Structure Learning of Similar Ising Models: Information-theoretic Bounds

Saurabh Sihag and Ali Tajer
Electrical, Computer, and Systems Engineering Department
Rensselaer Polytechnic Institute

***Abstract*—This paper considers the problem of estimating the structures of a pair of structurally similar graphs associated with two distinct Ising models. It is assumed that the graphs have the same number of nodes with unknown structures, with the additional side information that a known subset of nodes have identical structures (connectivity) in both graphs. The objective is the exact recovery of the structures of both graphs. The bounded degree and bounded edge sub-classes of Ising models are investigated, and necessary and sufficient conditions on the sample complexity for bounded probability of error under the two criteria are established. Furthermore, the results are compared with the conditions on the sample complexity of recovering the graphs independently. One major observation is that by judicially leveraging the information about the identical sub-graphs by jointly recovering both structures, the sample complexity reduces by a factor cp^2, where p is the number of nodes in the graph and c is some constant.**

I. Introduction

Graphical models provide structural representations of the conditional dependence among multiple random variables [1] and [2]. The nodes of the graphical models represent the random variables whose inter-dependence is encoded by the edges among them, such that the joint probability of the random variables captures the structure of the graph. Graphical models have applications in a wide range of domains, e.g., computer vision [3], genetics [4]–[6], social networks [7], and power systems [8]. In this paper, we focus on Ising models and consider the problem of recovering the graphical structures of a pair of graphical models with partially identical structures.

Graphical models with partially identical structures arise in various domains such as biological networks [4], physical infrastructures [9], and behavioral analysis [10]. The models in such applications consist of multiple layers of networks of information sources, in which the networks share some of their information sources, generating shared random variables. For instance, different gene networks that represent the subtypes of the same cancer may share similar edges across all subtypes and also have unique edges corresponding to each subtype [4].

Information-theoretic tools are effective for finding algorithm-independent guarantees on learning the structures of graphical models. The existing information-theoretic studies on graphical models include those of [11]–[13], which analyze the sample complexity for selecting the model of a given graph in various sub-classes of Ising models. Specifically, [11] establishes the necessary and sufficient conditions on the sample complexity for the exact recovery of the Ising models under bounded degree and bounded number of edges. These results are generalized in [14] to establish necessary conditions for set-based graphical model selection, in which the graph estimator outputs a set of potentially true graphs instead of a unique graph. Necessary conditions for recovering girth-bounded graphs and path-restricted graphs are analyzed in [12]. The problem of graphical model selection for various sub-classes of Ising models under the criterion of approximate recovery is investigated in [13], in which a certain number of missed edges or incorrectly included edges are tolerated in the estimated graph structure. Approximate recovery bounds on the sample complexity are characterized for Ising and Gaussian models without considering the effect of edge weights in [15]. The problems of structure recovery and inverse covariance matrix estimation for Gaussian models are studied in [16], where information-theoretic bounds on the sample complexity are delineated. Similarly, information-theoretic bounds are established for the class of power-law graphs in [17].

The problem of joint graphical model inference has been studied in [4], [5], [10], and [18]–[22]. Specifically, an empirical Bayes method is developed in [4] to identify interactions that are unique to each class and that are shared across all classes. Graphical Lasso-based algorithms are developed in [5] and [18]–[20] for joint inference of Gaussian graphical models. An optimization-based approach to the joint estimation of the graph structures using discrete data is studied in [10]. A Bayesian approach to jointly estimating Gaussian graphical models is investigated in [21], where the models with shared structure are identified from the data groups and their relative similarity is leveraged for inference.

All aforementioned studies on joint learning of graphical models focus on empirical frameworks for graph estimation or selection. In contrast, in this paper, we characterize infomation-theoretic necessary and sufficient conditions on the sample complexity under bounded probability of error in jointly recovering two partially identical graphs from edge-bounded and degree-bounded sub-classes of Ising models, and compare with the existing relevant results for single graphs studied in [11]. We observe that the graph decoder that jointly recovers a pair of partially identical graphs with p number of nodes saves at least cp^2 number of samples, where c is a positive constant. The specific subclass with this behavior is characterized by fixed maximum degree and maximum number of edges in a graph as p grows. We also observe in several cases that the gap between the results in this paper and the corresponding results from [11] scales at a similar rate as the sample complexity, albeit with significant savings in the sample complexity.

