Application of information-percolation method to reconstruction problems on graphs

Yury Polyanskiy and Yihong Wu*

Abstract

In this paper we propose a method of proving impossibility results based on applying strong data-processing inequalities to estimate mutual information between sets of variables forming certain Markov random fields. The end result is that mutual information between two "far away" (as measured by the graph distance) variables is bounded by the probability of the existence of an open path in a bond-percolation problem on the same graph. Furthermore, stronger bounds can be obtained by establishing mutual information comparison results with an erasure model on the same graph, with erasure probabilities given by the contraction coefficients.

As applications, we show that our method gives sharp threshold for partially recovering a rank-one perturbation of a random Gaussian matrix (spiked Wigner model), yields the best known upper bound on the noise level for group synchronization (obtained concurrently by Abbe and Boix), and establishes new impossibility result for community detection on the stochastic block model with k communities.

Contents

*Y.P. is with the Department of EECS, MIT, Cambridge, MA, email: yp@mit.edu. Y.W. is with the Department of Statistics and Data Science, Yale University, New Haven, CT, email: yihong.wu@yale.edu.

1 Introduction

As a generalization of ideas of Evans-Schulman [ES99], a method for upper-bounding the mutual information between sets of variables via the probability of the existence of a percolation path was proposed by the authors in [PW17, Theorem 5]. This allows one to reuse results on critical threshold for percolation to show the vanishing of mutual information. The original bound was stated for Bayesian networks (known as directed graphical models). In this paper we show that similar results can be obtained for certain Markov random fields (*undirected* graphical models) too, especially those arising in statistical reconstruction problems on graphs such as community detection and group synchronization.

Our original motivation was to improve the bound on the phase transition threshold in the $\mathbb{Z}_2$-synchronization on a 2D square grid which appeared in the work of Abbe, Massoulié, Montanari, Sly and Srivastava [AMM+18]. The possibility of such an improvement was anticipated by Abbe and Boix [Abb18], who presented their work in [AB18a], concurrently with the initial circulation of this work. The resulting improvement is stated below as Corollary 5 and concides with the result in [AB18a, AB18b].

The paper is organized as follows. First, we present the idea in its simplest form (binary labels and binary symmetric channels) in Section 2. Second, we extend the method in two different directions in Section 3 to general (non-binary) labels and channels, and in Section 4 to non-independent labels. To showcase our general results in Section 5 we consider three applications (of which the first two are chosen following [AB18a]): group synchronization, spiked Wigner model and stochastic block model with k blocks. For the latter our results strengthen (in some regime) the best known impossibility results on correlated (partial) recovery for $k = 3$. One of the main technical tools is the strong data processing inequality for mutual information, which is surveyed in the [PW17]; in Appendix A we provide a quick review emphasizing binary-input channels. We conclude with Appendix B comparing our results with the work of Abbe and Boix [AB18b].

2 Information–percolation bound (basic version)

We start by recalling some basic notions from information theory; cf. e.g. [CT06]. The mutual information $I(X;Y)$ between random variables X and Y with joint law P_{XY} is $I(X;Y) = D(P_{XY}\|P_X \otimes P_Y)$, where $D(P\|Q)$ is the Kullback-Leibler (KL) divergence between distributions P and Q, defined as $D(P\|Q) = \int dP \log \frac{dP}{dQ}$ if $P \ll Q$ and ∞ otherwise. In addition, the χ^2-divergence is defined as $\chi^2(P\|Q) = \int dP(\frac{dP}{dQ} - 1)^2$ if $P \ll Q$ and ∞ otherwise, and the squared Hellinger distance is $H^2(P,Q) = \int(\sqrt{\frac{dP}{d\mu}} - \sqrt{\frac{dQ}{d\mu}})^2 d\mu$ for any μ such that $P \ll \mu$ and $Q \ll \mu$. For discrete X, $I(X;Y) = H(X) - H(X|Y)$, the difference between the Shannon entropy of X and the conditional entropy of X given Y.

Two properties of mutual information are particularly useful for the present paper: (a) Chain rule: $I(X;Y,Z) = I(X;Y) + I(X;Z|Y)$, where $I(X;Z|Y)$ is the conditional mutual information. (b) *Data processing inequality* (DPI): whenever $W \to X \to Y$ forms a Markov chain, we have $I(W;Y) \le I(W;X)$. Furthermore, a quantitative version of the DPI is the *strong data processing inequality* (SDPI),

$$I(W;Y) \le \eta(P_{Y|X}) I(W;X) \tag{1}$$

where $\eta(P_{Y|X}) \in [0,1]$ is called the KL contraction coefficient of the channel. For example, if $P_{Y|X}$ is the binary symmetric channel (BSC) with flip probability δ, denoted by $\mathsf{BSC}(\delta)$, that is, $Y = X + Z$

mod $2 \triangleq X \oplus Z$, where $Z \sim \mathrm{Bern}(\delta)$ is independent of X, we have $\eta(\mathsf{BSC}(\delta)) = (1-2\delta)^2$. For more on SDPI, we refer the reader to the survey [PW17] and the references therein.

Let $\mathrm{ER}(G,p)$ denote the Erdös-Rényi random graph on the vertex set V, where each edge $e \in E$ is kept independently with probability p. Abbreviate $\mathrm{ER}(K_n,p)$ as $\mathrm{ER}(n,p)$, where K_n is the complete graph on $[n]$.

In this section we consider the following graphical model. Let $G=(V,E)$ be a simple undirected graph with finite or countably-infinite V. Let $\{X_v : v \in V\}$ be $\overset{\text{i.i.d.}}{\sim}\mathrm{Bern}(1/2)$ and Let $\{Z_e : e \in E\}$ be $\overset{\text{i.i.d.}}{\sim}\mathrm{Bern}(\delta)$. For each $e=(u,v) \in E$, let $Y_e = X_u \oplus X_v \oplus Z_e$. For any S, let $X_S = \{X_v : v \in S\}$.

Theorem 1. *For any subset $S \subset V$ and any vertex $v \in V$,*

$$I(X_v; X_S, Y_E) \le \mathrm{perc}_G(v,S) \log 2, \tag{2}$$

where $\mathrm{perc}_G(v,S) = \mathbb{P}[v \text{ is connected to } S \text{ in } \mathrm{ER}(G,\eta)]$, with

$$\eta \triangleq (1-2\delta)^2.$$

Remark 1. Notice that right-hand side of (2) can be seen as $I(X_v; X_S, \tilde{Y}_E)$ where for $e=(u,v)$, $\tilde{Y}_e$ is a random variable equal to $X_u \oplus X_v$ with probability η and $*$ (erasure) otherwise. This is not accidental – it can be shown via [PW17, Prop. 15, 16] that observations over the erasure channel $\mathsf{BEC}(\eta)$ lead to strictly larger mutual informations: $I(X_{S_1}; Y_E | X_{S_2}) \le I(X_{S_1}; \tilde{Y}_E | X_{S_2})$, regardless of the joint distribution P_{X_V}. This generalization is pursued in Section 4.

Proof. By the monotone convergence property of mutual information (and probability), it suffices to consider finite graph G.

Let $\bar{X}_V = \{X_v \oplus 1 : v \in V\}$. The symmetry of the problem shows that

$$(X_V, Y_E) \overset{d}{=} (\bar{X}_V, Y_E).$$

In particular, we have

$$I(X_v; Y_E) = 0 \tag{3}$$

for any v.

Fix V and $v \in V$. We induct on the number of edges $|E|$. For the base case of $E=\emptyset$, by the independence of $\{X_v\}$, we have

$$I(X_v; X_S) = \mathbf{1}_{\{v \in S\}} \log 2 = \mathrm{perc}_G(v,S) \log 2.$$

Next suppose (2) holds for all $G'=(V,E')$ with $|E'| < |E|$ and all S, i.e.

$$I(X_z; X_S, Y_{E'}) \le \mathrm{perc}_{G'}(z,S) \log 2. \tag{4}$$

We now show (2) holds for E. Fix S. Suppose there is no edge in E incident to any vertex in S. Then both sides of (2) are zero by (3). Otherwise, there exists an edge $e=(u,z) \in E$ incident to some vertex $z \in S$. Set $E' = E \setminus e$ and $G' = (V,E')$.

