Information Recovery from Pairwise Measurements: 
A Shannon-Theoretic Approach 

Yuxin Chen, Changho Suh, and Andrea J. Goldsmith * 

April 9, 2015 

Abstract 
This paper is concerned with jointly recovering $n$ node-variables $\{x_i\}_{1 \leq i \leq n}$ from a collection of pairwise difference measurements. Specifically, several noisy measurements of $x_i - x_j$ are acquired. 

This is represented by a graph with an edge set $\mathcal{E}$ such that $x_i - x_j$ is observed only if $(i,j) \in \mathcal{E}$. The distributions of the noisy measurements are modeled by a set of channels with given input/output transition measures. Using information-theoretic tools applied to the channel decoding problem, we develop a unified framework to characterize the region where information recovery is possible, which accommodates general graph structures, alphabet sizes, and channel transition measures. In particular, we isolate and highlight a family of minimum distance measures underlying the channel transition probabilities, which plays a central role in determining the recovery limits. We develop both lower and upper bounds on the fundamental recovery limits in terms of these minimum information distance measures; both bounds scale inversely with the min-cut size. For a very broad class of homogeneous graphs, our results are tight up to some explicit constant factor, and depend only on the graph edge sparsity irrespective of other second-order graphical metrics like the spectral gap. We develop consequences of our general theory for three applications, including the stochastic block model, the outlier model, and the haplotype assembly problem. Our theory leads to order-wise optimal recovery conditions for all these cases, and even improves upon existing results in some regimes. 

Index Terms: pairwise difference measurements, Kullback–Leibler divergence, Hellinger divergence, Rényi divergence, random graphs, geometric graphs 

1 Introduction 
In various data processing scenarios, one wishes to acquire information about a large collection of independent 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. Fortunately, such pairwise observations often carry a significant amount of information across all objects of interest. As a result, reliable joint information recovery becomes feasible as soon as a sufficiently large number of pairwise measurements are obtained. 

This paper explores a general class of pairwise measurements, which we term pairwise difference measurements. Consider $n$ variables $x_1, \cdots, x_n$, and suppose we obtain independent measurements of the differences $x_i - x_j$ over a few pairs $(i,j)$. This pairwise difference functional can be represented by a 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, we model the noisy measurements $y_{ij}$’s as the output of the following equivalent channel: 

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

as illustrated in Fig. 1. Here, the distribution of the output $y_{ij}$ is specified solely by the associated channel input $x_i - x_j$, with $p(\cdot | \cdot)$ representing the channel transition probability. The goal is to recover $\mathbf{x} = \{x_1, \cdots, x_n\}$ based on these channel outputs $y_{ij}$’s.

*Y. Chen is with the Department of Statistics, Stanford University, Stanford, CA 94305, USA (email: yxchen@stanford.edu). C. Suh is with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea (e-mail: chsuh@kaist.ac.kr). A. J. Goldsmith is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA (email: andrea@wsl.stanford.edu). This paper will be presented in part at the International Symposium on Information Theory (ISIT), 2015.

1Here, “−” represents some algebraic subtraction operation (broadly defined), as we detail in Section 2.
Figure 1: Equivalent channel model. The measurement models are represented by a graph $G$ with an edge set $E$ such that $x_i - x_j$ are measured if and only if $(i, j) \in E$. Each $x_i - x_j$ is independently passed through a channel with transition probability $p(y_{ij} | x_i - x_j)$, resulting in the noisy measurement $y_{ij}$.

As long as $G$ is connected, the ground truth $\mathbf{x}$ is uniquely determined by the pairwise difference functional $\{x_i - x_j \mid (i, j) \in E\}$, up to some global offset. Therefore, information recovery is identical to decoding the input of the equivalent channel $1$ based on $y_{ij}$’s. Intuitively, faithful decoding is feasible only when (i) the graph $G$ is sufficiently connected so that we have enough measurements involving each node variable, and (ii) the channel output distributions given any two distinct inputs are sufficiently separated and hence distinguishable; such separation is often quantified based on some distance metrics underlying the channel transition measures.

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

- **Community Detection and Graph Clustering.** Various real-world networks exhibit community structures $1$, 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 problem, sometimes referred to as graph clustering (e.g. $2$), is clearly a special instance of the pairwise difference model.

- **Alignment, Registration and Synchronization.** Consider $n$ views of a single scene from different angles and positions. 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 $3,4$, spectroscopy imaging and structural biology $5,6$, and multi-reference alignment $7$.

- **Joint Graph Matching.** Given $n$ isomorphic sets, one needs to globally map common features across all these sets. Existing pairwise matching methods $8,9,10$ allow us to map shared features between a pair of images in isolation. For a collection of images, a fundamental problem amounts to how to refine these noisy pairwise matches in order to obtain globally consistent matches. This problem arises in numerous applications in computer vision and graphics, solving jigsaw puzzles, etc.

- **Haplotype 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 $12,14$, which amounts to reconstructing the haplotypes based on disagreement between pairs of single reads $15,16$ – a special case of the pairwise measurement model with binary alphabet.

Many of these practical applications have witnessed a flurry of activity in algorithm development, which are based primarily on computational considerations. For instance, inspired by recent success in spectral methods $17$ and semidefinite relaxation $18$ (particularly those developed for matrix recovery problems),
many provably efficient algorithms have been proposed for graph clustering [2], joint matching [10,11], synchronization [6], and so on. These algorithms are shown to enjoy intriguing recovery guarantees under simple randomized models, although the choice of performance metrics is often studied in a model-specific manner. On the other hand, there have been several information theoretic results in place for a few applications, e.g. stochastic block model [19] and synchronization [20]. 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 that dictate the success of exact information recovery from pairwise difference measurements?

2. If so, can we characterize the fundamental limits – in terms of these channel distance metrics – such that perfect recovery is feasible only above the limits?

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

1.1 Main Contributions

The main contribution of this paper is towards a unified characterization of fundamental information recovery limits, by means of both information-theoretic and graph-theoretic tools. In particular, we single out and highlight a family of minimum channel distance measures (i.e. the minimum Kullback–Leibler (KL), Hellinger, and Rényi divergence), as well as two graphical metrics (i.e. the minimum cut and the cut-homogeneity exponent defined in Section 3), that play central roles in determining the feasibility of exact recovery. Based on these metrics, we develop a sufficient and a necessary condition for information recovery such that the minimax probability of error tends to zero as \( n \) grows. These results 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 factor) for a very broad class of homogeneous graphs, subsuming as special cases Erdős–Rényi models, random geometric graphs, and many other expander graphs. In addition, we develop tighter bounds for the Erdős–Rényi model – the most widely considered random graph model.

Our results demonstrate that the fundamental recovery limits, presented in terms of the minimum channel distance measures, scale with the cut-homogeneity exponent and are inversely proportional to the minimum cut size. To interpret this, the minimum channel distance together with the min-cut size captures the amount of information one has available to differentiate the two closest input signals. Somewhat surprisingly, for a broad class of homogeneous graphs, the minimum distance criterion depends solely on the vertex degrees of \( G \), which improves upon some prior results (e.g. [19]) that rely on additional second-order graphical metrics (e.g. the spectral gap or the Cheeger’s constant).

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), which 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 outlier 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.

1.2 Related Work

On the fundamental limits side, most prior works only focused on binary input and output alphabets. Among them, Abbe et al. [19] 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 the mean time, Si et al. [15] and Kamath et al. [16] determined the information-theoretic limits for a similar setup motivated from genome sequencing, which correspond to random graphs and (generalized) ring graphs,

3
respectively. A sufficient recovery condition for general graphs has also been derived in [19], although it was not guaranteed to be order optimal. Our preliminary work [20] explored the fundamental recovery limits under general alphabets and graph structures, but was restricted to the simplistic outlier 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 [15,16,19,20] can all be treated as a special type of “graphical channels” as defined by Abbe and Montanari [21,22], which refer 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 information recovery limits, which was previously out of reach. On the other hand, a recent interesting work [23] explored partial recovery under binary alphabets, accounting for a variety of graphs (including two-dimensional grids). A more general theory regarding partial recovery will be left for future work.

1.3 Terminology and Notation

Graph Terminology. Let \( \deg(v) \) represent the degree of a vertex \( v \). For any two vertex sets \( S_1 \) and \( S_2 \), denote by \( E(S_1, S_2) \) (resp. \( e(S_1, S_2) \)) the set (resp. the number) of edges with exactly one endpoint in \( S_1 \) and another in \( S_2 \). Below we introduce several widely used (random) graph models; see [24–26] and the references therein for in-depth discussion.

1. Erdős–Rényi graph. An Erdős–Rényi graph of \( n \) vertices, denoted by \( G_{n,p} \), 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 \( G_{n,r} \), is generated via a 2-step procedure: i) place \( n \) vertices uniformly and independently on the surface of a unit sphere\(^2\); 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 \( e(S, S^c) \geq h_\mathcal{G} |S| \) for all vertex set \( S \) satisfying \( |S| \leq n/2 \).

Divergence Measures. Our results are established upon a family of divergence measures including KL divergence and Hellinger divergence of order \( \alpha \) \([27,28]\). Formally, for any two probability measures \( P \) and \( Q \), the KL divergence of \( Q \) from \( P \) is defined and denoted as

\[
\text{KL}(P \parallel Q) := \int dP \log \left( \frac{dP}{dQ} \right).
\] (2)

whereas the Hellinger divergence of order \( \alpha \in (0,1) \) of \( Q \) from \( P \) is defined to be \([27]\)

\[
\text{Hel}_\alpha(P \parallel Q) := \frac{1}{1 - \alpha} \left[ 1 - \int (dP)^\alpha (dQ)^{1-\alpha} \right].
\] (3)

When \( \alpha = \frac{1}{2} \), this reduces to the so-called squared Hellinger distance\(^3\)

\[
\text{Hel}_\frac{1}{2}(P \parallel Q) = 2 - 2 \sqrt{dP} \sqrt{dQ} = \int (\sqrt{dP} - \sqrt{dQ})^2.
\] (4)

Besides, the \( \chi^2 \) divergence is defined as

\[
\chi^2(P \parallel Q) = \int \left( \frac{dP}{dQ} - 1 \right)^2 dQ.
\] (5)

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

\[
\text{KL}(p \parallel q) = \text{KL}(P \parallel Q), \quad \text{Hel}_\alpha(p \parallel q) = \text{Hel}_\alpha(P \parallel Q), \quad \text{and} \quad \chi^2(p \parallel q) = \chi^2(P \parallel Q).
\] (6)

\(^2\)We consider \( G_{n,r} \) on a unit sphere instead of \([0,1]^2\) to eliminate edge effects.

\(^3\)Several other sources introduce a prefactor of 1/2 in order to normalize the squared Hellinger distance, resulting in the definition \( \int \frac{1}{2}(\sqrt{dP} - \sqrt{dQ})^2 \). Here, we adopt the unnormalized version as defined in [29, Section 2.4].
More generally, the $f$-divergence of $Q$ from $P$ is defined as

$$ D_f (P \| Q) := \int f \left( \frac{dP}{dQ} \right) dQ $$

for any convex function $f(\cdot)$ such that $f(1) = 0$ [27, 28]. In particular, the Hellinger divergence $\text{Hel}_\alpha (P \| Q)$, the KL divergence, and the $\chi^2$ divergence are special cases of $f$-divergence generated by $f(x) = \frac{1}{\alpha} (1 - x^\alpha)$, $f(x) = x \log x$ (or $f(x) = x \log x - x + 1$), and $f(x) = (x - 1)^2$, respectively.

Finally, we introduce the Rényi divergence of positive order $\alpha$, where $\alpha \neq 1$, of a distribution $P$ from a distribution $Q$ as [30, 31]

$$ D_\alpha (P \| Q) : = - \frac{1}{1 - \alpha} \log \left( \int (dP)^\alpha (dQ)^{1-\alpha} \right) $$

$$ = - \frac{1}{1 - \alpha} \log (1 - (1 - \alpha) \text{Hel}_\alpha). $$

It follows from the elementary inequality $1 - x \leq e^{-x}$ that $D_\alpha (P \| Q) \geq \text{Hel}_\alpha (P \| Q)$. This together with the monotonicity of $D_\alpha$ [31, Theorem 3] gives

$$ \text{Hel}_\alpha (P \| Q) \leq D_\alpha (P \| Q) \leq \text{KL}(P \| Q), \quad 0 < \alpha < 1. $$

Other Notation. Let 1 and 0 be the all-one and all-zero vectors, respectively. We denote by $\text{supp}(x)$ and $\|x\|$ the support and the support size of $x$, respectively. The standard notion $f(n) = o(g(n))$ means $\lim_{n \to \infty} f(n)/g(n) = 0$; $f(n) = \Omega(g(n))$ or $f(n) \gtrsim g(n)$ mean $\exists$ a constant $c$ such that $f(n) \geq cg(n)$; $f(n) = \tilde{O}(g(n))$ or $f(n) \lesssim g(n)$ mean $\exists$ a constant $c$ such that $f(n) \leq cg(n)$; $f(n) = \Theta(g(n))$ or $f(n) \asymp g(n)$ mean $\exists$ constants $c_1$ and $c_2$ such that $c_1 g(n) \leq f(n) \leq c_2 g(n)$. Throughout this paper, we use $\log(\cdot)$ to represent the natural logarithm.

1.4 Organization

The remainder of the paper is organized as follows. We first present the formal problem setup in Section 2. The key channel distance measures as well as the graphical metrics are introduced and discussed in Section 3. Section 4 presents the non-asymptotic sufficient and necessary recovery conditions in the most general form, which accommodate general alphabets, graph structures, and channel characteristics. We then turn to more refined recovery conditions for the Erdős–Rényi model in Section 5. To illustrate the effectiveness of our framework, we develop some concrete consequences of the general theorems for a few specific examples in Section 6. Finally, Section 7 concludes the paper with a summary of our findings and a discussion of future directions. The proofs of the main results and the auxiliary lemmas are deferred to the Appendix.

2 Problem Formulation

Consider $n$ vertices $V = \{1, \cdots, n\}$, each represented by a node-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 “+”. For any $x_i, x_j \in \mathcal{X}$, the pairwise difference operation is defined as

$$ x_i - x_j := x_i + (-x_j), $$

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

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

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 $x_i - x_j \pmod{M}$ is a valid example of (11).
2. Relative rotation: set $x_i = R_i$ for some rotation matrix $R_i$ and let “+” denote matrix multiplication. Then $x_i - x_j$ stands for $R_i R_j^{-1}$, which represents the relative rotation between $i$ and $j$. Hence, it is a special case of (11).

3. Pairwise matching: if we set $x_i$ to be some permutation matrix $\Pi_i$ and let “+” be matrix multiplication, then the pairwise matching, captured by $\Pi_i \Pi_j^T$, also belongs to the pairwise difference model.