This research was supported in part by the U. S. National Science Foundation under the CAREER Award ECCS-1554482 and the grant DMS-1737976.

II. Graph Model

Consider a set of vertices $V \triangleq \{1,\dots,p\}$ connected by two distinct collection of edges denoted by $E_1 \subseteq V \times V$ and $E_2 \subseteq V \times V$, forming two distinct undirected graphs $\mathcal{G}_1 \triangleq (V, E_1)$, $\mathcal{G}_2 \triangleq (V, E_2)$. We use the convention $(u,v) \in E_i$ to show that an edge connects nodes $u, v \in V$ in graph $\mathcal{G}_i$. We also define $\mathcal{N}_i(u) \subseteq V$ as the set of nodes in the neighborhood of a node $u \in V$ in graph $\mathcal{G}_i$, i.e.,

$$\mathcal{N}_i(u) \triangleq \{w \in V : (u,w) \in E_i\} \ . \tag{1}$$

The degree of node u is denoted by d_u^i, where $d_u^i \triangleq |\mathcal{N}_i(u)|$. We leverage these two graphs to graphically represent two Ising graphical models.

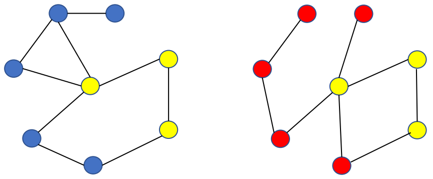

Fig. 1. Two graphs with partially identical structures. Yellow nodes in both graphs have identical internal edge structures.

Graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are assumed to share a common structure in a cluster of nodes denoted by $V_c \subseteq V$, i.e., the internal subgraph formed by a set of nodes V_c is identical in both graphs. An example of structurally similar graphs is shown in Fig. 1.

In the Ising model, each vertex $u \in V$ in graph $\mathcal{G}_i$ is associated with a binary random variable denoted by $X_i^u \in \mathcal{X} \triangleq \{-1, 1\}$. The joint probability density function (pdf) of the random variables $\mathbf{X}_i \triangleq [X_i^1, \dots, X_i^p]$ associated with graph $\mathcal{G}_i$ is given by

$$f_i(\mathbf{X}_i) \triangleq \frac{1}{Z_i} \exp\left(\sum_{u,v\in V} \lambda_i^{uv} X_i^u X_i^v \right) , \tag{2}$$

where

$$\lambda_i^{uv} \triangleq \left\{ \begin{array}{ll} \lambda, & \text{if } (u,v) \in E_i \\ 0, & \text{otherwise} \end{array} \right. , \tag{3}$$

and Z_i is the partition function given by

$$Z_i \triangleq \sum_{X_i \in \{-1,1\}^p} \exp\left(\sum_{u,v\in V} \lambda_i^{uv} X_i^u X_i^v \right) . \tag{4}$$

Parameter $\lambda \in \mathbb{R}^+$ captures the dependence among the random variables associated with the vertices in the graph. As illustrated in [11], as λ approaches to 0 or grows to infinity, recovering the structure of the graph from its samples becomes increasingly more difficult. Furthermore, corresponding to graph $\mathcal{G}_i$, we also define the maximum neighborhood weight according to

$$\zeta_i \triangleq \max_{w\in V} \sum_{u\in\mathcal{N}_i(w)} \lambda_i^{wu} \ . \tag{5}$$

III. Problem Formulation

In this section we formalize the similarity models and the joint structure recovery criteria.

A. Graph Similarity Models

Definition 1. *Two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are said to be $\eta-$similar, for some $\eta \in (0,1)$, if they share a cluster of nodes that have identical internal graphical structures in both graphs, and the size of the cluster is $|V_c| = \lfloor \eta p \rfloor$.*

For the convenience in notations, we define $q \triangleq \lfloor \eta p \rfloor$ and $\bar{q} \triangleq p - \lfloor \eta p \rfloor$. We denote the class of Ising models by $\mathcal{I}$, and the class of $\eta-$similar pairs of Ising models by $\mathcal{I}_\eta$. In this paper, we consider the following sub-classes of Ising models.