Next we apply the strong data processing inequality (SDPI) for BSC (see [PW17] for a survey on SDPIs): Note that $Y_e = X_u + X_z + Z_e$, where $Z_e \overset{\text{i.i.d.}}{\sim}\mathrm{Bern}(\delta)$ and independent of X_V. Since $z \in S$, conditioned on $(X_S, Y_{E'})$, we have the Markov chain: $X_v \to X_u \to Y_e$. Therefore

$$I(X_v; Y_e | X_S, Y_{E'}) \le \eta I(X_v; X_u | X_S, Y_{E'}).$$

Adding $I(X_v; X_S, Y_{E'})$ to both sides gives

$$I(X_v; X_S, Y_E) \le \eta I(X_v; X_S, X_u, Y_{E'}) + \bar{\eta} I(X_v; X_S, Y_{E'}).$$

Applying the induction hypothesis (4) to the RHS of the above display, we have:

$$I(X_v; X_S, Y_E) \le \underbrace{(\eta \cdot \mathrm{perc}_{G'}(v, S \cup \{u\}) + \bar{\eta}\mathrm{perc}_{G'}(v, S))}_{=\mathrm{perc}(G)(v,S)} \cdot \log 2$$

□

2.1 Simple example of tightness of the bound

Let G be a complete infinite d-ary tree rooted at node ρ. Let S_k denote the set of all nodes at depth k. Then, by results on broadcasting on trees [EKPS00], it is easy to see that[1]

$$\lim_{k\to\infty} I(X_\rho; Y_E, X_{S_k}) = \begin{cases} 0, & (1-2\delta)^2 d \le 1 \\ > 0, & (1-2\delta)^2 d > 1 \end{cases} \tag{5}$$

The bound in Theorem 1 is tight in this case in the sense that the right-hand side of (2) converges to zero if and only if the branching process with offspring distribution $\mathrm{Binom}(d, \eta)$ (with $\eta = (1-2\delta)^2$) is ultimately extinct almost surely, which occurs when $(1-2\delta)^2 d \le 1$ by standard results in branching process [AN72].

3 General version: information percolation

Consider a bipartite graph $G = (V, W, E)$ with parts V, W and edges E, with finite or countably-infinite V, W, E. For any subset $W' \subset W$ we will denote $G[W']$ the induced subgraph on vertices $V \cup W'$.

Let $\{X_v : v \in V\}$ be a collection of independent discrete random variables. Let $\{Y_w : w \in W\}$ be a collection of random variables conditionally independent given X_V and distributed each as

$$Y_w \sim P_{Y_w|X_{N(w)}} \qquad \forall w \in W\,, \tag{6}$$

where $N(w) \subset V$ denote the neighborhood of w in the bipartite graph G. Let $\eta_w \triangleq \eta_{KL}(P_{Y_w|X_{N(w)}})$ be the SDPI constant corresponding to this channel [PW17].

Let $\tilde{G}$ denote the subgraph $G[\tilde{W}]$ induced by the random subset $\tilde{W}$, where each vertex $w \in W$ is included in $\tilde{W}$ independently with probability η_w. For a pair of sets $S_1, S_2 \subset V$ we define the average number of vertices in S_1 that are connected to S_2:

$$\mathrm{perc}_G(S_1, S_2) \triangleq \sum_{v \in S_1} \mathbb{P}[v \text{ is connected to } S_2 \text{ in } \tilde{G}]\,.$$

We note the following simple identity: if w is such that $N(w) \cap S_2 \neq \emptyset$ then

$$\mathrm{perc}_G(S_1, S_2) = \eta_w \mathrm{perc}_{G[W\setminus w]}(S_1, S_2 \cup N(w)) + (1-\eta_w)\mathrm{perc}_{G[W\setminus w]}(S_1, S_2)\,. \tag{7}$$

[1] Indeed, due to the fact that X_v's are iid $\mathrm{Bern}(\frac{1}{2})$, we have $I(X_\rho; Y_E, X_{S_k}) = I(X_\rho; Z_{S_k})$, where for each $u \in S_k$, $Z_u = X_u + \sum_{e \in \text{ path from } \rho \text{ to } u} Y_e$. This is precisely the setting of broadcasting on trees, where the label at each node is obtained by passing that of its parent through $\mathsf{BSC}(\delta)$ independently. It was found in [EKPS00] that the cutoff of the total variation $d_{\mathrm{TV}}(P_{Z_{S_k}|X_\rho=+}, P_{Z_{S_k}|X_\rho=-})$ (and equivalently, the mutual information $I(X_\rho; Z_{S_k})$) happens at the threshold $(1-2\delta)^2 d \le 1$.

To recover the setting of the previous section, where the graph was simple, we can consider the bipartite graph which is the incidence graph between vertices and edges (in this case the degree of every $w \in W$ is two).

Theorem 2. *For any subsets S_1, S_2 of V, we have*

$$I(X_{S_1}; X_{S_2}|Y_W) \le \mathrm{perc}_G(S_1, S_2) \cdot \sup_{v \in V} H(X_v)\,. \tag{8}$$

Remark 2. Note that $I(X_{S_1}; Y_W) = 0$ does not hold even in the setting of the previous section, unless S_1 is a singleton (see (3)). Indeed, one may consider the graph $a - b - c$ in the context of Theorem 1. For $S_1 = \{a, c\}$, $I(X_{a,c}; Y_{ab,bc}) \ge I(X_a + X_c; Y_{ab} + Y_{bc}) \ge 1 - h(2\delta(1-\delta))$. Thus $I(X_{S_1}; X_{S_2}, Y_W) \ne I(X_{S_1}; X_{S_2}|Y_W)$ and the former does not satisfy the inequality in Theorem 2.

Proof. Again, because of the identity

$$I(X_{S_1}; X_{S_2}|Y_W) = I(X_{S_1}; X_{S_2}, Y_W) - I(X_{S_1}; Y_W)$$

and the continuity of mutual information and percolation probability we may only consider finite S_1, S_2, W.

We will prove (8) by induction on $|W|$. Assume that

$$H(X_v) \le H_1 \qquad \forall v \in V\,.$$

First, suppose that $W = \emptyset$. We have then:

$$I(X_{S_1}; X_{S_2}) = \sum_{i \in S_1 \cap S_2} H(X_i) \le |S_1 \cap S_2| H_1 = \mathrm{perc}_{G[W]}(S_1, S_2) H_1\,.$$

Next, suppose that we have shown (8) for all $G[W']$ with $|W'| < |W|$. Consider two cases:

Case 1. There does not exist $w \in W$ such that $N(w) \cap S_2 \ne \emptyset$. Then, we have

$$I(X_{S_1}; X_{S_2}|Y_W) \le I(X_{S_1}, Y_W; X_{S_2}) \le I(X_{S_1}, X_{S_0}; X_{S_2}) \le |S_1 \cap S_2| H_1\,,$$

where $S_0 = \cup_{w \in W} N(w)$ and the last equality is due to $S_0 \cap S_2 = \emptyset$. Similarly, we have

$$\mathrm{perc}_G(S_1, S_2) = |S_1 \cap S_2|$$

and (8) is established.

Case 2. There exists $w \in W$ such that $N(w) \cap S_2 \ne \emptyset$. Let $W' = W \setminus w$. Then we have

$$\begin{aligned} I(X_{S_1}; X_{S_2}, Y_{W'}, Y_w) &= I(X_{S_1}; X_{S_2}, Y_{W'}) + I(X_{S_1}; Y_w|X_{S_2}, Y_{W'}) \\ &\le I(X_{S_1}; X_{S_2}, Y_{W'}) + \eta_w I(X_{S_1}; X_{N(w)}|X_{S_2}, Y_{W'}) \\ &= (1-\eta_w) I(X_{S_1}; X_{S_2}, Y_{W'}) + \eta_w I(X_{S_1}; X_{N(w) \cup S_2}, Y_{W'})\,, \\ &= (1-\eta_w) I(X_{S_1}; X_{S_2}|Y_{W'}) + \eta_w I(X_{S_1}; X_{N(w) \cup S_2}|Y_{W'}) + I(X_{S_1}; Y_{W'}) \end{aligned} \tag{9}$$

where the inequality is an application of the SDPI, which is justified since given $X_{S_2}, Y_{W'}$ we still have the Markov chain: $X_{S_1} \to X_{N(w)} \to Y_w$, in view of the definition (6).

Subtracting $I(X_{S_1}; Y_W)$ from both sides of (9) we get

$$I(X_{S_1}; X_{S_2}|Y_W) \le (1-\eta_w) I(X_{S_1}; X_{S_2}|Y_{W'}) + \eta_w I(X_{S_1}; X_{N(w) \cup S_2}|Y_{W'}) + I(X_{S_1}; Y_{W'}) - I(X_{S_1}; Y_W) \tag{10}$$

$$\le (1-\eta_w) I(X_{S_1}; X_{S_2}|Y_{W'}) + \eta_w I(X_{S_1}; X_{N(w) \cup S_2}|Y_{W'})\,, \tag{11}$$

since $I(X_{S_1}; Y_{W'}) \le I(X_{S_1}; Y_W)$ by the monotonicity of the mutual information. From the induction hypothesis and (7) we conclude the proof of (8). □

4 General version: channel comparison

In the setting of Section 2, we have imposed the condition (3) which implies

$$I(X_v; X_S, Y_E) = I(X_v; X_S|Y_E) = I(X_v; Y_E|X_S)\,.$$

Consequently, Theorem 1 (giving a bound on the first quantity) and Theorem 2 (giving a bound on the second one) are equivalent when (3) holds. In fact, Theorem 2 holds in wider generality. Can we also bound the third quantity? It turns out the answer is yes, and in fact this generalization allows one to remove the most restrictive condition of Theorem 2 – the independence of X_v's. (However, the two theorems bound different quantities.) To focus ideas, we recommend revisiting Remark 1.