- Measurement Graph and Channel Model. The pairwise difference measurements are represented by a measurement graph $\mathcal{G}$ that comprises an edge set $\mathcal{E}$. 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 noisy channel, whose output $y_{ij}$ follows the conditional distribution

$$p \left( y_{ij} \bigg| x_i - x_j = l \right) = \mathbb{P}_l (y_{ij}), \quad 0 \leq l < M,$$

where $\mathbb{P}_l (\cdot)$ denotes the transition measure that maps a given input $l$ to the output space. It is assumed that the family $\{ \mathbb{P}_l \mid 0 \leq l < M \}$ satisfies the symmetry property$^4$.

$$\text{KL} (\mathbb{P}_{l+i} \parallel \mathbb{P}_i) = \text{KL} (\mathbb{P}_{l-i} \parallel \mathbb{P}_i), \quad 0 \leq l, i < M;$$

$$\text{Hel}_\alpha (\mathbb{P}_{l+i} \parallel \mathbb{P}_i) = \text{Hel}_\alpha (\mathbb{P}_{l-i} \parallel \mathbb{P}_i), \quad 0 \leq l, i < M, 0 < \alpha < 1;$$

$$\text{D}_\alpha (\mathbb{P}_{l+i} \parallel \mathbb{P}_i) = \text{D}_\alpha (\mathbb{P}_{l-i} \parallel \mathbb{P}_i), \quad 0 \leq l, i < M, 0 < \alpha < 1;$$

The output alphabet $\mathcal{Y}$ can either be continuous or discrete. In contrast to conventional information theory settings, no coding is employed across channel uses. In the discrete setting, these divergence measures can be efficiently estimated even under large alphabet – see, e.g. [32] and their subsequent work.

This paper centers on exact information recovery, that is, to reconstruct each input $x = \{x_1, \cdots, x_n\}$ up to global offset. This is all one can hope for since the input $x$ and its shifted version $x + l \cdot 1 = \{x_1 + l, \cdots, x_n + l\}$ will result in identical (and hence indistinguishable) output $y := \{y_{ij} \in \mathcal{Y} \mid (i, j) \in \mathcal{E}\}$. In light of this, we introduce the zero-one distance modulo a global offset factor as follows

$$\text{dist} (w, x) := 1 - \max_{0 \leq l < M} \mathbb{I} \{w = x + l \cdot 1\},$$

where $\mathbb{I}$ is the indicator function. Apparently, $\text{dist} (w, x) = 0$ holds for the set of all solutions that differ from $x$ only by a global shift term. Equipped with this distance metric, we define, for any recovery procedure $\psi : \mathcal{Y}^{|\mathcal{E}|} \rightarrow \mathcal{X}^n$, the probability of error as

$$P_e (\psi) := \max_{x \in \mathcal{X}^n} \mathbb{P} \left( \text{dist} (\psi (y), x) \neq 0 \mid x \right).$$

We aim to characterize the regime where the minimax probability of error $\inf_{\psi} P_e (\psi)$ tends to 0.

3 Preliminaries: Key Metrics

Before proceeding to the main results, we pause to introduce a few channel distance measures and graphical metrics, which will prove critical in the subsequent development of our theory.

3.1 Key Distance Metrics on Channel Transition Measures

There are several channel distance measures that dictate the hardnes of information recovery. Specifically, we emphasize the minimum KL, Hellinger, and Rényi divergence with respect to the channel transition measures defined as

$$\text{KL}^{\min} \quad := \quad \min_{l \neq l'} \text{KL} (\mathbb{P}_l \parallel \mathbb{P}_{l'});$$

$$\text{Hel}_\alpha^{\min} \quad := \quad \min_{l \neq l'} \text{Hel}_\alpha (\mathbb{P}_l \parallel \mathbb{P}_{l'});$$

$$\text{D}_\alpha^{\min} \quad := \quad \min_{l \neq l'} \text{D}_\alpha (\mathbb{P}_l \parallel \mathbb{P}_{l'}) = - \frac{1}{1 - \alpha} \log \left( 1 - (1 - \alpha) \text{Hel}_\alpha^{\min} \right).$$

$^4$Here, with a slight abuse of notation, we let $\mathbb{P}_i = \mathbb{P}_i \mod M$ for any $i \not\in \{0, 1, \cdots, M-1\}$.  

6
These minimum distance measures capture the distinguishability of outputs given minimally separated inputs. In addition, for any $\epsilon > 0$, we define
\[
m^{kl}(\epsilon) := \max_i \left\{ i \mid i \neq l, \ KL(\mathbb{P}_i \parallel \mathbb{P}_l) \leq (1 + \epsilon) KL^{\min} \right\}, \tag{22}
\]
which clearly satisfies
\[
1 \leq m^{kl} < M.
\]
This quantity reflects how many probability measure pairs under study are close to the minimum information divergence. As will be seen, inclusion of this quantity often leads to tighter bounds.

We conclude this part with a fact that links the KL divergence and squared Hellinger distance. It is well known that they are almost equivalent (modulo some small constant) when the two probability measures involved are close to each other—a regime where two measures are the hardest to differentiate. This is collected in the following fact; see [33, Proposition 2] for a different version.

**Fact 1.** Suppose that $P$ and $Q$ are two probability measures such that

\[
\frac{dP}{dQ} \leq R \quad \text{and} \quad \frac{dQ}{dP} \leq R
\]
hold uniformly over the probability space. Then, one has

\[
\max \{2 - 0.5 \log R, 1\} \cdot \text{Hel}_2^2 (P \parallel Q) \leq KL (P \parallel Q) \leq (2 + \log R) \cdot \text{Hel}_2^2 (P \parallel Q). \tag{23}
\]
Furthermore, if $R \leq 4.5$, one has

\[
(2 - 0.4 \log R) \cdot \text{Hel}_2^2 (P \parallel Q) \cdot \leq KL (P \parallel Q) \leq (2 + 0.4 \log R) \cdot \text{Hel}_2^2 (P \parallel Q). \tag{24}
\]

**Proof.** See Appendix [1] \(\square\)

### 3.2 Key Graphical Metrics

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

For any integer $m$, define

\[
\mathcal{N} (m) := \{ S \subset V : e(S, S^c) \leq m \}, \tag{25}
\]
which consists of the collection of cuts of about the same size. We are particularly interested in the peak growth rate of the cardinality of $\mathcal{N}$ as defined below

\[
\tau{k}^{\text{cut}} := \frac{1}{k} \log |\mathcal{N}(k \cdot \text{mincut})| \quad \text{and} \quad \tau^{\text{cut}} := \max_{k > 0} \tau{k}^{\text{cut}}. \tag{26}
\]

In the sequel, we will term $\tau^{\text{cut}}$ the **cut-homogeneity exponent**, which is illustrated through the following two extreme examples:

- **Complete graph $K_n$ on $n$ vertices.** In this case, $e(S, S^c) = |S| (n - |S|)$ and $\text{mincut} = n - 1$. Simple combinatorial argument suggests that $|\mathcal{N}(m)| \asymp n^m$, revealing that $\tau^{\text{cut}} = \max \frac{1}{k} \log |\mathcal{N}(k \cdot \text{mincut})| \asymp \log n$.

- **2 complete graphs $K_{n/2}$ connected by a single bridge.** In this graph, the min-cut size is $\text{mincut} = 1$ due to the existence of a bridge, but we still have $|\mathcal{N}(m)| \lesssim n^m$ when $k \geq n$. This indicates that $\tau^{\text{cut}} = \frac{1}{k} \log |\mathcal{N}(k \cdot \text{mincut})| \lesssim \frac{\log n}{n}$.

Intuitively, for various graphs one has $|\mathcal{N}(m)| \lesssim n^{\frac{m}{d_{\text{min}}}}$ when $m \geq d_{\text{min}}$, suggesting that $\tau^{\text{cut}} \lesssim \frac{\text{mincut} \log n}{d_{\text{min}}}$ when $k \gtrsim \frac{d_{\text{min}}}{\text{mincut}}$. In many homogeneous graphs (take a complete graph for an example) one has $\text{mincut} \asymp d_{\text{min}}$, and thus $\tau^{\text{cut}}$ can reach the order of $\log n$. In contrast, $\text{mincut} = o(d_{\text{min}})$ often holds in inhomogeneous graphs, and hence $\tau^{\text{cut}}$ tends to be lower. In light of this, $\tau^{\text{cut}}$ captures the extent of homogeneity for cut-size distributions.

Interestingly, for various homogeneous graphs of interest, $\tau^{\text{cut}}$ can be bounded above in a tight and simple manner, namely, $\tau^{\text{cut}} \lesssim \log n$, as asserted in the following lemma. We shall use $\mathcal{E}(u)$ throughout to denote the set of edges incident to a vertex $u$. 

7
Lemma 1. (1) **Homogeneous geometric graphs.** Suppose that $G$ 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 $(u, v) \in \mathcal{E}$,
\[ |\mathcal{E}(u) \cap \mathcal{E}(v)| \geq \rho \cdot \mincut; \]  

(b) for each $(u, v) \in \mathcal{E}$, denoting by $w^{(i)}$ the $i$th closest vertex to $v$ among the vertices in $\mathcal{E}(u) \cap \mathcal{E}(v)$, one has
\[ |\mathcal{E}(v) \setminus \mathcal{E}(w^{(i)})| \leq \frac{1}{2} \rho \cdot \mincut, \quad 1 \leq i \leq \kappa \rho \cdot \mincut. \]

Under the above two conditions, one has
\[ \tau_{\text{cut}} \leq \frac{8 \log n}{\kappa \rho}. \]  

(2) **Expander graphs.** If $G$ is an expander graph with edge expansion $h_G$, then
\[ \tau_{\text{cut}} \leq \frac{\mincut \log n}{h_G} + \log 2. \]

Proof. See Appendix E. \qed

To gain some insights, we emphasize a few concrete examples accommodated by this lemma. 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. Two examples are worth mentioning. The first is a random geometric graph $G_n, r$, 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 $w$ ($w \geq 5$) vertices apart. For both cases, $\kappa$ and $\rho$ are constants bounded away from 0, indicating that $\tau_{\text{cut}} \lesssim \log n$.

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 $h_G/\mincut = \Theta(1)$, we conclude from Lemma 1 that $\tau_{\text{cut}} \lesssim \log n$.

As a final remark, we are not aware of a graph obeying $\tau_{\text{cut}} = \omega(\log n)$. In all aforementioned examples, one always has $\tau_{\text{cut}} \lesssim \log n$.

4 Main Results: General Graphs

This section develops both sufficient and necessary conditions for exact information recovery. These conditions are presented in terms of the minimum information divergence measures (i.e. KL, Hellinger, and Rényi divergence), and are capable of accommodating general graph structures, channel characteristics, and input alphabets. In particular, as long as the alphabet size is not super-polynomial in $n$, our results are tight (up to some explicit universal constant) for a variety of homogeneous graphs, including the widely adopted random graph models like Erdős–Rényi graphs, random geometric graphs, and other expander graphs with good expansion properties.

4.1 Maximum Likelihood Decoding

We start by analyzing the performance of the maximum likelihood (ML) decoder $\psi_{\text{ml}}(y)$, which seeks a solution that maximizes the likelihood function:
\[ \psi_{\text{ml}}(y) := \arg \max_{x \in \mathbb{X}^n} \mathbb{P}\{y \mid x\}. \]

It is well-known that ML is able to minimize the Bayesian probability of error under uniform priors. We develop a sufficient condition in terms of the minimum Hellinger / Rényi divergence that enables perfect information recovery, using ML decoding. The condition is universal and holds for all graphs, and it depends only on the min-cut size and the cut-homogeneity exponent irrespective of other graphical metrics.
Theorem 1. Consider any connected graph $G$. For any $\delta > 0$ and any $0 < \alpha < 1$, the ML rule $\psi_{\text{ml}}$ achieves

$$P_e(\psi_{\text{ml}}) \leq \frac{1}{(2n)^{\delta} - 1}.$$ 

provided that

$$(1 - \alpha) \text{Hel}_{\alpha}^{\text{min}} \geq \frac{8\tau_{\text{cut}} + (\delta + 8) \log (2n) + 4 \log M}{\text{mincut}}$$

or

$$(1 - \alpha) D_{\alpha}^{\text{min}} \geq \frac{8\tau_{\text{cut}} + (\delta + 8) \log (2n) + 4 \log M}{\text{mincut}}$$

Proof. See Appendix A \hfill \Box

Remark 1. In general, we are unable to characterize the recovery limits in terms of the KL divergence, since the KL divergence cannot be well controlled for all measures, especially when $\frac{dP_l}{dP_k}$ $(l \neq j)$ becomes unbounded. In contrast, the Hellinger / Rényi divergence is generally stable and more convenient to analyze in this case. Note that the above sufficient condition holds for Hellinger / Rényi divergence of any order $\alpha$ $(0 < \alpha < 1)$.

Remark 2 (Multi-shot measurements). The recovery condition (33) based on the Rényi divergence is particularly handy when accounting for multi-shot measurements. For instance, suppose that one obtains $L$ i.i.d. measurements over each edge. We let $\text{Hel}_{\alpha}^{\text{min}}$ and $D_{\alpha}^{\text{min}}$ be the minimum divergence defined w.r.t. single-shot measurements, then the additivity of $D_{\alpha} (\cdot \| \cdot)$ over product distributions gives (31, Theorem 28)

$$D_{\alpha}^{\text{min}} = L D_{\alpha}^{\text{min}}_{\text{single}} \geq L \text{Hel}_{\alpha}^{\text{min}}_{\text{single}}.$$ 

To conclude, Theorem 1 continues to hold if one replaces $\text{Hel}_{\alpha}^{\text{min}}$ (resp. $D_{\alpha}^{\text{min}}$) with $L \text{Hel}_{\alpha}^{\text{min}}_{\text{single}}$ (resp. $L D_{\alpha}^{\text{min}}_{\text{single}}$).

Theorem 1 characterizes a non-asymptotic condition of $\text{Hel}_{\alpha}^{\text{min}}$ or $D_{\alpha}^{\text{min}}$ under which $\psi_{\text{ml}} (\cdot)$ admits perfect information recovery. The boundary condition scales as

$$\Theta \left( \frac{\tau_{\text{cut}} + \log n + \log M}{\text{mincut}} \right),$$

which is inversely proportional to the min-cut size. This makes sense since all information one has available to link the components across the minimum cut is conveyed through the cut set. The scaling (35) also reveals that the contribution of $\tau_{\text{cut}}$ will be negligible if it is below $o (\log n)$.

As will be seen later, this result is most useful when applied to those homogeneous graphs satisfying

$$\text{mincut} \approx d_{\text{min}} \approx d_{\text{max}},$$

for which the fundamental limits decay inversely with the vertex degree. To gain some insights about the metrics, we emphasize that the channel decoding model considered herein differs from the classical information theory setting in that the input is “uncoded”. Consequently, the fundamental bottleneck for minimax “uncoded” information recovery is presented by the minimum distance between the set of hypotheses, rather than the mutual information in an average sense. Notably, two hypotheses $x$ and $w$ are minimally separated when they differ only by one component $v$. The resulting pairwise inputs \{$(y_{ij} | (i, j) \in E)$\} contain about $\deg (v)$ copies of information for distinguishing $x$ and $w$, where each copy of information can be approximated through the information divergence metrics $\text{Hel}_{\alpha}^{\text{min}}$ or $D_{\alpha}^{\text{min}}$ (or $\text{KL}^{\text{min}}$ as adopted in Section 4.2). This offers an intuitive interpretation as to why the minimax limit scales inversely proportional to the vertex degree for homogeneous graphs.

