldsymbol{x}) := 1 - \max_{0 \le l < M} \mathbb{I}\{\boldsymbol{w} = \boldsymbol{x} + l \cdot \mathbf{1}\}, \tag{14}$$

where $\mathbb{I}$ is the indicator function. Apparently, $\operatorname{dist}(\boldsymbol{w}, \boldsymbol{x}) = 0$ holds for all $\boldsymbol{w}$ that differ from $\boldsymbol{x}$ only by a global offset.

[5]We assume the observation model is symmetric because this is the case in all motivating applications listed in Section I. We note, however, that all results and analyses immediately extend to the non-asymmetric case, provided that $\mathbb{P}_l(\cdot)$ is defined with respect to (y_{ij}, y_{ji}), that is, $\mathbb{P}_l(y_{ij}, y_{ji}) := p(y_{ij}, y_{ji} \big| x_i - x_j = l)$.

With this metric in place, we define, for any recovery procedure $\psi : \mathcal{Y}^{|\mathcal{E}|} \mapsto \mathcal{X}^n$, the *probability of error* as

$$P_{\mathsf{e}}(\psi) := \max_{\boldsymbol{x} \in \mathcal{X}^n} \mathbb{P}\left\{\operatorname{dist}(\psi(\boldsymbol{y}), \boldsymbol{x}) \neq 0 \;\middle|\; \boldsymbol{x}\right\}. \tag{15}$$

The aim is to characterize the regime where the minimax probability of error $\inf_\psi P_{\mathsf{e}}(\psi)$ is vanishing.

B. Key Separation Metrics on Channel Transition Measures

Before proceeding to the main results, we introduce a few channel separation measures that capture the resolutions of the measurements, which will be critical in subsequent development of our theory. Specifically, we isolate the minimum KL, Hellinger, and Rényi divergence with respect to the channel transition measures as follows[6]

$$\mathsf{KL}^{\min} := \min_{l \neq k} \mathsf{KL}(\mathbb{P}_l \,\|\, \mathbb{P}_k); \tag{16}$$

$$\mathsf{Hel}_\alpha^{\min} := \min_{l \neq k} \mathsf{Hel}_\alpha(\mathbb{P}_l \,\|\, \mathbb{P}_k); \tag{17}$$

$$D_\alpha^{\min} := \min_{l \neq k} D_\alpha(\mathbb{P}_l \,\|\, \mathbb{P}_k) \tag{18}$$

$$= -\frac{1}{1-\alpha} \log\left(1 - (1-\alpha)\,\mathsf{Hel}_\alpha^{\min}\right). \tag{19}$$

These minimum divergence measures essentially reflect the distinguishability of channel outputs given minimally separated inputs.[7] As will be seen later, the minimum Hellinger and Rényi divergence are crucial in developing sufficient recovery conditions, while the minimum KL divergence plays an important role in deriving minimax lower bounds. It is well known (see [39]–[41], [46] for various inequalities connecting them) that these measures are almost equivalent (modulo some small constant) when any two probability measures under study are close to each other—a regime where two measures are the hardest to differentiate. In particular, we underscore one fact that links the KL divergence and the squared Hellinger distance, which we shall use several times in the rest of the paper; see [47, Proposition 2] for an alternative version.

Fact 1: Suppose that P and Q are two probability measures such that

$$\frac{\mathrm{d}P}{\mathrm{d}Q} \le R \quad \text{and} \quad \frac{\mathrm{d}Q}{\mathrm{d}P} \le R$$

hold uniformly over the probability space. Then one has

$$\begin{aligned} \mathsf{KL}(P \,\|\, Q) &\ge \max\{2 - 0.5 \log R,\ 1\} \cdot \mathsf{Hel}_{\frac{1}{2}}(P \,\|\, Q); \\ \mathsf{KL}(P \,\|\, Q) &\le (2 + \log R) \cdot \mathsf{Hel}_{\frac{1}{2}}(P \,\|\, Q). \end{aligned} \tag{20}$$

Furthermore, if $R \le 4.5$, then one has

$$\begin{aligned} \mathsf{KL}(P \,\|\, Q) &\ge (2 - 0.4 \log R) \cdot \mathsf{Hel}_{\frac{1}{2}}(P \,\|\, Q); \\ \mathsf{KL}(P \,\|\, Q) &\le (2 + 0.4 \log R) \cdot \mathsf{Hel}_{\frac{1}{2}}(P \,\|\, Q). \end{aligned} \tag{21}$$

[6]Here and throughout, we assume that $\mathbb{P}_l$ is absolutely continuous with respect to $\mathbb{P}_k$ for any l and k.

[7]One natural question arises as to how to estimate such divergence metrics from measured data, which has become an active research topic. When both the input and output alphabet sizes are small, one can first estimate the entire probability measures via low-rank matrix recovery schemes, and then plug them in to calculate the divergence metrics. When the alphabet size is large or when the output is continuous-valued, one might resort to more careful functional estimation algorithms (e.g. [42], [45]).

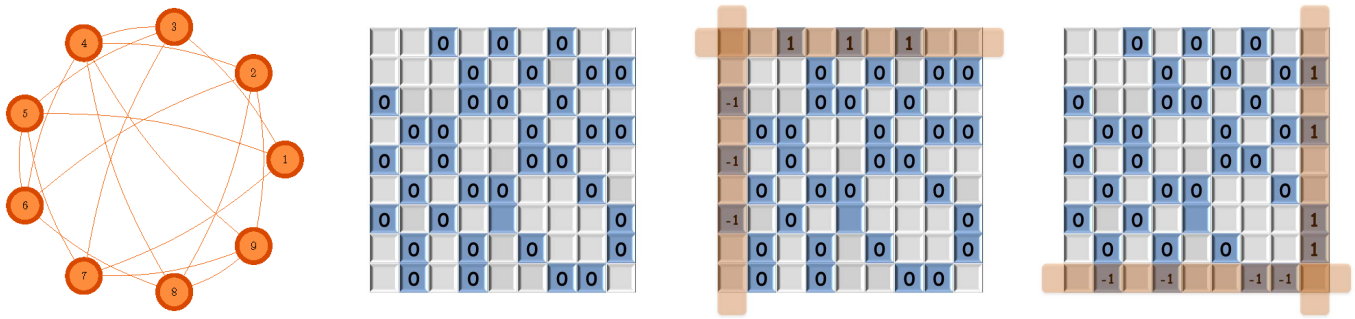

Fig. 3. The pairwise inputs $[x_i - x_j]_{1\le i,j\le n}$ under a realization of $\mathcal{G}_{n,p_{\text{obs}}}$ with $n = 8$ and $p_{\text{obs}} = 0.3$ as shown in (a). The input patterns are shown for (b) the ground truth $\boldsymbol{x} = \mathbf{0}$, (c) the hypothesis $\tilde{\boldsymbol{x}} = [1, 0, \cdots, 0]$, and (d) the hypothesis $\tilde{\boldsymbol{x}} = [0, \cdots, 0, 1]$. The blue parts represent the entries being measured, and the orange region constitutes the parts of pairwise inputs that disagree with the ground truth.

Proof: See Appendix H. □

We conclude this part with another quantity that will often prove useful in tightening our results. Specifically, for any $\zeta > 0$, we define

$$m^{\mathsf{kl}}(\zeta) := \max_{l} \left| \left\{ i \;\middle|\; i \neq l,\ \mathsf{KL}\left(\mathbb{P}_i \,\|\, \mathbb{P}_l\right) \le (1+\zeta)\, \mathsf{KL}^{\min} \right\} \right|. \tag{22}$$

It is self-evident that $1 \le m^{\mathsf{kl}}(\zeta) < M$ holds regardless of ζ. This quantity determines the number of distinct input pairs under study that result in nearly-minimal output separation.

III. Main Results: Erdős–Rényi Graphs

At an intuitive level, faithful decoding is feasible only when (i) the measurement graph $\mathcal{G}$ is sufficiently connected so that we have enough measurements involving each vertex variable, and (ii) the channel output distributions given any two distinct inputs are sufficiently separated and hence distinguishable. To develop a more quantitative understanding about these two factors, we start with the Erdős–Rényi model, a tractable yet the most widely adopted random graph model for numerous applications. Specifically, we suppose that the measurement graph $\mathcal{G}$ is drawn from $\mathcal{G}_{n,p_{\text{obs}}}$ for some edge probability $p_{\text{obs}} \gtrsim \log n/n$. As will be shown in Section IV, many properties and intuitions that we develop for this specific graph model hold in greater generality.

A. Maximum Likelihood Decoding

To begin with, we analyze the performance guarantees of the maximum likelihood (ML) decoder

$$\psi_{\text{ml}}(\boldsymbol{y}) := \arg\max_{\boldsymbol{x}\in\mathcal{X}^n} \mathbb{P}\{\boldsymbol{y} \mid \boldsymbol{x}\}. \tag{23}$$

It is well-known that the ML rule minimizes the Bayesian probability of error under uniform input priors. We develop a sufficient recovery condition in terms of the edge probability and the minimum information divergence, which characterizes the tradeoff between the degree of graph connectivity and the resolution of channel outputs.

Theorem 1: Fix $\delta > 0$, and suppose that $\mathcal{G} \sim \mathcal{G}_{n,p_{\text{obs}}}$. Then there exist some universal constants C, $c_1 > 0$ such that if

$$\sup_{0<\alpha<1} \left\{ (1-\alpha)\, \mathsf{Hel}_{\alpha}^{\min} \right\} \cdot (p_{\text{obs}} n) \ge (1+\delta)\log(2n) + 2\log(M-1), \tag{24}$$

then the ML decoder ψ_{ml} obeys

$$P_{\text{e}}(\psi_{\text{ml}}) \le \frac{1}{(2n)^{\max\left\{\frac{3}{4}\delta - \frac{1}{4}\delta^2,\ \frac{\delta-1}{2}\right\}} - 1} + \frac{3}{n^{10}-1} + Cn^{-c_1\delta n}.$$

Proof: See Appendix A. □

Remark 1: By definition (3), one has $(1-\alpha)\,\mathsf{Hel}_{\alpha}^{\min} \le 1$ for any $0 < \alpha < 1$. As a consequence, the condition (24) implies

$$p_{\text{obs}} > \frac{\log n}{n},$$

thus ensuring the connectivity of $\mathcal{G}_{n,p_{\text{obs}}}$ with probability approaching one.

Theorem 1 essentially implies that the ML rule is guaranteed to work with high probability as long as

$$\sup_{\alpha} \left\{ (1-\alpha)\, \mathsf{Hel}_{\alpha}^{\min} \right\} \ge (1+o(1)) \frac{\log n + 2\log M}{p_{\text{obs}} n}.$$

Our result is non-asymptotic in the sense that it holds for all parameters $(n, M, \mathsf{Hel}_{\alpha}^{\min})$ instead of limiting to the asymptotic regime with n tending to infinity. Recognizing that $p_{\text{obs}} n$ is exactly the average vertex degree d_{avg}, our recovery condition reads

$$\sup_{\alpha} \left\{ (1-\alpha)\, \mathsf{Hel}_{\alpha}^{\min} \right\} \cdot d_{\text{avg}} \gtrsim \log n, \tag{25}$$

provided that $M \lesssim \mathcal{O}(\text{poly}(n))$ and $\alpha \in (0,1)$ is some fixed constant independent of n.

We pause to develop some intuitive understanding about the condition (25). In contrast to classical information theory settings, the channel decoding model considered herein concerns "uncoded" channel input. Consequently, the recovery bottleneck for the ML rule is presented by the minimum output distance given two distinct hypotheses, rather than the mutual information that plays a crucial rule in coded transmission. To be more precise, two hypotheses $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ are the least separated when they differ only by one component, say, v.

As a concrete example, one can take $\boldsymbol{x} = [0, \cdots, 0]$ and $\tilde{\boldsymbol{x}} = [1, 0, \cdots, 0]$ as illustrated in Fig. 3. The resulting pairwise outputs $\left\{y_{ij} \mid (i,j) \in \mathcal{E}\right\}$ thus contain about $\deg(v)$ pieces of information for distinguishing $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$; see, e.g., the orange shaded region highlighted in Fig. 3. Since the information contained in each measurement can be quantified by certain divergence metrics, namely, $\mathsf{Hel}_\alpha^{\min}$ (or $\mathsf{KL}^{\min}$ as adopted in Section III-B), the total amount of information one has available to distinguish two minimally separated hypotheses is captured by

$$\mathsf{Hel}_\alpha^{\min} \cdot d_{\text{avg}} \quad \text{or} \quad \mathsf{KL}^{\min} \cdot d_{\text{avg}}. \tag{26}$$

Furthermore, there are at least n distinct hypotheses that are all minimally apart from the ground truth $\boldsymbol{x}$ (e.g. $\tilde{\boldsymbol{x}} = [1, 0, \cdots, 0]$, $\tilde{\boldsymbol{x}} = [0, 1, \cdots, 0]$, $\cdots$, $\tilde{\boldsymbol{x}} = [0, \cdots, 0, 1]$). Representation of these hypotheses calls for at least $\log n$ bits, and hence the information that one can exploit to distinguish $\boldsymbol{x}$ from them—i.e. (26)—needs to exceed $\log n$. This offers an intuitive interpretation of the recovery condition (25).

Careful readers will note that Theorem 1 is presented in terms of the Hellinger divergence rather than the KL divergence. We remark on these this technical matter as follows.

Remark 2: In general, we are unable to develop the recovery conditions in terms of the KL divergence. This arises partly because the KL divergence cannot be well controlled for all measures, especially when $\left\| \frac{\mathrm{d}\mathbb{P}_l}{\mathrm{d}\mathbb{P}_j} \right\|_\infty$ $(l \neq j)$ grows. In contrast, the Hellinger divergence is generally stable and more convenient to analyze in this case.

We conclude this part with an extension. Examining our analysis reveals that all arguments continue to hold even if the output distributions are *location-dependent*. Formally, suppose that the distribution of y_{ij} is parametrized by

$$p\left(y_{ij} \middle| x_i - x_j = l\right) = \mathbb{P}_l^{ij}\left(y_{ij}\right), \quad 0 \le l < M, \ (i,j) \in \mathcal{E}. \tag{27}$$

This leads to a modified version of the minimum divergence metric as follows

$$\begin{aligned} &\overline{\mathsf{Hel}}_\alpha^{\min} \\ &:= \min\left\{ \mathsf{Hel}_\alpha(\mathbb{P}_l^{ij} \parallel \mathbb{P}_k^{ij}) \,\middle|\, l \neq k,\ 0 \le l, k < M,\ (i,j) \in \mathcal{E} \right\}. \end{aligned} \tag{28}$$

With these modified metrics in place, the preceding sufficient recovery condition immediately extends to this generalized model.

Theorem 2: The recovery condition of Theorem 1 continues to hold under the transition probabilities (27), if $\mathsf{Hel}_\alpha^{\min}$ *is replaced by* $\overline{\mathsf{Hel}}_\alpha^{\min}$ *as defined in (28).*

B. Minimax Lower Bound

In order to assess the tightness of our recovery guarantee for the ML rule, we develop two necessary conditions that apply to any recovery procedure. Here and below, $H(x) := -x\log x - (1-x)\log(1-x)$ stands for the binary entropy function.

Theorem 3: Suppose that $\mathcal{G} \sim \mathcal{G}_{n, p_{\text{obs}}}$. *Fix any* $\zeta \ge 0$ *and* $\epsilon > 0$, *and assume that* $p_{\text{obs}} > \frac{c\log n}{n}$ *for some sufficiently large constant* $c > 0$.

(a) If

$$\mathsf{KL}^{\min} \cdot p_{\text{obs}} n \le \frac{(1-\epsilon)\left(\log n + \log m^{\mathsf{kl}}(\zeta)\right) - H(\epsilon)}{(1+\epsilon)(1+\zeta)}, \tag{29}$$

then $\inf_\psi P_{\mathsf{e}}(\psi) \ge \epsilon - n^{-10}$.

(b) Suppose that $\alpha \le \frac{1}{1+\epsilon}$ *and* $p_{\text{obs}} n > 2\epsilon\alpha \log n$. *If*

$$(1-\alpha)\,\mathsf{Hel}_\alpha^{\min} \cdot p_{\text{obs}} n < \frac{\epsilon\alpha\log n}{1+\zeta} - r_\epsilon \tag{30}$$

for some residual[8] r_ϵ, *then* $\inf_\psi P_{\mathsf{e}}(\psi) \ge n^{-\epsilon} - n^{-10}$.

Proof: See Appendices B and C. □

We remark that the two necessary recovery conditions in Theorem 3 concern two regimes of separate interest. Specifically, Condition (29) based on the KL divergence is most useful when investigating first-order convergence, namely, the situation where we only require the minimax probability of error to be asymptotically vanishing without specifying convergence rates. In comparison, Condition (30) based on the Hellinger distance is more convenient when we further demand exact recovery to occur with polynomially high probability (e.g. $1 - \frac{1}{n}$). In various "big-data" applications, the term "with high probability" might only refer to the case where the error probability decays at least at a polynomial rate.

On the other hand, while Condition (29) is not directly presented in terms of M, we can often capture the effect of the alphabet size through the surrogate m^{kl}, provided that $\log m^{\mathsf{kl}} \asymp \log M$. In fact, this arises in many scenarios of interest. As an example, see the random corruption model to be discussed in Section V-B, where $m^{\mathsf{kl}} = M - 1$.

C. Tightness of Theorems 1 and 3 and Minimal Sample Complexity

Encouragingly, the recovery conditions derived in Theorems 1 and 3 are often tight up to some small multiplicative constant. In the sequel, we will assume that $p_{\text{obs}} \gg \log n / n$, and will pay special attention to two of the most popular divergence metrics: the KL divergence and the squared Hellinger distance.

1) Consider the first-order convergence, that is, the regime where $\inf_\psi P_{\mathsf{e}}(\psi) \to 0$ $(n \to \infty)$. Combining Theorems 1 and 3(a) gives

$$\begin{aligned} &\inf_\psi P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\longrightarrow} 0 \\ &\quad \text{if } \mathsf{Hel}_{\frac{1}{2}}^{\min} \cdot p_{\text{obs}} n > (1+o(1))(2\log n + 4\log M), \\ &\inf_\psi P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\not\longrightarrow} 0 \\ &\quad \text{if } \mathsf{KL}^{\min} \cdot p_{\text{obs}} n < (1-o(1))\left(\log n + \log m^{\mathsf{kl}}\right). \end{aligned}$$

[8] More precisely, $r_\epsilon := \log 2 + \frac{2[\epsilon\alpha\log n - \log 2]^2}{n p_{\text{obs}}}$.