We proceed to describing the setting of the forthcoming more general result. Consider a bipartite graph $G = (V, W, E)$ with parts V, W and edges E, with finite or countably-infinite V, W, E. For any subset $W' \subset W$, we again denote by $G[W']$ the induced subgraph on vertices $V \cup W'$.

Let $\{X_v : v \in V\}$ be a collection of discrete random variables (not necessarily independent). Let $\{Y_w : w \in W\}$ and $\{\tilde{Y}_w : w \in W\}$ be two collection of random variables each conditionally independent given X_V and distributed as

$$Y_w \sim P_{Y_w|X_{N(w)}} \qquad \forall w \in W\,, \tag{12}$$

$$\tilde{Y}_w \sim Q_{Y_w|X_{N(w)}} \qquad \forall w \in W \tag{13}$$

where $N(w) \subset V$ denote the neighborhood of w in the bipartite graph.

We also recall the definition of the less noisy relation: stochastic matrix $Q_{\tilde{Y}|X}$ is less noisy than $P_{Y|X}$ if for every distribution $P_{U,X}$ we have

$$I(U;Y) \le I(U;\tilde{Y})$$

where mutual informations are computed under the joint distribution

$$P_{U,X,\tilde{Y},Y}(u,x,\tilde{y},y) = P_{U,X}(u,x)Q_{\tilde{Y}|X}(\tilde{y}|x)P_{Y|X}(y|x)\,.$$

See [vD97, Theorem 2], [PW17, Prop. 14] and [MP16, Theorem 2, Prop. 8] for various characterizations of the less noisy relation.

Theorem 3. *Assume that for every $w \in W$, the channel $Q_{\tilde{Y}_w|X_{N(w)}}$ is less noisy than $P_{Y_w|X_{N(w)}}$. Then for any subsets $S_1, S_2 \subset V$, we have*

$$I(X_{S_1}; Y_E|X_{S_2}) \le I(X_{S_1}; \tilde{Y}_E|X_{S_2})\,. \tag{14}$$

Remark 3. The connection between Theorems 3 and 2 arises from [PW17, Proposition 15]: the SDPI constant of the channel $P_{Y|X}$ satisfies $\eta_{\mathrm{KL}}(P_{Y|X}) \le 1-\delta$ if and only if $P_{Y|X}$ is more noisy than the erasure channel $Q_{\tilde{Y}|X}$ which outputs $\tilde{Y} = X$ with probability $1-\delta$ and $\tilde{Y} = *$ (erasure) otherwise.

Remark 4. One cannot replace the less noisy condition with "more capable", a weaker notion (see [KM75]). Indeed, it is known that erasure channel with probability of erasure $1 - h(\delta)$ is more capable than $\mathsf{BSC}(\delta)$. But then consider the example in Section 2.1. If the more capable variation of Theorem 3 were true, we would be able to reduce the probability of an open bond from $(1-2\delta)^2$ to $1-h(\delta)$ and thus contradict (5).

Proof. Conditioning on X_{S_2} we get a Markov chain $X_{S_1} \to X_V \to Y_E$. By [PW17, Prop. 14], the less noisy relation tensorizes. That is, the channel $X_V \to \tilde{Y}_E$ is less noisy than $X_V \to Y_E$. Consequently, we get (14). □

5 Applications to reconstruction problems

In this section we apply the information-percolation bound to various reconstruction problems on graphs, specifically, $\mathbb{Z}_2$-synchronization, spiked Wigner model, community detection on stochastic block model (SBM) with two, and more than two communities. The first three were considered earlier in [AB18a], while the fourth application is new and apparently not obtainable via methods of [AB18a, AB18b].

5.1 Group synchronization over $\mathbb{Z}/2\mathbb{Z}$

The problem of group synchronization refers to the following: Given a graph $G = (V, E)$, let $X_V = \{X_v\}_{v\in V}$ be a collection of independent random variables that are uniformly distributed on some compact group. The goal is to recover X_V (up to a global group action which is not identifiable) from pairwise measurements $Y_E = \{Y_{uv}\}_{(u,v)\in E}$, where Y_{uv} is a noisy observation of $X_u^{-1}X_v$. The paradigm of group synchronization arises in a various applications such as localization, imaging and computer vision (cf. the references in [AMM+18]).

The synchronization problem over the d-dimensional grid was studied in [AMM+18] for various groups, focusing on *correlated recovery*, i.e., achieving a reconstruction error that is strictly better than random guessing. The simplest problem is for the group $\mathbb{Z}/2\mathbb{Z}$, commonly known as $\mathbb{Z}_2$-synchronization, which precisely corresponds to the setting of Section 2. If the observation channel is $\mathsf{BSC}(\delta)$, it is shown in [AMM+18] that correlated recovery is impossible if $1 - 2\delta \le \frac{1}{2}$. Next, we apply the information-percolation method in Theorem 1 to improve the threshold $(1-2\delta)^2 \le \frac{1}{2}$; this result was first announced and proved independently in [AMM+18]. To prove the impossibility of the correlated recovery of X_V, it suffices to show that for any pair of vertices $u \neq v$, it is impossible to reconstruct the bit $T_{uv} = X_v \oplus X_u$ better than chance.

Corollary 4. *For any two (possibly non-adjacent) vertices $u, v \in V$, any estimator $\hat{T}_{uv} = \hat{T}_{uv}(Y_E)$ satisfies*

$$\mathbb{P}[\hat{T}_{uv} \neq T_{uv}] \ge \frac{1}{2} - \sqrt{\frac{1}{2\log e} I(X_u; X_v, Y_E)} \ge \frac{1}{2} - \sqrt{\frac{\log 2}{2\log e}\mathrm{perc}_G(v, u)} \tag{15}$$

Consequently,

$$\frac{1}{|V|^2}\sum_{u,v\in V} \mathbb{P}[\hat{T}_{uv} \neq T_{uv}] \ge \frac{1}{2} - o(1) \tag{16}$$

provided

$$\sum_{u,v\in V} I(X_u; X_v, Y_E) = o(|V|^2) \quad \text{or} \quad \sum_{u,v\in V} \mathrm{perc}_G(v, u) = o(|V|^2).$$

Remark 5. It is clear, from Theorem 2, that the result above extends to arbitrary channels $P_{Y_e|X_u,X_v}$ for $e = (u, v)$, arbitrary function $T = T(X_u, X_v)$ and arbitrary (discrete) X_v. The only general requirement we need to impose the validity of (3). The only change is that the first term $\frac{1}{2}$ in the right-hand side of (15) should be replaced with $1 - \max_s \mathbb{P}[T(X_u, X_v) = s]$ and $\log 2$ in the denominator inside the square root with $\max_v H(X_v)$. We put this corollary first, as it originally motivated the writing of this article.

Proof. It suffices to show (15) as the rest follows from Jensen's inequality. Next abbreviate T_{uv} as

T. Note that

$$\begin{aligned} I(T;Y_E) &\overset{\text{(a)}}{\leq} I(X_u, X_v; Y_E) = I(X_u; Y_E|X_v) + I(X_v; Y_E) \\ &\overset{\text{(b)}}{=} I(X_u; Y_E|X_v) \\ &\overset{\text{(c)}}{=} I(X_u; X_v, Y_E) \\ &\overset{\text{(d)}}{\leq} \mathrm{perc}_G(v,u) \log 2, \end{aligned}$$

where (a) is the data processing inequality for mutual information; (b) follows from (3); (c) follows from the assumption that $X_u \perp\!\!\!\perp X_v$; (d) follows from Theorem 1.

On the other hand, for any estimator $\hat{T} = \hat{T}(Y_E)$, let $p = \mathbb{P}[\hat{T} = T]$ and $q = \mathbb{Q}[\hat{T} = T]$, where $\mathbb{Q}$ denote the probability measure where Y_E and T are independent. Thus $q \leq P_{\max}(T) \triangleq \max_t \mathbb{P}[T = t]$. By the data processing inequality and the Pinsker inequality, we have

$$I(T;Y_E) \geq d(p\|q) \geq 2 \log e (p-q)^2.$$

Thus,

$$\mathbb{P}[\hat{T} = T] \leq P_{\max}(T) + \sqrt{\frac{\mathrm{perc}_G(v,u) \log 2}{2 \log e}}.$$

□

Using Kesten's result on 2D-square grid percolation [Kes80], we get:

Corollary 5. *Let G be an infinite 2D-grid and suppose the goal is to estimate $T_n = X_{0,0} \oplus X_{n,n}$ for large n given observations of all (infinitely many) edges Y_e. If*

$$(1-2\delta)^2 \leq \frac{1}{2}$$

then for any estimator $\hat{T}_n = \hat{T}_n(Y_E)$ we have $\mathbb{P}[\hat{T}_n \neq T_n] \to \frac{1}{2}$.