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

$$p \left(y_{ij} \bigg| x_l - x_j = l \right) = p_{ij}^{x_l} (y_{ij}), \quad 0 \leq l < M,$$

for a family \{$(p_{ij}^{x_l}) \ | \ 0 \leq l < M, 1 \leq i, j \leq n$\}. This leads to a modified version of the minimum divergence metric as follows

$$\text{Hel}_{\alpha}^{\text{min}} := \min \left\{ \text{Hel} \left( p_{ij}^{x_l} \| p_{ij}^{x_k} \right) \mid l \neq k, 1 \leq l, k \leq M, i \neq j \right\};$$

$$D_{\alpha}^{\text{min}} := \frac{1}{1 - \alpha} \log \left( 1 - (1 - \alpha) \text{Hel}_{\alpha}^{\text{min}} \right).$$
A sufficient condition under this generalized model is then given as follows.

**Theorem 2.** The recovery condition of Theorem 1 continues to hold under the transition probabilities (36), if $\text{Hel}_{\alpha}$ and $D_{\alpha}$ are replaced by $\text{Hel}_{\alpha}$ and $D_{\alpha}$, respectively.

### 4.2 Lower Bounds

Encouragingly, for a broad class of homogeneous graphs, the sufficient recovery condition for the ML decoder matches the fundamental lower limit in an order-wise tight sense, as demonstrated in the following theorem.

**Theorem 3 (KL Version).** Fix $0 < \epsilon \leq \frac{1}{2}$. For any graph $G$, if the KL divergence satisfies

$$
\text{KL}_{\min} \leq \max \left\{ \frac{(1 - \epsilon) \tau_{\text{cut}} - H(\epsilon)}{\mincut}, \frac{(1 - \epsilon) \log m^{kl} - H(\epsilon)}{1 + \epsilon} \mincut, \frac{(1 - \epsilon) (\log n + \log m^{kl}) - H(\epsilon)}{1 + \epsilon} \right\} \tag{39}
$$

then the minimax probability of error exceeds

$$
\inf_{\psi} P_e(\psi) \geq \epsilon.
$$

Here, $H(x) := -x \log x - (1 - x) \log (1 - x)$ stands for the binary entropy function.

**Proof.** The proof follows from the generalized Fano’s inequality [29, 34] and, in particular, an extended version that involves general $f$-divergence measures [35]. See Appendix B.

**Theorem 3** presents a general fundamental lower limit on $\text{KL}_{\min}$ that allows perfect recovery, which scales as $\Theta(\tau_{\text{cut}} + \log m^{kl} + \log n d_{\max})$.

For those homogeneous graphs obeying $\mincut \approx d_{\max}$ and $\tau_{\text{cut}} \ll \log n$ (cf. Lemma 1), the lower bound relies only on the first-order graph quantity, that is, the vertex degree. In the regime where the transition measures of the graphical channels are close to each other, Theorem 3 can also be easily translated to a Hellinger divergence version modulo some negligible terms.

Apart from this, we have derived another necessary condition based directly on Hellinger divergence, although it becomes weaker under those inhomogeneous graphs with $\mincut \ll d_{\max}$.

**Theorem 4 (Hellinger Version).** Consider any graph $G$, any $\epsilon > 0$, and $\alpha \leq \frac{1}{1 + \epsilon}$. Suppose that $d_{\max} \geq 2\epsilon n \log n - 2 \log 2$. If

$$
\text{Hel}_{\alpha} \leq \frac{\epsilon n \log n}{1 - \alpha} - r_{\epsilon}, \tag{40}
$$

for some residual $r_{\epsilon} = O\left(\frac{1}{d_{\max}}\right)$, then one has $\inf_{\psi} P_e(\psi) \geq n^{-\epsilon}$.

**Proof.** See Appendix C.

We remark that the two necessary recovery conditions in Theorems 3-4 concern two regimes of separate interest. Specifically, Condition (39) 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 small without specifying convergence rates. In contrast, Condition (40) based on the Hellinger distance is more convenient when we further expect 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. We will see more discussion in Section 5.

---

5 More precisely, $r_{\epsilon} := \log 2 + \frac{2(\epsilon n \log n - \log 2)^2}{(1 - \epsilon) d_{\max}}$. 

10
4.3 Implications

The following discussion concentrates on the case where $\frac{d_P}{d_j} = 1 + o(1)$ for all $0 \leq l, j < M$, which essentially corresponds to the most difficult situation for exact decoding. The following discussion will be devoted to the most popular divergence metrics – the KL divergence and the squared Hellinger distance. We recall from Fact 1 that

$$KL_{\text{min}} = 2(1 + o(1)) Hel_{\frac{1}{2}}$$

holds in this regime.

1. **Tightness.** Theorems 1-3 characterize a sufficient and a necessary condition as follows

$$\inf_{\psi} P_e(\psi) \xrightarrow{n \to \infty} 0 \quad \text{if } Hel_{\frac{1}{2}} > (1 + o(1)) \frac{16 \tau_{\text{cut}} + 16 \log n + 8 \log M}{\text{mincut}};$$

$$\inf_{\psi} P_e(\psi) \xrightarrow{n \to \infty} 0 \quad \text{if } Hel_{\frac{1}{2}} < (1 - o(1)) \max \left\{ \frac{\tau_{\text{cut}}}{2\text{mincut}}, \frac{\log m^{kl}}{2\text{mincut}}, \frac{\log n + \log m^{kl}}{2d_{\text{max}}} \right\}.$$

One can then easily verify that that the recovery condition we derive is optimal up to a factor $O\left( \max \left\{ \frac{\log M}{\log m^{kl}}, \frac{d_{\text{max}}}{\text{mincut}} \right\} \right)$. When specialized to the class of homogeneous graphs satisfying $d_{\text{max}} = (1 + o(1)) \text{mincut}$ (e.g. Erdős–Rényi Graphs, random geometric graphs, etc.), the recovery condition is within a multiplicative gap

$$O\left( \frac{\log M + \log n}{\log m^{kl} + \log n} \right)$$

from optimal. Consequently, our results are order-wise optimal as long as $M \lesssim \log n$ or $m^{kl} \approx M$.

2. **First-order v.s. second-order graphical metrics.** For a broad class of homogeneous graphs (e.g. Erdős–Rényi graphs, random geometric graphs, generalized rings, or many other expander graphs), one has $\text{mincut} \approx d_{\text{max}}$ and $\tau_{\text{cut}} \lesssim \log n$ (see Lemma 1). As a result, the boundary recovery condition reduces to

$$Hel_{\frac{1}{2}} \approx \frac{\log n}{d_{\text{max}}}$$

as long as $M \lesssim \log (n)$. In other words, for these graphs the recovery conditions we develop depend only on the vertex degree, which do not rely on second-order expansion properties like the spectral gap or Cheeger’s constant. This improves upon the sufficient recovery condition derived in [19 Theorem 4.3] that might be loose when $G$ does not enjoy strong expansion properties. This leads to the following observation: for various homogeneous graphs, the information-theoretic limits for graph-based decoding is determined solely by first-order graphical features. This is in stark contrast to the performance guarantees for many tractable algorithms (e.g. spectral method or semidefinite programming), whose success typically rely on strong second-order expansion properties.

3. **A Unified Non-asymptotic Framework.** Our framework can accommodate a variety of practical scenarios that respect the high-dimensional regime, where both the number $n$ of objects and the alphabet size $M$ tend to be large. 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 becomes unwieldy and the Chernoff information measure is not guaranteed to capture the minimax rate. By comparison, our results build upon alternative probability distance measures (particularly the Hellinger / Rényi divergence). This results in a simple unified framework that allows the non-asymptotic characterization of the minimax rate (modulo some constant factor) simultaneously for most settings.

---

*For instance, random geometric graphs typically have much worse expansion properties than Erdős–Rényi graphs.*
5 Main Results: Erdős–Rényi Graphs

This section is devoted to the Erdős–Rényi model, a tractable yet the most widely adopted random graph model for numerous applications. Specifically, we suppose that the measurement graph \( G \sim G_{n, p_{\text{obs}}} \) for some edge selection probability \( p_{\text{obs}} \geq c_0 \frac{\log n}{n} \). While our results in Section 4 are sufficient to characterize the minimax recovery limit in an order-wise optimal sense, we emphasize the importance of \( G_{n, p_{\text{obs}}} \) by developing tighter results for this special model.

5.1 ML Decoding and Minimax Lower Bounds

The following theorem characterizes a regime where the ML decoder is guaranteed to work under the Erdős–Rényi model.

**Theorem 5.** Suppose that \( G \sim G_{n, p_{\text{obs}}} \) with \( p_{\text{obs}} \geq c_0 \frac{\log n}{n} \) for some sufficiently large constant \( c_1 > 0 \). Fix \( \delta > 0 \) and \( 0 < \alpha < 1 \). For any constant \( \zeta > 0 \), there exist some universal constants \( C, c_1 > 0 \) such that if

\[
(1 - \alpha) H_{\alpha}^{\min} \geq \frac{(1 + \delta) \log (2n) + 2 \log (M - 1)}{(1 - \zeta) p_{\text{obs}} n},
\]

or

\[
(1 - \alpha) D_{\alpha}^{\min} \geq \frac{(1 + \delta) \log (2n) + 2 \log (M - 1)}{(1 - \zeta) p_{\text{obs}} n},
\]

then the ML decoder \( \psi_{\text{ml}} \) achieves

\[
P_e(\psi_{\text{ml}}) \leq \frac{1}{(2n)^{\max\{\frac{4}{4-\frac{1}{4}}\}} - 1} + \frac{3}{n^{10}} + C n^{-c_0 \delta n}.
\]

**Proof.** See Appendix D.

**Remark 3 (Multi-shot measurements).** Consider multi-shot measurements as discussed in Remark 2. Similarly, one can conclude that Theorem 5 continues to hold if one replaces \( H_{\alpha}^{\min} \) (resp. \( D_{\alpha}^{\min} \)) with \( L H_{\alpha, \text{single}}^{\min} \) and \( L D_{\alpha, \text{single}}^{\min} \), where \( L \) stands for the number of i.i.d. measurements taken on each edge.

In Theorem 5, we present a non-asymptotic sufficient condition for the ML rule to achieve exact information recovery, namely,

\[
(1 - \alpha) H_{\alpha}^{\min} \text{ or } (1 - \alpha) D_{\alpha}^{\min} = \Omega \left( \frac{\log n + \log M}{p_{\text{obs}} n} \right) = \Omega \left( \frac{\log n + \log M}{d_{\max}} \right).
\]

More precisely, we only require

\[
(1 - \alpha) H_{\alpha}^{\min} \text{ or } (1 - \alpha) D_{\alpha}^{\min} \geq (1 + o(1)) \frac{\log n + 2 \log M}{p_{\text{obs}} n}
\]

to ensure asymptotically zero probability of error, which improves upon Theorem 1 by an explicit constant factor. Similarly, the above sufficient condition extends to the case with location-dependent output distributions, as stated below.

**Theorem 6.** The recovery condition of Theorem 5 continues to hold under the transition probabilities (36), if \( H_{\alpha}^{\min} \) and \( D_{\alpha}^{\min} \) are replaced by \( L H_{\alpha, \text{single}}^{\min} \) and \( L D_{\alpha, \text{single}}^{\min} \) as defined in (37) and (38), respectively.

On the other hand, the boundary conditions presented in Theorems 3-4 rely on the maximum vertex degree of a graph. When specialized to the Erdős–Rényi graph \( G \sim G_{n, p_{\text{obs}}} \), the Chernoff-type inequality [38, Theorems 4.4-4.5] indicates that for any \( \epsilon > 0 \), one has

\[
d_{\max} \leq (1 + o(1)) n p_{\text{obs}}
\]

with probability exceeding \( 1 - n^{-10} \), provided that \( p_{\text{obs}} > c_0 \frac{\log n}{n} \) for some sufficiently large constant \( c > 0 \). By applying the parts of Theorems 3-4 that depend only on \( d_{\max} \), we immediately obtain the following corollary.
Corollary 1. Suppose that $\mathcal{G} \sim \mathcal{G}_{n, p_{\text{obs}}}$. Fix $\epsilon > 0$, and assume that $p_{\text{obs}} = \frac{c \log n}{n}$ for some sufficiently large constant $c > 0$.

(a) If

$$\text{KL}_{\min} \leq \frac{(1 - \epsilon) \left( \log n + \log m^{kl} \right) - H(\epsilon)}{(1 + \epsilon)^2 np_{\text{obs}}}.$$  \hspace{1cm} (44)

then one has $\inf_{\psi} P_\epsilon (\psi) \geq \epsilon - \frac{1}{n^{1/2}}$.

(b) For any $\alpha \leq \frac{1}{17}$ and small constant $\zeta > 0$, if

$$\text{Hel}_{\min} < \frac{\epsilon \alpha \log n}{(1 + \zeta)(1 - \alpha) np_{\text{obs}}} - r_\epsilon$$  \hspace{1cm} (45)

for some residual $r_\epsilon = O \left( \frac{1}{np_{\text{obs}}} \right)$, then one has $\inf_{\psi} P_\epsilon (\psi) \geq n^{-\epsilon} - \frac{1}{n^{1/2}}$.

5.2 Tightness of Theorem 5 and Corollary 1

Encouragingly, the recovery conditions derived in Theorem 5 and Corollary 1 are often tight up to some constant. Our discussion below is based on both KL divergence and the squared Hellinger distance, and concerns the situation where $p_{\text{obs}} = \omega \left( \log n/n \right)$.