When applied to the most challenging case where $\frac{\mathrm{d}\mathbb{P}_l}{\mathrm{d}\mathbb{P}_j} = 1 + o(1)$ for all $l \neq j$, these conditions read (with the assistance of Fact 1)

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\longrightarrow} 0 \quad \text{if } \mathsf{Hel}_{\frac{1}{2}}^{\min} \cdot p_{\mathrm{obs}} n > (1+o(1))(2\log n + 4\log M), \tag{31}$$

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\not\longrightarrow} 0 \quad \text{if } \mathsf{Hel}_{\frac{1}{2}}^{\min} \cdot p_{\mathrm{obs}} n < (1-o(1))\frac{\log n + \log m^{\mathsf{kl}}}{2}. \tag{32}$$

which are matching conditions modulo some multiplicative factor not exceeding

$$(1+o(1))\frac{4\log n + 8\log M}{\log n + \log m^{\mathsf{kl}}}. \tag{33}$$

2) We now move on to more stringent convergence by considering the regime where $\lim_{n\to\infty} \inf_{\psi} P_{\mathsf{e}}(\psi) \lesssim 1/n$. Putting Theorem 1 (with $\delta = 3$) and Theorem 3(b) (with $\alpha = 1/2$ and $\epsilon = 1$) together implies that

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\longrightarrow} 0 \quad \text{if } \mathsf{Hel}_{\frac{1}{2}}^{\min} \cdot p_{\mathrm{obs}} n > (1+o(1))(8\log n + 4\log M), \tag{34}$$

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\not\longrightarrow} 0 \quad \text{if } \mathsf{Hel}_{\frac{1}{2}}^{\min} \cdot p_{\mathrm{obs}} n < (1-o(1))\log n, \tag{35}$$

which holds regardless of $\frac{\mathrm{d}\mathbb{P}_l}{\mathrm{d}\mathbb{P}_j}$. These conditions do not impose constraints on the alphabet size, and are tight up to a multiplicative gap of

$$(1+o(1))\left(8 + \frac{4\log M}{\log n}\right). \tag{36}$$

Remark 3: The multiplicative factor (33) is small when either the alphabet size $M = O(\mathrm{poly}\log(n))$ (in which case this factor is 4) or when $M \asymp m^{\mathsf{kl}}$ (in which case this factor is at most 8). Similarly, the multiplicative factor (36) is the smallest when the alphabet size is $O(\mathrm{poly}\log(n))$. However, both results might become loose when $\log M$ and $\frac{\log M}{\log m^{\mathsf{kl}}}$ grow. For many practical applications, the alphabet size is typically much smaller than n, in which case our results are tight within a reasonable constant factor.

In summary, we have characterized the fundamental recovery condition under the Erdős–Rényi model, which reads

$$\mathsf{Hel}_{\frac{1}{2}}^{\min} \cdot d_{\mathrm{avg}} \gtrsim \log n \tag{37}$$

as long as the alphabet size M is not super-polynomial[9] in n. Put another way, in order to allow exact recovery, the *sample complexity*—i.e. the total number of edges of $\mathcal{G}$ (which is around $nd_{\mathrm{avg}}/2$)—necessarily obeys

$$\text{minimum sample complexity} \asymp \frac{n\log n}{\mathsf{Hel}_{\frac{1}{2}}^{\min}}. \tag{38}$$

Interestingly, these simple characterizations as well as the underlying intuitions carry over to many more homogeneous graphs, as will be seen in the next section.

Before concluding this section, we remark on the possibility of improving the preconstant. There are a few cases where the tight preconstants have been settled, including the stochastic block model [27], [48] and the censor block model [19], [48], provided that the alphabet size (or the number of communities) is $M = 2$. Asymptotically, the necessary and sufficient recovery condition reads

$$\sup_{0<\alpha<1}\left\{(1-\alpha)\mathsf{Hel}_{\alpha}^{\min}\right\} > \frac{\log n}{p_{\mathrm{obs}} n} \quad \text{or}$$

$$\text{minimum sample complexity} > \frac{n\log n}{2\sup_{0<\alpha<1}\left\{(1-\alpha)\mathsf{Hel}_{\alpha}^{\min}\right\}},$$

thus justifying the tightness of the sufficient recovery condition we derive. In fact, when the alphabet size is a constant, the fundamental divergence measure that dictates the information limits is often some variant of the minimum Chernoff information[10] or the Hellinger divergence (when optimized over α). Nevertheless, it is not clear whether such findings extend to the large alphabet settings. Part of the reason is that the minimum Chernoff information or Hellinger divergence do not necessarily capture the precise error exponent when testing many hypotheses. We leave to future work the investigation of tight preconstants in the large alphabet scenarios.

IV. Main Results: General Graphs

We now broaden our scope by exploring general measurement graphs beyond the simple Erdős–Rényi model, with emphasis on the family of homogeneous graphs.

A. Preliminaries: Key Graphical Metrics

Our theory relies on several widely encountered graphical metrics including the minimum vertex degree, the average vertex degree, the maximum vertex degree, and the size of the minimum cut, which we denote by $d_{\min}$, d_{avg}, $d_{\max}$, and mincut, respectively. This subsection introduces a few other not-so-common graphical quantities that prove crucial in presenting our results.

For any integer m, define

$$\mathcal{N}(m) := \left\{\mathcal{S} \subset \mathcal{V} \mid e\left(\mathcal{S}, \mathcal{S}^{\mathrm{c}}\right) \leq m\right\}, \tag{39}$$

which comprises all cuts of size at most m. We are particularly interested in the peak growth rate of the cardinality of $\mathcal{N}$ as defined below

$$\tau_k^{\mathrm{cut}} := \frac{1}{k}\log\left|\mathcal{N}(k\cdot\mathsf{mincut})\right| \quad \text{and} \quad \tau^{\mathrm{cut}} := \max_{k>0} \tau_k^{\mathrm{cut}}. \tag{40}$$

[9]When the alphabet size is super-polynomial in n, our upper and lower bounds are within a factor of $O\left(\frac{\log M}{\log n}\right)$ from optimal. We note, however, that in all our motivating applications, the alphabet size M is typically much smaller than $\exp(\Theta(n))$ and hence the regime with super-polynomial alphabet size is of little practical relevance.

[10]Note that the Chernoff information is defined to be $-\log\left\{1 - \sup_{0<\alpha<1}(1-\alpha)\mathsf{Hel}_{\alpha}\right\}$.

In the sequel, we will term τ^{cut} the ***cut-homogeneity exponent***. In fact, if we rewrite

$$\tau_k^{\mathrm{cut}} := \mathsf{mincut} \cdot \left\{ \frac{1}{k \cdot \mathsf{mincut}} \log \left| \mathcal{N}\left(k \cdot \mathsf{mincut}\right) \right| \right\}, \quad (41)$$

then we see that τ^{cut} relies on two factors: (i) the cut-set distribution exponents $\left\{\frac{1}{k}\log|\mathcal{N}(k)|\right\}_{k>0}$ and (2) the size of the minimum cut, both of which are important in capturing the degree of homogeneity of the cut-set distribution. This metric is best illustrated through the following two extreme examples:

- *Complete graph K_n on n vertices*. This homogeneous graph obeys $e\left(\mathcal{S}, \mathcal{S}^{\mathrm{c}}\right) = |\mathcal{S}|\left(n - |\mathcal{S}|\right)$ and $\mathsf{mincut} = n-1$. A simple combinatorial argument yields $|\mathcal{N}(m)| \asymp \binom{n}{m/n} \asymp n^{m/n}$, revealing that

$$\begin{aligned} \tau^{\mathrm{cut}} &= \max_k \frac{1}{k} \log \left| \mathcal{N}\left(k \cdot \mathsf{mincut}\right) \right| \\ &\asymp \max_k \frac{1}{k} \frac{k \cdot \mathsf{mincut}}{n} \log n \asymp \log n. \end{aligned}$$

- *Two complete subgraphs $K_{n/2}$ connected by a single bridge*. In this graph, the min-cut size is $\mathsf{mincut} = 1$ due to the existence of a bridge, but we still have $|\mathcal{N}(m)| \lesssim n^{m/n}$ when $m \geq n$. A little algebra gives

$$\begin{aligned} \tau^{\mathrm{cut}} &= \max_k \frac{1}{k} \log \left| \mathcal{N}\left(k \cdot \mathsf{mincut}\right) \right| \\ &\lesssim \max_k \frac{1}{k} \frac{k \cdot \mathsf{mincut}}{n} \log n \asymp \frac{\log n}{n}. \end{aligned}$$

Interestingly, for various homogeneous graphs of interest, τ^{cut} can be bounded above in a tight and simple manner, namely, $\tau^{\mathrm{cut}} \lesssim \log n$. This is formally stated in the following lemma, which accounts for homogeneous geometric graphs and expander graphs. In words, a graph is said to be a homogeneous geometric graph if it satisfies two properties: (i) each connected pair of vertices shares sufficiently many neighbors; (ii) when two vertices are geometrically close, they share a large fraction of neighbors. Here and throughout, we shall use $\mathcal{V}(u)$ to denote the set of neighbors of a vertex u.

Lemma 1:

(1) Homogeneous Geometric Graphs: Suppose that $\mathcal{G}$ is connected and is embedded in some Euclidean space. Assume that there exist two numerical constants $\rho > 0$ and $0 < \kappa < \frac{1}{2}$ such that

(a) for each $(u, v) \in \mathcal{E}$,

$$|\mathcal{V}(u) \cap \mathcal{V}(v)| \geq \rho \cdot \mathsf{mincut}; \quad (42)$$

(b) for each $(u, v) \in \mathcal{E}$, denoting by $w^{(i)}$ the i^{th} closest vertex to v among the vertices in $\mathcal{V}(u) \cap \mathcal{V}(v)$, one has

$$\left| \mathcal{V}(v) \setminus \mathcal{V}(w^{(i)}) \right| \leq \frac{1}{2}\rho \cdot \mathsf{mincut}, \quad 1 \leq i \leq \kappa\rho \cdot \mathsf{mincut}. \quad (43)$$

Under the above two conditions, one has

$$\tau^{\mathrm{cut}} \leq \frac{8}{\kappa\rho} \log(2n). \quad (44)$$

(2) Expander Graphs: If $\mathcal{G}$ is an expander graph with edge expansion $h_{\mathcal{G}}$, then

$$\tau^{\mathrm{cut}} \leq \frac{\mathsf{mincut}}{h_{\mathcal{G}}} \log n + \log 2. \quad (45)$$

Proof: See Appendix E. □

We highlight a few concrete examples covered by this lemma.

- The following instances of homogeneous geometric graphs are worth mentioning. The first is a random geometric graph $\mathcal{G}_{n,r}$, provided that $r^2 > c\log n$ for some sufficiently large $c > 0$. The second is a generalized ring in which two vertices are connected as long as they are at most a few vertices apart. For both cases, κ and ρ are constants bounded away from zero, indicating that

$$\tau^{\mathrm{cut}} \lesssim \log n.$$

- Another situation concerns those expander graphs with good expansion properties, including but not limited to Erdős–Rényi graphs, random regular graphs, and small world graphs. Since the expansion properties of these graphs obey $h_{\mathcal{G}}/\mathsf{mincut} = \Theta(1)$, we conclude from Lemma 1 that

$$\tau^{\mathrm{cut}} \lesssim \log n.$$

As a final remark, we are not aware of a graph for which τ^{cut} exceeds the order of $\log n$. In all aforementioned examples, one always has $\tau^{\mathrm{cut}} \lesssim \log n$. In-depth study about the upper limit on τ^{cut} might lead to further simplification of our results, which we leave for future work.

B. ML Decoding and Minimax Lower Bounds

This section presents recovery conditions based on the minimum information divergence and certain graphical metrics, which accommodate general graph structures, channel characteristics, and input alphabets. We defer detailed discussion of our results to Section IV-C.

To begin with, the following theorem—whose proof can be found in Appendix D—characterizes a regime where the ML decoder is guaranteed to work.

Theorem 4: Consider any connected graph $\mathcal{G}$. For any $\delta > 0$ and any $0 < \alpha < 1$, the ML rule ψ_{ml} achieves

$$P_{\mathrm{e}}(\psi_{\mathrm{ml}}) \leq \frac{1}{(2n)^{\delta} - 1},$$

provided that

$$\begin{aligned} &\sup_{0<\alpha<1} \left\{ (1-\alpha)\, \mathsf{Hel}_{\alpha}^{\min} \right\} \cdot \mathsf{mincut} \\ &\quad \geq 8\tau^{\mathrm{cut}} + (\delta + 8)\log(2n) + 4\log M. \end{aligned} \quad (46)$$

Remark 4: Theorem 4 continues to hold if $\mathsf{Hel}_{\alpha}^{\min}$ is replaced by $D_{\alpha}^{\min}$.

The sufficient recovery condition given in Theorem 4 is universal and holds for all graphs, and depends only on the min-cut size and the cut-homogeneity exponent irrespective of other graphical metrics. Similarly, the above sufficient condition extends to the scenario with location-dependent output distributions, as stated below.

Theorem 5: The recovery condition of Theorem 4 continues to hold under the transition probabilities (27), if $\mathsf{Hel}_{\alpha}^{\min}$ *and* $D_{\alpha}^{\min}$ *are replaced by* $\overline{\mathsf{Hel}}_{\alpha}^{\min}$ *and* $\overline{D}_{\alpha}^{\min}$*, respectively, where* $\overline{D}_{\alpha}^{\min} := -\frac{1}{1-\alpha}\log\left(1-(1-\alpha)\,\overline{\mathsf{Hel}}_{\alpha}^{\min}\right)$.

Next, we present a fundamental lower limit on $\mathsf{KL}^{\min}$ that admits perfect information recovery, based on the same graphical metrics in addition to the maximum vertex degree.

Theorem 6 (KL Version): Fix $\zeta \geq 0$ *and* $0 < \epsilon \leq 1/2$*. For any graph* $\mathcal{G}$*, if the KL divergence satisfies*

$$\mathsf{KL}^{\min}\cdot\mathsf{mincut} \leq \max\left\{(1-\epsilon)\,\tau^{\mathrm{cut}} - H(\epsilon),\ \frac{(1-\epsilon)\log m^{\mathsf{kl}}(\zeta) - H(\epsilon)}{1+\zeta}\right\} \tag{47}$$

or

$$\mathsf{KL}^{\min}\cdot d_{\max} \leq \frac{(1-\epsilon)\left(\log n + \log m^{\mathsf{kl}}(\zeta)\right) - H(\epsilon)}{1+\zeta}, \tag{48}$$

then the minimax probability of error exceeds $\inf_{\psi} P_{\mathsf{e}}(\psi) \geq \epsilon$.

Proof: See Appendix B. □

Notably, the conditions (47) and (48) do not imply each other. The first condition (47)—which characterizes the effects of cut-set distributions and alphabet size—is dominant for inhomogeneous graphs where $\mathsf{mincut} \ll d_{\max}$ (e.g. the graph formed by connecting two $K_{n/2}$ with a single bridge as described in Section IV-A). In comparison, the other condition (48) becomes tighter as $\frac{\mathsf{mincut}}{d_{\max}}$ grows, which is particularly useful when accounting for the family of homogeneous graphs where $d_{\max} \asymp \mathsf{mincut}$.

Finally, we complement the above KL version by another lower bound developed directly based on the Hellinger divergence, although it becomes loose for those inhomogeneous graphs obeying $\mathsf{mincut} \ll d_{\max}$. This is particularly useful when investigating the scenario that demands high-probability recovery (e.g. with success probability at least $1 - n^{-1}$). The proof can be found in Appendix C.

Theorem 7 (Hellinger Version): Consider any graph $\mathcal{G}$*, any* $\epsilon > 0$*, and* $\alpha \leq \frac{1}{1+\epsilon}$*. Suppose that* $d_{\max} \geq 2\epsilon\alpha\log n$*. If*

$$(1-\alpha)\,\mathsf{Hel}_{\alpha}^{\min}\cdot d_{\max} \leq \epsilon\alpha\log n - r_{\epsilon} \tag{49}$$

for some residual[11] r_{ϵ}*, then* $\inf_{\psi} P_{\mathsf{e}}(\psi) \geq n^{-\epsilon}$.

Remark 5: For any fixed $\epsilon > 0$*, one can optimize (49) over all* $0 < \alpha \leq \frac{1}{1+\epsilon}$ *to derive a tighter condition.*

C. Interpretation and Discussion

We now discuss the messages conveyed by the aforementioned results, for which we emphasize a broad family of homogeneous graphs before turning to the most general graphs. In what follows, our discussion assumes $\frac{\mathrm{d}\mathbb{P}_l}{\mathrm{d}\mathbb{P}_j} = \mathcal{O}(1)$ for all $0 \leq l, j < M$, in which case one has (by invoking Fact 1)

$$\mathsf{KL}^{\min} \asymp \mathsf{Hel}_{\frac{1}{2}}^{\min}. \tag{50}$$

[11] More precisely, $r_{\epsilon} := \log 2 + \frac{2[\epsilon\alpha\log n - \log 2]^2}{d_{\max}}$.

Operating upon such assumptions enables us to significantly simplify the presentation, while still capturing the regime that is statistically the most challenging (compared to its complement regime where $\frac{\mathrm{d}\mathbb{P}_l}{\mathrm{d}\mathbb{P}_j} \gg 1$).

1) Homogeneous Graphs: Our recovery conditions are most useful when applied to homogeneous graphs. Formally speaking, we term $\mathcal{G}$ a *homogeneous graph* if it satisfies

$$\mathsf{mincut} \asymp d_{\mathrm{avg}} \asymp d_{\max}, \tag{51}$$

which subsumes as special cases the widely adopted Erdős–Rényi graphs, random geometric graphs, small world graphs, rings, grids, and many other expander graphs. A few implications are in order.