5.2 Spiked Wigner model

Consider the following statistical model for PCA:

$$Y = \sqrt{\frac{\lambda}{n}} X X^\top + W \tag{17}$$

where $X = (X_1, \ldots, X_n) \in \{\pm 1\}^n$ consists of independent Rademacher entries, and W is a Wigner matrix which is symmetric consisting of independent standard normal off-diagonal entries. This ensemble is known as the spiked Wigner model (rank-one perturbation of the Wigner ensemble). Observing the matrix Y, the goal is to achieve correlated recovery, i.e., to reconstruct X (up to a global sign flip) better than chance, that is, find $\hat{X} = \hat{X}(Y) \in \{\pm 1\}^n$, such that

$$\liminf_{n\to\infty} \frac{1}{n} \mathbb{E}[|\langle X, \hat{X} \rangle|] > 0. \tag{18}$$

It is known that for fixed λ, if $\lambda > 1$, spectral method (taking the signs of the the first eigenvector of Y) achieves correlated recovery [BBAP05]. Conversely, if $\lambda < 1$, correlated recovery is information-theoretically impossible.

As the next result shows, applying Theorem 1 together with classical results on Erdös-Rényi graphs immediately yields the optimal threshold previously obtained in [DAM16, Theorem 4.3]. Here, $o(1)$ is any vanishing factor so this result is the best possible.

Corollary 6. *Correlated recovery in the sense of (18) is impossible if*

$$\lambda \le 1 + o(1). \tag{19}$$

Proof. Note that (18) is equivalent to

$$\limsup_{n\to\infty} \frac{1}{n^2}\mathbb{E}\left[\left\|XX^\top - \hat{X}\hat{X}^\top\right\|_{\mathrm{F}}^2\right] < 2. \tag{20}$$

It is clear that the diagonal entries of Y are independent of X and hence the problem reduces to the setting in Section 2 with G being the complete graph on n vertices and $Y_{ij} = \sqrt{\frac{\lambda}{n}}X_iX_j + W_{ij}$ for $i < j$. Applying Theorem 1 together with Corollary 4, we conclude that: for any $i < j$,

$$\inf_{\hat{T}_{ij}(\cdot)} \mathbb{P}\left[X_iX_j \neq \hat{T}_{ij}(Y)\right] \ge \frac{1}{2} - O(\mathbb{P}\left[i \text{ and } j \text{ are connected in } \mathrm{ER}(n,\eta)\right]).$$

where $\eta = \eta(N(-\sqrt{\frac{\lambda}{n}},1), N(\sqrt{\frac{\lambda}{n}},1)) = \frac{\lambda}{n}(1+o(1))$ in view of (50). Summing over $i \neq j$, we conclude that for any $\hat{X} = \hat{X}(Y) \in \{\pm 1\}^n$,

$$\begin{aligned}
\mathbb{E}\left\|XX^\top - \hat{X}\hat{X}^\top\right\|_{\mathrm{F}}^2 &= 4\sum_{i\neq j}\mathbb{P}\left[X_iX_j \neq \hat{X}_i\hat{X}_j\right] \\
&\ge 2n^2 - 2\sum_{i\in[n]}\mathbb{E}\left[\text{size of the connected component in } \mathrm{ER}(n,\eta) \text{ containing } i\right] \\
&\ge 2n^2 - n\mathbb{E}\left[C_{\max}\right],
\end{aligned}$$

where $C_{\max}$ denotes the size of the largest connected component in the Erdős-Rényi graph $\mathrm{ER}(n,\eta)$. Existing results in the random graph theory show that $\mathbb{E}[C_{\max}] = o(n)$ whenever $\eta = \frac{1}{n}(1+o(1))$, which implies the impossibility of (20). Specifically, let $\eta = \frac{1}{n^2}(n+s)$, where $s = o(n)$ by assumption. By monotonicity, it suffices to consider the case of $s = \omega(n^{2/3})$. By a result of Łuczak [Luc90, Lemma 3] (see also [JLR00, Theorem 5.12]), we have $C_{\max} \le c_0 s$ with probability at least $1 - c_1 n^{1/3} s^{-1/2}$ for some universal constants c_0, c_1. Since $C_{\max} \le n$, this shows $\mathbb{E}[C_{\max}] = o(n)$, completing the proof. □

Remark 6 (Channel universality)**.** Consider a more general observation model than (17): Let $P(\cdot|\theta)$ be a family of conditional distributions parametrized by $\theta \in \mathbb{R}$, with conditional density $p_\theta(\cdot)$ with respect to some reference measure μ. Given $M = \sqrt{\frac{\lambda}{n}}XX^\top$, we observe the matrix $Y = (Y_{ij})$, where each Y_{ij} is obtained by passing M_{ij} through the same channel independently, with the conditional distribution given by $P_{Y_{ij}|M_{ij}} = P(\cdot|M_{ij})$. The spiked Wigner model corresponds to the Gaussian channel $P(\cdot|\theta) = N(\theta,1)$.

Under appropriate regularity conditions on the channel, the sharp threshold (19) is replaced by the following:

$$\lambda \le \frac{1}{J_0} + o(1) \tag{21}$$

where $J_\theta \triangleq \int (\frac{\partial p_\theta}{\partial \theta})^2 \frac{1}{p_\theta} d\mu$ is the Fisher information. This follows from the relationship between the contraction coefficient and the Fisher information. To see why this is true intuitively, note that $M_{ij} \in \{\pm\epsilon\}$, with $\epsilon \triangleq \sqrt{\frac{\lambda}{n}}$. Using the characterization (45) of the contraction coefficient for binary-input channels, we have $\eta = \sup_{\beta\in[0,1]} \mathrm{LC}_\beta(p_\epsilon \| p_{-\epsilon})$, where LC_β is an f-divergence[2] with $f(x) =$

[2]Recall an f-divergence is defined as $D_f(P\|Q) = \mathbb{E}_P[f(\frac{dP}{dQ})]$ for convex f with $f(1) = 0$ [Csi69].

$f_\beta(x) = \beta\bar{\beta}\frac{(x-1)^2}{\beta x+\bar{\beta}}$. By the local expansion of f-divergence, we have $D_f(P_{\theta-\delta}\|P_\theta) = \frac{f''(1)J_\theta}{2}\delta^2(1+o(1))$ as $\delta \to 0$. Note that $f''_\beta(1) = 2\beta\bar{\beta}$, maximized at $\beta = \frac{1}{2}$. It follows that $\eta = \frac{\lambda J_0 + o(1)}{n}$. Thus the same percolation bound used in Corollary 6 shows that (21) implies the impossibility of correlated reconstruction. In the positive direction, it was suggested in [LKZ15, Section II-C] that spectral method applied to the score matrix succeeds provided that $\lambda > \frac{1}{J_0}$. In fact, the full mutual information $I(M;Y)$ also undergoes a phase transition at this point, see [KXZ16] and [BDM+16].

5.3 Community detection: two communities

Consider a complete graph K_n and $X_v \overset{\text{i.i.d.}}{\sim} \text{Bern}(1/2)$. Unlike the group-synchronization case, we have the following observation channel: for each edge $e = (u, v)$ we have

$$Y_e = \begin{cases} \text{Bern}(p), & X_u = X_v \\ \text{Bern}(q), & X_u \neq X_v \end{cases} \tag{22}$$

In other words, Y is the adjacency matrix of a random graph (known as the stochastic block model), in which any pair of vertices are connected with probability p if they are from the same community (with the same labels) or with probability q otherwise.

Given the matrix $Y = (Y_{ij})$, the goal is to achieve correlated recovery, that is, estimating the labels up to a global flip better than random guess. In other words, construct $\hat{X} = \hat{X}(Y) \in \{0,1\}^n$, such that

$$\limsup_{n\to\infty} \frac{1}{n}\mathbb{E}[\min\{d(\hat{X}, X), n - d(\hat{X}, X)\}] < \frac{1}{2}, \tag{23}$$

where d denotes the Hamming distance. Equivalently, the goal is to estimate $\mathbf{1}_{\{X_i = X_j\}}$ for any pair i, j on the basis of Y with probability of error asymptotically (as $n \to \infty$) not tending to $1/2$. The exact region when this is impossible is known [MNS15, MNS13]: for $p = \frac{a}{n}$ and $q = \frac{b}{n}$ with fixed a, b, correlated recovery is possible if and only if

$$\frac{(a-b)^2}{2(a+b)} > 1.$$

Appying the information-percolation method (namely Theorem 2) we get the following slightly suboptimal result (see Fig. 1).