- **Degree-bounded class** $\mathcal{I}_\eta^d$: This class contains all the $\eta-$similar pair of graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, each with a maximum degree d. Clearly, in this class, we have $\zeta_i = \lambda d$, where ζ_i is defined in (5).
- **Edge-bounded class** $\mathcal{I}_\eta^k$: This class contains all the $\eta-$similar pair of graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, each with at most k edges in each graph and at most $\lfloor \gamma k \rfloor, \gamma \in [0,1]$, number of edges lie in the shared sub-graph with nodes in V_c. We set $\bar{\gamma} \triangleq 1 - \gamma$. We also assume that the maximum neighborhood weight is bounded by ζ, i.e., $\zeta_i \leq \zeta$ for $i \in \{1,2\}$.

B. Recovery Criteria

We collect n independent samples $\mathbf{X}_i$, generated according to f_i, from graph $\mathcal{G}_i$, for $i \in \{1,2\}$. We denote the collection of n samples from $\mathcal{G}_i$ by $\mathbf{X}_i^n$. The objective is to jointly estimate graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ from samples $\mathbf{X}_1^n$ and $\mathbf{X}_2^n$[†]. We denote the graph decoder $\psi : \mathcal{X}^{n\times p} \times \mathcal{X}^{n\times p} \rightarrow \mathcal{C}$ as the function that maps the data to graphs in class $\mathcal{C}$. To capture the accuracy of decoder, corresponding to any generic class of pairs of graph $\mathcal{C}$, we define $\mathsf{P}(\mathcal{C})$ as the maximal probability of error in exact recovery over the class $\mathcal{C}$, i.e.,

$$\mathsf{P}(\mathcal{C}) \triangleq \max_{(\mathcal{G}_1,\mathcal{G}_2)\in\mathcal{C}} \mathbb{P}[\psi(\mathbf{X}_1^n, \mathbf{X}_2^n) \neq (\mathcal{G}_1, \mathcal{G}_2)] \ , \tag{6}$$

where the probability is computed with respect to distributions f_1 and f_2. Under this setting, a total of $2n$ samples are used for joint model selection, corresponding to which we establish performance guarantees on $\mathsf{P}(\mathcal{C})$ and analyze their scaling behavior with respect to various parameters, i.e., λ, p, d, and k. Let $n_{\rm s}$ be the number of samples required for recovering the structure of a single graph with the same performance guarantees (c.f. [11]). Hence,

$$D \triangleq 2(n_{\rm s} - n) \tag{7}$$

quantifies the gap between the sample complexities of jointly recovering $\mathcal{G}_1$ and $\mathcal{G}_2$ and recovering them independently, while achieving the same level of reliability in decoding. In the next section, we provide necessary and sufficient conditions on the sample complexity n for an arbitrary target reliability level. These results can also be leveraged to analyze the scaling behavior of D in different regimes.

IV. Main Results

In this section, we provide the necessary and sufficient conditions on the sample complexity n. The sufficient conditions

[†] The results in this paper can be generalized to settings with more than two graphs, and different numbers of samples from each graph. For clarity, we only analyze the setting with two graphs, and equal number of samples per graph.

TABLE I

SUMMARY OF MAIN RESULTS (NON-EXPONENTIAL SCALING) FOR EXACT RECOVERY OF ISING MODELS.

Graph Class	Parameters	Scaling behavior		Comparison with [11]	
		Necessary Conditions	Sufficient Conditions	Necessary Conditions	Sufficient Conditions
Bounded degree $\mathcal{I}_\eta^d$	$\lambda = O\left(\frac{1}{d}\right)$	$\Omega(d^2 \log p)$	$\Omega(d^3 \log p)$	$D \sim \Omega(d^2)$	$D \sim O(d^3 \log p)$
	$\lambda = O\left(\frac{1}{p}\right)$ d fixed	$\Omega(p^2 \log p)$	$\Omega(p^2 \log p)$	$D \sim \Omega(p^2)$	$D \sim O(p^2 \log p)$
Bounded edge $\mathcal{I}_\eta^k$	$\lambda = O\left(\frac{1}{\sqrt{k}}\right)$	$\Omega(k \log p)$	$\Omega(k^2 \log p)$	$D \sim \Omega(k)$	$D \sim O(k^2 \log p)$
	$\lambda = O\left(\frac{1}{p}\right)$ k fixed	$\Omega(p^2 \log p)$	$\Omega(p^2 \log p)$	$D \sim \Omega(p^2)$	$D \sim O(p^2 \log p)$

are established based on the analysis of a maximum likelihood (ML) decoder. The necessary conditions established are algorithm-independent and serve as performance benchmarks for any designed algorithm.