1. Consider the first-order convergence, that is, the regime where $\inf_{\psi} P_\epsilon (\psi) \to 0 \ (n \to \infty)$. Combining Theorem 5 and Corollary 1(a) suggest the following: when $n$ tends to infinity, one has

$$\inf_{\psi} P_\epsilon (\psi) \xrightarrow{n \to \infty} 0 \quad \text{if} \quad \text{Hel}_{\min} > (1 + o(1)) \frac{2 \log n + 4 \log M}{p_{\text{obs}} n};$$

$$\inf_{\psi} P_\epsilon (\psi) \xrightarrow{n \to \infty} 0 \quad \text{if} \quad \text{KL}_{\min} < (1 - o(1)) \frac{\log n + \log m^{kl}}{p_{\text{obs}} n}.$$  \hspace{1cm} (46)

If we further suppose the hard case where $\frac{\partial P_\epsilon}{\partial \psi} = 1 + o(1)$ for all $l \neq j$, then these taken collectively with Fact 4 reduce to

$$\inf_{\psi} P_\epsilon (\psi) \xrightarrow{n \to \infty} 0 \quad \text{if} \quad \text{Hel}_{\min} > (1 + o(1)) \frac{2 \log n + 4 \log M}{p_{\text{obs}} n};$$

$$\inf_{\psi} P_\epsilon (\psi) \xrightarrow{n \to \infty} 0 \quad \text{if} \quad \text{Hel}_{\min} < (1 - o(1)) \frac{\log n + \log m^{kl}}{2p_{\text{obs}} n}.$$  \hspace{1cm} (47)

In other words, when $M \preceq \text{poly} (n)$ or $m^{kl} = M^{\Theta(1)}$, the fundamental limits for exact recovery is given by

$$\text{Hel}_{\min} \gtrsim \frac{\log n}{p_{\text{obs}} n}.$$  \hspace{1cm} (48)

In general, our minimum distance bounds are tight modulo a factor of

$$(1 + o(1)) \frac{4 \log n + 8 \log M}{\log n + \log m^{kl}}.$$  \hspace{1cm} (49)

2. Consider the second-order convergence and, particularly, the regime where $\inf_{\psi} P_\epsilon (\psi) \to \frac{1}{n} \ (n \to \infty)$. Putting Theorem 5 (with $\delta = 3$) and Corollary 1(b) (with $\alpha = \frac{1}{2}$ and $\epsilon = 1$) together implies that

$$\inf_{\psi} P_\epsilon (\psi) < \frac{1}{n} \quad \text{if} \quad \text{Hel}_{\min} > (1 + o(1)) \frac{8 \log n + 4 \log M}{p_{\text{obs}} n},$$

$$\inf_{\psi} P_\epsilon (\psi) > \frac{1}{n} \quad \text{if} \quad \text{Hel}_{\min} < (1 - o(1)) \frac{\log n + \log m^{kl}}{p_{\text{obs}} n},$$

which holds for all situations. This implies that the fundamental transition boundary for exact recovery is

$$\text{Hel}_{\min} \gtrsim \frac{\log n}{p_{\text{obs}} n},$$

as long as $M \preceq \text{poly} (n)$. Furthermore, our results are within a multiplicative gap of

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

from optimal for all possible alphabet size.
6 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, and even improve upon existing results in certain regimes.

6.1 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 [2,39–43] for a highly incomplete list of references.

We will focus on the regime where $\frac{\alpha, \beta}{\log n}$ and $\alpha > \beta$, which consists of all but the densest community structures. Treating the SBM as a graphical channel over a complete measurement graph ($p_{obs} = 1$) 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)

$$P_0 = Bern \left( \frac{\alpha \log n}{n} \right), \quad \text{and} \quad P_1 = Bern \left( \frac{\beta \log n}{n} \right).$$

This allows us to compute

$$\text{Hel}_{\min} = \left( \sqrt{\frac{\alpha \log n}{n}} - \sqrt{\frac{\beta \log n}{n}} \right)^2 + \left( \sqrt{1 - \frac{\alpha \log n}{n}} - \sqrt{1 - \frac{\beta \log n}{n}} \right)^2 \leq \frac{1 + o(1)}{\alpha} \left( \sqrt{\alpha} - \sqrt{\beta} \right)^2 \log n \frac{n}{n}.$$

In addition, the relation between KL divergence and $\chi^2$ divergence (e.g. [31 Equation (7)]) suggests that

$$\text{KL}_{\min} \leq \text{KL} \left( \frac{\beta \log n}{n} \mid \frac{\alpha \log n}{n} \right) \leq \chi^2 \left( \frac{\beta \log n}{n} \mid \frac{\alpha \log n}{n} \right) \leq \frac{\left( \frac{\beta \log n}{n} - \frac{\alpha \log n}{n} \right)^2}{\frac{\log n}{n} \left( 1 - \frac{\alpha \log n}{n} \right)} = 1 + o(1),$$

$$\left( \sqrt{\alpha} - \sqrt{\beta} \right) \leq \frac{\alpha - \beta}{\sqrt{\alpha}},$$

where (a) follows from the identity $\chi^2 (p \parallel q) = \frac{(p-q)^2}{q} + \frac{(p-q)^2}{1-q} = \frac{(p-q)^2}{q(1-q)}$. With these two estimates in place, Theorem 5 and Corollary 1 immediately give

$$\inf_{\psi} P_e (\psi) \rightarrow 0 \quad \text{if} \quad \left( \sqrt{\alpha} - \sqrt{\beta} \right)^2 \geq 2 \left( 1 + o(1) \right),$$

$$\inf_{\psi} P_e (\psi) \rightarrow 0 \quad \text{if} \quad \left( \alpha - \beta \right)^2 \leq \left( 1 - o(1) \right) \alpha.$$

In fact, the precise phase transition for exact cluster recovery has only been determined last year [40,41], which asserts that

$$\inf_{\psi} P_e (\psi) \rightarrow 0 \quad \text{if} \quad \left( \sqrt{\alpha} - \sqrt{\beta} \right)^2 > 2,$$

$$\inf_{\psi} P_e (\psi) \rightarrow 0 \quad \text{if} \quad \left( \sqrt{\alpha} - \sqrt{\beta} \right)^2 < 2.$$
This in turn justifies that the sufficient recovery condition we develop is precise. When it comes to the necessary condition, one can verify that the condition \( [56] \) is more stringent than \( (\alpha - \beta)^2 < 4 (\alpha + \beta) \). In comparison, under the hypothesis \( \alpha > \beta \), the boundary of our condition \( [54] \) is sandwiched between the curves \( (\alpha - \beta)^2 \leq \frac{1}{2} (1 - o(1)) (\alpha + \beta) \) and \( (\alpha - \beta)^2 \leq (1 - o(1)) (\alpha + \beta) \). These taken collectively indicate that our theory leads to recovery guarantees that are tight up to a small constant factor.

Three remarks are in order. To begin with, our results do not impose the assumptions that each cluster has size exactly \( n/2 \) while obtaining the same sufficient recovery condition. Secondly, we are able to accommodate all values of \( \alpha, \beta \) up to \( o \left( \frac{n}{\log n} \right) \), which is in contrast to \( [46] \) that concentrates on the sparsest possible regime (i.e. \( \alpha, \beta \asymp 1 \)). Leaving out these technical matters, 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 distance metric that dictates the recovery boundary for SBMs.

Finally, we became aware of the very recent work \( [44] \) that characterizes the fundamental limits for the generalized SBM, that is, the model where \( n \) nodes are partitioned into multiple clusters. The limits are determined by an \( f \)-divergence called Chernoff-Hellinger divergence, and largely depend on cluster sizes. While the current work does not take into account pre-determined cluster sizes, we can derive, for a certain case, a sufficient and a necessary condition for minimax recovery that capture the worst-case cluster size situations. For instance, imagine that the edge density across clusters \( i \) and \( j \) is given by \( Q_{ij} \log n/n \) and hence \( \mathbb{P}_1 = \text{Bern} \left( \frac{Q_{ij} \log n}{n} \right) \) (cf. Definition 13). Then, our general theory reveals that

\[
\inf_{\psi} P_e (\psi) \xrightarrow{n \to \infty} 0 \quad \text{if} \quad \min_{i \neq j} \left( \sqrt{Q_i} - \sqrt{Q_j} \right)^2 \geq 2 (1 + o(1)) ,
\]

\[
\inf_{\psi} P_e (\psi) \xrightarrow{n \to \infty} 0 \quad \text{if} \quad \min_{i \neq j} \frac{(Q_i - Q_j)^2}{Q_i} \leq (1 - o(1)) .
\]

More delicate analyses that accommodate prescribed cluster sizes for the general pairwise measurement model are left for future work.

### 6.2 Outlier Model

We now turn to another model called the outlier model, which subsumes as special cases several applications including alignment, synchronization, and joint matching (e.g. \([6,10,11]\)).

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

\[
y_{ij} = \begin{cases} x_i - x_j , & \text{with probability } p_{\text{true}}, \\ \text{Unif}_M , & \text{else,} \end{cases}
\]

where \( \text{Unif}_M \) represents the uniform distribution over \( \{0, \ldots , M - 1\} \), \( p_{\text{true}} \) stands for the non-corruption rate, and \( - \) is the general subtraction operation defined in Section 2 over the discrete alphabet. In words, approximately a fraction \( 1 - p_{\text{true}} \) of measurements act as random outliers and contain no useful information. Note that under this outlier model, one has

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

The following corollary, which is an immediate consequence of Theorem \([4] \) and Corollary \([1] \) presents concrete fundamental recovery limits for the outlier model. For ease of presentation, we restrict our discussion to the Erdős–Rényi model, but remark that all results immediately 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.

\[ \text{To see this, observe that } (\sqrt{\alpha} - \sqrt{\beta})^2 < 2 \text{ is identical to } (\alpha - \beta)^2 < 2 (\sqrt{\alpha} + \sqrt{\beta})^2 , \text{ which is more stringent than} \]

\[ (\alpha - \beta)^2 < 4 (\alpha + \beta) \text{ due to the elementary inequality } (a + b)^2 \leq 2 (a^2 + b^2) . \]

\[ \text{Examining the proof of our main theorems indicates that we can relax this assumption and work with the general case where the edge density across clusters } i \text{ and } j \text{ is given by } Q_{ij} \frac{\log n}{n} , \text{ but we omit it here as it is not the main focus of the present paper.} \]

---

\[ 7 \text{To see this, observe that } (\sqrt{\alpha} - \sqrt{\beta})^2 < 2 \text{ is identical to } (\alpha - \beta)^2 < 2 (\sqrt{\alpha} + \sqrt{\beta})^2 , \text{ which is more stringent than} \]

\[ (\alpha - \beta)^2 < 4 (\alpha + \beta) \text{ due to the elementary inequality } (a + b)^2 \leq 2 (a^2 + b^2) . \]

\[ \text{Examining the proof of our main theorems indicates that we can relax this assumption and work with the general case where the edge density across clusters } i \text{ and } j \text{ is given by } Q_{ij} \frac{\log n}{n} , \text{ but we omit it here as it is not the main focus of the present paper.} \]
Corollary 2. Consider the outlier model (59), and fix $\epsilon > 0$. When $G \sim G_{n, p_{\text{obs}}}$ with $p_{\text{obs}} > c_1 \log n \over n$ for some sufficiently large constant $c_1 > 0$,

$$
\inf_{\psi} P_e(\psi) \xrightarrow{n \to \infty} 0 \quad \text{if} \quad \frac{1}{M} \left( \sqrt{1 - p_{\text{true}} + M p_{\text{true}}} - \sqrt{1 - p_{\text{true}}} \right)^2 \geq (1 + \epsilon) \frac{\log n + 2 \log M}{p_{\text{obs}} n},
$$

(60)

$$
\inf_{\psi} P_e(\psi) \xrightarrow{n \to \infty} 0 \quad \text{if} \quad p_{\text{true}} \leq \max \left\{ \frac{(1 - \epsilon)(\log n + \log M)}{p_{\text{obs}} n \log \left( 1 + \frac{1}{M} - p_{\text{true}} \right)}, \frac{M}{M - 1} \left( \frac{\log n}{p_{\text{obs}} n} - 1 \right) \right\}.
$$

(61)

To establish this corollary, we start by considering the graph $G_{\text{true}}$ that is formed by all edges where $y_{ij} = x_i - x_j$. It is self-evident that $G_{\text{true}} \sim G_{n, (p_{\text{true}} + \frac{1}{M} - p_{\text{true}}) p_{\text{obs}}}$, and thus $(1 - \frac{1}{M} - p_{\text{true}} + \frac{1}{M}) p_{\text{obs}} > \frac{\log n}{n}$ 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 $\text{KL}^{\text{min}}$ and $\text{Hel}^{\text{min}}$, which we gather in the following lemma.

Lemma 2. Consider the outlier model (59). For any $0 \leq p_{\text{true}} < 1$, one has

$$
\text{KL}^{\text{min}} = p_{\text{true}} \log \left( 1 + \frac{p_{\text{true}} M}{1 - p_{\text{true}}} \right) \leq \frac{p_{\text{true}}^2 M}{1 - p_{\text{true}}}
$$

(62)

and

$$
\text{Hel}^{\text{min}}_2 = \frac{2}{\pi} \left( \sqrt{1 - p_{\text{true}} + M p_{\text{true}}} - \sqrt{1 - p_{\text{true}}} \right)^2 \geq \frac{p_{\text{true}}^2 M}{2(1 - p_{\text{true}} + M p_{\text{true}})}.
$$

(63)

Proof. See Appendix [F].

To give the readers a better sense, we depict in Fig. 2 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 comparison, we apply the general theory once again with the assistance of the inequalities in Lemma 2 to obtain

$$
\inf_{\psi} P_e(\psi) \xrightarrow{n \to \infty} 0 \quad \text{if} \quad p_{\text{true}} \geq 2(1 + \epsilon) \sqrt{\frac{(1 - p_{\text{true}} + M p_{\text{true}})(\log n + 2 \log M)}{p_{\text{obs}} n M}},
$$

(64)

$$
\inf_{\psi} P_e(\psi) \xrightarrow{n \to \infty} 0 \quad \text{if} \quad p_{\text{true}} \leq \max \left\{ (1 - \epsilon) \sqrt{\frac{(1 - p_{\text{true}})(\log n + \log M)}{p_{\text{obs}} n M}}, \log n \right\}.
$$

(65)
6.2.1 Tightness under Binary Alphabet

Consider the consequences of our preceding results when \( M = 2 \), which has also been studied \[19\]. When \( p_{\text{obs}} \gg \frac{\log n}{n} \), our results \((64)\) and \((65)\) assert that

\[
\inf_{\psi} P_e(\psi) \xrightarrow{n \to \infty} 0 \text{ if } p_{\text{true}} \geq (1 + o(1)) \frac{2 \log n}{p_{\text{obs}} n}, \tag{66}
\]

\[
\inf_{\psi} P_e(\psi) \xrightarrow{n \to \infty} 0 \text{ if } p_{\text{true}} \leq (1 - o(1)) \frac{\log n}{2 p_{\text{obs}} n}. \tag{67}
\]

From this, we see that our bounds are within a factor gap \( 2 + o(1) \) from optimal, which holds uniformly over all possible values of \((p_{\text{obs}}, p_{\text{true}})\) as long as \( p_{\text{obs}} \gg \frac{\log n}{n} \). This constant gap is illustrated in Fig. 2(a) as well.