Proposition 7. *For the binary stochastic block model with edge probabilities p and q, for any $i \neq j \in [n]$, the following bound holds non-asymptotically:*

$$I(X_i; X_j, Y_E) \le \mathbb{P}\left[i \text{ and } j \text{ are connected in } \text{ER}(n, \eta)\right] \tag{24}$$

where $\eta = p + q - 2pq + 2\sqrt{p(1-p)q(1-q)}$. Furthermore, if $p = \frac{a}{n}$ and $q = \frac{b}{n}$, then correlated recovery (i.e., (23)) is impossible if

$$(\sqrt{a} - \sqrt{b})^2 < 1 + o(1). \tag{25}$$

Proof. The mutual information bound (24) follows from Theorem 1 and the exact expression for the contraction coefficients in (46), which satisfies

$$\eta_{\text{KL}}(\text{Bern}(a/n), \text{Bern}(b/n)) = \frac{(\sqrt{a} - \sqrt{b})^2 + o(1)}{n}, \tag{26}$$

where the $o(1)$ terms is uniform in (a, b) in view (49). The remaining proof is the same as Corollary 6 using the behavior of the giant component of the Erdös-Rényi graph. □

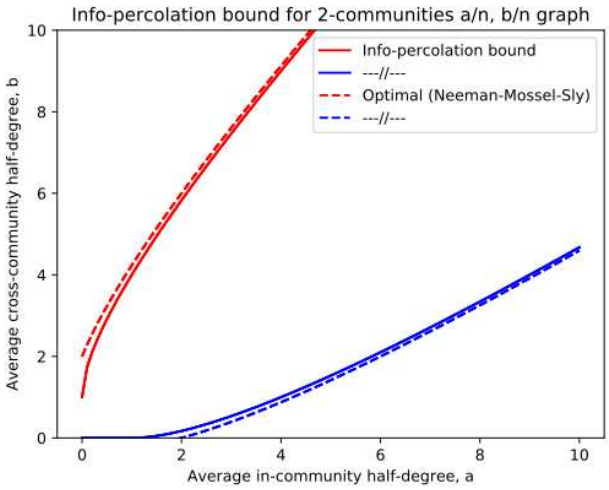

Figure 1: Comparing optimal (Mossel-Neeman-Sly [MNS15]) region with the percolation bound.

5.4 Community detection: k communities

In the setting of the previous section, suppose now that $X_v \overset{\text{i.i.d.}}{\sim} \text{Unif}[k]$, with the same observation channel (22) with $p = \frac{a}{n}$ and $q = \frac{b}{n}$. This is the stochastic block model with k equal-sized communities, and the notion of correlated recovery is extended as follows: for any $x, \hat{x} \in [k]^n$, define the following error metric:

$$d(x, \hat{x}) \triangleq \min_{\pi \in S_k} \frac{1}{n} \sum_{i \in [n]} \mathbf{1}_{\{x_i \neq \pi(\hat{x}_i)\}} \tag{27}$$

that is, the number of classification errors up to a global permutation of labels. We say that correlated recovery is possible if there exists a (sequence of) estimator $\hat{X} \in [k]^n$ that outperforms random guessing, i.e.,

$$\limsup_{n \to \infty} \mathbb{E}[d(X, \hat{X})] < \frac{k-1}{k}. \tag{28}$$

For $k \geq 3$, the sharp threshold is not known. In terms of the impossibility result, the best known sufficient condition is [BMNN16, Theorem 1]

$$\frac{(a-b)^2}{a + (k-1)b} < \frac{2k \log(k-1)}{k-1}. \tag{29}$$

Now, it turns out that applying Theorem 1 would only yield a k-independent bound (25). To get an improved estimate, instead, we use the comparison theorem with the erasure model in Theorem 3 and then show the impossibility of reconstruction on the corresponding erasure model. The threshold is given by (30) in the next proposition and the numerical comparison with the bound of (29) is shown in Fig. 2. For $k = 3$, (30) improves over (29) in some regime but not for $k = 4$. For large k, (30) is suboptimal by a logarithmic factor.

Proposition 8. *Correlated recovery in the sense of (28) is impossible if*

$$(\sqrt{a} - \sqrt{b})^2 \leq \frac{k}{2}. \tag{30}$$

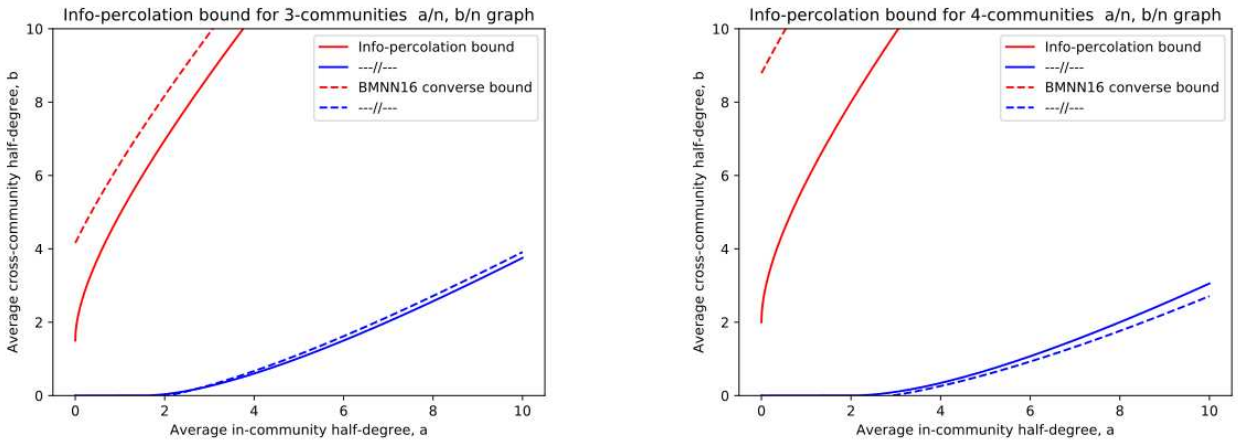

Figure 2: Comparing the inner (impossibility) bound of [BMNN16] with Prop. 8 for $k = 3$ and $k = 4$ communities. For $k = 3$, Prop. 8 improves the state of the art.

Proof. We start by setting up the mutual comparison with the corresponding model per Theorem 3. Let $\eta = \frac{(\sqrt{a}-\sqrt{b})^2+o(1)}{n}$ be given in (26). Define the corresponding erasure model on the same graph: for each $(u, v) \in \binom{n}{[2]}$, let $\tilde{Y}_{uv} = \mathbf{1}_{\{X_u=X_v\}}$ with probability η and $\tilde{Y}_{uv} =?$ with probability $1 - \eta$ independently. Equivalently, the reconstruction problem under the erasure model can be phrased as follows. Let $G = ([n], E)$ denote an Erdös-Rényi graph $\mathrm{ER}(n, \eta)$ independent of X. Then for each $(u, v) \in E$, we observe a deterministic function $\tilde{Y}_{uv} = \mathbf{1}_{\{X_u=X_v\}}$. By Theorem 3 and Remark 3, we have the following comparison result: for any $S \subset [n]$,

$$I(X_S; Y) \le I(X_S; \tilde{Y}). \tag{31}$$

By symmetry, $I(X_S; \tilde{Y})$ only depends on $|S|$. Next we assume $S = [m]$ and show that for any fixed m,

$$I(X_S; \tilde{Y}) = o(1), \quad n \to \infty$$

under the condition that $(\sqrt{a} - \sqrt{b})^2 \le \frac{k}{2}$.

By the chain rule, we have

$$\begin{aligned} I(X_S; \tilde{Y}) &= I(X_1; \tilde{Y}) + I(X_2; \tilde{Y}|X_1) + \dots I(X_m; \tilde{Y}|X_1, \dots, X_{m-1}) \\ &= \sum_{u=2}^{m} I(X_u; X_1, \dots, X_{u-1}, \tilde{Y}), \end{aligned} \tag{32}$$

where we used the fact that X_i's are independent and $I(X_1; \tilde{Y}) = 0$.

Next using the local tree structure of G, we show that for each u, $I(X_u; X_1, \dots, X_{u-1}, \tilde{Y}) = o(1)$. Condition on the realization of G. Fix t to be specified later. Let G_u^t denote the t-hop neighborhood of u. Let R be the boundary of G_u^t, i.e., the set of vertices that are at distance t to u. For any v whose distance to u exceeds t, R forms a cut separating u and v in the sense that any path from u to v passes through S. Then for any set of vertices U outside the t-hop neighborhood of r, we have

$$I(X_u; X_U, \tilde{Y}_E) \le I(X_u; X_R, \tilde{Y}_E) = I(X_u; X_R, \tilde{Y}_{\le t}), \tag{33}$$

where $\tilde{Y}_{\le t} \triangleq \tilde{Y}_{E(G_u^t)}$. Indeed, the first inequality follows from the fact that $X_u \to X_R \to X_{S'}$ forms a Markov chain conditioned on $\tilde{Y}_E$, and the second inequality follows from the independence of X_u and $Y_{E(G)\setminus E(G_u^t)}$ conditioned on the $(X_R, Y_{\le t})$.