A. *Sufficient Conditions*

In this section, we provide sufficient conditions on the number of samples for model selection of different classes of Ising models. We also analyze the scaling behaviors of the results with parameters λ, p, k, and d.

Theorem 1 (Class $\mathcal{I}_\eta^d$)**.** *Consider a pair of $\eta-$similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{I}_\eta^d$. If the sample size n satisfies*

$$n \geq r_1 \max\{B_1, 2B_2\}\ , \tag{8}$$

where we have defined

$$r_1 \triangleq \frac{2d(3\exp(2\lambda d)+1)}{\sinh^2(\lambda/4)}\ , \tag{9}$$

$$B_1 \triangleq \left(2\log q + \log 4d + \log\frac{1}{\delta}\right)\ , \tag{10}$$

$$B_2 \triangleq \left(\log\frac{8d\bar{q}}{\delta} + \log\left(\binom{\bar{q}}{2} + pq\right)\right)\ , \tag{11}$$

then there exists a graph decoder $\psi : \mathcal{X}^{n\times p} \times \mathcal{X}^{n\times p} \to \mathcal{I}_\eta^d$ that achieves $\mathsf{P}(\mathcal{I}_\eta^d) \leq \delta$.

Next, we elaborate on the scaling behavior shown in Theorem 1 in different regimes. Note that even though the terms r_1B_1 and r_1B_2 in (8) have similar scaling behavior in λ, p, and d for constant η, r_1B_1 dominates the sample complexity when the size of shared cluster is large enough, i.e., $\frac{q}{\bar{q}} \gg 1$. This captures the effect of the size of the shared cluster on sample complexity. Furthermore, depending on the relationship between λ and d, we have the following distinct behaviors for the sampling complexity saving D in different regimes. In all these regimes, it is assumed that the maximum degree d is increasing with the graph size p.

1. $\lambda = \Theta(1)$: In this regime, both r_1B_1 and r_1B_2 scale as $e^d d \log\frac{pd}{\delta}$. Also, in comparison with the existing results for a single graph, we conclude that D scales as $O(e^d \log p)$. Furthermore, when we have $d = \omega(\log(\log p))$ and δ fixed, both the bound on sample complexity and D scale exponentially in d.
2. $\lambda = O\left(\frac{1}{d}\right)$: In this regime, as $d \to \infty$, we have $\sinh(\lambda/4) = O(\lambda)$. Therefore, the sufficient condition from Theorem 1 can be simplified to $n \geq c_1 \max\{d^2, \lambda^{-2}\} d \log p/\delta$, where c_1 is a positive constant. For a constant δ, the bound on the sample complexity has an asymptotic scaling behavior given by $\Omega(d^3 \log p)$. In comparison with the existing result for a single graph, we conclude that D scales at most at the rate of $d^3 \log p$, i.e., the gain in the number of samples for joint model selection over independent model selection of graphs using an ML based decoder scales as $O(d^3 \log p)$.
3. $\lambda = \Theta(d)$: In this regime, the terms r_1B_1 and r_1B_2 scale as $e^{\lambda d} \log p$ for a constant δ. Furthermore, if we have $\lambda d = \omega(\log(\log p))$, then the scaling behavior is simplified to $e^{\lambda d}$. Also, in comparison with the corresponding results for single graphs, we conclude that D scales as $O(e^{\lambda d} \log p)$.