1) For all homogeneous graphs, one has

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\longrightarrow} 0 \quad \text{if } \mathsf{Hel}_{\frac{1}{2}}^{\min}\cdot d_{\mathrm{avg}} \gtrsim \tau^{\mathrm{cut}} + \log n + \log M, \tag{52}$$

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\not\longrightarrow} 0 \quad \text{if } \mathsf{Hel}_{\frac{1}{2}}^{\min}\cdot d_{\mathrm{avg}} \lesssim \tau^{\mathrm{cut}} + \log n + \log m^{\mathsf{kl}}. \tag{53}$$

In general, these results are within a multiplicative gap

$$\mathsf{gap} \lesssim \frac{\tau^{\mathrm{cut}} + \log n + \log M}{\tau^{\mathrm{cut}} + \log n + \log m^{\mathsf{kl}}}$$

from optimal, which are orderwise tight when either $M \lesssim \mathrm{poly}(n)$ or $\log m^{\mathsf{kl}} \asymp \log M$. In particular, as long as the alphabet size is not super-polynomial in n, we arrive at the fundamental recovery condition for this class of graphs:

$$\mathsf{Hel}_{\frac{1}{2}}^{\min}\cdot d_{\mathrm{avg}} \gtrsim \log n + \tau^{\mathrm{cut}}. \tag{54}$$

2) In comparison to the recovery guarantee developed for the Erdős–Rényi model, the condition (54) includes one extra correction term τ^{cut} concerning the cut-set distribution. To provide some intuition about τ^{cut}, suppose that the ground truth is $\boldsymbol{x} = \mathbf{0}$ and consider an alternative hypothesis $\tilde{\boldsymbol{x}}$ whose non-zero entries are all identical. If we denote by $\mathcal{S}$ the vertex set corresponding to the support of $\tilde{\boldsymbol{x}}$, then it is straightforward to see that all measurements that can help distinguish $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ reside in the cut set $\mathcal{E}(\mathcal{S}, \mathcal{S}^{\mathrm{c}})$. By definition, τ_k^{cut} determines the total number of distinct cuts whose size is within some fixed range. Since τ_k^{cut} is defined in a logarithmic and normalized manner, this in turn specifies how many bits are needed to represent all these cuts and, hence, all hypotheses associated with them. As a consequence, τ^{cut} presents another information-theoretic requirement.
3) While our results fall short of a general upper bound on τ^{cut}, we note that $\tau^{\mathrm{cut}} \lesssim \log n$ holds for a broad class of interesting models studied in the literature (and in fact all models that we are aware of), including but not limited to various homogeneous geometric graphs and expander graphs (cf. Lemma 1). As a consequence, the recovery condition (54) for these graphs further simplifies to

$$\mathsf{Hel}_{\frac{1}{2}}^{\min}\cdot d_{\mathrm{avg}} \gtrsim \log n, \tag{55}$$

TABLE I
SUMMARY OF KEY RESULTS FOR ALL GRAPH MODELS ($M \lesssim \text{poly}(n)$)

Measurement graphs	Fundamental recovery conditions
Erdős–Rényi graphs	$\mathsf{Hel}_{1/2}^{\min} \cdot d_{\text{avg}} \gtrsim \log n$
homogeneous geometric graphs, expander graphs	$\mathsf{Hel}_{1/2}^{\min} \cdot d_{\text{avg}} \gtrsim \log n$
homogeneous graphs	$\mathsf{Hel}_{1/2}^{\min} \cdot d_{\text{avg}} \gtrsim \log n + \tau^{\text{cut}}$
general graphs	$\mathsf{Hel}_{1/2}^{\min} \cdot \mathsf{mincut} \gtrsim g(n) + \tau^{\text{cut}} \quad (1 \lesssim g(n) \lesssim \log n)$

which coincides with the one under the special Erdős–Rényi model. Following the intuition given in Section III-A, one must rely on around d_{avg} measurements to distinguish two minimally separated hypotheses—i.e. those that differ by a single component—and hence the information bottleneck constitutes around $\mathsf{Hel}_{\frac{1}{2}}^{\min} \cdot d_{\text{avg}}$ bits, which needs to be at least $\log n$ bits in order to encode n minimally apart hypotheses.

4) The condition (55) in turn leads to an interesting observation: for a variety of homogeneous graphs, the information-theoretic limits for graph-based decoding are determined solely by the *edge sparsity*, as opposed to the performance guarantees for many tractable algorithms (e.g. spectral methods or semidefinite programming) whose success typically rely on strong *second-order* expansion properties.

Finally, by combining Theorem 4 and Theorem 7 (with $\epsilon = 1$), we arrive at the following criterion concerning "high-probability" recovery: for various homogeneous graphs that obey $\tau^{\text{cut}} \lesssim \log n$, the probability of error $P_{\text{e}}(\psi) \leq n^{-1}$ is possible if and only if

$$\mathsf{Hel}_{\frac{1}{2}}^{\min} \cdot d_{\text{avg}} \gtrsim \log n. \tag{56}$$

In contrast to the preceding discussion, this statement holds regardless of how $\frac{\mathrm{d}\mathbb{P}_i}{\mathrm{d}\mathbb{P}_j}$ scales.

2) General Graphs: We now move on to discussing the results in their full generality. One distinguishing feature from the family of homogeneous graphs is that the recovery boundary is dictated by the size of the minimum cut rather than the graph edge sparsity. For the convenience of the reader, we summarize all key results in Table I.

1) **Tightness under general graphs**. The recovery conditions presented in Theorems 4 and 6 can be summarized as follows

$$\inf_{\psi} P_{\text{e}}(\psi) \overset{n\to\infty}{\longrightarrow} 0 \quad \text{if } \mathsf{Hel}_{\frac{1}{2}}^{\min} \cdot \mathsf{mincut} \gtrsim \tau^{\text{cut}} + \log M + \log n, \tag{57}$$

$$\inf_{\psi} P_{\text{e}}(\psi) \overset{n\to\infty}{\not\longrightarrow} 0 \quad \text{if } \mathsf{Hel}_{\frac{1}{2}}^{\min} \cdot \mathsf{mincut} \lesssim \tau^{\text{cut}} + \log m^{\mathsf{kl}} + \frac{\mathsf{mincut}}{d_{\max}} \log n. \tag{58}$$

These are within a multiplicative gap from optimal, satisfying that

$$\mathsf{gap} \lesssim \frac{\tau^{\text{cut}} + \log M + \log n}{\tau^{\text{cut}} + \log m^{\mathsf{kl}} + \frac{\mathsf{mincut}}{d_{\max}} \log n}.$$

Recognizing that $\tau^{\text{cut}} \gtrsim 1$, we see that the derived bounds are orderwise optimal when $\log m^{\mathsf{kl}} \asymp \log M \asymp \log n$ (e.g. in the random corruption model presented in Section V-B under large alphabet). Even for the loosest case, the gap is at most logarithmic (i.e. $\mathcal{O}(\log n + \log M)$).

2) **Information bottleneck**. In contrast to (54) and (55), the amount of information one has available to differentiate two minimally separated hypotheses is approximately given by $\mathsf{Hel}_{\frac{1}{2}}^{\min} \cdot \mathsf{mincut}$ instead of $\mathsf{Hel}_{\frac{1}{2}}^{\min} \cdot d_{\text{avg}}$. This makes sense since the two hypotheses that are most difficult to differentiate are no longer those that differ by one component. Instead, the most challenging task lies in linking the variables across the minimum cut, which can convey at most $\mathsf{Hel}_{\frac{1}{2}}^{\min} \cdot \mathsf{mincut}$ bits of information, forming the most fragile component for simultaneous recovery.

3) **A unified non-asymptotic framework.** Our framework can accommodate a variety of practical scenarios that respect the high-dimensional regime: the alphabet size might be growing with n while the channel divergence metrics might be decaying. Furthermore, our problem falls under the category of multi-hypothesis testing in the presence of exponentially many hypotheses, where each hypothesis is *not* necessarily formed by i.i.d. sequences. Under such a setting, the conventional Sanov bound [49] based on the Chernoff information measure [50] becomes unwieldy. In contrast, our results build upon alternative probability divergence measures (particularly the Hellinger / Rényi divergence). This results in a simple unified framework that enables non-asymptotic characterization of the minimax limits (modulo some constant factor) simultaneously for most settings.

In general, the current approach is unable to close the worst-case gap $\mathcal{O}(\log n + \log M)$, which could be large when either n or M are exceedingly large. In order to improve the recovery conditions, one alternative is to derive a tighter lower bound on the graphical metric τ^{cut}. For instance, our bounds become orderwise tight whenever $\tau^{\text{cut}} \gtrsim \log n$, which arise in various graphs beyond the family of homogeneous graphs. We leave this for future investigation. In addition, our general lower bounds are developed based on Fano's inequality, since Fano's

inequality allows us to accommodate a set of input hypotheses that have significant overlaps. Unfortunately, Fano's inequality typically relies on the KL divergence between input hypotheses, which is in general not capable of capturing the right error exponent for hypothesis testing. It would be interesting to develop a variant of the Fano-type inequality based directly on the Chernoff information measures.

V. Consequences for Specific Applications

In this section, we apply our general theory to a few concrete examples that have been studied in prior literature. As will be seen, our general theorems lead to order-wise tight characterization for all these canonical examples.

A. Stochastic Block Model

We start by analyzing the stochastic block model (SBM), which is a generative way to model community structure. In the standard SBM, nodes are partitioned into two disjoint clusters (so one can assign labels $x_i \in \{0, 1\}$ for each node). Each pair of nodes is connected with probability $\frac{\alpha \log n}{n}$ or $\frac{\beta \log n}{n}$ depending on whether they fall within the same cluster or not. The goal is to infer the underlying clusters that produce the network. Of particular interest is exact recovery of the entire clusters, which has received considerable attention; see [4], [27], [28], [48], and [51]–[58] for a highly incomplete list of references.

We focus on the regime where $\alpha, \beta = o(n/\log n)$ and $\alpha > \beta$, which subsumes all but the densest community structures. Treating the SBM as a graphical channel over a complete measurement graph (i.e. $p_{\text{obs}} = 1$) with outputs being either 0 or 1—which encodes whether two nodes belong to the same cluster or not, we see that (cf. Definition 13)

$$\mathbb{P}_0 = \mathsf{Bern}\left(\frac{\alpha \log n}{n}\right), \quad \text{and} \quad \mathbb{P}_1 = \mathsf{Bern}\left(\frac{\beta \log n}{n}\right).$$

This allows us to compute

$$\begin{aligned}\mathsf{Hel}_{\frac{1}{2}}^{\min} &= \left(\sqrt{\frac{\alpha \log n}{n}} - \sqrt{\frac{\beta \log n}{n}}\right)^2 \\ &\quad + \left(\sqrt{1 - \frac{\alpha \log n}{n}} - \sqrt{1 - \frac{\beta \log n}{n}}\right)^2 \\ &= (1 + o(1))(\sqrt{\alpha} - \sqrt{\beta})^2 \frac{\log n}{n}.\end{aligned}$$

In addition, it follows from the relation between KL divergence and χ^2 divergence (e.g. [44, eq. (7)]) that

$$\begin{aligned}\mathsf{KL}^{\min} &\le \mathsf{KL}\left(\frac{\beta \log n}{n} \,\|\, \frac{\alpha \log n}{n}\right) \le \chi^2\left(\frac{\beta \log n}{n} \,\|\, \frac{\alpha \log n}{n}\right) \\ &\overset{\text{(a)}}{=} \frac{\left(\frac{\beta \log n}{n} - \frac{\alpha \log n}{n}\right)^2}{\frac{\alpha \log n}{n}\left(1 - \frac{\alpha \log n}{n}\right)} = \frac{(1 + o(1))(\alpha - \beta)^2}{\alpha} \frac{\log n}{n},\end{aligned} \tag{59}$$

where (a) follows from the identity $\chi^2(p \,\|\, q) = \frac{(p-q)^2}{q} + \frac{(p-q)^2}{1-q} = \frac{(p-q)^2}{q(1-q)}$. With these two estimates in place, Theorem 1 and Corollary 3 immediately give

$$\begin{aligned}&\inf_{\psi} P_{\text{e}}(\psi) \overset{n\to\infty}{\longrightarrow} 0 \ \text{ if } (\sqrt{\alpha} - \sqrt{\beta})^2 \ge 2(1 + o(1)), \\ &\inf_{\psi} P_{\text{e}}(\psi) \overset{n\to\infty}{\not\longrightarrow} 0 \ \text{ if } (\alpha - \beta)^2 \le (1 - o(1))\alpha.\end{aligned} \tag{60}$$

In fact, precise phase transition for exact cluster recovery has only been determined last year [27], [28]. There results assert that

$$\inf_{\psi} P_{\text{e}}(\psi) \overset{n\to\infty}{\longrightarrow} 0 \ \text{ if } (\sqrt{\alpha} - \sqrt{\beta})^2 > 2, \tag{61}$$

$$\inf_{\psi} P_{\text{e}}(\psi) \overset{n\to\infty}{\not\longrightarrow} 0 \ \text{ if } (\sqrt{\alpha} - \sqrt{\beta})^2 < 2, \tag{62}$$

justifying that the sufficient condition we develop is precise. When it comes to the necessary condition, one can verify that the condition (62) is more stringent than[12] $(\alpha - \beta)^2 < 4(\alpha + \beta)$. In comparison, the boundary of our condition (60) is sandwiched between the curves $(\alpha - \beta)^2 \le \frac{1}{2}(\alpha + \beta)$ and $(\alpha - \beta)^2 \le \alpha + \beta$. These taken collectively indicate that our theory is tight up to a small constant factor.

Several remarks are in order. To begin with, our results accommodate all values of α, β up to $o(n/\log n)$, which is broader than [27] that concentrates on the sparsest possible regime (i.e. $\alpha, \beta \asymp 1$). Leaving out this technical matter, a more interesting observation is that the achievability bound we develop for the ML rule matches the fundamental recovery limit in a precise manner, which seems to imply that the squared Hellinger distance is the right metric that dictates the recovery limits for the SBMs.

When finishing up this paper, we became aware of a very recent work [59] that characterizes the fundamental limits for the generalized SBM, that is, the model where n nodes are partitioned into multiple clusters. Extending our framework so as to accommodate the SBM in its full generality is a topic of future work.

B. Random Corruption Model

We now turn to another model called the *random corruption model*, which subsumes as special cases several applications including alignment, synchronization, and joint matching (e.g. [6], [10], [12], [14]).

Suppose that the measurements y_{ij}'s are independently corrupted following a distribution

$$y_{ij} = \begin{cases} x_i - x_j, & \text{with probability } p_{\text{true}}, \\ \mathsf{Unif}_M, & \text{else}, \end{cases} \tag{63}$$

where Unif_M is the uniform distribution over $\{0, \cdots, M-1\}$, p_{true} stands for the non-corruption rate, and "$-$" is some general subtraction operation defined in Section II. In words, a fraction $1 - p_{\text{true}}$ of measurements act as *random outliers* and

[12] To see this, observe that $(\sqrt{\alpha} - \sqrt{\beta})^2 < 2$ is identical to $(\alpha - \beta)^2 < 2(\sqrt{\alpha} + \sqrt{\beta})^2$, which is more stringent than $(\alpha - \beta)^2 < 4(\alpha + \beta)$ due to the elementary inequality $(a + b)^2 \le 2(a^2 + b^2)$.

contain no useful information. Note that under this random corruption model, one has

$$m^{\mathsf{kl}}(\epsilon) \equiv M-1, \quad \forall \epsilon \geq 0.$$

The following corollary—an immediate consequence of Theorem 1 and Corollary 3—presents concrete recovery limits for the random corruption model. For ease of presentation, we restrict our discussion to the Erdős–Rényi model, but remark that all results extend to homogeneous geometric graphs and other expander graphs (up to some constant factors) if one replaces $p_{\text{obs}}n$ with the average vertex degree.

Corollary 1: Fix $\epsilon > 0$. Consider the random corruption model (63), and assume $\mathcal{G} \sim \mathcal{G}_{n,p_{\text{obs}}}$ with $p_{\text{obs}} > \frac{c_1 \log n}{n}$ for some sufficiently large $c_1 > 0$. Then, one has

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\longrightarrow} 0$$
$$\text{if } \frac{1}{M}\left(\sqrt{1-p_{\text{true}}+Mp_{\text{true}}}-\sqrt{1-p_{\text{true}}}\right)^2 \geq (1+\epsilon)\frac{\log n+2\log M}{p_{\text{obs}}n}, \tag{64}$$

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\not\longrightarrow} 0$$
$$\text{if } p_{\text{true}} \leq \max\left\{\frac{(1-\epsilon)(\log n+\log M)}{p_{\text{obs}}n\log\left(1+\frac{p_{\text{true}}M}{1-p_{\text{true}}}\right)}, \frac{M}{M-1}\left(\frac{\log n}{p_{\text{obs}}n}-\frac{1}{M}\right)\right. \tag{65}$$

To establish this corollary, we start by considering the graph $\mathcal{G}_{\text{true}}$ that comprises all edges where $y_{ij} = x_i - x_j$. It is self-evident that $\mathcal{G}_{\text{true}} \sim \mathcal{G}_{n,\left(p_{\text{true}}+\frac{1-p_{\text{true}}}{M}\right)p_{\text{obs}}}$, and thus $\left(\left(1-\frac{1}{M}\right)p_{\text{true}}+\frac{1}{M}\right)p_{\text{obs}} > \frac{\log n}{n}$ is necessary to ensure connectivity (otherwise there will be no basis to link the node variables across disconnected components). Apart from this, everything boils down to calculating $\mathsf{KL}^{\min}$ and $\mathsf{Hel}^{\min}$, which we gather in the following lemma.

Lemma 2: Consider the random corruption model (63). For any $0 \leq p_{\text{true}} < 1$, one has

$$\mathsf{KL}^{\min} = p_{\text{true}}\log\left(1+\frac{p_{\text{true}}M}{1-p_{\text{true}}}\right);$$
$$\mathsf{Hel}^{\min}_{\frac{1}{2}} = \frac{2}{M}\left(\sqrt{1-p_{\text{true}}+Mp_{\text{true}}}-\sqrt{1-p_{\text{true}}}\right)^2. \tag{66}$$

More simply, these metrics can be bounded as

$$\mathsf{KL}^{\min} \leq \frac{p_{\text{true}}^2 M}{1-p_{\text{true}}} \text{ and } \mathsf{Hel}^{\min}_{\frac{1}{2}} \geq \frac{p_{\text{true}}^2 M}{2(1-p_{\text{true}}+Mp_{\text{true}})}. \tag{67}$$

Proof: See Appendix F. □

To illustrate these guarantees numerically, we depict in Fig. 4 an example of the preceding recovery conditions. In the sequel, we will discuss the tightness and implications of the above result for specific regimes, ranging from small alphabet to large alphabet. For convenience of theoretical comparison, we supply an alternative form obtained by applying the general theory but using the bounds (67):

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\longrightarrow} 0$$
$$\text{if } p_{\text{true}} \geq 2(1+\epsilon)\cdot \tag{68}$$
$$\sqrt{\frac{(1-p_{\text{true}}+Mp_{\text{true}})(\log n+2\log M)}{p_{\text{obs}}nM}}, \tag{69}$$

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\not\longrightarrow} 0$$
$$\text{if } p_{\text{true}} \leq \max\left\{(1-\epsilon)\sqrt{\frac{(1-p_{\text{true}})(\log n+\log M)}{p_{\text{obs}}nM}}, \frac{\log n}{p_{\text{obs}}n}\right\}. \tag{70}$$

1) Tightness Under Binary Alphabet: We start with the case where $M = 2$, which was also studied by [29]. When $p_{\text{obs}} \gg \frac{\log n}{n}$, our results (69) and (70) assert that

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\longrightarrow} 0 \text{ if } p_{\text{true}} \geq (1+o(1))\sqrt{\frac{2\log n}{p_{\text{obs}}n}}, \tag{71}$$

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\not\longrightarrow} 0 \text{ if } p_{\text{true}} \leq (1-o(1))\sqrt{\frac{\log n}{2p_{\text{obs}}n}}. \tag{72}$$

As a result, our bounds are within a factor $2 + o(1)$ from optimal, which holds for all possible values of $(p_{\text{obs}}, p_{\text{true}})$. This constant gap is illustrated in Fig. 4(a) as well.

In contrast, the bounds presented in [29] fall short of a uniform constant factor gap accommodating different parameter configurations. Adopting our notation, [29, Th. 4.1 and 4.2] reduce to[13]:

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\longrightarrow} 0 \quad \text{if } p_{\text{true}} > \sqrt{\frac{2\log n}{p_{\text{obs}}n}},$$
$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\not\longrightarrow} 0 \quad \text{if } p_{\text{true}} < \sqrt{\frac{2(1-3\tau/2)\log n}{p_{\text{obs}}n}},$$

where $0 < \tau < \frac{2}{3}$ is some numerical value so that $p_{\text{obs}} \leq 2n^{\tau-1}$. Hence, their bounds are tight up to a factor

$$g(\tau) = \frac{1+o(1)}{\sqrt{1-3\tau/2}},$$

which approaches 1 in the sparse graph regime as $\tau \to 0$ (e.g. $p_{\text{obs}} \asymp \frac{\log n}{n}$). On the other hand, it does not deliver meaningful conditions for the case where $\tau \geq \frac{2}{3}$ (i.e. $2n^{-\frac{1}{3}} \leq p_{\text{obs}} \leq 1$). In comparison, our bounds are looser for sparse graphs ($\tau < \frac{1}{2}$ or $p_{\text{obs}} < \frac{2}{\sqrt{n}}$) where $g(\tau) \leq 2$, but tighter for dense graphs ($\tau \geq \frac{1}{2}$ or $p_{\text{obs}} \geq \frac{2}{\sqrt{n}}$) where $g(\tau) \geq 2$.