In contrast, the bounds presented in \[19\] fall short of a uniform constant factor gap accommodating different parameter configurations. Adopting our notation, \[19, \text{Theorems 4.1 - 4.2}\] reduce to

\[
\inf_{\psi} P_e(\psi) \xrightarrow{n \to \infty} 0 \text{ if } p_{\text{true}} > \sqrt{\frac{(1 + o(1))! \log n}{p_{\text{obs}} n}},
\]

\[
\inf_{\psi} P_e(\psi) \xrightarrow{n \to \infty} 0 \text{ if } p_{\text{true}} < \sqrt{\frac{(1 - o(1))! 2 (1 - \frac{1}{\sqrt{2}}) \log n}{p_{\text{obs}} n}},
\]

where \( 0 < \tau < \frac{2}{3} \) is some numerical value satisfying \( p_{\text{obs}} \leq 2n\tau^{-1} \). Hence, their bounds are tight up to the following multiplicative factor

\[
g(\tau) = \frac{1 + o(1)}{\sqrt{1 - \frac{2}{\tau}}},
\]

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

Notably, when \( \tau \to 0 \) or \( p_{\text{obs}} \gg \frac{\log n}{n} \), the fundamental limit approaches \( \sqrt{\frac{(1 + o(1))! 2 \log n}{p_{\text{obs}} n}} \) in an accurate manner \[19\]. 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.

6.2.2 From Small Alphabet to Large Alphabet

The recovery conditions given in Corollary 2 can be further divided and simplified for two respective regimes, depending on whether \( M_{p_{\text{true}}} \ll 1 \) or \( M_{p_{\text{true}}} \gg 1 \). By substituting each of these two hypotheses into \((64)\), deriving the corresponding minimum \( p_{\text{true}} \) 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

\[
\inf_{\psi} P_e(\psi) \xrightarrow{n \to \infty} 0 \text{ if } p_{\text{true}} \geq 2 (1 + o(1)) \sqrt{\frac{\log n + 2 \log M}{p_{\text{obs}} n M}}, \tag{68}
\]

\[
\inf_{\psi} P_e(\psi) \nrightarrow 0 \text{ if } p_{\text{true}} \leq (1 - o(1)) \sqrt{\frac{\log n + \log M}{p_{\text{obs}} n M}}. \tag{69}
\]

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

\[
\inf_{\psi} P_e(\psi) \xrightarrow{n \to \infty} 0 \text{ if } p_{\text{true}} \geq 4 (1 + o(1)) \frac{\log n + 2 \log M}{p_{\text{obs}} n}, \tag{70}
\]

\[
\inf_{\psi} P_e(\psi) \nrightarrow 0 \text{ if } p_{\text{true}} \leq \frac{\log n}{p_{\text{obs}} n}. \tag{71}
\]

\(^9\)Note that \( p_{\text{true}} = 1 - 2\varepsilon \) and \( d = n p_{\text{obs}} \) for the notation \( \varepsilon \) and \( d \) defined in \[19\], respectively.
That being said, the fundamental recovery boundary presented in terms of \( p_{\text{true}} \) exhibits contrasting features in two separate regimes (for both Erdős–Rényi graphs and homogeneous geometric graphs), as illustrated in Fig. 2(b). Some interpretation are in order.

1. Information-limited regime \((M = o\left(\frac{d_{\min}}{\log n}\right))\). The maximum amount of information that can be conveyed through each pairwise measurement is captured by the KL divergence. In this small-alphabet regime, simple calculation reveals that \( KL_{\text{min}} \approx p_{\text{true}}^2 M \) (see Lemma 2), indicating that each measurement will carry more information as \( M \) increases. As a result, the fundamental recovery boundary decay with \( M \) as well as the vertex degree, both at square-root rates. In particular, for Erdős–Rényi graphs, the bounds in this regime are tight up to a factor of 2 (resp. a factor of \( \sqrt{\frac{3}{2}} = \sqrt{6} \)) in the presence of a constant alphabet size (resp. in the worst-case situation as long as \( M \leq n \)).

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 \( G \) 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 \( n \).

### 6.3 Haplotype Assembly

The pairwise measurement model can also be applied to analyze the haplotype assembly problem discussed in Section 1. As formulated and discussed in [15, 16], consider \( n \) SNPs on a chromosome, represented by a sequence \( \{x_1, \ldots, x_n\} \in \{0,1\}^n \) such that a major (resp. minor) allele is denoted by 0 (resp. 1). Employing a certain type of next generation sequencing technologies, one obtains a collection of independent paired reads such that for any \((i,j) \in \mathcal{E}\),

\[
y^{(k)}_{ij} = \begin{cases} x_i \oplus x_j, & \text{w.p. } 1 - \theta, \\ x_i \oplus x_j \oplus 1, & \text{w.p. } \theta, \end{cases}
\]

(72)

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

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 \( w > 0 \). This is captured by a generalized ring graph, denoted by \( G_{\text{ring}} = (V, \mathcal{E}_{\text{ring}}) \), such that

\[
(i,j) \in \mathcal{E}_{\text{ring}} \text{ iff } |i - j| \leq w.
\]

Besides, the number of reads \( L_{i,j} \) taken between \( i \) and \( j \) is assumed to be dependent on their separation, which obeys \( L_{i,j} = L |p_{|i-j|}| \)

(73)

for some parameters \( L \) and \( \{p_l | 1 \leq l \leq w\} \).

Additionally, a random and geometry-free measurement model has been investigated in [15] as well. The fundamental limits 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 3.** Consider the pairwise measurement model (72), and suppose that \( \theta \) and \( p_l \)'s are bounded away from 0.
(1) Suppose that $G \sim \mathcal{G}_{\text{ring}}$. There exist some universal constants $c_1 > c_2 > 0$ such that

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

\[
\inf_{\psi} P_{e}(\psi) \not\overset{n \to \infty}{\longrightarrow} 0 \quad \text{if } (1 - 2\theta)^2 \leq c_2 \frac{\log n}{L}.
\]

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

\[
\inf_{\psi} P_{e}(\psi) \overset{n \to \infty}{\longrightarrow} 0 \quad \text{if } (1 - 2\theta)^2 > c_4 \frac{\log n}{Lnp_{\text{obs}}},
\]

\[
\inf_{\psi} P_{e}(\psi) \not\overset{n \to \infty}{\longrightarrow} 0 \quad \text{if } (1 - 2\theta)^2 \leq c_5 \frac{\log n}{Lnp_{\text{obs}}}.
\]

Proof. For the sufficient condition, everything comes down to calculating the minimum Rényi divergence. For each $(i,j) \in \mathcal{E}$, letting $D_{1/2}^{i,j}$ be the Rényi divergence between the output distributions given two different input values, one gets

\[
-D_{1/2}^{i,j} \overset{(i)}{=} L_{i,j} \left\{ -2 \log \left( 1 - \frac{1}{2} \text{Hel}(\theta \| 1 - \theta) \right) \right\} \overset{(ii)}{=} L_{i,j} \text{Hel}(\theta \| 1 - \theta),
\]

where (i) follows from the additivity of Rényi divergence [31, Theorem 2.8], and (ii) follows since $1 - x \leq e^{-x}$. Furthermore,

\[
\frac{1}{2} \text{Hel}(\theta \| 1 - \theta) = (\sqrt{\theta} - \sqrt{1 - \theta})^2 = \frac{(1 - 2\theta)^2}{(\sqrt{\theta} + \sqrt{1 - \theta})^2} \approx (1 - 2\theta)^2.
\]

Recall that $L_{i,j} = L_{p_{\text{obs}}} \geq L$ when $G \sim \mathcal{G}_{\text{ring}}$, whereas $L_{i,j} = L$ when $G \sim \mathcal{G}_{n,p_{\text{obs}}}$. These taken together with Theorem 2 and Lemma 1 (resp. Theorem 6) establish the sufficient recovery condition for $\mathcal{G}_{\text{ring}}$ (resp. $\mathcal{G}_{n,p_{\text{obs}}}$).

Regarding the necessary condition, by replacing all $p_l$ with $\max_j p_l$ in (73), we obtain a statistically easier model to deal with; any impossibility result for the new model will also hold for the original model. We then proceed to compute the KL divergence for the new model, for which a little algebra yields

\[
\text{KL}_{\text{min}} \leq \text{KL}(\theta \| 1 - \theta) \leq L \chi^2(\theta \| 1 - \theta) = L \frac{(1 - 2\theta)^2}{\theta(1 - \theta)} \approx L(1 - 2\theta)^2,
\]

where the inequality follows from [31, Equation (7)]. Substituting (80) into Theorem 3 concludes the proof.

In [15][16], the fundamental limits were determined based on coverage (or sample complexity) as a metric, that is, the total number of reads required for perfect haplotype assembly. Recognizing that $nL$ (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

\[
nLw = \frac{n \log n}{(1 - 2\theta)^2}, \quad \text{when } G \sim \mathcal{G}_{\text{ring}};
\]

\[
\binom{n}{2}p_{\text{obs}} = \frac{n \log n}{(1 - 2\theta)^2}, \quad \text{when } G \sim \mathcal{G}_{n,p_{\text{obs}}}.
\]

Consequently, for the generalized ring graph, our results match the sample complexity limit characterized in [16], which is proportional to

\[
\frac{n \log n}{1 - e^{-\text{KL}(0.5 \| \theta)}} \geq \frac{n \log n}{\text{KL}(0.5 \| \theta)} = \frac{n \log n}{\frac{1}{2} (1 - 2\theta)^2 + o((1 - 2\theta)^2)}.
\]

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 $\Theta(n \log n)$ given in [15].

Notably, our results are not restricted to the classical large-sample asymptotics where $\theta$ is fixed and $n$ grows to infinity. This strengthens [15][16] by covering the regime where $\theta$ is close to $1/2$ (i.e. the situation where the parity reads are very noisy). As a final remark, while our results are tight in capturing the right scaling w.r.t. the read error rate $\theta$ as well as the number $n$ of SNPs, our derivation above is not tight in characterizing the behavior w.r.t. $p_l$ (or the notation $W$ given in [16]).
7 Concluding Remarks

This paper investigates simultaneous recovery of multiple node-variables based on noisy graph-based measurements, under the pairwise difference model that spans numerous applications. 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 that the feasibility of exact information recovery is captured by the minimum information divergence measures associated with the equivalent channel rather than the average mutual information. 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 limits will provide a general benchmark for evaluating the performance of practical algorithms over many applications.

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 improve in the partial recovery setting, namely, the situation where one only needs to guarantee 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. Finally, it would be of great interest to investigate computational tractability of information recovery, that is, whether there exists a nontrivial computational gap away from the statistical limits.

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, and Amir Dembo for his valuable instruction on large / moderate deviation theory.

A Proof of Theorem 1

Note that $\psi_{\text{ml}}$ distinguishes the null hypothesis $\mathbf{x} = \mathbf{x}^* = \{x_i^*\}_{1 \leq i \leq n}$ from the alternative hypothesis $\mathbf{x} = \mathbf{w} = \{w_i\}_{1 \leq i \leq n}$ 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 $\mathbf{x} = 0$, but note that the proof automatically works for all other ground truth values.

Let’s divide the set of all alternative hypotheses into several classes $A_k$ so that for each $k \geq 1$,

$$A_k := \{\mathbf{w} \neq 0 : |E \cap \text{supp}(\mathbf{w} \oplus \mathbf{w})| < k \cdot \text{mincut} \}. \quad (83)$$

Here, we employ the notation

$$E \cap \text{supp}(\mathbf{w} \oplus \mathbf{w}) := \{(i, j) \in E \mid w_i - w_j \neq 0, \ i > j \}.$$ For any $\mathbf{w} \in A_k$, if we let $S_l$ represent the set of vertices taking the value $l$, then one has

$$\sum_{l=0}^{M-1} n(S_l, S_l^c) = 2 |E \cap \text{supp}(\mathbf{w} \oplus \mathbf{w})| < 2k \cdot \text{mincut}. \quad (84)$$

On the other hand, consider the case where $k = 1$. All $\mathbf{w} \in A_1$ are equivalent to 0 up to some global offset. This is because for any non-trivial cut $(S_l, S_l^c)$, one must have $|E \cap \text{supp}(\mathbf{w} \oplus \mathbf{w})| \geq n(S_l, S_l^c) \geq \text{mincut}$, which violates the feasibility constraint $|E \cap \text{supp}(\mathbf{w} \oplus \mathbf{w})| < \text{mincut}$. In the following lemma, we link the cardinality of each hypothesis class $A_k$ with the cut-homogeneity exponent $\tau_{\text{cut}}$ defined in (26).

**Lemma 3.** The hypothesis class $A$ given in (83) satisfies

$$\frac{\log |A_k|}{k} < 2 \log M + 2 \log (2k \cdot \text{mincut}) + 4\tau_{\text{cut}}$$

$$\leq 2 \log M + 4 \log (2n) + 4\tau_{\text{cut}}. \quad (85)$$
Proof. See Appendix [C].

We are now in position to characterize the recovery ability of \( \psi_m \). Let \( P_w(\cdot) \) denote the measure given \( x = w \). Apparently, there are \( \frac{1}{2} \sum_{i=1}^{M-1} e(S_i, S_i^c) \) different locations \((i, j)\) satisfying \( i > j \) and \( w_i - w_j \neq 0 \), where \( e(S, S^c) \) is defined in Section 1.3. With this in mind, it follows from the Chernoff bound that

\[
\mathbb{P}_0 \left\{ \log \frac{dP_w(y)}{d\mathbb{P}_0(y)} > 0 \right\} = \mathbb{P}_0 \left\{ \sum_{(i,j) \in E, \ i > j} \alpha \log \frac{dP_w(y_{i,j})}{d\mathbb{P}_0(y_{i,j})} > 0 \right\} \leq \prod_{(i,j) \in E, \ i > j} \mathbb{E}_0 \left[ e^{\alpha \log \frac{dP_w(y_{i,j})}{d\mathbb{P}_0(y_{i,j})}} \right]
\]

\[
= \prod_{(i,j) \in E, \ i > j} [1 - (1 - \alpha) \operatorname{He}_{\alpha} (P_w(y_{i,j}) \| \mathbb{P}_0(y_{i,j}))]
\]

\[
= \exp \left\{ - (1 - \alpha) \sum_{(i,j) \in E, \ i > j} D_{\alpha} (P_w(y_{i,j}) \| \mathbb{P}_0(y_{i,j})) \right\}
\]

\[
\leq \exp \left\{ - (1 - \alpha) \frac{\sum_{i=0}^{M-1} e(S_i, S_i^c)}{2} D_{\alpha}^{\text{min}} \right\},
\]

where (86) follows from the definition of Hellinger divergence. When restricted to the hypotheses in \( \mathcal{A}_k \setminus \mathcal{A}_{k-1} \) for any \( k \geq 2 \), we know from the definition of \( \mathcal{A}_k \) that

\[
(k - 1) \text{mincut} \leq \sum_{i=0}^{M-1} e(S_i, S_i^c) < \text{mincut}.
\]

It then follows from the union bound that

\[
\mathbb{P}_0 \left\{ \exists w \in \mathcal{A}_k \setminus \mathcal{A}_{k-1} : \ log \frac{dP_w(y)}{d\mathbb{P}_0(y)} > 0 \right\} \leq |\mathcal{A}_k| \exp \left\{ - (1 - \alpha) \frac{\sum_{i=0}^{M-1} e(S_i, S_i^c)}{2} D_{\alpha}^{\text{min}} \right\}
\]

\[
\leq \exp \left\{ - (k - 1) \left( (1 - \alpha) \text{mincut} \cdot D_{\alpha}^{\text{min}} - \frac{k}{k - 1} \frac{\log |\mathcal{A}_k|}{k} \right) \right\}
\]

\[
\leq \exp \left\{ - (k - 1) \left( (1 - \alpha) \text{mincut} \cdot D_{\alpha}^{\text{min}} - \frac{2 \log |\mathcal{A}_k|}{k} \right) \right\}. \tag{89}
\]

This suggests that under the condition

\[
(1 - \alpha) \text{mincut} \cdot D_{\alpha}^{\text{min}} \geq (\delta + 8) \log (2n) + 8 \tau_{\text{cut}} + 4 \log (M),
\]

one has

\[
P_e (\psi_m) \leq \sum_{k \geq 2} \mathbb{P}_0 \left\{ \exists w \in \mathcal{A}_k \setminus \mathcal{A}_{k-1} : \ log \frac{dP_w(y)}{d\mathbb{P}_0(y)} > 0 \right\}
\]

\[
\leq \sum_{k \geq 2} \exp \left\{ - (k - 1) \left( (1 - \alpha) \text{mincut} \cdot D_{\alpha}^{\text{min}} - (4 \log M + 8 \log (2n) + 8 \tau_{\text{cut}}) \right) \right\}
\]

\[
\leq \sum_{k \geq 1} \exp (-k \cdot \delta \log (2n))
\]

\[
\leq \frac{1}{(2n)^{\delta^2}} \cdot \frac{1}{1 - (2n)^{-\delta}} = \frac{1}{(2n)^{\delta^2} - 1},
\]

where (90) follows by combining (89) and Lemma 3.