Theorem 2 (Class $\mathcal{I}_\eta^k$)**.** *Consider a pair of $\eta-$similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{I}_\eta^k$. If the sample size n satisfies*

$$n \geq r_2 \max\{2B_3, B_4\}\ , \tag{12}$$

for sufficiently large p, where we have defined

$$r_2 \triangleq \frac{3\exp(2\zeta)+1}{\sinh^2(\lambda/4)}\ , \tag{13}$$

$$B_3 \triangleq 2(k'+1)\log p + \log\frac{1}{\delta}\ , \tag{14}$$

$$B_4 \triangleq 2(\lfloor \gamma k \rfloor + 1)\log q + \log\frac{1}{\delta}\ , \tag{15}$$

$$k' \triangleq \min\left\{k, \binom{\bar{q}}{2} + \bar{q}q\right\}\ , \tag{16}$$

then there exists a graph decoder $\psi : \mathcal{X}^{n\times p} \times \mathcal{X}^{n\times p} \to \mathcal{I}_\eta^k$ that achieves $\mathsf{P}(\mathcal{I}_\eta^k) \leq \delta$.

For the class of $\mathcal{I}_\eta^k$ graphs, from the results in Theorem 2 it can be readily verified that r_2B_4 dominates r_2B_3 when $\frac{q}{\bar{q}} \gg 1$, which illustrates the effect of η. Note that k' in (16) reflects the maximum number of edges that can exist in the non-shared cluster of the graphs. Depending on the relationship between λ and k, we have the following distinct behaviors for the sampling complexity saving D in different regimes. In all these regimes it is assumed that k is increasing with the graph size p.

1. $\lambda = \Theta(1)$: When η is large enough, such that, $\frac{\bar{q}}{q} \ll 1$ and $\gamma k \gg \binom{\bar{q}}{2} + \bar{q}q$, the sample complexity is dominated by $r_2 B_4$ which scales as $\Omega(e^{\zeta} k \log p)$. Also, when we have $k' = k$, the bound on sample complexity scales as $\Omega(e^{\zeta} k \log p)$. Therefore, in this regime, the bound on sample complexity is always dominated by a term that has a scaling behavior given by $\Omega(e^{\zeta} k \log p)$ for fixed δ. Furthermore, in comparison with the results for single graphs, we conclude that D scales as $O(e^{\zeta} k \log p)$.
2. $\lambda = O\left(\frac{1}{\sqrt{k}}\right)$: When $k \to \infty$, we have $\sinh(\lambda/4) = O(\lambda)$. Therefore, in this regime, the bound on sample complexity scales according to $\Omega(e^{\zeta} k^2 \log p/\delta)$. If ζ is a constant or scales as $\zeta = O(\lambda\sqrt{k})$, then the bound on sample complexity scales as $\Omega(k^2 \log p)$ for fixed δ. Furthermore, under controlled ζ, we have $D = O(k^2 \log p)$, which implies that the gain in number of samples scales with at most $k^2 \log p$.
3. $\lambda = \Theta(\sqrt{k})$: In this regime, when both λ and k are increasing with p, the bound on the sample complexity scales as $\Omega(e^{\zeta} \log p)$. Also, by comparing with the result for single graphs, we conclude that $D = O(e^{\zeta} \log p)$. When we have $\zeta \geq \lambda\sqrt{k}$ and $\lambda\sqrt{k} = \omega(\log(\log p))$, both the bound on sample complexity and D scales exponentially in $e^{\lambda\sqrt{k}}$.

B. Necessary Conditions

Next, we provide the necessary conditions for exact recovery for the different sub-classes of Ising models. We also provide remarks on the scaling behavior of the sample complexity in terms of different parameters and compare the results with the sufficient conditions established in Theorem 1 and Theorem 2.

Theorem 3 (Class $\mathcal{I}_{\eta}^{d}$)**.** *Consider a pair of $\eta-$similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{I}_{\eta}^{d}$. If the sample size n satisfies*

$$n \leq (1-\delta) \max\left\{A_1, A_2, A_3\right\} \ , \tag{17}$$

where we have defined

$$A_1 \triangleq \frac{1}{2\lambda \tanh(\lambda)} \log\left(\frac{q^2}{8} + \frac{\bar{q}^4}{16}\right) \ , \tag{18}$$

$$A_2 \triangleq \frac{\exp(\lambda d)}{8\lambda d \exp(\lambda)} \log\left(\frac{qd}{4} + \frac{(\bar{q}d)^2}{16}\right) \ , \tag{19}$$

$$A_3 \triangleq \frac{qd}{16p} \log\frac{q}{8d} + \frac{\bar{q}d}{8p} \log\frac{\bar{q}}{8d} \ , \tag{20}$$

then for any graph decoder $\psi : \mathcal{X}^{n\times p} \times \mathcal{X}^{n\times p} \to \mathcal{I}_{\eta}^{d}$, we have

$$\mathsf{P}(\mathcal{I}_{\eta}^{d}) \geq \delta - \frac{1}{\log p} \ , \tag{21}$$

for sufficiently large p.