By [PW16, Proposition 12], since X_u only takes k values, we can bound the mutual information by the total variation as follows:

$$I(X_u; X_R, \tilde{Y}_{\le t}) \le \log(k-1) T(X_u; X_R, \tilde{Y}_{\le t}) + h(T(X_u; X_R, \tilde{Y}_{\le t})) \tag{34}$$

where $h(x) \triangleq x \log \frac{1}{x} + (1-x) \log \frac{1}{1-x}$, and

$$T(X_u; X_R, \tilde{Y}_{\le t}) \triangleq \mathbb{E}[d_{\mathrm{TV}}(P_{X_R, \tilde{Y}_{\le t}|X_u}, P_{X_R, \tilde{Y}_{\le t}})] \le \max_{x,x' \in [k]} d_{\mathrm{TV}}(P_{X_R, \tilde{Y}_{\le t}|X_u=x}, P_{X_R, \tilde{Y}_{\le t}|X_u=x'}) \tag{35}$$

where the last inequality follows from the convexity of the total variation.

Now choose $t = t_n$ such that $t = \omega(1)$ and $t = o(\log n)$. We show that

$$\tau \triangleq \max_{x,x' \in [k]} d_{\mathrm{TV}}(P_{X_R, \tilde{Y}_{\le t}|X_u=x}, P_{X_R, \tilde{Y}_{\le t}|X_u=x'}) = o(1). \tag{36}$$

To this end, let T_u^t denote a depth-t Galton-Watson (GW) tree rooted at u with offspring distribution $\mathrm{Poi}(d)$, with $d \triangleq n\eta$ is at most a constant by assumption. By the locally tree-like property of the Erdös-Rényi graph (see, e.g., [MNS15, Proposition 4.2] with $p = q$), there exists a coupling between T_u^t and G_u^t such that $\mathbb{P}\left[G_u^t = T_u^t\right] = 1 - o(1)$. In the sequel we condition on the event of $G_u^t = T_u^t$ In particular, by standard results in branching process [AN72], the expected number of ith progeny is d^i and hence the expect size of the t-neighborhood of u is $\frac{d^{t+1}-1}{d-1}$. By the Markov inequality, the size of the t-neighborhood of u is at most $M \triangleq (Cd)^t = n^{o(1)}$ with probability $1 - o(1)$. In other words, the majority of v are outside the t-neighborhood of u. Next we conditioned on the event $G_u^t = T_u^t$ and abbreviate T_u^t as T. For each $x \neq x'$, we construct a coupling $\{X_v^+, X_v^- : v \in V(T)\}$ and $\{Y_e : e \in E(T)\}$ so that $(X_{V(T)}^+, Y_{E(T)})$ and $(X_{V(T)}^-, Y_{E(T)})$ are distributed as the law of $(X_{V(T)}, Y_{E(T)})$ conditioned on the root $X_u = x$ and $X_u = x'$, respectively. The coupling is defined inductively as follows: First set $X_u^+ = x$ and $X_u^- = x'$. Next we generate each layer of observations recursively as follows: Given all the X_v's and Y_e's up to depth k, draw $Y_e = \mathrm{Bern}(1/k)$ independently for all edges between the kth and the $(k+1)$th layer. For each edge $e = (i, j)$ so that i is on kth layer and j is on $(k+1)$th layer, if $X_i^+ = X_i^-$, we couple all observations on the subtree rooted at i together, that is, set $X_j^+ = X_j^- = X_i^+$ if $Y_e = 1$ and $X_j^+ = X_j^- = R$ if $Y_e = 0$ where R is drawn uniformly at random from $[k] \setminus \{X_i^+\}$; if $X_i^+ \neq X_i^-$, we proceed as follows:

- if $Y_e = 1$, set $X_j^+ = X_i^+$ and $X_j^- = X_i^-$.
- if $Y_e = 0$, with probability $\frac{k-2}{k-1}$, set $X_j^+ = X_j^- = R$ with R drawn uniformly at random from $[k] \setminus \{X_i^+, X_i^-\}$, and with probability $\frac{1}{k-1}$ set $X_j^+ = X_i^-$, and $X_j^- = X_i^+$.

Note that for each i and each of its child j, we have

$$\mathbb{P}\left[X_j^+ \neq X_j^- | X_i^+ \neq X_i^-\right] = \mathbb{P}\left[Y_e = 1\right] + \mathbb{P}\left[Y_e = 0\right] \frac{1}{k-1} = \frac{2}{k}.$$

Thus, the number of uncoupled pairs (X_i^+, X_i^-) evolves as a GW tree with offspring distribution $\mathrm{Poi}(\frac{2d}{k})$, which dies out if $\frac{2d}{k} \le 1$ (see, e.g., [AN72, Theorem 1]), in which case we have $d_{\mathrm{TV}}(P_{X_{V(T)}, Y_{E(T)}|X_u=x}, P_{X_{V(T)}, Y_{E(T)}|X_u=x'}) \le \mathbb{P}\left[X_R^+ \neq X_R^-\right] \to 0$, as $t \to \infty$. This completes the proof of (36).

Combining (34)–(36), we have

$$I(X_u; X_1, \ldots, X_{u-1}, \tilde{Y}) \le \log(k-1)\tau + h(\tau) + (1 - \mathbb{P}\left[E \cap E'\right]) \log k$$

where $E = \{G_u^t = T_u^t, |V(T_u^t)| \le M\}$, $M = (Cd)^t = n^{o(1)}$, and E' denotes the event that $1, \ldots, u-1$ are all outside the t-hop neighborhood of u. We have already shown that $\tau = o(1)$ and $\mathbb{P}[E] = 1 - o(1)$. Furthermore, by symmetry $\mathbb{P}[E'] = \frac{M-1}{n-1} \cdots \frac{M-u}{n-u} \ge (\frac{M-m}{n-m})^m = 1 - o(1)$. To summarize, we have shown that $I(X_u; X_1, \ldots, X_{u-1}, \tilde{Y}) = o(1)$ and, in view of (32),

$$I(X_S; \tilde{Y}) = o(1) \tag{37}$$

for $S = [m]$ and hence any $S \in \binom{[n]}{m}$.

Finally, using (37) for appropriately chosen m, we show the impossibility of the correlated recovery (28). First of all, note that for any fixed $x, \hat{x} \in [k]^n$ and any $m \in [n]$ we have

$$d(x, \hat{x}) \ge \mathbb{E}_S[d(x_{\mathsf{S}}, \hat{x}_{\mathsf{S}})] \tag{38}$$

where $\mathsf{S} \sim \mathrm{Unif}(\binom{[n]}{m})$ and recall that for any S, we have $d(x_S, \hat{x}_S) = \frac{1}{|S|} \min_{\pi \in S_k} \sum_{i \in S} \mathbf{1}_{\{x_i \neq \pi(\hat{x}_i)\}}$ per (27). The inequality (38) simply follows from

$$\begin{aligned} d(x, \hat{x}) &= \min_{\pi \in S_k} \mathbb{P}_{I \sim \mathrm{Unif}([n])} \left[x_I \neq \hat{x}_{\pi(I)}\right] \\ &= \min_{\pi \in S_k} \mathbb{E}_{\mathsf{S} \sim \mathrm{Unif}(\binom{[n]}{m})} \mathbb{P}_{I \sim \mathrm{Unif}(\mathsf{S})} \left[x_I \neq \hat{x}_{\pi(I)}\right] \\ &= \mathbb{E}_{\mathsf{S}} \min_{\pi \in S_k} \mathbb{P}_{I \sim \mathrm{Unif}(\mathsf{S})} \left[x_I \neq \hat{x}_{\pi(I)}\right] \\ &\ge \mathbb{E}_{\mathsf{S}}[d(x_{\mathsf{S}}, \hat{x}_{\mathsf{S}})]. \end{aligned}$$

Fix a constant m independent of n. For any estimator $\hat{X} = \hat{X}(Y) \in [k]^n$, applying (38) yields

$$\mathbb{E}[d(X_{\mathsf{S}}, \hat{X}_{\mathsf{S}})] \le \mathbb{E}[d(X, \hat{X})], \tag{39}$$

where S is a random uniform m-set independent of $X, \hat{X}$.