Notably, when $p_{\text{obs}} \asymp \frac{\log n}{n}$, the fundamental limit approaches $\sqrt{\frac{2\log n}{p_{\text{obs}}n}}$ in an accurate manner [29]. This again corroborates the tightness of our achievability bound, implying that the squared Hellinger distance is the right quantity to control in the sparsest possible regime.

[13]Note that $p_{\text{true}} = 1-2\varepsilon$ and $d = np_{\text{obs}}$ for the notation ε and d defined in [29], respectively.

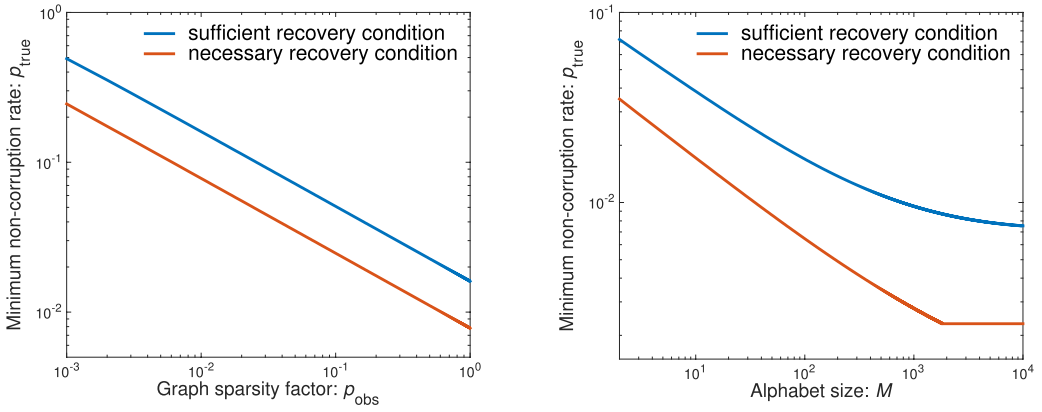

Fig. 4. The sufficient and the necessary conditions given in Corollary 1 when $n = 10^5$. The results are shown for: (a) $M = 2$; (b) $p_{\text{obs}} = 0.05$.

2) From Small Alphabet to Large Alphabet: The recovery conditions given in Corollary 1 can be further divided into and simplified for two respective regimes, depending on whether $Mp_{\text{true}} \lesssim 1$ or $Mp_{\text{true}} \gtrsim 1$. By substituting each of these two hypotheses into (69), deriving the corresponding minimum p_{true} for the respective case, and then checking the compatibility of $p_{\text{true}}M$ with the hypotheses, one immediately deduces:

1) When $M = o\left(\frac{p_{\text{obs}}n}{\log n}\right)$, one has

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\longrightarrow} 0 \quad \text{if } p_{\text{true}} \geq 2(1+o(1))\sqrt{\frac{\log n + 2\log M}{p_{\text{obs}}nM}}, \tag{73}$$

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\not\longrightarrow} 0 \quad \text{if } p_{\text{true}} \leq (1-o(1))\sqrt{\frac{\log n + \log M}{p_{\text{obs}}nM}}; \tag{74}$$

2) When $M = \omega\left(\frac{p_{\text{obs}}n}{\log n}\right)$, one has

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\longrightarrow} 0 \quad \text{if } p_{\text{true}} \geq \frac{4(1+o(1))(\log n + 2\log M)}{p_{\text{obs}}n}, \tag{75}$$

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\not\longrightarrow} 0 \text{ if } p_{\text{true}} \leq \frac{\log n}{p_{\text{obs}}n}. \tag{76}$$

That being said, the recovery boundary presented in terms of p_{true} exhibits contrasting features in two separate regimes, as illustrated in Fig. 4(b). Some interpretations are in order.

1) **Information-limited regime** ($M = o\left(\frac{d_{\min}}{\log n}\right)$). The amount of information that can be conveyed through each pairwise measurement is captured by the divergence measure. In this small-alphabet regime, a little algebra gives $\mathsf{KL}^{\min} \approx p_{\text{true}}^2 M$ (see Lemma 2), which is increasing in M. As a result, the alphabet size limits the amount of information that we can harvest, and the fundamental recovery boundary improves with M. For Erdős–Rényi graphs, the recovery conditions are tight up to a factor of 2 in the presence of a constant alphabet size, and up to a factor of $2\sqrt{\frac{3}{2}}$ for all $M \ll d_{\min}/\log n$.
2) **Connectivity-limited regime** ($M = \omega\left(\frac{d_{\min}}{\log n}\right)$). When M further increases and enters this regime, the information carried by each measurement saturates and no longer scales as $p_{\text{true}}^2 M$. In this regime, the measurement graph $\mathcal{G}$ presents a fundamental connectivity bottleneck. In fact, if $p_{\text{true}} = o\left(\frac{\log n}{d_{\min}}\right)$, then there will be at least one vertex that is not connected with a single useful measurement, and hence there will be absolutely no basis to infer the value of this isolated vertex. Our bounds in this regime are order-wise optimal as long as the alphabet size is not super-polynomial in n.

C. Haplotype Assembly

The pairwise measurement model can also be applied to analyze the haplotype assembly problem discussed in Section I. As formulated in [20] and [21], consider n SNPs on a chromosome, represented by a sequence $\{x_1, \cdots, x_n\} \in \{0,1\}^n$ such that a major (resp. minor) allele is denoted by 0 (resp. 1). Employing certain sequencing technologies, one obtains a collection of *independent* paired reads such that for any $(i,j) \in \mathcal{E}$,

$$y_{ij}^{(k)} = \begin{cases} x_i \oplus x_j, & \text{w.p. } 1-\theta, \\ x_i \oplus x_j \oplus 1, & \text{w.p. } \theta. \end{cases} \tag{77}$$

Here, $y_{ij}^{(k)}$ stands for the k^{th} noisy read of the parity between the i^{th} and the j^{th} SNPs, and $0 < \theta < 1/2$ denotes the read error rate. We assume that the reads taken on each edge are independent.

A realistic measurement graph that respects current sequencing technologies is the one in which measurements are obtained only when the i^{th} and the j^{th} SNPs are geometrically close, i.e., $|i-j| \leq w$ for some constant[14] $w > 0$. This

[14] As discussed in [21], the separation between two DNA reads (called the *insert size*) is typically bounded within a small range, with the median insert size not exceeding a few times the separation between adjacent SNPs.

is captured by a *generalized ring graph*, denoted by $\mathcal{G}_{\text{ring}} = \left(\mathcal{V}, \mathcal{E}_{\text{ring}}\right)$, such that

$$(i, j) \in \mathcal{E}_{\text{ring}} \quad \text{iff} \quad |i - j| \le w. \tag{78}$$

The number $L_{i,j}$ of reads taken between i and j is assumed to be dependent on their separation, i.e.[15]

$$L_{i,j} = L p_{|i-j|} \tag{79}$$

for some parameters L and $\{p_l \mid 1 \le l \le w\}$.

Additionally, a random and geometry-free measurement model has been investigated in [20] as well. The fundamental limit under this model is orderwise equivalent to that under an Erdős–Rényi graph with $L_{i,j} \equiv L$ for all $(i, j) \in \mathcal{E}$. For the sake of completeness, we derive consequences for both models as follows.

Corollary 2: Consider the model (77), and assume that θ and p_l are bounded away from 0.

(1) Suppose that $\mathcal{G} \sim \mathcal{G}_{\text{ring}}$. There exist some universal constants $c_1 > c_2 > 0$ such that

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\longrightarrow} 0 \quad \text{if } (1-2\theta)^2 > c_1 \frac{\log n}{L}, \tag{80}$$

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\not\longrightarrow} 0 \quad \text{if } (1-2\theta)^2 < c_2 \frac{\log n}{L}. \tag{81}$$

(2) Suppose that $\mathcal{G} \sim \mathcal{G}_{n,p_{\text{obs}}}$ and $p_{\text{obs}} > \frac{c_3 \log n}{n}$ for some sufficiently large constant $c_3 > 0$. Then there exist some universal constants $c_4, c_5 > 0$ such that

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\longrightarrow} 0 \quad \text{if } (1-2\theta)^2 > \frac{c_4 \log n}{L n p_{\text{obs}}}, \tag{82}$$

$$\inf_{\psi} P_{\mathsf{e}}(\psi) \overset{n\to\infty}{\not\longrightarrow} 0 \quad \text{if } (1-2\theta)^2 < \frac{c_5 \log n}{L n p_{\text{obs}}}. \tag{83}$$

Proof: For the sufficient condition, we only need to calculate the Rényi divergence. For each $(i, j) \in \mathcal{E}$, letting $D_{1/2}^{i,j}$ be the Rényi divergence of order $1/2$ between the distributions of $\{y_{ij}^{(k)}\}_{1 \le k \le L_{i,j}}$ given two distinct inputs (i.e. 0 and 1), one obtains

$$\begin{aligned} -D_{1/2}^{i,j} &\overset{\text{(i)}}{=} L_{i,j} \left\{ -2 \log \left(1 - \frac{1}{2} \mathsf{Hel}(\theta \,\|\, 1-\theta) \right) \right\} \\ &\overset{\text{(ii)}}{\ge} L_{i,j} \mathsf{Hel}(\theta \,\|\, 1-\theta), \end{aligned} \tag{84}$$

where (i) follows from additivity of Rényi divergence [44, Th. 2.8], and (ii) follows since $1 - x \le e^{-x}$. Furthermore,

$$\begin{aligned} \frac{1}{2}\mathsf{Hel}(\theta \,\|\, 1-\theta) &= \left(\sqrt{\theta} - \sqrt{1-\theta} \right)^2 \\ &= \frac{(1-2\theta)^2}{(\sqrt{\theta} + \sqrt{1-\theta})^2} \asymp (1-2\theta)^2. \end{aligned} \tag{85}$$

Recall that $L_{i,j} = L p_{|i-j|} \asymp L$ when $\mathcal{G} \sim \mathcal{G}_{\text{ring}}$, whereas $L_{i,j} = L$ when $\mathcal{G} \sim \mathcal{G}_{n,p_{\text{obs}}}$. These taken together with Theorem 5 and Lemma 1 (resp. Theorem 2) establish the sufficient condition for $\mathcal{G}_{\text{ring}}$ (resp. $\mathcal{G}_{n,p_{\text{obs}}}$).

[15]Careful readers will note that this assumption is different from the model adopted in [20] and [21], where the total number of reads is fixed with the reads independently generated. Nevertheless, the model considered here (which significantly simplifies presentation) is sufficient to capture the right scaling of the performance limits, since these two models are orderwise equivalent due to measure concentration.

For the necessary condition, by replacing all p_l with $\max_l p_l$ in (79), we obtain a new model such that any sufficient recovery condition for the original model holds for this new model as well. We then move on to compute the KL divergence for the new model:

$$\begin{aligned} \mathsf{KL}^{\min} &\le L\mathsf{KL}(\theta \,\|\, 1-\theta) \overset{\text{(a)}}{\le} L\chi^2(\theta \,\|\, 1-\theta) \\ &= L\frac{(1-2\theta)^2}{\theta(1-\theta)} \asymp L(1-2\theta)^2, \end{aligned} \tag{86}$$

where (a) follows from [44, eq. (7)]. Substitution into Theorem 6 finishes the proof. □

We now compare our results with prior results. The fundamental limits given in [20] and [21] were based on coverage (or sample complexity) as a metric, that is, the total number of reads required for perfect haplotype assembly. Recognizing that nLw (resp. $\binom{n}{2} p_{\text{obs}} L$) captures the order of the total number of paired reads for $\mathcal{G}_{\text{ring}}$ (resp. $\mathcal{G}_{n,p_{\text{obs}}}$), we see that the minimal sample complexity obeys

$$nLw \asymp \frac{n \log n}{(1-2\theta)^2}, \quad \text{when } \mathcal{G} \sim \mathcal{G}_{\text{ring}}; \tag{87}$$

$$\binom{n}{2} L p_{\text{obs}} \asymp \frac{n \log n}{(1-2\theta)^2}, \quad \text{when } \mathcal{G} \sim \mathcal{G}_{n,p_{\text{obs}}}. \tag{88}$$

Consequently, for the generalized ring graph, our results match the sample complexity limits characterized in [21] in an orderwise sense, which is proportional to

$$\frac{n \log n}{1 - e^{-\mathsf{KL}(0.5 \| \theta)}} \asymp \frac{n \log n}{\mathsf{KL}(0.5 \,\|\, \theta)} = \frac{n \log n}{\left(\frac{1}{2} + o(1)\right)(1-2\theta)^2}.$$

On the other hand, for the Erdős–Rényi graphs, the minimum sample complexity scales as $\frac{n \log n}{(1-2\theta)^2}$, which coincides with the orderwise limits $\Theta(n \log n)$ derived in [20].

Notably, our results are not restricted to the classical large-sample asymptotics where θ is fixed while n grows to infinity. This strengthens [20], [21] by accommodating the regime where $\theta - 1/2 = o(1)$, which characterizes the non-asymptotic tradeoff between n and the read quality. As a final remark, while our results are tight in capturing the right scaling w.r.t. the read error rate as well as the number of SNPs, our derivation is not tight in characterizing the behavior w.r.t. p_l (or the notation W given in [21]).

VI. Concluding Remarks

This paper investigates simultaneous recovery of multiple node variables based on noisy graph-based measurements, under the pairwise difference model. The problem formulation spans numerous applications including image registration, graph matching, community detection, and computational biology. We develop a unified framework in understanding all problems of this kind based on representing the available pairwise measurements as a graph, and then representing the noise on the measurements using a general channel with a given input/output transition measure. This framework accommodates large alphabets, general channel transition probabilities, and general graph structures in a non-asymptotic manner. Our results underscore the interplay between the minimum channel divergence measures and the minimum cut size of

the measurement graph. Moreover, for various homogeneous graphs, the recovery criterion relies almost only on the first-order graphical metrics independent of other second-order metrics like the spectral gap. We expect that such fundamental recovery criterion will provide a general benchmark for evaluating the performance of practical algorithms over many applications.

For concreteness, we restrict our attention to the pairwise difference model in this paper, but we remark that the analysis framework is somewhat generic and applies to a broader family of pairwise measurements. For instance, consider a more general invertible pairwise relation, denoted by $x_i \ominus x_j$, that satisfies

$$\begin{cases} x_1 \ominus x_2 \neq x_1 \ominus x_3, & \forall x_2 \neq x_3; \\ x_1 \ominus x_2 \neq x_4 \ominus x_2, & \forall x_1 \neq x_4. \end{cases} \tag{89}$$

As an example, the addition operator defined as $x_i \ominus x_j := ax_i + bx_j \ (\mathsf{mod}\ M)$ falls within this class as long as both (a, M) and (b, M) are coprime. Interestingly, most of the analyses carry over to such models and reappear suitably generalized. Details concerning full generalization of our results are left for future work.

While our paper centers on the minimax recovery involving all possible input configurations, there exists another family of applications where the inputs fall within a more restricted class (e.g. the class of inputs whose components are spread out over the entire alphabet). In addition, it would be interesting to establish how the fundamental limits can be improved under the partial recovery setting, namely, the situation where one only demands reconstruction of a (large) fraction of input variables. Even in the exact recovery situation, it remains to be seen whether the universal pre-constants can be further tightened. Moving away from the statistical guarantees, another important issue is the computational feasibility of information recovery. It has recently been demonstrated by [19] that the information and computation limits meet for many homogeneous graphs (e.g. rings, lines, small-world graphs, grids). It would be of great interest to see whether there exists any computational gap away from the statistical limits for a more general family of graphs.

Appendix A
Proof of Theorem 1

Suppose that both the ground truth and the null hypothesis are $\boldsymbol{x} = \boldsymbol{x}^*$. Consider the class of alternative hypotheses parametrized by k $(1 \le k \le n)$ as follows

$$\mathcal{H}_k := \left\{ \boldsymbol{x} \mid \left\| \boldsymbol{x} - \boldsymbol{x}^* \right\|_0 = n - k \right\}, \tag{90}$$

which comprises at most $\binom{n}{k}(M-1)^{n-k}$ distinct hypotheses. For notational convenience, denote by $\mathbb{P}_{\boldsymbol{w}}(\cdot)$ (resp. $\mathbb{P}_{\boldsymbol{0}}(\cdot)$) the probability measure of $\boldsymbol{y}$ conditional on the alternative hypothesis $\boldsymbol{x} = \boldsymbol{w}$ (resp. the null hypothesis $\boldsymbol{x} = \boldsymbol{x}^*$). We let $P_{\mathrm{e},\mathcal{H}_k}$ represent the probability of error when restricted to the class $\mathcal{H}_k$ of alternative hypotheses. For simplicity of presentation, we will assume $\boldsymbol{x}^* = \boldsymbol{0}$ in what follows, but all steps apply to other choices of $\boldsymbol{x}^*$.

For any $\boldsymbol{w} \in \mathcal{H}_k$, denote by $\mathcal{S}_i$ $(0 \le i < M)$ the set of vertices v obeying $w_v = i$, and let $n_i = |\mathcal{S}_i|$. Apparently, there are $\frac{1}{2}\sum_{i=1}^{M-1} e(\mathcal{S}_i, \mathcal{S}_i^{\mathrm{c}})$ distinct locations (l, j) satisfying $l > j$ and $w_l - w_j \neq 0$, where $e(\mathcal{S}, \mathcal{S}^{\mathrm{c}})$ denotes the number of cut edges as defined in Section I-C. With this in mind, it follows from the Chernoff bound that

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{0}}\left\{ \log \frac{\mathrm{d}\mathbb{P}_{\boldsymbol{w}}(\boldsymbol{y})}{\mathrm{d}\mathbb{P}_{\boldsymbol{0}}(\boldsymbol{y})} > 0 \,\middle|\, \mathcal{E} \right\} \\
&= \mathbb{P}_{\boldsymbol{0}}\left\{ \sum_{(l,j)\in\mathcal{E},\, l>j} \alpha \log \frac{\mathrm{d}\mathbb{P}_{\boldsymbol{w}}(y_{lj})}{\mathrm{d}\mathbb{P}_{\boldsymbol{0}}(y_{lj})} > 0 \,\middle|\, \mathcal{E} \right\} \\
&\le \prod_{(l,j)\in\mathcal{E},\, l>j} \mathbb{E}_{\boldsymbol{0}}\left[e^{\alpha \log \frac{\mathrm{d}\mathbb{P}_{\boldsymbol{w}}(y_{lj})}{\mathrm{d}\mathbb{P}_{\boldsymbol{0}}(y_{lj})}} \right] \\
&= \prod_{(l,j)\in\mathcal{E},\, l>j} \left[1 - (1-\alpha)\, \mathsf{Hel}_\alpha\left(\mathbb{P}_{\boldsymbol{w}}(y_{lj}) \,\|\, \mathbb{P}_{\boldsymbol{0}}(y_{lj}) \right) \right] \qquad (91) \\
&= \exp\left(-(1-\alpha) \sum_{(l,j)\in\mathcal{E},\, l>j} D_\alpha\left(\mathbb{P}_{\boldsymbol{w}}(y_{lj}) \,\|\, \mathbb{P}_{\boldsymbol{0}}(y_{lj}) \right) \right) \qquad (92) \\
&\le \exp\left(-(1-\alpha) \frac{\sum_{i=0}^{M-1} e(\mathcal{S}_i, \mathcal{S}_i^{\mathrm{c}})}{2} D_\alpha^{\min} \right), \qquad (93)
\end{aligned}$$

where (91) follows from the definition of the Hellinger divergence, (92) comes from the definition (9), and (93) arises since $D_\alpha\left(\mathbb{P}_{\boldsymbol{w}}(y_{lj}) \,\|\, \mathbb{P}_{\boldsymbol{0}}(y_{lj}) \right) \neq 0$ if and only if $w_l - w_j \neq 0$.