Next, we analyze the scaling behavior of the results in Theorem 3. We start by noting that the terms A_1, A_2, and A_3 have different scaling behaviors in terms of λ, p, and d. Depending on the combination of parameters, each of the three terms A_1, A_2, and A_3 will be the dominant (maximum) term in a specific regime. This leads to different scaling behaviors in different regimes, as described next. In the following regimes, we assume that d is increasing with p.

1. $\lambda = \Theta(1)$: In this regime, A_1 and A_3 scale according to $\Omega(\log p)$ and A_2 scales according to $\Omega(e^{d} \log p)$. Clearly, A_2 dominates the bound on sample complexity. Also, in comparison with the existing result for a single graph, we conclude that D scales at least at the rate e^{d}. The scaling behaviors of the necessary condition on the sample complexity and D are consistent with the corresponding results from Theorem 1.
2. $\lambda = O\left(\frac{1}{d}\right)$: In this regime, A_1 scales according to $\Omega(\frac{\log p}{\lambda \tanh \lambda})$, and A_2 scales according to $\Omega(e^{\lambda d} \frac{\log(pd)}{d})$. When $\lambda = O\left(\frac{1}{d}\right)$, the bound A_2 does not scale exponentially in d. As $p \to \infty$, we have $\tanh \lambda = O(\lambda)$. Therefore, A_1 dominates the bound on sample complexity and the necessary condition in Theorem 3 reduces to $n \leq c_3 \max\{d^2, \lambda^{-2}\} \log p$ for some constant c_3. This implies that the bound on the sample complexity scales according to $\Omega(d^2 \log p)$. In comparison with the corresponding necessary condition for a single graph, we conclude that as $p \to \infty$, $D = \Omega(d^2)$. Also, the necessary conditions on the sample complexity in this regime match the sufficient conditions established in Theorem 1 within a factor of d.
3. $\lambda = \Theta(d)$: In this regime, A_2 scales according to $O(e^{\lambda d} \log(pd))$ and $\Omega(e^{\lambda d} \frac{\log pd}{\lambda d})$. Therefore, the bound on the sample complexity is dominated by $\exp(\lambda d) \log p$. When $\lambda d = \omega(\log(\log p))$, we futher observe that A_2 scales exponentially in λd and dominates the bound on sample complexity. In comparison with the corresponding necessary condition for a single graph, we conclude that when η is fixed, the difference between the two terms scales at least at the rate $\exp(\lambda d)$ in this regime, which is at the same rate as the bound on sample complexity if $\lambda d = \omega(\log(\log p))$. Also, the scaling behaviors of the necessary condition on the sample complexity and D are consistent with the corresponding results from Theorem 1.

Theorem 4 (Class $\mathcal{I}_{\eta}^{k}$)**.** *Consider a pair of $\eta-$similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{I}_{\eta}^{k}$, in which $\gamma \leq \frac{\eta p}{4k}$. If the sample size n satisfies*

$$n \leq (1-\delta) \max\{A_1, A_4\} \ , \tag{22}$$

where A_1 is defined in (18) *and*

$$A_4 \triangleq \frac{\log\left(\frac{\bar{\gamma}^2 k^2/16 + \gamma k/4}{4}\right)}{32\lambda\sqrt{k}\exp(2\lambda)\sinh(\lambda)} \times \left(\frac{\sqrt{\gamma}}{\exp\left(\lambda\sqrt{\lfloor \gamma k \rfloor}\right)} + \frac{\sqrt{\bar{\gamma}}}{\exp\left(\lambda\sqrt{\lfloor \bar{\gamma} k \rfloor}\right)}\right)^{-1} \ , \tag{23}$$

then for any graph decoder $\psi : \mathcal{X}^{n\times p} \times \mathcal{X}^{n\times p} \to \mathcal{I}_{\eta}^{k}$, we have

$$\mathsf{P}(\mathcal{I}_{\eta}^{k}) \geq \delta - \frac{1}{\log p} \ . \tag{24}$$

Close scrutiny of A_1 and A_4 indicates that only A_4 depends on k. As a result, depending on the value of k, we can specify which of the terms A_1 and A_4 becomes the dominant one.