By the data processing inequality, we have for any fixed S,

$$I(X_S; \hat{X}_S) \le I(X_S; Y) \overset{(31)}{\le} I(X_S; \tilde{Y}) \overset{(37)}{=} o(1).$$

By Pinsker's inequality, we have $d_{\mathrm{TV}}(P_{X_S, \hat{X}_S}, P_{X_S} \otimes P_{\hat{X}_S}) \le \sqrt{2 I(X_S; \hat{X}_S)} = o(1)$. Note that the loss function d defined in (27) is bounded by one. Thus

$$\mathbb{E}[d(X_S, \hat{X}_S)] \ge \mathbb{E}[d(X_S, Z_S)] - d_{\mathrm{TV}}(P_{X_S, \hat{X}_S}, P_{X_S} \otimes P_{\hat{X}_S}) = \mathbb{E}[d(X_S, Z_S)] + o(1), \tag{40}$$

where Z_S has the same distribution as $\hat{X}_S$ and is independent of X_S. By Lemma 9 at the end of this subsection, we have

$$\mathbb{E}[d(X_S, Z_S)] \ge \left(\frac{k-1}{k} - m^{-1/3}\right)(1 - k! e^{-2m^{1/3}}). \tag{41}$$

Combining (39), (40) and (41), sending $n \to \infty$ followed by $m \to \infty$, we arrive at

$$\liminf_{n \to \infty} \mathbb{E}[d(X, \hat{X})] \ge \frac{k-1}{k}.$$

This completes the proof of the proposition. □

Lemma 9. *Let X be uniformly distributed on $[k]^m$ and Z is independent of X with an arbitrary distribution on $[k]^m$. For the loss function in (27), we have*[3]

$$d(X,Z) \geq \frac{k-1}{k} - m^{-1/3} \tag{42}$$

with probability at least $1-(k!e^{-2m^{1/3}})$.

Proof. For each fixed π, the Hamming distance $d_H(X,\pi(Z)) \sim \mathrm{Binom}(m, \frac{k-1}{k})$. From Hoeffding's inequality we have

$$\mathbb{P}[d_H(X,\pi(Z) < \frac{k-1}{k} - \delta] \leq e^{-2m\delta^2},$$

and from the union bound

$$\mathbb{P}[\min_{\pi} d_H(X,\pi(Z) < \frac{k-1}{k} - \delta] \leq k! e^{-2m\delta^2}.$$

Setting $\delta = m^{-1/3}$ completes the proof. □

Remark 7. In the above proof we considered the problem of reconstructing the root X_u variable of a Galton-Watson tree with the average degree d, where the vertex variables are iid and unifrom on $[k]$, and the edge variables are given $Y_{i,j} = \mathbf{1}_{\{X_i = X_j\}}$ for each edge i,j. The reconstruction of X_u is based on the values of all $Y_{i,j}$ and all vertex variables at an arbitrary deep layer of the tree. We have shown the reconstruction is impossible (unable to outperform random guessing) if $d \leq \frac{k}{2}$. At the same time, clearly reconstruction is possible if $d \geq k$ (in which case there is an arbitrarily long path of edges with $Y_{i,j} = 1$ starting from the root). So what is the exact threshold? A work in progress [GP19] shows a much improved bound, namely that reconstruction is impossible if

$$d < f(k) \triangleq \left(\frac{\log k - \log(k-1)}{\log k}\frac{k-1}{k} + \frac{1}{k}\right)^{-1} = k - (1+o(1))\frac{k}{\log k}.$$

Using this bound in place of $d < k/2$ it follows that correlated recovery in a k-SBM is not possible if

$$(\sqrt{a} - \sqrt{b})^2 < f(k). \tag{43}$$

This improves (29) for all $k \geq 3$ in some range of a, b. The work [GP19] presents further improvements to (43) based on applying SDPIs directly to an equivalent Potts model on a tree.

A Contraction coefficients of some binary-input channels

Consider an arbitrary channel $P_{Y|X}$. Denote the contraction coefficient, defined as the best constant in (1), by $\eta_{\mathrm{KL}}(P_{Y|X})$. It has an equivalent characterization:

$$\eta_{\mathrm{KL}}(P_{Y|X}) = \sup_{\pi_X \neq \pi'_X} \frac{D(Q_Y \| Q'_Y)}{D(\pi_X \| \pi'_X)}, \tag{44}$$

where Q_Y and Q'_Y are the distributions induced by π_X and π'_X, respectively.

[3] Note that for any fixed k, m and any string $x, z \in [k]^m$, we can always outperform random matching, i.e., $d(x,z) < \frac{k-1}{k}$. The point of (42) is that this improvement is negligible for large m.

Consider a binary input channel $P_{Y|X}$, where $P_{Y|X=0} = P$ and $P_{Y|X=1} = Q$. Then we can write $\eta_{\mathrm{KL}}(P_{Y|X}) = \eta_{\mathrm{KL}}(P, Q)$, for convenience. The following representation is given in [PW17, Proof of Theorem 21] in terms of the Le Cam divergence:

$$\eta_{\mathrm{KL}}(P, Q) = \sup_{\beta \in [0,1]} \underbrace{\beta\bar{\beta} \int \frac{(P-Q)^2}{\beta P + \bar{\beta} Q}}_{\triangleq \mathrm{LC}_\beta(P\|Q)}, \tag{45}$$

where we denote $\bar{\beta} = 1 - \beta$. For example, for a binary-input binary-output channel, direct calculation gives

$$\eta_{\mathrm{KL}}(\mathrm{Bern}(p), \mathrm{Bern}(q)) = p + q - 2pq - 2\sqrt{p\bar{p}q\bar{q}} \tag{46}$$

$$\leq (\sqrt{p} - \sqrt{q})^2 + 2\sqrt{pq}(p+q) \tag{47}$$

In particular, for the $\mathsf{BSC}(\delta)$ we have $q = 1 - p = \delta$ and $\eta_{\mathrm{KL}}(\mathsf{BSC}(\delta)) = (1-2\delta)^2$.

It is further shown in [PW17, Theorem 21] that squared Hellinger distance determines the contraction coefficient of binary-input channel up to a factor of two:

$$\frac{H^2(P, Q)}{2} \leq \eta(\{P, Q\}) \leq H^2(P, Q). \tag{48}$$

Thus, we have

$$\eta_{\mathrm{KL}}(\mathrm{Bern}(a/n), \mathrm{Bern}(b/n)) \leq \frac{(\sqrt{a} - \sqrt{b})^2 + o(1)}{n}, \quad n \to \infty \tag{49}$$

$$\eta_{\mathrm{KL}}(N(-\delta, 1), N(\delta, 1)) \leq \delta^2(1 + o(1)), \quad \delta \to 0. \tag{50}$$

For binary-input channels, the SDPI constant can be related to the following χ^2-mutual information:

$$I_{\chi^2}(X; Y) \triangleq \chi^2(P_{XY} \| P_X \otimes P_Y) \tag{51}$$

and notice that if $X \sim \mathrm{Bern}(1/2)$ then

$$I_{\chi^2}(X; Y) = \mathrm{LC}_{1/2}(P\|Q) = \int \frac{(P-Q)^2}{2(P+Q)}\,.$$

Hence from (45) we have

$$\eta_{\mathrm{KL}}(P_{Y|X}) \geq I_{\chi^2}(X; Y)\,. \tag{52}$$

Furthermore, under a symmetry assumption, (52) holds with the equality as we show next.

A binary-input channel $P_{Y|X}$ is called symmetric (often called a BMS channel in the information theory literature [RU08]) if there exists a measurable involution $T : \mathcal{Y} \to \mathcal{Y}$ such that $P_{Y|X=0}(T^{-1}A) = P_{Y|X=1}(A)$ for all measurable subsets $A \subset \mathcal{Y}$. For such a channel, we have that

$$\eta_{KL}(P_{Y|X}) = I_{\chi^2}(X; Y)\,, \qquad X \sim \mathrm{Bern}(1/2)\,, \tag{53}$$

Indeed, for the special case of $\mathsf{BSC}(\delta)$, both sides are equal to $(1-2\delta)^2$ by an explicit calculation. In general, a well-known decomposition result (cf. [RU08, Lemma 4.28]) shows that any BMS $P_{Y|X}$ can be represented as a mixture of BSC's. Namely, we can equivalently think of the action of the channel $P_{Y|X}$ as first generating a random variable $\Delta \in [0, 1]$ according to a fixed distribution P_Δ,

passing X through $\mathsf{BSC}(\Delta)$ to obtain $\tilde{Y}$, and then outputting both Δ and $\tilde{Y}$. With this model, we have $Y = (\Delta, \tilde{Y})$ and with $X \sim \mathrm{Bern}(1/2)$

$$I_{\chi^2}(X;Y) = I_{\chi^2}(X;\Delta,\tilde{Y}) = \mathbb{E}[(1-2\Delta)^2]\,.$$

Next, fix two distributions $\pi = \mathrm{Bern}(a)$ and $\pi' = \mathrm{Bern}(a')$. Let $Q = Q_{\Delta,\tilde{Y}}$ and $Q' = Q'_{\Delta,\tilde{Y}}$ be the corresponding distributions produced at the output of $P_{Y|X}$. Note that conditioned on $\Delta = \delta$ we have by the SDPI (44) for the $\mathsf{BSC}(\delta)$:

$$D(Q_{\tilde{Y}|\Delta=\delta}\|Q'_{\tilde{Y}|\Delta=\delta}) \le (1-2\delta)^2 D(\pi\|\pi')\,.$$

Since the marginal distribution of Δ is the same under Q and Q', taking expectation over Δ yields

$$D(Q\|Q') = \mathbb{E}_{\Delta\sim P_\Delta}[D(Q_{\tilde{Y}|\Delta}\|Q'_{\tilde{Y}|\Delta})] \le \mathbb{E}[(1-2\Delta)^2]D(\pi\|\pi')\,.$$

Therefore, from (44) we get that

$$\eta_{\mathrm{KL}}(P_{Y|X}) \le \mathbb{E}[(1-2\Delta^2)] = I_{\chi^2}(X;Y)\,.$$

Together with (52) this completes the proof of (53).