Finally, recognizing that \( D_{\alpha}^{\text{min}} \geq \operatorname{He}_{\alpha}^{\text{min}} \) concludes the proof for the condition based on \( \operatorname{He}_{\alpha}^{\text{min}} \).

B Proof of Theorem 3

Without loss of generality, assume that the minimum KL divergence can be approached by the following pairs of indices

\[
\text{KL} (P_1 \| P_0) = \text{KL}^{\text{min}},
\]

\[
\text{KL} (P_l \| P_0) \leq (1 + \epsilon) \text{KL}^{\text{min}}, \quad 2 \leq l \leq m^l,
\]

21
and suppose that both the ground truth and the null hypothesis are $x = x^* = 0$. We would like to ensure that the observation $y$ conditional on $x = 0$ is distinguishable from the observation $y$ under any alternative hypothesis $x \neq 0$.

(1) To begin with, recall the definition

$$
\mathcal{N}(k \cdot \text{mincut}) := \{S \subseteq V : e(S, S^c) \leq k \cdot \text{mincut}\}.
$$

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

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

This produces a collection of $|\mathcal{N}(k \cdot \text{mincut})|$ distinct alternative hypotheses, denoted by $B_k$. For each $w \in B_k$, the distributions $P_w$ and $P_0$ disagree only in those locations residing in the associated cut set, which amounts to at most $k \cdot \text{mincut}$ components. It then follows from the statistical independent assumption of $y_{ij}$ that

$$
\text{KL}(P_w || P_0) = e(S, S^c) \text{KL}_{\min} \leq k \cdot \text{mincut} \cdot \text{KL}_{\min}.
$$

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

$$
\frac{1}{|B_{k_0}|} \sum_{w \in B_{k_0}} \text{KL}(P_w || P_0) \leq (1 - \epsilon) \log |\mathcal{N}(k_0 \cdot \text{mincut})| - H(\epsilon),
$$

then one necessarily has $\inf_{\psi} P_{\psi}(\psi) \geq \epsilon$. With (91) in mind, we see that the condition (92) can be guaranteed when

$$
\text{KL}_{\min} \leq \frac{(1 - \epsilon) \tau^{\text{cut}} - H(\epsilon)}{\text{mincut}},
$$

which can further be ensured if

$$
\text{KL}_{\min} \leq \frac{(1 - \epsilon) \tau^{\text{cut}} - H(\epsilon)}{\text{mincut}}.
$$

(2) Next, suppose that the minimum cut is attained by $(S_{mc}, S^c_{mc})$. Consider another class $C$ of hypotheses consisting of $m^{kl}$ hypotheses. The $l$th candidate $w^{(l)}$ is given by

$$
\forall 1 \leq l \leq m^{kl} : w^{(l)}_i = \begin{cases} 
1, & \text{if } i \in S_{mc}, \\
0, & \text{otherwise,}
\end{cases}
$$

all of which obey

$$
\text{KL}(w^{(l)} || 0) \leq (1 + \epsilon) \text{mincut} \cdot \text{KL}_{\min}.
$$

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

$$
\frac{1}{m^{kl}} \sum_{l=1}^{m^{kl}} \text{KL}(w^{(l)} || 0) \leq (1 - \epsilon) \log m^{kl} - H(\epsilon).
$$

Observe from (93) that (94) can be ensured under the condition

$$
\text{KL}_{\min} \leq \frac{(1 - \epsilon) \log m^{kl} - H(\epsilon)}{(1 + \epsilon) \text{mincut}}.
$$

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

$$
\mathcal{H}_l := \{x : \|x\|_0 = 1, x \in \{0, 1\}^n\}, \quad 1 \leq l < M,
$$

where each class $\mathcal{H}_l$ is composed of $n$ distinct alternative hypotheses. This guarantees that for any $w \in \mathcal{H}_l$, the distributions of $y_{ij} \mid x = 0$ differ from that of $y_{ij} \mid x = w$ in at most $d_{\max}$ locations.
For any hypothesis class \(\mathcal{H}\) and any \(0 < \epsilon < \frac{1}{2}\), the Fano-type inequality [29, Equation (2.70)] suggests that \(\inf_{\psi} P_\epsilon(\psi) \geq \epsilon\) occurs in the regime where

\[
\frac{1}{|\mathcal{H}|} \sum_{w \in \mathcal{H}} \text{KL}(P_w \| P_0) \leq \frac{|\mathcal{H}| + 1}{|\mathcal{H}|} \left\{ (1 - \epsilon) \log |\mathcal{H}| - H(\epsilon) \right\}.
\]

(96)

By picking \(\mathcal{H}\) to be \(\mathcal{H} = \bigcup_{i=1}^{m^k} \mathcal{H}_i\) that obeys \(|\mathcal{H}| = m^k n\), we can see from definition of \(m^k\) that

\[
\frac{1}{|\mathcal{H}|} \sum_{w \in \mathcal{H}} \text{KL}(P_w \| P_0) \leq (1 + \epsilon) d_{\text{max}} \text{KL}_{\text{min}}
\]

and hence (96) holds under the condition

\[
\text{KL}_{\text{min}} \leq \frac{(1 - \epsilon) (\log n + \log m^k) - H(\epsilon)}{(1 + \epsilon) d_{\text{max}}}.
\]

(97)

Putting the above results together establishes Theorem 3.

C Proof of Theorem 4

In similar spirit of Theorem 3, assume that the minimum Hellinger divergence is achieved by the following pair of indices

\(\text{Hel}\_\alpha(P_1 \| P_0) = \text{Hel}\_\alpha^n\),

and let the ground truth and the null hypothesis be \(x = x^* = 0\). We focus on the set \(\mathcal{H}_i\) of configurations with binary alphabet as defined in (95).

For any class \(\mathcal{H}\) of alternative hypotheses, the minimax lower bound [35, Theorem II.1] suggests that every f-divergence \(D_f(\cdot)\) obeys

\[
\sum_{w \in \mathcal{H}} D_f(P_w \| P_0) \geq f(\langle |\mathcal{H}| (1 - P_\epsilon) \rangle) + (\langle |\mathcal{H}| - 1 \rangle f \left( \frac{|\mathcal{H}| P_\epsilon}{|\mathcal{H}| - 1} \right),
\]

where \(P_w\) stands for the probability measure of \(y = [y_{ij}]_{(i,j) \in \mathcal{E}}\) conditional on \(x = w\). When specialized to the Hellinger divergence of order \(\alpha\) (which corresponds to \(f(x) = \frac{1}{1 - \alpha} (1 - x^{\alpha})\)), the above inequality leads to

\[
(1 - \alpha) \sum_{w \in \mathcal{H}} \text{Hel}_\alpha(P_w \| P_0) \geq 1 - |\mathcal{H}|^{\alpha} (1 - P_\epsilon)^{\alpha} + (|\mathcal{H}| - 1) \left\{ 1 - \left( \frac{|\mathcal{H}| P_\epsilon}{|\mathcal{H}| - 1} \right)^{\alpha} \right\}
\]

\[
= |\mathcal{H}| - |\mathcal{H}|^{\alpha} (1 - P_\epsilon)^{\alpha} - (|\mathcal{H}| - 1)^{1 - \alpha} |\mathcal{H}|^{\alpha} P_\epsilon^{\alpha}
\]

\[
\geq |\mathcal{H}| - |\mathcal{H}|^{\alpha} - |\mathcal{H}| P_\epsilon^{\alpha}
\]

or, equivalently,

\[
P_\epsilon^{\alpha} \geq 1 - \frac{(1 - \alpha) \sum_{w \in \mathcal{H}} \text{Hel}_\alpha(P_w \| P_0)}{|\mathcal{H}|} = \frac{1}{|\mathcal{H}|^{1 - \alpha}}.
\]

(98)

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

\[
1 - (1 - \alpha) \text{Hel}_\alpha(P^n \| Q^n) = \int (dP^n)^{\alpha} (dQ^n)^{1 - \alpha} = \left( \int (dP)^{\alpha} (dQ)^{1 - \alpha} \right)^n
\]

\[
= (1 - (1 - \alpha) \text{Hel}_\alpha(P \| Q))^n.
\]

(99)
If all hypotheses $w \in H$ satisfy $\|w - x^\star\|_0 \leq k$, then $P_w$ and $P_0$ are different over at most $kd_{\text{max}}$ locations. Thus, if the divergence measure at each of these locations is identical and equal to some given value $h_\alpha$, then it follows from the statistical independence among $y_{ij}$ that

$$1 - (1 - \alpha) \text{Hel}_\alpha (P_w \| P_0) \geq (1 - (1 - \alpha) h_\alpha)^{kd_{\text{max}}}.$$ 

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

$$P_\alpha^\alpha \geq \left(1 - (1 - \alpha) h_\alpha\right)^{kd_{\text{max}}} - \frac{1}{|H|^{1 - \alpha}} \geq e^{-((1 - \alpha) h_\alpha + (1 - \alpha)^2 k_{\text{max}})kd_{\text{max}}} - \frac{1}{|H|^{1 - \alpha}},$$

which follows from the inequality that $\log (1 - x) \geq -x - x^2$ for any $0 \leq x \leq \frac{1}{2}$.

As a result, if the following condition holds

$$e^{-((1 - \alpha) h_\alpha + (1 - \alpha)^2 k_{\text{max}})kd_{\text{max}}} - \frac{1}{|H|^{1 - \alpha}} \geq \xi^\alpha$$

or, equivalently,

$$(1 - \alpha) h_\alpha \left[1 + (1 - \alpha) h_\alpha\right] \leq -\frac{\log \left(\xi^\alpha + |H|^{-(1 - \alpha)}\right)}{kd_{\text{max}}},$$

then for any test procedure, the minimax probability of error must obey $\inf_\psi P_\psi \geq \xi$. Solving the quadratic inequality (100) and utilizing the fact $\sqrt{1 + 4x} - 1 \geq 2x - 4x^2$ ($x \geq 0$) shows that (100) is guaranteed to hold as soon as

$$(1 - \alpha) h_\alpha \leq -\frac{\log \left(\xi^\alpha + |H|^{-(1 - \alpha)}\right)}{kd_{\text{max}}} - \frac{2 \log^2 \left(\xi^\alpha + |H|^{-(1 - \alpha)}\right)}{(kd_{\text{max}})^2}. \tag{101}$$

Finally, setting $\xi = n^{-\epsilon}$ and $H = H_1$ (cf. Definition (95)), one has $|H| = n$, $k = 1$ and $h_\alpha = \text{Hel}_\alpha^\min$. In the regime where

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

we have

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

The condition (101) is then guaranteed to hold if

$$(1 - \alpha) \text{Hel}_\alpha^{\min} \leq -\frac{\log \left(2\xi^\alpha\right)}{d_{\text{max}}} - \frac{2 \log^2 \left(2\xi^\alpha\right)}{d_{\text{max}}^2}, \tag{102}$$

which can be ensured if

$$\alpha \text{Hel}_\alpha^{\min} \leq \frac{\alpha \log n - \log 2}{(1 - \alpha) d_{\text{max}}} - \frac{2 \left[\alpha \log n - \log 2\right]^2}{(1 - \alpha) d_{\text{max}}^2}, \tag{103}$$

where we use $\xi^\alpha = n^{-\epsilon \alpha}$. Besides, the condition $(1 - \alpha) h_\alpha \leq \frac{1}{2}$ becomes $\text{Hel}_\alpha^{\min} \leq \frac{1}{2(1 - \alpha)}$, which can be ensured under (103) together with the condition

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

as claimed.
D Proof of Theorem 5

Suppose that the null hypothesis is \( \mathbf{x} = \mathbf{x}^* \). Consider the class of alternative hypotheses with any \( k \) (\( 1 \leq k \leq n \)) as follows

\[
\mathcal{H}_k := \{ \mathbf{x} \mid \| \mathbf{x} - \mathbf{x}^* \|_0 = n - k \},
\]

which comprises at most \( \binom{n}{k} (M - 1)^{n-k} \) distinct hypotheses. For notational convenience, denote by \( \mathbb{P}_w(\cdot) \) (resp. \( \mathbb{P}_0(\cdot) \)) the probability measure of \( \mathbf{y} \) conditional on the alternative hypothesis \( \mathbf{x} = \mathbf{w} \) (resp. the null hypothesis \( \mathbf{x} = \mathbf{x}^* \)). We let \( P_{e,\mathcal{H}_k} \) represent the probability of error when restricted to the class of alternative hypotheses \( \mathcal{H}_k \).

For any \( \mathbf{w} \in \mathcal{H}_k \), denote by \( S_i \) \( 0 \leq i < M \) the set of vertices \( v \) obeying \( v_0 - x^*_i = i \), and let \( n_i = |S_i| \). For any \( 0 < \alpha < 1 \), it follows from (88) that

\[
\mathbb{P}_0 \left\{ \log \frac{d\mathbb{P}_w(\mathbf{y})}{d\mathbb{P}_0(\mathbf{y})} > 0 \right\} \leq \exp \left( - (1 - \alpha) \sum_{i=0}^{M-1} e(S_i, S^*_i) D_{\alpha}^{\min} \right),
\]

For the time being, we claim that there exists some numerical value \( \zeta > 0 \) such that

\[
\sum_{i=0}^{M-1} e(S_i, S^*_i) \geq (1 - \zeta) p_{\text{obs}} \sum_{i=0}^{M-1} n_i (\alpha - n_i) = (1 - \zeta) p_{\text{obs}} \left( n^2 - \sum_{i=0}^{M-1} n_i^2 \right)
\]

holds for any partition \( \{S_i \mid 0 \leq i < M\} \). This claim will be proved later on in Lemma 4.