Additionally, define the quantity

$$N_{\boldsymbol{w}} := \frac{1}{2}\left| \cup_i \left\{ (l, j) \in \left(\mathcal{S}_i, \mathcal{S}_i^{\mathrm{c}}\right) \right\} \right| = \frac{1}{2}\sum_{i=0}^{M-1} |\mathcal{S}_i|\,(n - |\mathcal{S}_i|),$$

then it follows from the definition of $\mathcal{G}_{n,p_{\mathrm{obs}}}$ that

$$\frac{\sum_{i=0}^{M-1} e(\mathcal{S}_i, \mathcal{S}_i^{\mathrm{c}})}{2} \sim \mathsf{Binomial}\left(N_{\boldsymbol{w}}, p_{\mathrm{obs}}\right).$$

Unconditioning on $\mathcal{E}$ in the inequality (93) gives

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{0}}\left\{ \log \frac{\mathrm{d}\mathbb{P}_{\boldsymbol{w}}(\boldsymbol{y})}{\mathrm{d}\mathbb{P}_{\boldsymbol{0}}(\boldsymbol{y})} > 0 \right\} \\
&\le \mathbb{E}\left[\exp\left(-(1-\alpha) \frac{\sum_{i=0}^{M-1} e(\mathcal{S}_i, \mathcal{S}_i^{\mathrm{c}})}{2} D_\alpha^{\min} \right) \right] \\
&= \sum_{l=0}^{N_{\boldsymbol{w}}} \binom{N_{\boldsymbol{w}}}{l} p_{\mathrm{obs}}^l (1-p_{\mathrm{obs}})^{N_{\boldsymbol{w}}-l} \exp\left\{ -l \cdot (1-\alpha)\, D_\alpha^{\min} \right\} \\
&= (1-p_{\mathrm{obs}})^{N_{\boldsymbol{w}}} \sum_{l=0}^{N_{\boldsymbol{w}}} \binom{N_{\boldsymbol{w}}}{l} \left(\frac{p_{\mathrm{obs}}}{1-p_{\mathrm{obs}}} \right)^l \\
&\qquad \cdot \exp\left\{ -l \cdot (1-\alpha)\, D_\alpha^{\min} \right\} \qquad (94) \\
&\overset{(a)}{=} (1-p_{\mathrm{obs}})^{N_{\boldsymbol{w}}} \left(1 + \frac{p_{\mathrm{obs}}}{1-p_{\mathrm{obs}}} \exp\left\{ -(1-\alpha)\, D_\alpha^{\min} \right\} \right)^{N_{\boldsymbol{w}}} \\
&= \left(1 - p_{\mathrm{obs}} + p_{\mathrm{obs}} \exp\left\{ -(1-\alpha)\, D_\alpha^{\min} \right\} \right)^{N_{\boldsymbol{w}}} \\
&\overset{(b)}{\le} \exp\left\{ -N_{\boldsymbol{w}}\, p_{\mathrm{obs}} \left(1 - \exp\left(-(1-\alpha)\, D_\alpha^{\min} \right) \right) \right\}
\end{aligned}$$

$$\overset{\text{(c)}}{=} \exp\left\{-N_{\boldsymbol{w}}\, p_{\text{obs}}\,(1-\alpha)\,\mathsf{Hel}_{\alpha}^{\min}\right\}$$
$$= \exp\left\{-p_{\text{obs}}\,(1-\alpha)\,\mathsf{Hel}_{\alpha}^{\min}\frac{1}{2}\left(n^2-\sum_{i=0}^{M-1}n_i^2\right)\right\}, \qquad (95)$$

where (a) follows from the binomial theorem, (b) relies on the elementary inequality $1-x \le e^{-x}$, (c) comes from the definition (18), and the last line follows since

$$N_{\boldsymbol{w}} = \frac{1}{2}\sum_{i=0}^{M-1}|\mathcal{S}_i|\,(n-|\mathcal{S}_i|) = \frac{1}{2}\sum_{i=0}^{M-1}n_i\,(n-n_i)$$
$$= \frac{1}{2}\left(n^2-\sum_{i=0}^{M-1}n_i^2\right).$$

It remains to control $\sum_{i=0}^{M-1}n_i^2$. Recognize that the input is unique only up to global offset, that is, for any l, the inputs $\boldsymbol{w}$ and $\boldsymbol{w}-l\cdot\mathbf{1}$ result in the same pairwise inputs $\left[w_i-w_j\right]_{1\le i,j\le n}$. Therefore, we assume without loss of generality that[16]

$$k = n_0 \ge \max\{n_1, n_2, \cdots, n_{M-1}\}. \qquad (96)$$

Letting $\rho := \left\lfloor \frac{n}{k} \right\rfloor$, we claim that $\sum_{i=0}^{M-1}n_i^2$ under the constraint $n_i \le k$ is maximized by the configuration

$$\begin{cases} n_0 = n_1 = \cdots = n_{\rho-1} = k, \\ n_\rho = n - k\rho, \\ n_{\rho+1} = \cdots = n_{M-1} = 0, \end{cases}$$

which we will prove by contradiction. Without loss of generality, suppose that the maximizing solution is $n_0 \ge n_1 \ge \cdots \ge n_{M-1}$, and denote by $\tilde{\rho}$ the smallest index such that $n_{\tilde{\rho}} \le k-1$. If $\tilde{\rho} \le \rho-1$, then by replacing $\left(n_{\tilde{\rho}}, n_{\tilde{\rho}+1}\right)$ with $\left(n_{\tilde{\rho}}+1, n_{\tilde{\rho}+1}-1\right)$, we obtain a strictly better feasible solution since

$$\left(n_{\tilde{\rho}}+1\right)^2 + \left(n_{\tilde{\rho}+1}-1\right)^2 = n_{\tilde{\rho}}^2 + n_{\tilde{\rho}+1}^2 + 2\left(n_{\tilde{\rho}} - n_{\tilde{\rho}+1}\right) + 2$$
$$> n_{\tilde{\rho}}^2 + n_{\tilde{\rho}+1}^2.$$

This results in contradiction, and hence $\tilde{\rho} = \rho$. Similarly, we cannot have $n_\rho < n - k\rho$, since replacing $\left(n_\rho, n_{\rho+1}\right)$ with $\left(n_\rho+1, n_{\rho+1}-1\right)$ leads to a strictly better solution. Consequently, for all $\{n_i : 0 \le i < M\}$ satisfying (96), one has

$$\sum_{i=0}^{M-1}n_i^2 \le \left\lfloor \frac{n}{k} \right\rfloor \cdot k^2 + \left(n - k\left\lfloor \frac{n}{k} \right\rfloor\right)^2, \qquad (97)$$

leaving us two cases below to deal with.

Case 1: Suppose that $k \le n/2$. The inequality $n - k\left\lfloor \frac{n}{k} \right\rfloor \le k$ leads to

$$\sum_{i=0}^{M-1}n_i^2 \le \left\lfloor \frac{n}{k} \right\rfloor \cdot k^2 + \left(n - k\left\lfloor \frac{n}{k} \right\rfloor\right)k = nk.$$

[16]Otherwise, if $n_i = \max\{n_1, n_2, \cdots, n_{M-1}\}$ instead, we can always enforce a global shift i on $\boldsymbol{w}$ to yield $\boldsymbol{w} - i\mathbf{1}$ in order to satisfy this condition without affecting the output distribution.

This combined with (95) yields

$$\mathbb{P}_{\mathbf{0}}\left\{\log\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{w}}(\boldsymbol{y})}{\mathrm{d}\mathbb{P}_{\mathbf{0}}(\boldsymbol{y})} > 0\right\}$$
$$\le \exp\left(-\frac{p_{\text{obs}}\left(n^2-nk\right)}{2}(1-\alpha)\,\mathsf{Hel}_{\alpha}^{\min}\right).$$

Employing the union bound over $\mathcal{H}_k$ we obtain

$$P_{\mathsf{e},\mathcal{H}_k} \le \binom{n}{k}(M-1)^{n-k}$$
$$\cdot \exp\left(-\frac{p_{\text{obs}}\left(n^2-nk\right)}{2}(1-\alpha)\,\mathsf{Hel}_{\alpha}^{\min}\right)$$
$$= \exp\left(\log\binom{n}{k} + (n-k)\log(M-1)\right.$$
$$\left. -\frac{p_{\text{obs}}\left(n^2-nk\right)}{2}(1-\alpha)\,\mathsf{Hel}_{\alpha}^{\min}\right)$$

Under the assumption (24), one has

$$(1-\alpha)\,\mathsf{Hel}_{\alpha}^{\min}\cdot p_{\text{obs}}n \ge \log n + 2\log(M-1)$$

for some $0 < \alpha < 1$, which further gives

$$\frac{p_{\text{obs}}\left(n^2-nk\right)}{2}(1-\alpha)\,\mathsf{Hel}_{\alpha}^{\min}$$
$$\ge \frac{(n-k)\log n}{2} + (n-k)\log(M-1).$$

Putting the above computation together yields

$$P_{\mathsf{e},\mathcal{H}_k} \le \exp\left(\log\binom{n}{k} - \frac{(n-k)\log n}{2}\right)$$
$$\overset{\text{(i)}}{\le} 2^n\cdot n^{-\frac{1}{2}(n-k)} \overset{\text{(ii)}}{\le} 2^n\cdot n^{-\frac{1}{4}n} \qquad (98)$$
$$\le C_1 n^{-c_1 n} \qquad (99)$$

for some universal constants $C_1, c_1 > 0$, where (i) uses the fact that $\binom{n}{k} \le 2^n$, (ii) holds since $k \le n/2$, and the last inequality follows since $2^n \ll n^{-\Theta(n)}$. This approaches zero (super)-exponentially fast.

Case 2: We now move on to the case where $k > n/2$. In this regime one has $\left\lfloor \frac{n}{k} \right\rfloor = 1$, and thus (97) gives

$$\sum\nolimits_{i=0}^{M-1}n_i^2 \le k^2 + (n-k)^2 = n^2 - 2k(n-k),$$

This taken collectively with (95) implies that

$$\mathbb{P}_{\mathbf{0}}\left\{\log\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{w}}(\boldsymbol{y})}{\mathrm{d}\mathbb{P}_{\mathbf{0}}(\boldsymbol{y})} > 0\right\} \le \exp\left(-p_{\text{obs}}(1-\alpha)\,k\,(n-k)\,\mathsf{Hel}_{\alpha}^{\min}\right).$$

Apply the union bound over $\mathcal{H}_k$ to deduce that

$$P_{\mathsf{e},\mathcal{H}_k} \le \binom{n}{k}(M-1)^{n-k}$$
$$\cdot \exp\left(-p_{\text{obs}}\,(1-\alpha)\,k\,(n-k)\,\mathsf{Hel}_{\alpha}^{\min}\right)$$
$$= \exp\left(\log\binom{n}{k} + (n-k)\log(M-1)\right. \qquad (100)$$
$$\left. - p_{\text{obs}}\,(1-\alpha)\,k\,(n-k)\,\mathsf{Hel}_{\alpha}^{\min}\right). \qquad (101)$$

For any constant $\delta > 0$, if the minimum Hellinger divergence obeys

$$(1-\alpha)\,\mathsf{Hel}_{\alpha}^{\min}\cdot p_{\mathrm{obs}} n \geq (1+\delta)\log(2n) + 2\log(M-1)$$

for some $0 < \alpha < 1$, then in the regime where $k > n/2$ one has

$$p_{\mathrm{obs}} k\,(n-k)\,(1-\alpha)\,\mathsf{Hel}_{\alpha}^{\min} \geq \frac{(1+\delta)\,k\,(n-k)\log(2n)}{n} + (n-k)\log(M-1).$$

Substitution into (100) gives

$$P_{\mathsf{e},\mathcal{H}_k} \leq \exp\left(\log\binom{n}{k} - \frac{(1+\delta)\,k\,(n-k)}{n}\log(2n)\right),$$

which can be further divided into two cases.

(i) If $k/n > 1-\delta/4$ and $k/n > 1/2$, then the error probability is bounded by

$$\begin{aligned} P_{\mathsf{e},\mathcal{H}_k} &\leq \exp\Big((n-k)\log(2n) - \frac{(1+\delta)\,k}{n}(n-k)\log(2n)\Big) \\ &= \exp\left((n-k)\log(2n)\cdot(1-(1+\delta)\,k/n)\right) \\ &\leq \exp\left((n-k)\log(2n)\right) \cdot\left(1-(1+\delta)\max\left\{1-\frac{\delta}{4},\ \frac{1}{2}\right\}\right) \\ &\leq \exp\left(-\tilde{\delta}\,(n-k)\log(2n)\right), \end{aligned}$$

where $\tilde{\delta} := \max\left\{\frac{3}{4}\delta - \frac{1}{4}\delta^2,\ \frac{\delta-1}{2}\right\}$.

(ii) If $\frac{k}{n} = 1-\tau$ for some $\frac{\delta}{4} \leq \tau \leq \frac{1}{2}$, then

$$P_{\mathsf{e},\mathcal{H}_k} \leq \exp\left(nH(\tau) - (1+\delta)\,n\,(1-\tau)\,\tau\log(2n)\right) \tag{102}$$

$$\leq \exp\left(-n\left((1-\tau)\,\tau\log(2n) - H(\tau)\right)\right) \leq C_2\exp\left(-c_2\delta n\log n\right). \tag{103}$$

for some universal constants $c_2, C_2 > 0$, where (102) makes use of the fact [60, Example 11.1.3] that $\frac{1}{n}\log\binom{n}{k} \leq H(k/n) = H(\tau)$, with $H(\tau)$ denoting the binary entropy function.

Putting the above inequalities together and applying the union bound reveal that

$$\begin{aligned} P_{\mathsf{e}} &\leq \sum_{k=\lceil\frac{n}{M}\rceil}^{n/2} P_{\mathsf{e},\mathcal{H}_k} + \sum_{k=n/2+1}^{\left(1-\frac{\delta}{4}\right)n} P_{\mathsf{e},\mathcal{H}_k} + \sum_{k=\left(1-\frac{\delta}{4}\right)n}^{n-1} P_{\mathsf{e},\mathcal{H}_k} \\ &\leq \frac{n}{2}\cdot C_1 n^{-c_1 n} + \frac{n}{2}\cdot C_2\exp\left(-c_2\delta n\log n\right) + \sum_{k=\left(1-\frac{\delta}{4}\right)n}^{n-1}\exp\left(-\tilde{\delta}\,(n-k)\log(2n)\right) \\ &\leq \frac{n}{2}\cdot C_1 n^{-c_1 n} + \frac{n}{2}\cdot C_2\exp\left(-c_2\delta n\log n\right) + \frac{1}{(2n)^{\tilde{\delta}}}\frac{1}{1-(2n)^{-\tilde{\delta}}} \\ &\leq C_0 e^{-c_0\delta n\log n} + \frac{1}{(2n)^{\tilde{\delta}}-1}. \end{aligned}$$

with $c_0, C_0 > 0$ denoting some universal constants.

Appendix B
Proof of Theorems 3(a) and 6

This section is mainly devoted to proving Theorem 6, which subsumes Theorem 3(a) as a special case. Without loss of generality, assume that the minimum KL divergence can be approached by the following pairs of indices

$$\begin{aligned} &\mathsf{KL}\left(\mathbb{P}_1 \,\|\, \mathbb{P}_0\right) = \mathsf{KL}^{\min}, \\ &\mathsf{KL}\left(\mathbb{P}_l \,\|\, \mathbb{P}_0\right) \leq (1+\zeta)\,\mathsf{KL}^{\min}, \quad 2 \leq l \leq m^{\mathsf{kl}}(\zeta), \end{aligned}$$

and suppose that both the ground truth and the null hypothesis are $\boldsymbol{x} = \boldsymbol{x}^* = \boldsymbol{0}$. We would like to ensure that the observation $\boldsymbol{y}$ conditional on $\boldsymbol{x} = \boldsymbol{0}$ is distinguishable from the observation $\boldsymbol{y}$ under any alternative hypothesis $\boldsymbol{x} \neq \boldsymbol{0}$.