B Comparison with [AB18b]

The first bound on $\mathbb{Z}_2$-synchronization threshold over a 2D-square grid was obtained in [AMM+18] by leveraging a standard coupling technique, in which the action of the $\mathsf{BSC}(\delta)$ is modeled as passing a bit uncorrupted with probability $1-2\delta$ or rerandomizing it otherwise. A natural argument then shows that on an arbitrary lattice the $\mathbb{Z}_2$-synchronization is impossible whenever $(1-2\delta)$ is smaller than the bond-percolation threshold of the lattice.

The present work sprang from the remark of E. Abbe [Abb18], suggesting that an improved estimate on this threshold is possible. In the previous work [PW17] of the authors, a general technique is developed for showing vanishing of mutual information in a network of $\mathsf{BSC}(\delta)$-channels whenever $(1-2\delta)^2$ is below the vertex-percolation threshold. While the Bayesian network setup of [PW17] is not directly applicable to the setting of group synchronization, the method (of induction on the number of edges) does apply. This lead us to Theorem 1, which was disseminated slightly prior to the talk [AB18a] presenting a similar result (subsequently published as [AB18b]). Both our Theorem 1 and [AB18b] yield the same threshold for $\mathbb{Z}_2$-synchronization on a 2D-square grid, cf. Corollary 4.

The main result of [AB18b] is the following. Consider the setting of Theorem 2 and assume in addition

1. that each label X_v is binary and unbiased: $X_v \sim \mathrm{Bern}(1/2)$;
2. that each $w \in W$ has degree 2;
3. that each channel $P_{Y_w|X_{N(w)}}$ has the following special form

$$Y_w \sim Q_w(\cdot|X_u \oplus X_v)\,,$$

where $N(w) = \{u, v\}$ and $Q_w(\cdot|\cdot)$ is a binary-input symmetric channel (BMS).

Then

$$I_{\chi^2}(X_u; X_S, Y_W) \le \mathrm{perc}_G(v, S)\,,$$

where I_{χ^2} was defined in (51) and $\mathrm{perc}_G(v, S)$ is a probability of existence of an open path from u to S if each vertex $w \in W$ is retained with probability $I_{\chi^2}(X_{N(w)}; Y_w)$.

In view of (53) and the bound $I(X_u; X_S, Y_W) \le \log(1+I_{\chi^2}(X_u; X_S, Y_W)) \le \log e \cdot I_{\chi^2}(X_u; X_S, Y_W)$, we see that indeed the result of [AB18b] is a special case of Theorem 2.

Notably the proof in [AB18b] also proceeds by induction on the number of edges (i.e. on the size of W), similar to our proofs of Theorems 1-2 and [PW17, Theorem 5]. Indeed, suppose that the result has been shown for W' and $W = W' \cup \{w_0\}$. Suppose also $N(w_0) = \{u_0, v_0\}$, and in addition that $Q_{w_0} = \mathsf{BSC}(\delta_0)$ (this assumption is easy to remove by a separate argument). Then [AB18b] exploits the extra structure imposed by the assumptions above and directly computes

$$I_{\chi^2}(X_u; X_v, Y_W) = I_0 + (I_1 - I_0)(1 - 2\delta_0)^2 h((1 - 2\delta_0))^2\,, \tag{54}$$

where $h : [0, 1] \to [0, 1]$ is a non-decreasing function that depends only on W' and $\{Q_w, w \in W'\}$, $I_0 = I_{\chi^2}(X_u; X_v, Y_{W'})$ and $I_1 = I_{\chi^2}(X_u; X_v, Y_{W'}, \tilde{Y}_{w_0})$, with $\tilde{Y}_{w_0} = X_{u_0} \oplus X_{v_0}$ denoting the noiseless observation. It is then easy to see that (54) grows slower (in terms of $(1-2\delta_0)^2$) than the percolation probability.

Acknowledgement

Y. Wu is grateful to Jiaming Xu for discussions pertaining to Remark 6. Y. Polyanskiy thanks Emmanuel Abbe for introducing him to and sharing his results on the group synchronization over a 2D grid.

References

[AB18a] E. Abbe and E. Boix. Broadcasting and synchronizing bits on graphs. Presentation at *Workshop on Combinatorial Statistics*, May 2018.

[AB18b] Emmanuel Abbe and Enric Boix. An information-percolation bound for spin synchronization on general graphs. *arXiv preprint arXiv:1806.03227*, 2018.

[Abb18] E. Abbe. Personal communication, April 2018.

[AMM+18] Emmanuel Abbe, Laurent Massoulié, Andrea Montanari, Allan Sly, and Nikhil Srivastava. Group synchronization on grids. *to appear in Mathematical Statistics and Learning*, 2018. arXiv preprint arXiv:1706.08561.

[AN72] Krishna B Athreya and Peter E Ney. *Branching Processes*. Springer-Verlag, 1972.

[BBAP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[BDM+16] Jean Barbier, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, and Lenka Zdeborová. Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 424–432. Curran Associates, Inc., 2016.

[BMNN16] Jess Banks, Cristopher Moore, Joe Neeman, and Praneeth Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *Conference on Learning Theory*, pages 383–416, 2016.

[Csi69] I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Publ. Math. Inst. Hungar. Acad. Sci., Ser. A*, 8:85–108, 1969.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory, 2nd Ed.* Wiley-Interscience, New York, NY, USA, 2006.

[DAM16] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the two-groups stochastic block model. *Information and Inference: A Journal of the IMA*, 6(2):125–170, 2016.

[EKPS00] William Evans, Claire Kenyon, Yuval Peres, and Leonard J Schulman. Broadcasting on trees and the Ising model. *Ann. Appl. Probab.*, 10(2):410–433, 2000.

[ES99] William S Evans and Leonard J Schulman. Signal propagation and noisy circuits. *IEEE Trans. Inf. Theory*, 45(7):2367–2373, 1999.

[GP19] Y. Gu and Y. Polyanskiy. Nonlinear log-Sobolev inequalities for the Potts channel with applications to reconstruction problems. in preparation, May 2019.

[JLR00] Svante Janson, Tomasz Łuczak, and Andrzej Rucinski. *Random graphs*. John Wiley & Sons, 2000.

[Kes80] Harry Kesten. The critical probability of bond percolation on the square lattice equals 1/2. *Communications in mathematical physics*, 74(1):41–59, 1980.

[KM75] J Körner and K Marton. Comparison of two noisy channels. *Topics in information theory*, pages 411–423, 1975.

[KXZ16] Florent Krzakala, Jiaming Xu, and Lenka Zdeborová. Mutual information in rank-one matrix estimation. In *2016 IEEE Information Theory Workshop (ITW)*, pages 71–75. IEEE, 2016.

[LKZ15] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Mmse of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing*, pages 680–687, 2015.

[Łuc90] Tomasz Łuczak. Component behavior near the critical point of the random graph process. *Random Structures & Algorithms*, 1(3):287–310, 1990.

[MNS13] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, pages 1–44, 2013.

[MNS15] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015.

[MP16] A. Makur and Y. Polyanskiy. Comparison of channels: criteria for domination by a symmetric channel. *arXiv:1609.06877*, September 2016.

[PW16] Yury Polyanskiy and Yihong Wu. Dissipation of information in channels with input constraints. *IEEE Trans. Inf. Theory*, 62(1):35–55, January 2016. also arXiv:1405.3629.

[PW17] Y. Polyanskiy and Y. Wu. Strong data-processing inequalities for channels and Bayesian networks. In *Convexity, Concentration and Discrete Structures, part of The IMA Volumes in Mathematics and its Applications*, volume 161. Springer-Verlag, New York, 2017. also arXiv:1508.06025.

[RU08] Tom Richardson and Ruediger Urbanke. *Modern coding theory*. Cambridge university press, 2008.

[vD97] Marten van Dijk. On a special class of broadcast channels with confidential messages. *IEEE Trans. Inform. Theory*, 43(2):712–714, March 1997.