Recognize that the input is unique only up to global additive offset, that is, for any \( l \), the inputs \( \mathbf{w} \) and \( \mathbf{w} - l \cdot 1 \) result in the same pairwise inputs \( [w_i - w_j]_{1 \leq i,j \leq n} \). Therefore, we assume without loss of generality that

\[
k = n_0 \geq \max \{ n_1, n_2, \cdots, n_{M-1} \}.
\]

Letting \( \rho := \left\lfloor \frac{n_0}{k} \right\rfloor \), we claim that \( \sum_{i=0}^{M-1} n_i^2 \) under the constraint \( n_i \leq k \) is maximized by the configuration

\[
\begin{align*}
n_0 &= n_1 = \cdots = n_{\rho-1} = k, \\
n_\rho &= n - k\rho, \\
n_{\rho+1} &= \cdots = n_{M-1} = 0,
\end{align*}
\]

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

\[
(n_{\tilde{\rho}} + 1)^2 + (n_{\tilde{\rho}+1} - 1)^2 = n_{\tilde{\rho}}^2 + n_{\tilde{\rho}+1}^2 + 2(n_{\tilde{\rho}} - n_{\tilde{\rho}+1}) + 2 > n_{\tilde{\rho}}^2 + n_{\tilde{\rho}+1}^2.
\]

This results in contradiction, and hence \( \tilde{\rho} = \rho \). Similarly, we cannot have \( n_\rho < n - k\rho \), since replacing \( (n_\rho, n_{\rho+1}) \) with \( (n_\rho + 1, n_{\rho+1} - 1) \) leads to a strictly better solution. Consequently, for all \( \{n_i : 0 \leq i < M\} \) satisfying (107), one has

\[
\sum_{i=0}^{M-1} n_i^2 \leq \left\lfloor \frac{n}{k} \right\rfloor \cdot k^2 + \left( n - k \left\lfloor \frac{n}{k} \right\rfloor \right) \cdot k = nk,
\]

leaving us two cases below to deal with.

**Case 1.** Suppose that \( k \leq n/2 \). The inequality \( n - k \left\lfloor \frac{n}{k} \right\rfloor \leq k \) leads to

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

This combined with (105) and the claim (106) gives

\[
\mathbb{P}_0 \left\{ \log \frac{d\mathbb{P}_w(\mathbf{y})}{d\mathbb{P}_0(\mathbf{y})} > 0 \right\} \leq \exp \left( - \frac{(1 - \zeta) p_{\text{obs}} (1 - \alpha) (n^2 - nk)}{2} D_{\alpha}^{\min} \right).
\]

\[\text{12} \text{Otherwise, if } n_i = \max \{ n_1, n_2, \cdots, n_{M-1} \} \text{ instead, we can always enforce a global shift } i \text{ on } \mathbf{w} \text{ to yield } \mathbf{w} - i \cdot 1 \text{ in order to satisfy this condition without changing the output distribution.}\]
Substituting it into (110) yields that

\[ P_{e, \mathcal{H}_k} \leq \binom{n}{k} (M-1)^{n-k} \exp \left( -\frac{(1-\zeta) \log n}{2} \right) \]

for some universal constants \( C_1, c_1 > 0 \), where the second inequality uses the fact that \( \binom{n}{k} \leq 2^n \), and the last inequality follows since \( 2^n \ll n^{-\Theta(n \log n)} \). This approaches 0 (super)-exponentially fast.

**Case 2.** Consider the case where \( k > n/2 \). In this regime we have \( \left\lfloor \frac{n}{2} \right\rfloor = 1 \), and thus

\[
\sum_{i=0}^{M-1} n_i^2 \leq k^2 + (n-k)^2 = n^2 + 2k^2 - 2nk,
\]

This taken collectively with (105) and the claim (106) implies that

\[
P_0 \left\{ \frac{\log \frac{\partial^2 \mu_1 (y)}{\partial \nu_1 (y)}}{\partial \nu_1 (y)} > 0 \right\} \leq \exp \left( -\frac{(1-\zeta) \log n}{2} \right) \exp \left( -\frac{(1-\zeta) \log n}{2} \right).
\]

Apply the union bound over \( \mathcal{H}_k \) to get

\[
P_{e, \mathcal{H}_k} \leq \binom{n}{k} (M-1)^{n-k} \exp \left( -\frac{(1-\zeta) \log n}{2} \right) \]

For any constant \( \delta > 0 \), if the minimum Rényi divergence obeys

\[
P_{e, \mathcal{H}_k} \leq \exp \left( \log \left( \frac{n}{k} \right) + (n-k) \log (M-1) - \frac{(1-\zeta) \log n}{2} \right) \exp \left( \log \left( \frac{n}{k} \right) + (n-k) \log (M-1) - \frac{(1-\zeta) \log n}{2} \right) \]

then in the regime where \( k > n/2 \) one has

\[
(1-\zeta) \log n \geq \frac{(1+\delta) \log n}{n} + (n-k) \log (M-1) \]

Substituting it into (110) yields that

\[
P_{e, \mathcal{H}_k} \leq \exp \left( \log \left( \frac{n}{k} \right) + \frac{(1+\delta) \log n}{n} \right).
\]
We shall separate all vertices into two types as follows: See Appendix H.


(i) If \( \frac{k}{n} > 1 - \frac{\delta}{4} \), then the error probability is bounded by

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

\[
\leq \exp\left( (n-k) \log(2n) \cdot \left( 1 - \max\left( (1 + \delta) \left( 1 - \frac{\delta}{2}\right) \right) \right) \right)
\]

\[
\leq \exp\left( -\delta (n-k) \log(2n) \right),
\]

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

(ii) If \( \frac{k}{n} = 1 - \tau \) for some \( \frac{4}{7} \leq \tau \leq \frac{1}{2} \), then

\[
P_{e,\mathcal{H}_k} \leq \exp\left( n\mathcal{H}(\tau) - (1 + \delta) n (1 - \tau) \tau \log(2n) \right)
\]

\[
\leq \exp\left( -n (1 - \tau) \tau \log(2n) - \mathcal{H}(\tau) \right)
\]

\[
\leq C_2 \exp\left( -c_2 \delta n \log(2n) \right).
\]

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

\[
P_e \leq \sum_{k=\lceil \frac{n}{2} \rceil}^{n/2} P_{e,\mathcal{H}_k} + \sum_{k=n/2+1}^{n-1} P_{e,\mathcal{H}_k} + \sum_{k=(1-\delta)n}^{n-1} P_{e,\mathcal{H}_k}
\]

\[
\leq \frac{n}{2} \cdot C_1 n^{-c_1 n \log n} + \frac{n}{2} \cdot C_2 \exp\left( -c_2 \delta n \log n \right) + \sum_{k=(1-\delta)n}^{n-1} \exp\left( -\delta (n-k) \log(2n) \right)
\]

\[
\leq \frac{n}{2} \cdot C_1 n^{-c_1 n \log n} + \frac{n}{2} \cdot C_2 \exp\left( -c_2 \delta n \log n \right) + \frac{1}{(2n)^{\delta - 1}}
\]

\[
\leq C_0 e^{-c_0 n \log n} + \frac{1}{(2n)^{\delta - 1}}.
\]

with \( c_0, C_0 > 0 \) denoting some universal constants.

To finish up, we need to establish the claim \([106]\), which can be guaranteed with high probability as stated in the following lemma.

Lemma 4. For any constant \( \zeta \in (0,1) \), there exists a numerical value \( c(\zeta) \) that depends only on \( \zeta \) such that

\[
e(S,S^c) \geq (1 - \zeta) |S| (n - |S|) p_{\text{obs}}, \quad \forall S \subseteq V
\]

holds with probability at least \( 1 - \frac{\delta}{n^{\zeta} - 1} \), provided that \( p_{\text{obs}} \geq \frac{c(\zeta) \log n}{n} \).

Proof. See Appendix \([1]\) \( \square \)

The recovery condition based on \( \text{Hel}_\alpha^{\text{min}} \) is an immediate consequence from the inequality \( D_\alpha^{\text{min}} \geq \text{Hel}_\alpha^{\text{min}} \).

E Proof of Lemma \([1]\)

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

\[
e(S,S^c) \leq k \cdot \text{mincut}.
\]

(112)

We shall separate all vertices into two types as follows:

- Type-1 vertex: any vertex whose cut-edge degree is at least \( \frac{1}{2} k \rho \cdot \text{mincut} \);
- Type-2 vertex: any vertex whose cut-edge degree is less than \( \frac{1}{2} k \rho \cdot \text{mincut} \).
Figure 3: An example of the cut \((S, S^c)\) in a geometric graph. Here, \(S\) consists of all black vertices, while \(S^c\) contains all white vertices. The blue solid edges represent the cut edges.

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

To develop some intuitive understanding of the above notions, we depict in Fig. 3 an example of a cut \((S, S^c)\) in a geometric graph, where \(S^c\) 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. 3 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 \(112\), there exist no more than \(2k^{\kappa \rho}\) type-1 vertices. Simple combinatorial arguments thus suggest that there are at most \(\left(\frac{n}{2k^{\kappa \rho}}\right)^{\frac{n}{2k^{\kappa \rho}}}\) different ways to pick \(\frac{2k^{\kappa \rho}}{2k^{\kappa \rho}}\) vertices, \(2^{\frac{n}{2k^{\kappa \rho}}}\) ways to color all these type-1 vertices, and \(\left(k \cdot \text{mincut}\right)^{2k^{\kappa \rho}}\) different combinations of cut-edge degrees among them. Taken together these counting arguments imply that there exist no more than \(\left(\frac{n}{2k^{\kappa \rho}}\right)^{\frac{n}{2k^{\kappa \rho}}}\) different feasible ways to select the set of type-1 vertices as well as assign colors and cut-edge degrees for them, if one is required to satisfy the cut size constraint \(112\).

We claim that for any cut \((S, S^c)\) obeying \(112\), 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 \(S\) or \(S^c\)),

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. The following discussion concentrates only on black type-1 vertices without loss of generality.

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

\[
\left(1 - \frac{1}{2^k}\right) \rho \cdot \text{mincut},
\]

[13]Here, we use the fact that \(k \cdot \text{mincut} \leq n^2\), and hence \((k \cdot \text{mincut})^{2k^{\kappa \rho}} \leq n^{2k^{\kappa \rho}}\).
namely, \( v \) is connected with no more than \( (1 - \frac{1}{2} \kappa) \rho \cdot \text{mincut} \) white vertices. For any of its neighbors \( u \) (i.e. \( u, v \in E \)), 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 \( \text{[113]} \) of \( v \), the number of black vertices that \( u \) is linked with is at least

\[
|E(u) \cap E(v)| - \left(1 - \frac{1}{2} \kappa\right) \rho \cdot \text{mincut} \\
\geq \rho \cdot \text{mincut} - \left(1 - \frac{1}{2} \kappa\right) \rho \cdot \text{mincut} = \frac{1}{2} \kappa \rho \cdot \text{mincut},
\]

where the inequality follows from Assumption \( \text{[27]} \). This means that \( u \) must be a type-1 vertex (cf. definition of type-1 vertices) and its color must have been revealed, resulting in contradiction. In summary, all vertices with unknown colors around such a \( v \) 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 \( (1 - \frac{1}{2} \kappa) \rho \cdot \text{mincut} \). Again, consider any of its neighbors \( u \) whose color remains unknown, which must be incident to fewer than \( \frac{1}{2} \kappa \rho \cdot \text{mincut} \) cut edges since by construction it is a type-2 vertex. This already suggests the following fact: if there are at least \( \frac{1}{2} \kappa \rho \cdot \text{mincut} \) vertices falling in \( E(u) \cap E(v) \) 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 \( \frac{1}{2} \kappa \rho \cdot \text{mincut} \) vertices of opposite color. As a result, we can uniquely determine the color of \( u \) unless

- (P1) the colors of fewer than \( \kappa \rho \cdot \text{mincut} \) vertices\(^{14} \)

in \( E(u) \cap E(v) \) 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 \( \text{[27]} \) and the condition \( \kappa < \frac{1}{2} \), any white neighbor \( w \) of \( u \) must be connected with at least

\[
|E(u) \cap E(w)| - \frac{1}{2} \kappa \rho \cdot \text{mincut} = \left(1 - \frac{1}{2} \kappa\right) \rho \cdot \text{mincut} \geq \frac{1}{2} \kappa \rho \cdot \text{mincut}
\]

black vertices falling within \( E(u) \cap E(w) \), 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 \( E(w) \) 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 \( E(u) \cap E(v) \), then the color of \( u \) and all remaining vertices in \( E(u) \cap E(v) \) can be uniquely determined.

Now let \( w \) be the uncolored vertex in \( E(u) \cap E(v) \) that is the nearest to \( v \), which by (P1) must be within the \( (\kappa \rho \cdot \text{mincut}) \) closest vertices to \( v \) in \( E(u) \cap E(v) \). From Assumption \( \text{[28]} \), we see that \( w \) must be connected to all but \( \frac{1}{2} \rho \cdot \text{mincut} \) neighbors surrounding \( v \) and, as a result, be connected to at least

\[
|\text{cut-edge}(v)| - |E(v) \setminus E(w)| \geq \left(1 - \frac{1}{2} \kappa\right) \rho \cdot \text{mincut} - \frac{1}{2} \rho \cdot \text{mincut} \\
= \frac{1}{2} (1 - \kappa) \rho \cdot \text{mincut} \geq \frac{1}{2} \kappa \rho \cdot \text{mincut}.
\]

white vertices since \( \kappa < \frac{1}{2} \), where \( \text{cut-edge}(v) \) represents the set of cut edges incident to \( v \). 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 \( \frac{1}{2} \kappa \rho \cdot \text{mincut} \leq (1 - \frac{1}{2} \kappa) \rho \cdot \text{mincut} \), 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.