(1) To begin with, recall the definition

$$\mathcal{N}(k\cdot\mathsf{mincut}) := \left\{\mathcal{S}\subseteq\mathcal{V}\colon e\left(\mathcal{S},\mathcal{S}^{\mathrm{c}}\right) \leq k\cdot\mathsf{mincut}\right\}.$$

For each vertex set $\mathcal{S} \in \mathcal{N}(k\cdot\mathsf{mincut})$, we generate one representative hypothesis $\boldsymbol{w}$ such that

$$w_i = \begin{cases} 1, & \text{if } i\in\mathcal{S}, \\ 0, & \text{otherwise.} \end{cases}$$

This produces a collection of $|\mathcal{N}(k\cdot\mathsf{mincut})|$ distinct alternative hypotheses, denoted by $\mathcal{B}_k$. For each $\boldsymbol{w} \in \mathcal{B}_k$, the distributions $\mathbb{P}_{\boldsymbol{w}}$ and $\mathbb{P}_{\boldsymbol{0}}$ disagree only over those locations residing in the associated cut set, which amounts to at most $k\cdot\mathsf{mincut}$ components. It then follows from the independence assumption of y_{ij} that

$$\mathsf{KL}\left(\mathbb{P}_{\boldsymbol{w}} \,\|\, \mathbb{P}_{\boldsymbol{0}}\right) = e\left(\mathcal{S},\mathcal{S}^{\mathrm{c}}\right)\mathsf{KL}^{\min} \leq k\cdot\mathsf{mincut}\cdot\mathsf{KL}^{\min}. \tag{104}$$

Suppose that $k_0 := \arg\max_{k\geq 1}\tau_k^{\mathrm{cut}}$ and fix $0 < \epsilon \leq \frac{1}{2}$. Applying the Fano-type inequality [41, eq. (2.70)] suggests that if

$$\frac{1}{|\mathcal{B}_{k_0}|}\sum_{\boldsymbol{w}\in\mathcal{B}_{k_0}}\mathsf{KL}\left(\mathbb{P}_{\boldsymbol{w}} \,\|\, \mathbb{P}_{\boldsymbol{0}}\right) \leq (1-\epsilon)\log\left|\mathcal{N}(k_0\cdot\mathsf{mincut})\right| - H(\epsilon), \tag{105}$$

then one necessarily has $\inf_{\psi} P_{\mathrm{e}}(\psi) \geq \epsilon$. With (104) and the definition (40) in mind, we see that (105) would follow from

$$\mathsf{KL}^{\min}\cdot k_0\mathsf{mincut} \leq (1-\epsilon)\,k_0\tau^{\mathrm{cut}} - H(\epsilon),$$

which can further be ensured if

$$\mathsf{KL}^{\min}\cdot\mathsf{mincut} \leq (1-\epsilon)\,\tau^{\mathrm{cut}} - H(\epsilon).$$

(2) Next, suppose that the minimum cut is attained by $\left(\mathcal{S}_{\mathsf{mc}},\mathcal{S}_{\mathsf{mc}}^{\mathrm{c}}\right)$. Consider another class $\mathcal{C}$ of hypotheses consisting of m^{kl} hypotheses. The lth candidate $\boldsymbol{w}^{(l)}$ is given by

$$\forall 1 \leq l \leq m^{\mathsf{kl}}: \quad w_i^{(l)} = \begin{cases} l, & \text{if } i\in\mathcal{S}_{\mathsf{mc}}, \\ 0, & \text{otherwise,} \end{cases}$$

all of which obey

$$\mathsf{KL}\left(\boldsymbol{w}^{(l)} \| \boldsymbol{0}\right) \leq (1+\zeta)\,\mathsf{mincut}\cdot\mathsf{KL}^{\min}. \tag{106}$$

Applying the Fano inequality once again, we get $\inf_\psi P_{\mathrm{e}}(\psi) \geq \epsilon$ as long as

$$\frac{1}{m^{\mathsf{kl}}} \sum_{l=1}^{m^{\mathsf{kl}}} \mathsf{KL}\left(\boldsymbol{w}^{(l)} \| \mathbf{0}\right) \leq (1-\epsilon) \log m^{\mathsf{kl}} - H(\epsilon). \quad (107)$$

Observe from (106) that (107) can be ensured under the condition

$$\mathsf{KL}^{\min} \cdot (1+\zeta)\, \mathsf{mincut} \leq (1-\epsilon) \log m^{\mathsf{kl}} - H(\epsilon).$$

(3) Finally, consider the set of configurations with *binary* alphabet having support size 1, i.e. the following $M-1$ classes of hypotheses

$$\mathcal{H}_l := \left\{\boldsymbol{x} \mid \|\boldsymbol{x}\|_0 = 1,\ \boldsymbol{x} \in \{0, l\}^n\right\}, \quad 1 \leq l < M, \quad (108)$$

where each class $\mathcal{H}_l$ is composed of n distinct alternative hypotheses. This guarantees that for any $\boldsymbol{w} \in \mathcal{H}_l$, the distribution of $\left\{y_{ij}\right\} \mid \boldsymbol{x} = \mathbf{0}$ differ from that of $\left\{y_{ij}\right\} \mid \boldsymbol{x} = \boldsymbol{w}$ in at most $d_{\max}$ locations.

For any hypothesis class $\mathcal{H}$ and any $0 < \epsilon < \frac{1}{2}$, the Fano-type inequality [41, eq. (2.70)] suggests that $\inf_\psi P_{\mathsf{e}}(\psi) \geq \epsilon$ occurs as long as

$$\frac{1}{|\mathcal{H}|} \sum_{\boldsymbol{w} \in \mathcal{H}} \mathsf{KL}\left(\mathbb{P}_{\boldsymbol{w}} \| \mathbb{P}_{\mathbf{0}}\right) \leq \frac{|\mathcal{H}|+1}{|\mathcal{H}|} \{(1-\epsilon) \log |\mathcal{H}| - H(\epsilon)\}. \quad (109)$$

By picking $\mathcal{H}$ to be $\mathcal{H} = \bigcup_{l=1}^{m^{\mathsf{kl}}} \mathcal{H}_l$—which obeys $|\mathcal{H}| = m^{\mathsf{kl}} n$—we can see from definition of m^{kl} that

$$\frac{1}{|\mathcal{H}|} \sum_{\boldsymbol{w} \in \mathcal{H}} \mathsf{KL}\left(\mathbb{P}_{\boldsymbol{w}} \| \mathbb{P}_{\mathbf{0}}\right) \leq (1+\zeta)\, d_{\max} \mathsf{KL}^{\min}$$

and hence (109) would hold under the condition

$$\mathsf{KL}^{\min} \cdot (1+\zeta)\, d_{\max} \leq (1-\epsilon)\left(\log n + \log m^{\mathsf{kl}}\right) - H(\epsilon). \quad (110)$$

Putting the above results together establishes Theorem 6.

We now specialize to Theorem 3(a), which follows immediately from (110). Specifically, for an Erdős–Rényi graph $\mathcal{G} \sim \mathcal{G}_{n, p_{\mathrm{obs}}}$, the Chernoff-type inequality [61, Th. 4.4 and 4.5] indicates that for any $\epsilon > 0$,

$$d_{\max} \leq (1+\epsilon)\, n p_{\mathrm{obs}} \quad (111)$$

holds with probability exceeding $1 - n^{-10}$, provided that $p_{\mathrm{obs}} > \frac{c \log n}{n}$ for some sufficiently large constant $c > 0$. Substitution into (110) immediately leads to Theorem 3(a).

Appendix C
Proof of Theorems 3(b) and 7

We start with the proof of Theorem 7, which accounts for a much broader context than Theorem 3(b). In similar spirit of Theorem 6, assume that the minimum Hellinger divergence is achieved by the following pair of indices

$$\mathsf{Hel}_\alpha\left(\mathbb{P}_1 \| \mathbb{P}_0\right) = \mathsf{Hel}_\alpha^{\min},$$

and let the ground truth and the null hypothesis be $\boldsymbol{x} = \boldsymbol{x}^* = \mathbf{0}$.

For any class $\mathcal{H}$ of alternative hypotheses, the minimax lower bound [62, Th. II.1] suggests that every f-divergence $D_f(\cdot)$ obeys

$$\sum_{\boldsymbol{w} \in \mathcal{H}} D_f\left(\mathbb{P}_{\boldsymbol{w}} \| \mathbb{P}_{\mathbf{0}}\right) \geq f\left(|\mathcal{H}|\,(1-P_{\mathrm{e}})\right) + (|\mathcal{H}|-1)\, f\left(\frac{|\mathcal{H}|\, P_{\mathrm{e}}}{|\mathcal{H}|-1}\right),$$

where $\mathbb{P}_{\boldsymbol{w}}$ is the probability measure of $\left[y_{ij}\right]_{(i,j)\in\mathcal{E}}$ conditional on $\boldsymbol{x} = \boldsymbol{w}$. When specialized to the Hellinger divergence of order α (which corresponds to $f(x) = \frac{1}{1-\alpha}(1-x^\alpha)$), the above inequality leads to

$$\begin{aligned}
&(1-\alpha) \sum_{\boldsymbol{w} \in \mathcal{H}} \mathsf{Hel}_\alpha\left(\mathbb{P}_{\boldsymbol{w}} \| \mathbb{P}_{\mathbf{0}}\right) \\
&\geq 1 - |\mathcal{H}|^\alpha (1-P_{\mathrm{e}})^\alpha + (|\mathcal{H}|-1)\left\{1 - \left(\frac{|\mathcal{H}|\, P_{\mathrm{e}}}{|\mathcal{H}|-1}\right)^\alpha\right\} \\
&= |\mathcal{H}| - |\mathcal{H}|^\alpha (1-P_{\mathrm{e}})^\alpha - (|\mathcal{H}|-1)^{1-\alpha} |\mathcal{H}|^\alpha P_{\mathrm{e}}^\alpha \\
&\geq |\mathcal{H}| - |\mathcal{H}|^\alpha - |\mathcal{H}|\, P_{\mathrm{e}}^\alpha.
\end{aligned}$$

Put another way,

$$P_{\mathrm{e}}^\alpha \geq 1 - \frac{(1-\alpha)\sum_{\boldsymbol{w}\in\mathcal{H}} \mathsf{Hel}_\alpha\left(\mathbb{P}_{\boldsymbol{w}} \| \mathbb{P}_{\mathbf{0}}\right)}{|\mathcal{H}|} - \frac{1}{|\mathcal{H}|^{1-\alpha}}. \quad (112)$$

Notably, for any product measures $P^n = P \times P \times \cdots \times P$ and $Q^n = Q \times Q \times \cdots \times Q$, the Hellinger divergence satisfies the decoupling equality

$$\begin{aligned}
1 - (1-\alpha)\, \mathsf{Hel}_\alpha\left(P^n \| Q^n\right) &= \int \left(\mathrm{d}P^n\right)^\alpha \left(\mathrm{d}Q^n\right)^{1-\alpha} \\
&= \left(\int (\mathrm{d}P)^\alpha (\mathrm{d}Q)^{1-\alpha}\right)^n \quad (113) \\
&= \left(1 - (1-\alpha)\, \mathsf{Hel}_\alpha\left(P \| Q\right)\right)^n. \quad (114)
\end{aligned}$$

If all hypotheses $\boldsymbol{w} \in \mathcal{H}$ satisfy $\|\boldsymbol{w} - \boldsymbol{x}^*\|_0 \leq k$, then $\mathbb{P}_{\boldsymbol{w}}$ and $\mathbb{P}_{\mathbf{0}}$ are different over at most $k d_{\max}$ locations. Thus, if the divergence measure at each of these locations is identical and equal to some given value h_α, then it follows from the independence assumption of y_{ij} that

$$1 - (1-\alpha)\, \mathsf{Hel}_\alpha\left(\mathbb{P}_{\boldsymbol{w}} \| \mathbb{P}_{\mathbf{0}}\right) \geq \left(1 - (1-\alpha)\, h_\alpha\right)^{k d_{\max}}.$$

This together with (112) suggests that: as long as $(1-\alpha)\, h_\alpha \leq \frac{1}{2}$, one necessarily has

$$\begin{aligned}
P_{\mathrm{e}}^\alpha &\geq \left(1 - (1-\alpha)\, h_\alpha\right)^{k d_{\max}} - |\mathcal{H}|^{-(1-\alpha)} \\
&\geq e^{-\left((1-\alpha)h_\alpha + (1-\alpha)^2 h_\alpha^2\right) k d_{\max}} - |\mathcal{H}|^{-(1-\alpha)},
\end{aligned}$$

which results from the inequality that $\log(1-x) \geq -x - x^2$ for any $0 \leq x \leq 1/2$.

As a consequence, if the following condition holds

$$e^{-\left((1-\alpha)h_\alpha + (1-\alpha)^2 h_\alpha^2\right) k d_{\max}} - |\mathcal{H}|^{-(1-\alpha)} \geq \xi^\alpha$$

or, equivalently,

$$(1-\alpha)\, h_\alpha\left[1 + (1-\alpha)\, h_\alpha\right] \leq -\frac{\log\left(\xi^\alpha + |\mathcal{H}|^{-(1-\alpha)}\right)}{k d_{\max}}, \quad (115)$$

then the minimax probability of error must exceed $\inf_\psi P_{\mathrm{e}} \geq \xi$. Solving the quadratic inequality (115) and utilizing the fact $\sqrt{1+4x}-1 \geq 2x-4x^2$ $(x \geq 0)$, we see that (115) would follow as long as

$$(1-\alpha)\, h_\alpha \leq -\frac{\log\left(\xi^\alpha + |\mathcal{H}|^{-(1-\alpha)}\right)}{k d_{\max}} - \frac{2\log^2\left(\xi^\alpha + |\mathcal{H}|^{-(1-\alpha)}\right)}{(k d_{\max})^2}. \tag{116}$$

Finally, setting $\xi = n^{-\epsilon}$ and $\mathcal{H} = \mathcal{H}_1$ (cf. Definition (108)), one has $|\mathcal{H}| = n$, $k = 1$ and $h_\alpha = \mathsf{Hel}_\alpha^{\min}$. In the regime where

$$\epsilon \leq \frac{1-\alpha}{\alpha} \quad \Longleftrightarrow \quad \alpha \leq \frac{1}{1+\epsilon},$$

we have

$$\xi^\alpha = n^{-\epsilon\alpha} \geq n^{-(1-\alpha)} = |\mathcal{H}|^{-(1-\alpha)}.$$

The condition (116) is then guaranteed to hold if

$$(1-\alpha)\, \mathsf{Hel}_\alpha^{\min} \leq -\frac{\log(2\xi^\alpha)}{d_{\max}} - \frac{2\log^2(2\xi^\alpha)}{d_{\max}^2}, \tag{117}$$

which would follow if

$$(1-\alpha)\, \mathsf{Hel}_\alpha^{\min} \leq \frac{\epsilon\alpha\log n - \log 2}{d_{\max}} - \frac{2\left[\epsilon\alpha\log n - \log 2\right]^2}{d_{\max}^2}, \tag{118}$$

where we have used $\xi^\alpha = n^{-\epsilon\alpha}$. Besides, the condition $(1-\alpha)\, h_\alpha \leq \frac{1}{2}$ becomes $(1-\alpha)\, \mathsf{Hel}_\alpha^{\min} \leq \frac{1}{2}$, which can be ensured under (118) together with the condition

$$\frac{\epsilon\alpha\log n}{d_{\max}} \leq \frac{1}{2}$$

as claimed.

Finally, recall that when $\mathcal{G} \sim \mathcal{G}_{n,p_{\mathrm{obs}}}$, one has $d_{\max} \leq (1+\epsilon)p_{\mathrm{obs}}n$ as long as $np_{\mathrm{obs}}/\log n$ is sufficiently large. Plugging this into the preceding bound completes the proof of Theorem 3(b).

Appendix D
Proof of Theorem 4

Note that ψ_{ml} distinguishes the null hypothesis $\boldsymbol{x} = \boldsymbol{x}^* = \{x_i^*\}_{1\leq i\leq n}$ from the alternative hypothesis $\boldsymbol{x} = \boldsymbol{w} = \{w_i\}_{1\leq i\leq n}$ only based on those components (i, j) where

$$x_i^* - x_j^* \neq w_i - w_j,$$

and its recovery capability depends only on the distinction of output distributions over these locations. For ease of presentation, we will suppose in the rest of the proof that both the ground truth and the null hypothesis are $\boldsymbol{x} = \boldsymbol{0}$, but note that the proof carries over to all other ground truth values.

Let's divide the set of all alternative hypotheses into several classes $\mathcal{A}_k$ so that for each $k \geq 1$,

$$\mathcal{A}_k := \{\boldsymbol{w} \neq \boldsymbol{0} : \ |\mathcal{E} \cap \mathrm{supp}(\boldsymbol{w} \ominus \boldsymbol{w})| < k \cdot \mathsf{mincut}\}, \tag{119}$$

where we employ the notation

$$\mathcal{E} \cap \mathrm{supp}(\boldsymbol{w} \ominus \boldsymbol{w}) := \left\{(i,j) \in \mathcal{E} \mid w_i - w_j \neq 0, \quad i > j\right\}.$$

Apparently, any cut set cannot contain more than n^2 edges, and hence $\mathcal{A}_k = \emptyset$ for any $k \geq n^2/\mathsf{mincut}$. For any $\boldsymbol{w} \in \mathcal{A}_k$, if we let $\mathcal{S}_l$ represent the set of vertices taking the value l, then by definition of $\mathcal{A}_k$ one has

$$\sum_{l=0}^{M-1} e\left(\mathcal{S}_l, \mathcal{S}_l^{\mathrm{c}}\right) = 2\,|\mathcal{E} \cap \mathrm{supp}(\boldsymbol{w} \ominus \boldsymbol{w})| < 2k \cdot \mathsf{mincut}. \tag{120}$$

On the other hand, consider the case where $k = 1$. All $\boldsymbol{w} \in \mathcal{A}_1$ are equivalent to $\boldsymbol{0}$ up to some global offset. This is because for any non-trivial cut $(\mathcal{S}_l, \mathcal{S}_l^{\mathrm{c}})$, one must have $|\mathcal{E} \cap \mathrm{supp}(\boldsymbol{w} \ominus \boldsymbol{w})| \geq e\left(\mathcal{S}_l, \mathcal{S}_l^{\mathrm{c}}\right) \geq \mathsf{mincut}$, which violates the feasibility constraint $|\mathcal{E} \cap \mathrm{supp}(\boldsymbol{w} \ominus \boldsymbol{w})| < \mathsf{mincut}$. In the following lemma, we link the cardinality of each hypothesis class $\mathcal{A}_k$ with the cut-homogeneity exponent τ^{cut} defined in (40).

Lemma 3: For any $k \leq n^2/\mathsf{mincut}$, the hypothesis class $\mathcal{A}_k$ defined in (119) satisfies

$$\frac{\log|\mathcal{A}_k|}{k} < 2\log M + 2\log(2k \cdot \mathsf{mincut}) + 4\tau^{\mathrm{cut}} \leq 2\log M + 4\log(2n) + 4\tau^{\mathrm{cut}}. \tag{121}$$

Proof: See Appendix G. □

We are now in position to characterize the recovery ability of ψ_{ml}. Let $\mathbb{P}_{\boldsymbol{w}}(\cdot)$ denote the measure given $\boldsymbol{x} = \boldsymbol{w}$. For any $0 < \alpha < 1$, it follows from (93) that

$$\mathbb{P}_{\boldsymbol{0}}\left\{\log\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{w}}(\boldsymbol{y})}{\mathrm{d}\mathbb{P}_{\boldsymbol{0}}(\boldsymbol{y})} > 0\right\} \leq \exp\left(-(1-\alpha)\,\frac{\sum_{i=0}^{M-1} e\left(\mathcal{S}_i, \mathcal{S}_i^{\mathrm{c}}\right)}{2} D_\alpha^{\min}\right), \tag{122}$$

When restricted to the hypotheses in $\mathcal{A}_k \backslash \mathcal{A}_{k-1}$ for any $2 \leq k \leq n^2/\mathsf{mincut}$, we know from the definition of $\mathcal{A}_k$ that

$$(k-1)\,\mathsf{mincut} \leq |\mathcal{E} \cap \mathrm{supp}(\boldsymbol{w} \ominus \boldsymbol{w})| = \frac{\sum_{i=0}^{M-1} e\left(\mathcal{S}_i, \mathcal{S}_i^{\mathrm{c}}\right)}{2} < k\mathsf{mincut}.$$

It then follows from the union bound that

$$\mathbb{P}_{\boldsymbol{0}}\left\{\exists \boldsymbol{w} \in \mathcal{A}_k \backslash \mathcal{A}_{k-1} : \ \log\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{w}}(\boldsymbol{y})}{\mathrm{d}\mathbb{P}_{\boldsymbol{0}}(\boldsymbol{y})} > 0\right\}$$

$$\leq |\mathcal{A}_k| \exp\left(-(1-\alpha)\,\frac{\sum_{i=0}^{M-1} e\left(\mathcal{S}_i, \mathcal{S}_i^{\mathrm{c}}\right)}{2} D_\alpha^{\min}\right) \tag{123}$$

$$\leq \exp\left(-(k-1)\left((1-\alpha)\, D_\alpha^{\min}\mathsf{mincut} - \frac{k}{k-1}\frac{\log|\mathcal{A}_k|}{k}\right)\right)$$

$$\leq \exp\left(-(k-1)\left((1-\alpha)\, D_\alpha^{\min}\mathsf{mincut} - \frac{2\log|\mathcal{A}_k|}{k}\right)\right) \tag{124}$$

$$\leq \exp\left\{-(k-1)\left((1-\alpha)\, D_\alpha^{\min} \cdot \mathsf{mincut} - \left(4\log M + 8\log(2n) + 8\tau^{\mathrm{cut}}\right)\right)\right\}, \tag{125}$$

where (125) results from Lemma 3. This suggests that if there is some $0 < \alpha < 1$ obeying

$$(1-\alpha)\, D_\alpha^{\min} \cdot \mathsf{mincut} \geq (\delta+8)\log(2n) + 8\tau^{\mathrm{cut}} + 4\log M,$$

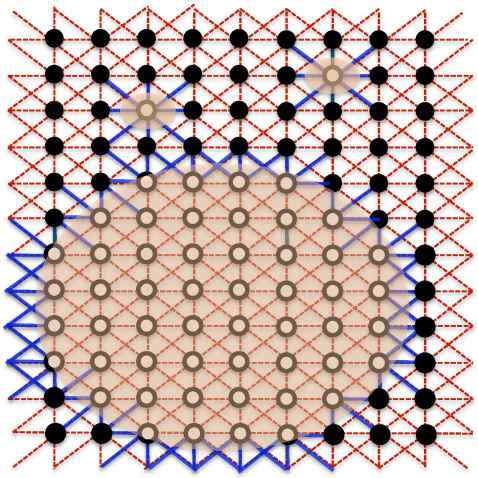

Fig. 5. An example of the cut $(\mathcal{S}, \mathcal{S}^{\mathrm{c}})$ in a geometric graph. Here, $\mathcal{S}$ consists of all black vertices, while $\mathcal{S}^{\mathrm{c}}$ contains all white vertices. The blue solid edges represent the cut edges.

then one achieves

$$\begin{aligned} P_{\mathrm{e}}\left(\psi_{\mathrm{ml}}\right) &\leq \sum_{k=2}^{\frac{n^2}{\mathsf{mincut}}} \mathbb{P}_{\mathbf{0}}\left\{\exists \boldsymbol{w} \in \mathcal{A}_k \backslash \mathcal{A}_{k-1}: \ \log \frac{\mathrm{d}\mathbb{P}_{\boldsymbol{w}}(\boldsymbol{y})}{\mathrm{d}\mathbb{P}_{\mathbf{0}}(\boldsymbol{y})} > 0\right\} \\ &\leq \sum_{k\geq 2} \exp\Big\{-(k-1)\Big((1-\alpha) D_{\alpha}^{\min} \cdot \mathsf{mincut} \\ &\qquad - \left(4\log M + 8\log(2n) + 8\tau^{\mathrm{cut}}\right)\Big)\Big\} \\ &\leq \sum_{k\geq 1} \exp\left(-k\cdot\delta\log(2n)\right) \\ &\leq \frac{1}{(2n)^{\delta}} \cdot \frac{1}{1-(2n)^{-\delta}} = \frac{1}{(2n)^{\delta}-1}. \end{aligned}$$

To finish up, recognizing that $D_{\alpha}^{\min} \geq \mathsf{Hel}_{\alpha}^{\min}$ immediately establishes the recovery condition in terms of $\mathsf{Hel}_{\alpha}^{\min}$.

Appendix E
Proof of Lemma 1

(1) Define the *cut-edge degree* of a vertex v to be the number of edges in $\mathcal{E}(\mathcal{S}, \mathcal{S}^{\mathrm{c}})$ that v is incident to. Consider any cut $(\mathcal{S}, \mathcal{S}^{\mathrm{c}})$ with size

$$e\left(\mathcal{S}, \mathcal{S}^{\mathrm{c}}\right) \leq k \cdot \mathsf{mincut}. \tag{126}$$

We shall separate all vertices into two types as follows:

- *Type-1 vertex*: any vertex whose cut-edge degree is at least $\frac{1}{2}\kappa\rho \cdot \mathsf{mincut}$;
- *Type-2 vertex*: any vertex whose cut-edge degree is less than $\frac{1}{2}\kappa\rho \cdot \mathsf{mincut}$.

For ease of presentation, we will color all vertices in $\mathcal{S}$ black and all vertices in $\mathcal{S}^{\mathrm{c}}$ white; each feasible coloring scheme thus corresponds to one valid cut $(\mathcal{S}, \mathcal{S}^{\mathrm{c}})$ in $\mathcal{N}(k \cdot \mathsf{mincut})$.

To develop some intuitive understanding of the above notions, we depict in Fig. 5 an example of a cut $(\mathcal{S}, \mathcal{S}^{\mathrm{c}})$ in a geometric graph, where $\mathcal{S}^{\mathrm{c}}$ consists of all vertices residing within the shaded area, and the blue solid edges indicate the cut edges. Typically, type-1 vertices, which are incident to many cut edges, are lying on or close to the boundary of the cut. In Fig. 5, these correspond to those vertices lying around the boundary of the shaded area in addition to those singleton white vertices. In contrast, type-2 vertices often refer to those staying away from the cut boundary (e.g. those white nodes in the center of the shaded area). It may be useful to keep this figure in mind when reading about the subsequent proof.

To prove Lemma 1, we start by examining how many combinations of type-1 vertices are feasible and how many ways there are to color them. By definition, for any cut obeying (126), the number V_1 of type-1 vertices satisfies $V_1 \leq \frac{2e(\mathcal{S},\mathcal{S}^{\mathrm{c}})}{\frac{1}{2}\kappa\rho\mathsf{mincut}} \leq \frac{4k}{\kappa\rho}$ (note that each edge is incident to two vertices and might be counted twice). Simple combinatorial arguments thus suggest that there are at most $n^{4k/\kappa\rho}$ distinct ways to pick these type-1 vertices, and then no more than $2^{V_1} \leq 2^{4k/\kappa\rho}$ ways to color all these type-1 vertices, and finally at most $\binom{2e(\mathcal{S},\mathcal{S}^{\mathrm{c}})}{V_1} \leq (2k\cdot\mathsf{mincut})^{V_1} \leq (2k\cdot\mathsf{mincut})^{4k/\kappa\rho}$ different combinations of cut-edge degrees among them. Taken together these counting arguments imply that there exist no more than[17]

$$n^{4k/\kappa\rho}\,(2k\cdot\mathsf{mincut})^{4k/\kappa\rho}\,2^{4k/\kappa\rho} < (2n)^{8k/\kappa\rho}$$

distinct ways to select the set of type-1 vertices as well as assign colors and cut-edge degrees for each of them, if one is required to satisfy the cut size constraint (126).

We claim that for any cut $(\mathcal{S}, \mathcal{S}^{\mathrm{c}})$ obeying (126), once the following three pieces of information are gathered:

(i) which vertices are type-1 vertices,
(ii) the cut-edge degrees of these type-1 vertices,
(iii) the colors of these type-1 vertices (i.e. whether they belong to $\mathcal{S}$ or $\mathcal{S}^{\mathrm{c}}$),

then the colors of all remaining vertices (and hence all information about this cut) can be uniquely determined. Following the preceding pictorial interpretation, the whole point of this claim is to demonstrate that as long as some appropriate conditions regarding the cut boundary is known, then one can figure out all remaining cut information. To establish this claim, we shall consider the following two cases separately. Without loss of generality, the following discussion concentrates only on *black* type-1 vertices.

- **Case 1.** Consider any vertex v whose color has been revealed to be black, and whose cut-edge degree does *not* exceed

$$\left(1-\frac{1}{2}\kappa\right)\rho\cdot\mathsf{mincut}, \tag{127}$$

namely, v is connected with no more than $\left(1-\frac{1}{2}\kappa\right)\rho\cdot\mathsf{mincut}$ white vertices. For any of its neighbors u (i.e. $(u,v)\in\mathcal{E}$), if the color of u has not been revealed, then we claim that it must be black. To see this, suppose instead that u is white, then from the above connectivity assumption (127) of v, the number of black vertices that

[17]Here, we use the fact that $2k\cdot\mathsf{mincut} \leq n^2$, and hence $(2k\cdot\mathsf{mincut})^{4k/\kappa\rho} \leq n^{8k/\kappa\rho}$.

u is linked with is at least

$$|\mathcal{V}(u)\cap\mathcal{V}(v)|-\left(1-\frac{1}{2}\kappa\right)\rho\cdot\mathsf{mincut}$$
$$\geq \rho\mathsf{mincut}-\left(1-\frac{1}{2}\kappa\right)\rho\cdot\mathsf{mincut}=\frac{1}{2}\kappa\rho\mathsf{mincut},$$

where the inequality follows from Assumption (42). This means that u must be a type-1 vertex (cf. definition of type-1 vertices) and its color must have been revealed, thus resulting in contradiction. In summary, all neighbors of v with unknown colors are necessarily black.

- **Case 2.** Consider any vertex v whose color has been revealed to be black, and whose cut-edge degree is known to be larger than $\left(1-\frac{1}{2}\kappa\right)\rho\cdot\mathsf{mincut}$. Again, consider any of its neighbors u whose color remains unknown, which must be incident to fewer than $\frac{1}{2}\kappa\rho\cdot\mathsf{mincut}$ cut edges since by construction it is a type-2 vertex. This already suggests the following fact: if there are at least $\frac{1}{2}\kappa\rho\cdot\mathsf{mincut}$ vertices falling in $\mathcal{V}(u)\cap\mathcal{V}(v)$ known to be white (resp. black), then the color of u must be white (resp. black), since by definition a type-2 vertex cannot be connected to $\frac{1}{2}\kappa\rho\cdot\mathsf{mincut}$ vertices of opposite color. As a result, we can uniquely determine the color of u unless
  - (**P1**) the colors of fewer than $\kappa\rho\cdot\mathsf{mincut}$ vertices[18] in $\mathcal{V}(u)\cap\mathcal{V}(v)$ have been revealed.

  This remaining situation is the subject of the discussion below.

  Suppose that the true color of u is black. Recall that u is a type-2 vertex and hence it is connected to fewer than $\frac{1}{2}\kappa\rho$ white vertices. From Assumption (42) and the condition $\kappa<\frac{1}{2}$, any white neighbor w of u must be connected with at least

$$|\mathcal{V}(u)\cap\mathcal{V}(w)|-\frac{1}{2}\kappa\rho\cdot\mathsf{mincut}\geq\left(1-\frac{1}{2}\kappa\right)\rho\mathsf{mincut}$$
$$\geq\frac{1}{2}\kappa\rho\cdot\mathsf{mincut}$$

  black vertices falling within $\mathcal{V}(u)\cap\mathcal{V}(w)$, and hence w must be a type-1 vertex and its color has necessarily been identified. Similarly, if u is white, then the colors of *all* black vertices surrounding u must have been revealed. As a result, all vertices in $\mathcal{V}(u)$ with unknown colors must be of the same color as u. That being said, as long as one can identify the color of one extra vertex in $\mathcal{V}(u)\cap\mathcal{V}(v)$, then the color of u and all remaining vertices in $\mathcal{V}(u)\cap\mathcal{V}(v)$ can be uniquely determined.

  Now let w be the uncolored vertex in $\mathcal{V}(u)\cap\mathcal{V}(v)$ that is the nearest to v, which by (P1) must be within the ($\kappa\rho\mathsf{mincut}$) closest vertices to v in $\mathcal{V}(u)\cap\mathcal{V}(v)$. From Assumption (43), we see that w must be connected to all but $\frac{1}{2}\rho\cdot\mathsf{mincut}$ neighbors surrounding v and, as a result, be connected to at least

$$\text{cut-degree}(v)-|\mathcal{V}(v)\setminus\mathcal{V}(w)|$$
$$\geq\left(1-\frac{1}{2}\kappa\right)\rho\cdot\mathsf{mincut}-\frac{1}{2}\rho\cdot\mathsf{mincut}$$
$$=\frac{1}{2}(1-\kappa)\rho\cdot\mathsf{mincut}\geq\frac{1}{2}\kappa\rho\cdot\mathsf{mincut}$$

  white vertices since $\kappa<\frac{1}{2}$, where cut-degree(v) represents the cut-edge degree of v. Therefore, if w is black, then it has to be a type-1 vertex, which is contradictory, and we have determined it to be white.

Putting the above two cases together indicates that all vertices that are connected to the set of type-1 vertices can be uniquely colored, and we shall use $\mathcal{V}_{\text{new}}$ to denote them. If there still exist uncolored vertices, a nonempty subset of them must be connected to $\mathcal{V}_{\text{new}}$. Since all vertices in $\mathcal{V}_{\text{new}}$ are type-2 vertices and have cut-degrees not exceeding $\frac{1}{2}\kappa\rho\mathsf{mincut}\leq\left(1-\frac{1}{2}\kappa\right)\rho\mathsf{mincut}$, repeating the arguments in Case 1 allows us to determine the color of all vertices surrounding $\mathcal{V}_{\text{new}}$. This step further shrinks the size of the uncolored set. Repeating this argument until all vertices are colored, we establish the claim. All in all, we have thus demonstrated that the number of feasible coloring schemes is bounded above by $(2n)^{8k/\kappa\rho}$, which in turn justifies

$$\tau_k^{\text{cut}}\leq\frac{8\log(2n)}{\kappa\rho},\quad\forall k\geq 1.$$

(2) If $\mathcal{G}$ is an expander graph with edge expansion $h_{\mathcal{G}}$, then for any vertex set $\mathcal{S}$ with $|\mathcal{S}|\leq\frac{n}{2}$, one has

$$|\mathcal{S}|\leq e\left(\mathcal{S},\mathcal{S}^{\text{c}}\right)/h_{\mathcal{G}}\tag{128}$$

from the definition of $h_{\mathcal{G}}$. For any $d>0$, if one requires that

$$e\left(\mathcal{S},\mathcal{S}^{\text{c}}\right)\leq kd,\tag{129}$$

then the above inequality leads to

$$|\mathcal{S}|\leq kd/h_{\mathcal{G}},$$

indicating that there are at most $2\binom{n}{\lfloor\frac{kd}{h_{\mathcal{G}}}\rfloor}\leq 2n^{kd/h_{\mathcal{G}}}$ feasible cuts $(\mathcal{S},\mathcal{S}^{\text{c}})$ satisfying (129). Setting $d=\mathsf{mincut}$ immediately leads to

$$|\mathcal{N}(k\cdot\mathsf{mincut})|\leq 2n^{k\mathsf{mincut}/h_{\mathcal{G}}},$$
$$\Rightarrow\tau_k^{\text{cut}}=\frac{1}{k}\log|\mathcal{N}(k\cdot\mathsf{mincut})|\leq\frac{\mathsf{mincut}\log n}{h_{\mathcal{G}}}+\frac{\log 2}{k}$$

for all $k\geq 1$, as claimed.

Appendix F
Proof of Lemma 2

We begin with explicit expressions of the divergence measures. For any $k\neq l$ and $p\in[0,1]$, one has

$$\mathsf{KL}\left(p\delta_k+(1-p)\text{Unif}_M\parallel p\delta_l+(1-p)\text{Unif}_M\right)$$
$$=\left(p+\frac{1-p}{M}\right)\log\left(\frac{p+\frac{1-p}{M}}{\frac{1-p}{M}}\right)+\frac{1-p}{M}\log\left(\frac{\frac{1-p}{M}}{p+\frac{1-p}{M}}\right)\tag{130}$$

$$=p\log\left(\frac{(M-1)p+1}{1-p}\right),\tag{131}$$

[18] Otherwise there are either $\frac{1}{2}\kappa\rho\mathsf{mincut}$ white vertices or $\frac{1}{2}\kappa\rho\mathsf{mincut}$ black colors in $\mathcal{E}(u)\cap\mathcal{E}(v)$ with their colors revealed.

where δ_k denotes the Dirac measure on the point k, and (130) follows since the two distributions under study differ only at two points $x = k$ and $x = l$. Similarly, one obtains (cf. Definition (4))

$$\begin{aligned}&\mathsf{Hel}_{\frac{1}{2}}\left(p\delta_k + (1-p)\mathrm{Unif}_M \parallel p\delta_l + (1-p)\mathrm{Unif}_M\right)\\ &= 2\left(\sqrt{p+\frac{1-p}{M}} - \sqrt{\frac{1-p}{M}}\right)^2 \qquad (132)\\ &= \frac{2}{M}\left(\sqrt{(M-1)\,p+1} - \sqrt{1-p}\right)^2. \qquad (133)\end{aligned}$$

When applied to the random corruption model, these suggest

$$\mathsf{KL}^{\min} = p_{\mathrm{true}}\log\left(1+\frac{p_{\mathrm{true}}M}{1-p_{\mathrm{true}}}\right) \le \frac{p_{\mathrm{true}}^2 M}{1-p_{\mathrm{true}}}, \quad (134)$$

and

$$\mathsf{Hel}_{\frac{1}{2}}^{\min} = \frac{2}{M}\left(\sqrt{1-p_{\mathrm{true}}+Mp_{\mathrm{true}}} - \sqrt{1-p_{\mathrm{true}}}\right)^2. \quad (135)$$

It remains to control the Hellinger divergence. To this end, the elementary identity $a-b = \frac{a^2-b^2}{a+b}$ gives

$$\begin{aligned}&\left(\sqrt{1-p_{\mathrm{true}}+Mp_{\mathrm{true}}} - \sqrt{1-p_{\mathrm{true}}}\right)^2\\ &= \left(\frac{p_{\mathrm{true}}M}{\sqrt{1-p_{\mathrm{true}}+Mp_{\mathrm{true}}} + \sqrt{1-p_{\mathrm{true}}}}\right)^2\\ &\ge \left(\frac{p_{\mathrm{true}}M}{2\sqrt{1-p_{\mathrm{true}}+Mp_{\mathrm{true}}}}\right)^2 = \frac{p_{\mathrm{true}}^2M^2}{4\left(1-p_{\mathrm{true}}+Mp_{\mathrm{true}}\right)},\end{aligned}$$

indicating that $\mathsf{Hel}_{\frac{1}{2}}^{\min} \ge \frac{p_{\mathrm{true}}^2 M}{2(1-p_{\mathrm{true}}+Mp_{\mathrm{true}})}$ as claimed.