\(^{14}\) Otherwise there are either \( \frac{1}{2} \kappa \rho \cdot \text{mincut} \) white vertices or \( \frac{1}{2} \kappa \rho \cdot \text{mincut} \) black colors in \( E(u) \cap E(v) \) with their colors revealed.
All in all, we have thus demonstrated that the number of feasible coloring schemes is bounded above by
\[ n^{8/k \kappa \rho}, \] which in turn justifies
\[ \tau_{cut}^k \leq \frac{8 \log n}{k \rho}, \quad \forall k \geq 1. \]

(2) If \( G \) is an expander graph with edge expansion \( h_G \), then for any vertex set \( S \) with \( |S| \leq \frac{n}{2} \), one has
\[ |S| \leq \frac{e(S, S^c)}{h_G} \tag{114} \]
from the definition of \( h_G \). For any \( d > 0 \), if one requires that
\[ e(S, S^c) \leq kd, \tag{115} \]
then the above inequality leads to
\[ |S| \leq \frac{kd}{h_G}, \]
indicating that there are at most \( 2\left(\frac{n}{2}\right) \leq 2n^{kd/h_G} \) feasible cuts \((S, S^c)\) satisfying (115). Setting \( d = \mincut \) immediately leads to
\[ |N(k \cdot \mincut)| \leq 2 n^{\mincut/h_G}, \]
\[ \Rightarrow \tau_{cut}^k = \frac{1}{k} \log |N(k \cdot \mincut)| \leq \frac{\mincut \log n}{h_G} + \frac{\log 2}{k}, \quad \forall k \geq 1 \]
as claimed.

\section*{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
\[ \text{KL} (p \delta_k + (1-p) \text{Unif}_M \parallel p \delta_l + (1-p) \text{Unif}_M) \]
\[ = \left( p + \frac{1-p}{M} \right) \log \left( \frac{p + \frac{1-p}{M}}{\frac{1-p}{M}} \right) + \frac{1-p}{M} \log \left( \frac{1-p}{p + \frac{1-p}{M}} \right) \tag{116} \]
\[ = p \log \left( \frac{(M-1)p+1}{1-p} \right), \tag{117} \]
where \( \delta_k \) denotes the Dirac measure on the point \( k \), and (116) follows since the two distributions under study differ only at two points \( x = k \) and \( x = l \). Similarly, one obtains (cf. Definition 4)
\[ \text{Hel}_{\frac{1}{2}} (p \delta_k + (1-p) \text{Unif}_M \parallel p \delta_l + (1-p) \text{Unif}_M) \]
\[ = 2 \left( \sqrt{p + \frac{1-p}{M}} - \sqrt{\frac{1-p}{M}} \right)^2 = \frac{2}{M} \left( \sqrt{(M-1)p+1} - \sqrt{1-p} \right)^2. \tag{118} \]

When applied to the outlier model, these suggest
\[ \text{KL}^{\text{min}} = p_{\text{true}} \log \left( \frac{1 + \frac{p_{\text{true}} M}{1 - p_{\text{true}}} \right) \leq \frac{p_{\text{true}} M}{1 - p_{\text{true}}}, \tag{119} \]
and
\[ \text{Hel}^{\text{min}}_{\frac{1}{2}} = \frac{2}{M} \left( \sqrt{1 - p_{\text{true}} + M p_{\text{true}}} - \sqrt{1 - p_{\text{true}}} \right)^2. \tag{120} \]

It remains to control the Hellinger divergence. To this end, the elementary identity \( a - b = \frac{a^2 - b^2}{a + b} \) gives
\[ \left( \sqrt{1 - p_{\text{true}} + M p_{\text{true}}} - \sqrt{1 - p_{\text{true}}} \right)^2 = \left( \frac{p_{\text{true}} M}{\sqrt{1 - p_{\text{true}} + M p_{\text{true}}} + \sqrt{1 - p_{\text{true}}}} \right)^2 \]
\[ \geq \left( \frac{p_{\text{true}} M}{2 \sqrt{1 - p_{\text{true}} + M p_{\text{true}}}} \right)^2 = \frac{p_{\text{true}} M^2}{4 (1 - p_{\text{true}} + M p_{\text{true}})}, \]
indicating that \( \text{Hel}^{\text{min}}_{\frac{1}{2}} \geq \frac{p_{\text{true}} M^2}{2(1 - p_{\text{true}} + M p_{\text{true}})} \) as claimed.
G Proof of Lemma 3

Consider any hypothesis $x = w \in A_k$, which obeys $|E \cap \text{supp}(w \oplus w)| < k \cdot \text{mincut}$. Denote by $S_l$ the set of vertices that takes the value $l$ ($0 \leq l < M$), let $\mathcal{I} := \{ l \mid S_l \neq \emptyset \}$ represent the indices of those non-empty ones. Our proof proceeds by evaluating the following quantities:

1. How many different choices of $\mathcal{I} \neq \emptyset$ are admissible?
2. For each given $\mathcal{I} \neq \emptyset$, how many combinations of cut-set sizes \( \{e(S_l, S_{l}^{c}) \mid l \in \mathcal{I} \neq \emptyset \} \) are feasible?
3. For each given cut-set size $N_l$, how many cuts $(S_l, S_{l}^{c})$ are compatible with the constraint $e(S_l, S_{l}^{c}) \leq N_l$?

Clearly, multiplying all these quantities together gives rise to an upper bound on $|A_k|$.

We now compute the above quantities separately.

- To begin with, our assumption on the min-cut size ensures that $e(S_l, S_{l}^{c}) \geq \text{mincut}$

for each non-empty $S_l$. This together with the feasibility constraint

$$2 |E \cap \text{supp}(w \oplus w)| = \sum_{l=0}^{M-1} e(S_l, S_{l}^{c}) \leq 2k \cdot \text{mincut}$$

guarantees that the number of non-empty $S_l$’s cannot exceed $2k$. Consequently, there exist at most $(M \choose 2k) \leq M^{2k}$ possible combinations of $\mathcal{I} \neq \emptyset$.

- Secondly, from (122), the total cut-set size is bounded above by $2k \cdot \text{mincut}$. Therefore, for any given $\mathcal{I} \neq \emptyset$, there are no more than

$$\left( \frac{2k \cdot \text{mincut}}{|\mathcal{I} \neq \emptyset|} \right)^{|\mathcal{I} \neq \emptyset|} \leq (2k \cdot \text{mincut})^{2k}$$

feasible ways to assign cut-set sizes $e(S_l, S_{l}^{c})$ for all $l \in \mathcal{I} \neq \emptyset$.

- Thirdly, suppose that for each $l \in \mathcal{I} \neq \emptyset$,

$$e(S_l, S_{l}^{c}) = c_l \cdot \text{mincut}$$

for some numerical values $c_l \geq 1$. From the definition (26), the number of feasible choices of $(S_l, S_{l}^{c})$ compatible with (123) is bounded above by

$$|\mathcal{N}(c_l \cdot \text{mincut})| \leq |\mathcal{N}([c_l] \cdot \text{mincut})| \leq \exp ([c_l] \cdot \tau_{\text{cut}}) \leq \exp (2c_l \cdot \tau_{\text{cut}}).$$

Recognize that the constraint (122) requires

$$\sum_{l} c_l < 2k.$$

As a result, when the cut sizes $e(S_l, S_{l}^{c})$ are given, the total number of valid partitions $\{S_l \mid 0 \leq l < M\}$ cannot exceed

$$\prod_{l=0}^{M-1} \exp (2c_l \cdot \tau_{\text{cut}}) < \exp (4k \cdot \tau_{\text{cut}}).$$

(124)

Putting the above combinatorial bounds together implies that

$$|A_k| < M^{2k} (2k \cdot \text{mincut})^{2k} \exp (4k \cdot \tau_{\text{cut}}).$$

Using the inequality $k \cdot \text{mincut} \leq n^2$ we conclude the proof.
H Proof of Lemma 4

Fix \( \zeta \in (0,1) \). Since \( e(S, S^c) = e(S^c, S) \), it suffices to consider the case where \( |S| \leq \frac{n}{2} \).

1. Consider the case where \( \frac{4 \log n}{\rho_{\text{obs}}} \leq |S| \leq \frac{n}{2} \). If \( |S| = s \), then we recognize that

\[
e(S, S^c) \sim \text{Binomial}(s(n-s), \rho_{\text{obs}}).
\]

Applying the Chernoff-type bound [38, Theorem 4.4] yields

\[
\mathbb{P}\{ e(S, S^c) \leq (1 - \zeta) \rho_{\text{obs}} s (n-s) \} \leq \exp\left(-\frac{\zeta^2}{2} \rho_{\text{obs}} s (n-s) \right).
\]

This together with the union bound implies that: for any \( \frac{4 \log n}{\rho_{\text{obs}}} \leq s \leq \frac{n}{2} \), one has

\[
\mathbb{P}\{ \exists S : |S| = s \text{ and } e(S, S^c) \leq (1 - \zeta) \rho_{\text{obs}} |S| (n-|S|) \} \leq \sum_{1 \leq s \leq \frac{n}{2}} \exp\left(-\frac{\zeta^2}{4} \rho_{\text{obs}} s (n-s) \right).
\]

where (125) results from the assumption that \( s \geq \frac{4 \log n}{\rho_{\text{obs}}^2} \). If \( \rho_{\text{obs}} \geq \frac{80 \log n}{c_4 n^2} \), then using the union bound we get

\[
\mathbb{P}\{ \exists S : \frac{4 \log n}{\rho_{\text{obs}}^2} \leq |S| \leq \frac{n}{2} \text{ and } e(S, S^c) \leq (1 - \zeta) \rho_{\text{obs}} |S| (n-|S|) \} \leq \frac{1}{n^{10}} \frac{1}{n^{-10}} = \frac{1}{n^{20}}.
\]

2. Consider the case where \( 1 \leq |S| \leq \min\left\{ \frac{4 \log n}{\rho_{\text{obs}}^2}, \frac{1}{12} \zeta n \right\} \). If \( |S| = s \), then we can see that

\[
e(S, S) \sim \text{Binomial}(s^2, \rho_{\text{obs}}).
\]

Applying the Chernoff-type bound [38, Theorem 4.4] suggests that if \( \zeta \rho_{\text{obs}} s (n-s) \geq 6 \rho_{\text{obs}} s^2 \) or, equivalently, \( \zeta \geq \frac{6s}{n-\frac{n}{2}} \), then

\[
\mathbb{P}\{ e(S, S) \geq \zeta \rho_{\text{obs}} s (n-s) \} \leq 2^{-\zeta \rho_{\text{obs}} s (n-s)}.
\]

Employing the union bound we get: for any \( 1 \leq s \leq \min\left\{ \frac{4 \log n}{\rho_{\text{obs}}^2}, \frac{1}{12} \zeta n \right\} \), one has

\[
\mathbb{P}\{ \exists S : |S| = s \text{ and } e(S, S) \geq \zeta \rho_{\text{obs}} s (n-s) \} \leq \frac{n}{s} \exp\left(-\zeta \rho_{\text{obs}} s (n-s) \log 2 \right) \leq \exp\left(-s \left( \zeta \rho_{\text{obs}} (n-s) \log 2 - \log n \right) \right) \leq \exp\left(-s \left( \frac{\zeta \log 2}{2} \rho_{\text{obs}} n - \log n \right) \right),
\]

where the last inequality relies on the assumption that \( s \leq \frac{n}{2} \). If \( \rho_{\text{obs}} \geq \frac{22}{5} \log 2 \frac{\log n}{n} \), then

\[
\mathbb{P}\{ \exists S : |S| = s \text{ and } e(S, S) \geq \zeta \rho_{\text{obs}} s (n-s) \} \leq \exp(-10s \log n),
\]

32
and hence
\[\mathbb{P}\left\{ \exists S : 1 \leq |S| \leq \min \left\{ \frac{4 \log n}{\frac{12}{n} \zeta n}, \frac{4 \log n}{n} \zeta n \right\} \quad \text{and} \quad e(S, S) \geq \zeta_{\text{obs}} |S| (n - |S|) \right\} \]

\[\leq \sum_{1 \leq s \leq \frac{12}{n}} \exp \left( -10s \log n \right) \]

\[\leq \frac{1}{n^{10} - 1}.
\]

On the other hand, the Chernoff bound \[39\] Theorem 4.5 together with the union bound indicates that
\[\mathbb{P}\{d_{\text{min}} \leq (1 - \zeta) p_{\text{obs}} n\} \leq n \exp \left( -\frac{\zeta^2}{2} p_{\text{obs}} n \right) \leq \frac{1}{n^{10}},\]

provided that \( p_{\text{obs}} \geq \frac{22 \log n}{n} \). Putting the above bounds together reveals that on an event of high probability,
\[e(S, S^c) \geq |S| d_{\text{min}} - e(S, S) \]

\[\geq (1 - \zeta) |S| p_{\text{obs}} n - \zeta_{\text{obs}} s (n - s) \]

\[\geq (1 - 2\zeta) |S| p_{\text{obs}} (n - |S|)\]

holds simultaneously for all \( S \) with \( 1 \leq |S| \leq \min \left\{ \frac{4 \log n}{\frac{12}{n} \zeta n}, 1 \right\} \).

Finally, we note that the assumption \( \zeta \geq \frac{6a}{n - s} \) is guaranteed as long as \( s \leq \frac{1}{12} \zeta n \), and that when \( p_{\text{obs}} \geq \frac{48 \log n}{\zeta^2 n} \), one has \( \min \left\{ \frac{4 \log n}{\frac{12}{n} \zeta n}, 1 \right\} = \frac{4 \log n}{\zeta_n} \), completing the proof.

I 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 \( f_2(x) = (\sqrt{x} - 1)^2 \), respectively. That said, one can write

\[\text{KL}(P \| Q) = \mathbb{E}_Q \left[ f_1 \left( \frac{dP}{dQ} \right) \right] \quad \text{and} \quad \text{Hel}_1(P \| Q) = \mathbb{E}_Q \left[ f_2 \left( \frac{dP}{dQ} \right) \right].\]

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

\[(2 - 0.5 |\log x|) f_2(x) \leq f_1(x) \leq (2 + |\log x|) f_2(x), \quad \forall x > 0.\]

This immediately establish that

\[\text{KL}(P \| Q) = \mathbb{E}_Q \left[ f_1 \left( \frac{dP}{dQ} \right) \right] \leq (2 + \log R) \mathbb{E}_Q \left[ f_2 \left( \frac{dP}{dQ} \right) \right] = (2 + \log R) \text{Hel}_1(P \| Q)\]

and

\[\text{KL}(P \| Q) = \mathbb{E}_Q \left[ f_1 \left( \frac{dP}{dQ} \right) \right] \geq (2 - 0.5 \log R) \mathbb{E}_Q \left[ f_2 \left( \frac{dP}{dQ} \right) \right] = (2 - 0.5 \log R) \text{Hel}_1(P \| Q).\]

These together with the well known inequality \[29\] Lemma 2.4

\[\text{KL}(P \| Q) \geq \text{Hel}_1(P \| Q)\]

establish \[23\].

Similarly, from the inequality

\[(2 - 0.4 |\log x|) f_2(x) \leq f_1(x) \leq (2 + 0.4 |\log x|) f_2(x), \quad \forall x \in (0, 4.5],\]

one can show that

\[\max \{2 - 0.4 \log R, 1\} \cdot \text{Hel}_1(P \| Q) \leq \text{KL}(P \| Q) \leq (2 + 0.4 \log R) \cdot \text{Hel}_1(P \| Q) \tag{128}\]

as long as \( R \leq 4.5 \), as claimed.
References