Appendix G
Proof of Lemma 3

Consider any hypothesis $\boldsymbol{x} = \boldsymbol{w} \in \mathcal{A}_k$, which obeys $|\mathcal{E}\cap\mathrm{supp}(\boldsymbol{w}\ominus\boldsymbol{w})| < k\cdot\mathsf{mincut}$. Denote by $\mathcal{S}_l$ the set of vertices that take the value l $(0 \le l < M)$, and let $\mathcal{I}_{\neg\emptyset} := \{l \mid \mathcal{S}_l \neq \emptyset\}$ represent the indices of those non-empty ones. Our proof proceeds by evaluating the following quantities:

1) How many distinct choices of $\mathcal{I}_{\neg\emptyset}$ are admissible?
2) For each given $\mathcal{I}_{\neg\emptyset}$, how many combinations of cut-set sizes $\left\{e\left(\mathcal{S}_l,\mathcal{S}_l^{\mathrm{c}}\right) \mid l\in\mathcal{I}_{\neg\emptyset}\right\}$ are feasible?
3) For each given cut-set size N_l, how many cuts $\left(\mathcal{S}_l,\mathcal{S}_l^{\mathrm{c}}\right)$ are compatible with the constraint $e\left(\mathcal{S}_l,\mathcal{S}_l^{\mathrm{c}}\right) \le N_l$?

Clearly, multiplying all these quantities together gives rise to an upper bound on $|\mathcal{A}_k|$.

We now compute the above quantities separately.

- To begin with, our assumption on the min-cut size ensures that
$$e\left(\mathcal{S}_l,\mathcal{S}_l^{\mathrm{c}}\right) \ge \mathsf{mincut} \quad (136)$$
for each non-empty $\mathcal{S}_l$. This together with the feasibility constraint
$$2\,|\mathcal{E}\cap\mathrm{supp}(\boldsymbol{w}\ominus\boldsymbol{w})| = \sum_{l=0}^{M-1} e\left(\mathcal{S}_l,\mathcal{S}_l^{\mathrm{c}}\right) \le 2k\cdot\mathsf{mincut} \quad (137)$$
guarantees that the number of non-empty $\mathcal{S}_l$'s cannot exceed $2k$. Consequently, there exist at most M^{2k} possible combinations of $\mathcal{I}_{\neg\emptyset}$.
- Secondly, from (137), the total cut-set size is bounded above by $2k\cdot\mathsf{mincut}$. Therefore, for any given $\mathcal{I}_{\neg\emptyset}$, there are no more than
$$\binom{2k\mathsf{mincut}}{|\mathcal{I}_{\neg\emptyset}|} \le (2k\cdot\mathsf{mincut})^{2k}$$
feasible ways to assign cut-set sizes $e\left(\mathcal{S}_l,\mathcal{S}_l^{\mathrm{c}}\right)$ for all $l\in\mathcal{I}_{\neg\emptyset}$.
- Thirdly, suppose that for each $l\in\mathcal{I}_{\neg\emptyset}$,
$$e\left(\mathcal{S}_l,\mathcal{S}_l^{\mathrm{c}}\right) = c_l\cdot\mathsf{mincut} \quad (138)$$
for some numerical values $c_l \ge 1$. From the definition (40), the number of feasible choices of $\left(\mathcal{S}_l,\mathcal{S}_l^{\mathrm{c}}\right)$ compatible with (138) is bounded above by
$$\begin{aligned}|\mathcal{N}(c_l\cdot\mathsf{mincut})| &\le |\mathcal{N}(\lceil c_l\rceil\,\mathsf{mincut})| \le \exp\left(\lceil c_l\rceil\,\tau^{\mathrm{cut}}\right)\\ &\le \exp\left(2c_l\tau^{\mathrm{cut}}\right).\end{aligned}$$
Recognize that the constraint (137) requires
$$\sum_l c_l < 2k.$$
As a result, when the cut sizes $e\left(\mathcal{S}_l,\mathcal{S}_l^{\mathrm{c}}\right)$ are given, the total number of valid partitions $\{\mathcal{S}_l \mid 0\le l<M\}$ cannot exceed
$$\prod_{l=0}^{M-1}\exp\left(2c_l\tau^{\mathrm{cut}}\right) < \exp\left(4k\tau^{\mathrm{cut}}\right). \quad (139)$$

Putting the above combinatorial bounds together implies that

$$|\mathcal{A}_k| < M^{2k}\,(2k\cdot\mathsf{mincut})^{2k}\exp\left(4k\tau^{\mathrm{cut}}\right).$$

Using the inequality $k\mathsf{mincut} \le n^2$ we conclude the proof.

Appendix H
Proof of Fact 1

Recall that KL divergence and Hellinger divergence are both f-divergence associated with the *non-negative* convex functions $f_1(x) = x\log x - x + 1$ and $f_2(x) = \left(\sqrt{x}-1\right)^2$, respectively. That said, one can write

$$\mathsf{KL}(P\parallel Q) = \mathbb{E}_Q\left[f_1\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right]$$

and

$$\mathsf{Hel}_{\frac{1}{2}}(P\parallel Q) = \mathbb{E}_Q\left[f_2\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right].$$

One can verify that the function f_1 can be uniformly bounded above using f_2 in the following way:

$$\left(2-0.5\,|\log x|\right)f_2(x) \le f_1(x) \le \left(2+|\log x|\right)f_2(x), \qquad \forall x>0.$$

This immediately establish that

$$\begin{aligned}\mathsf{KL}\left(P \parallel Q\right) &= \mathbb{E}_Q\left[f_1\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right] \\ &\le \left(2+\log R\right)\mathbb{E}_Q\left[f_2\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right] \\ &= \left(2+\log R\right)\mathsf{Hel}_{\frac{1}{2}}\left(P \parallel Q\right)\end{aligned}$$

and

$$\begin{aligned}\mathsf{KL}\left(P \parallel Q\right) &= \mathbb{E}_Q\left[f_1\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right] \\ &\ge \left(2-0.5\log R\right)\mathbb{E}_Q\left[f_2\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right] \\ &= \left(2-0.5\log R\right)\mathsf{Hel}_{\frac{1}{2}}\left(P \parallel Q\right).\end{aligned}$$

These together with the well known inequality [41, Lemma 2.4]

$$\mathsf{KL}\left(P \parallel Q\right) \ge \mathsf{Hel}_{\frac{1}{2}}\left(P \parallel Q\right)$$

establish (20).

Similarly, from the inequality

$$\left(2-0.4\left|\log x\right|\right) f_2\left(x\right) \le f_1\left(x\right) \le \left(2+0.4\left|\log x\right|\right) f_2\left(x\right)$$

for all $x \in (0, 4.5]$, one can show that

$$\mathsf{KL}\left(P \parallel Q\right) \le \max\left\{2-0.4\log R,\ 1\right\} \cdot \mathsf{Hel}_{\frac{1}{2}}\left(P \parallel Q\right)$$

and

$$\mathsf{KL}\left(P \parallel Q\right) \le \left(2+0.4\log R\right) \cdot \mathsf{Hel}_{\frac{1}{2}}\left(P \parallel Q\right)$$

as long as $R \le 4.5$, as claimed.

ACKNOWLEDGMENTS

Y. Chen would like to thank Emmanuel Abbe, Afonso Bandeira, Amit Singer for inspiring discussion on synchronization and graphical channels, Jiaming Xu for stimulating discussion on hypothesis testing, Amir Dembo for his valuable instruction on large and moderate deviation theory. The authors thank Jonathan Scarlett and I-Hsiang Wang for helpful discussions and suggestions.

REFERENCES

[1] Y. Chen and A. J. Goldsmith, "Information recovery from pairwise measurements," in *Proc. Int. Symp. Inf. Theory*, Jun. 2014, pp. 2012–2016.
[2] Y. Chen, C. Suh, and A. J. Goldsmith, "Information recovery from pairwise measurements: A Shannon-theoretic approach," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 2336–2340.
[3] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, nos. 3–5, pp. 75–174, 2010.
[4] A. Jalali, Y. Chen, S. Sanghavi, and H. Xu. (2011). "Clustering partially observed graphs via convex optimization." [Online]. Available: http://arxiv.org/abs/1104.4803
[5] Y. Chen, S. Sanghavi, and H. Xu, "Improved graph clustering," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6440–6455, Oct. 2014.
[6] Y. Chen and E. Candes, "Nonconvex joint alignment," to be published.
[7] D. J. Crandall, A. Owens, N. Snavely, and D. P. Huttenlocher, "SfM with MRFs: Discrete-continuous optimization for large-scale structure from motion," in *Proc. IEEE CVPR*, Jun. 2011, pp. 3001–3008.
[8] C. Zach, M. Klopschitz, and M. Pollefeys, "Disambiguating visual relations using loop constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1426–1433.
[9] M. Cucuringu, A. Singer, and D. Cowburn, "Eigenvector synchronization, graph rigidity and the molecule problem," *Inf. Inference*, vol. 1, no. 1, pp. 21–67, 2012.
[10] L. Wang and A. Singer. (2012). "Exact and stable recovery of rotations for robust synchronization." [Online]. Available: https://arxiv.org/abs/1211.2441
[11] A. Bandeira, M. Charikar, A. Singer, and A. Zhu, "Multireference alignment using semidefinite programming," in *Proc. ITCS*, 2014, pp. 459–470.
[12] Y. Chen, L. Guibas, and Q. Huang, "Near-optimal joint object matching via convex relaxation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2014, pp. 100–108.
[13] D. Pachauri, R. Kondor, and V. Singh, "Solving the multi-way matching problem by permutation synchronization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1860–1868.
[14] Q.-X. Huang and L. Guibas, "Consistent shape maps via semidefinite programming," *Comput. Graph. Forum*, vol. 32, no. 5, pp. 177–186, 2013.
[15] J. Yan, Y. Li, W. Liu, H. Zha, X. Yang, and S. M. Chu, "Graduated consistency-regularized optimization for multi-graph matching," in *European Conference on Computer Vision*. Zurich, Switzerland: Springer, 2014, pp. 407–422.
[16] S. Browning and B. Browning, "Haplotype phasing: Existing methods and new developments," *Nature Rev. Genetics*, vol. 12, no. 10, pp. 703–714, 2011.
[17] D. He, A. Choi, K. Pipatsrisawat, A. Darwiche, and E. Eskin, "Optimal algorithms for haplotype assembly from whole-genome sequence data," *Bioinformatics*, vol. 26, no. 12, pp. i183–i190, 2010.
[18] N. Donmez and M. Brudno, "Hapsembler: An assembler for highly polymorphic genomes," in *Research in Computational Molecular Biology*. Springer, 2011, pp. 38–52.
[19] Y. Chen, G. Kamath, C. Suh, and D. Tse, "Community recovery in graphs with locality," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2016, pp. 689–698.
[20] H. Si, H. Vikalo, and S. Vishwanath, "Haplotype assembly: An information theoretic view," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2014, pp. 182–186.
[21] G. Kamath, E. Sasoglu, and D. Tse. (2015). "Optimal haplotype assembly from high-throughput mate-pair reads." [Online]. Available: https://arxiv.org/abs/1502.01975
[22] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.
[23] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *J. Mach. Learn. Res.*, vol. 99, pp. 2057–2078, Jul. 2010.
[24] E. J. Candes and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.
[25] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, Jun. 2011.
[26] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM J. Optim.*, vol. 21, no. 2, pp. 572–596, 2011.
[27] E. Abbe, A. S. Bandeira, and G. Hall. (2014). "Exact recovery in the stochastic block model." [Online]. Available: https://arxiv.org/abs/1405.3267
[28] E. Mossel, J. Neeman, and A. Sly. (2014). "Consistency thresholds for binary symmetric block models." [Online]. Available: https://arxiv.org/abs/1407.1591
[29] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer. (2014). "Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery." [Online]. Available: https://arxiv.org/abs/1404.4749
[30] E. Abbe and A. Montanari, "Conditional random fields, planted constraint satisfaction and entropy concentration," *Theory Comput.*, vol. 11, no. 17, pp. 413–443, 2015.
[31] E. Abbe and A. Montanari, "The mutual information of a class of graphical channels," in *Proc. Allerton*, Oct. 2013, pp. 20–25.
[32] A. Globerson, T. Roughgarden, D. Sontag, and C. Yildirim. (2014). "Tight error bounds for structured prediction." [Online]. Available: https://arxiv.org/abs/1409.5834
[33] A. Globerson, T. Roughgarden, D. Sontag, and C. Yildirim, "How hard is inference for structured prediction?" in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1–10.
[34] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.

[35] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1644–1656, Apr. 2013.
[36] S. Chen, R. Varma, A. Sandryhaila, and J. Kovačević, "Discrete signal processing on graphs: Sampling theory," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6510–6523, Dec. 2015.
[37] R. Durrett, *Random Graph Dynamics*. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[38] M. Penrose, *Random Geometric Graphs*, vol. 5. Oxford, U.K.: Oxford Univ. Press, 2003.
[39] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4394–4412, Oct. 2006.
[40] F. Topsoe, "Some inequalities for information divergence and related measures of discrimination," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1602–1609, Jul. 2000.
[41] A. B. Tsybakov and V. Zaiats, *Introduction to Nonparametric Estimation*, vol. 11. New York, NY, USA: Springer, 2009.
[42] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," in *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2835–2885, May 2014.
[43] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, vol. 1. 1961, pp. 547–561.
[44] T. Van Erven and P. Harremos, "Rényi divergence and Kullback–Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.
[45] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, no. 6, p. 066138, 2004.
[46] I. Sason and S. Verdú. (2015). "Bounds among f-divergences." [Online]. Available: https://arxiv.org/abs/1508.00335
[47] S. Dragomir, "Upper and lower bounds for Csiszar f-divergence in terms of Hellinger discrimination and applications," in *Proc. Nonlinear Anal. Forum*, vol. 7. 2002, pp. 1–14.
[48] B. Hajek, Y. Wu, and J. Xu. (2014). "Achieving exact cluster recovery threshold via semidefinite programming." [Online]. Available: https://arxiv.org/abs/1412.6156
[49] I. Csiszàr, "Sanov property, generalized I-projection and a conditional limit theorem," *Ann. Probab.*, vol. 12, no. 3, pp. 768–793, 1984.
[50] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, vol. 2. New York, NY, USA: Springer, 1998.
[51] E. Mossel, J. Neeman, and A. Sly. (2012). "Stochastic block models and reconstruction." [Online]. Available: https://arxiv.org/abs/1202.1499
[52] Y. Chen and J. Xu. (2014). "Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices." [Online]. Available: https://arxiv.org/abs/1402.1267
[53] A. Montanari and S. Sen. (2015). "Semidefinite programs on sparse random graphs and their application to community detection." [Online]. Available: http://arxiv.org/abs/1504.05910
[54] A. S. Bandeira. (2015). "Random Laplacian matrices and convex relaxations." [Online]. Available: http://arxiv.org/abs/1504.03987
[55] J. Lei and L. Zhu. (2014). "A generic sample splitting approach for refined community recovery in stochastic block models." [Online]. Available: https://arxiv.org/abs/1411.1469
[56] S.-Y. Yun and A. Proutiere. (2014). "Accurate community detection in the stochastic block model via spectral algorithms." [Online]. Available: http://arxiv.org/abs/1412.7335
[57] P. Chin, A. Rao, and V. Vu. (2015). "Stochastic block model and community detection in the sparse graphs: A spectral algorithm with optimal rate of recovery." [Online]. Available: https://arxiv.org/abs/1501.05021
[58] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. (2015). "Achieving optimal misclassification proportion in stochastic block model." [Online]. Available: https://arxiv.org/abs/1505.03772
[59] E. Abbe and C. Sandon. (2015). "Community detection in general stochastic block models: Fundamental limits and efficient recovery algorithms." [Online]. Available: https://arxiv.org/abs/1503.00609
[60] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 2006.
[61] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[62] A. Guntuboyina, "Lower bounds for the minimax risk using f-divergences, and applications," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2386–2399, Apr. 2011.

Yuxin Chen (S'09) received the B.S. in Microelectronics with High Distinction from Tsinghua University in 2008, the M.S. in Electrical and Computer Engineering from the University of Texas at Austin in 2010, the M.S. in Statistics from Stanford University in 2013, and the Ph.D. in Electrical Engineering from Stanford University in 2015. He is currently a postdoctoral scholar in the Department of Statistics at Stanford University. His research interests include high-dimensional structured estimation, convex and nonconvex optimization, information theory, and network science.

Changho Suh (S'10–M'12) is an Ewon Associate Professor in the School of Electrical Engineering at Korea Advanced Institute of Science and Technology (KAIST) since 2012. He received the B.S. and M.S. degrees in Electrical Engineering from KAIST in 2000 and 2002 respectively, and the Ph.D. degree in Electrical Engineering and Computer Sciences from UC-Berkeley in 2011. From 2011 to 2012, he was a postdoctoral associate at the Research Laboratory of Electronics in MIT. From 2002 to 2006, he had been with the Telecommunication R&D Center, Samsung Electronics.

Dr. Suh received the 2015 Hadong Young Engineer Award from the Institute of Electronics and Information Engineers, the 2013 Stephen O. Rice Prize from the IEEE Communications Society, the David J. Sakrison Memorial Prize from the UC-Berkeley EECS Department in 2011, and the Best Student Paper Award of the IEEE International Symposium on Information Theory in 2009.

Andrea J. Goldsmith (S'90–M'93–SM'99–F'05) is the Stephen Harris professor of Electrical Engineering at Stanford University. She was previously on the faculty of Electrical Engineering at Caltech. Her research interests are in information theory and communication theory, and their application to wireless communications and related fields. Dr. Goldsmith co-founded and served as CTO for two wireless companies: Accelera, Inc., which develops software-defined wireless network technology for cloud-based management of WiFi access points, and Quantenna Communications, Inc., which develops high-performance WiFi chipsets. She has also held industry positions at Maxim Technologies, Memorylink Corporation, and AT&T Bell Laboratories. She is a Fellow of the IEEE and of Stanford, and has received several awards for her work, including the IEEE ComSoc Edwin H. Armstrong Achievement Award as well as Technical Achievement Awards in Communications Theory and in Wireless Communications, the National Academy of Engineering Gilbreth Lecture Award, the IEEE ComSoc and Information Theory Society Joint Paper Award, the IEEE ComSoc Best Tutorial Paper Award, the Alfred P. Sloan Fellowship, and the Silicon Valley/San Jose Business Journal's Women of Influence Award. She is author of the book "Wireless Communications" and co-author of the books "MIMO Wireless Communications" and "Principles of Cognitive Radio," all published by Cambridge University Press, as well as an inventor on 28 patents. She received the B.S., M.S. and Ph.D. degrees in Electrical Engineering from U.C. Berkeley.

Dr. Goldsmith has served as editor for the IEEE TRANSACTIONS ON INFORMATION THEORY, the *Journal on Foundations and Trends in Communications and Information Theory and in Networks*, the IEEE TRANSACTIONS ON COMMUNICATIONS, and the *IEEE Wireless Communications Magazine* as well as on the Steering Committee for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. She participates actively in committees and conference organization for the IEEE Information Theory and Communications Societies and has served on the Board of Governors for both societies. She has also been a Distinguished Lecturer for both societies, served as President of the IEEE Information Theory Society in 2009, founded and chaired the student committee of the IEEE Information Theory society, and chaired the Emerging Technology Committee of the IEEE Communications Society. At Stanford she received the inaugural University Postdoc Mentoring Award, served as Chair of Stanfords Faculty Senate in 2009, and currently serves on its Faculty Senate, Budget Group, and Task Force on Women and Leadership.