Hence, we have the following regimes and scaling behaviors under each. In the following regimes, it is assumed that k is increasing in p.

1. $\lambda = \Theta(1)$: In this regime, A_1 scales at the rate $\log p$ and A_4 scales at the rate $e^{\sqrt{k}} \log p$. Therefore, A_4 dominates the sample complexity. Hence, D scales as $\Omega(e^{\sqrt{k}})$ in this regime. Furthermore, when we have $\zeta \geq \lambda\sqrt{k}$ and $e^{\sqrt{k}} = \omega(\log p)$, both A_4 and D scale exponentially in $\sqrt{k}$, and the scaling behaviors of the necessary conditions and D are consistent with those derived from the sufficient conditions in Theorem 2.
2. $\lambda = O\left(\frac{1}{\sqrt{k}}\right)$: In this regime, A_1 scales according to $\Omega\left(\frac{\log p}{\lambda \tanh \lambda}\right)$ and dominates A_4, which scales according to $O(\lambda^{-1} \log k)$. Therefore, as $k \to \infty$, A_1 dominates the sample complexity and the necessary condition in Theorem 4 reduces to $n \leq c_4 \max\{k, \lambda^{-2}\} \log p$ for some constant c_4. Furthermore, A_1 has an asymptotic scaling behavior given by $\Omega(k \log p)$. In comparison with the corresponding lower bound for a single graph, we conclude that $D = \Omega(k)$ in this regime. Furthermore, the scaling behaviors of the sufficient condition and D derived from the results in Theorem 2 match the corresponding results from Theorem 4 within a factor k.
3. $\lambda = \Theta(\sqrt{k})$: In this regime, A_4 scales as $e^{\lambda\sqrt{k}}$ and A_1 scales as $\frac{\log p}{\lambda \tanh \lambda}$. Therefore, as λ increases with k and p, A_4 dominates the sample complexity. Hence, the difference D scales exponentially in $\lambda\sqrt{k}$, which is same as the scaling behavior of the necessary condition on sample complexity. Also, the scaling behaviors of the sufficient conditions and D are consistent with those derived from the necessary conditions in Theorem 4.

C. Exact Sample Complexity

Next, we compare the results of Theorems $1 - 4$ jointly in order to specify a certain regime under which we have the same scaling behavior of the sample complexity, i.e., the necessary and sufficient conditions scale at the same rate for both sub-classes $\mathcal{I}_\eta^d$ and $\mathcal{I}_\eta^k$.

Corollary 1 (Exact Recovery). *When the maximum degree d and the maximum number of edges k are fixed, and we have $\lambda = O(1/p)$, then the exact scaling behavior of the sample complexity in classes $\mathcal{I}_\eta^d$ and $\mathcal{I}_\eta^k$ is $\Omega(p^2 \log p)$, as the size of the graph, p grows.*

The main results with non-exponential scaling behavior are summarized in Table 1. In all the results, it is assumed that $\eta > 0$ is fixed and therefore, it does not affect the scaling behavior of the sample complexity.

V. Conclusion

In this paper, we have considered the problem of joint model selection of partially identical graphs in various sub-classes of Ising models. Structural similarity between any two graphs implies potentially redundant information. Under the criteria of exact recovery of the structure of the graphs, we have characterized necessary and sufficient conditions on the sample complexity of joint model selection for various sub-classes of Ising models. We have also analyzed the scaling behavior of the sample complexity for joint model selection presented in this paper and compared the results with that for model selection of single graphs in the existing literature.

References

[1] S. L. Lauritzen, *Graphical Models.* Clarendon Press, May 1996, vol. 17.

[2] J. Pearl, *Causality: Models, Reasoning, and Inference.* Oxford: Cambridge University Press, 2009.

[3] C. S. Won and H. Derin, "Unsupervised segmentation of noisy and textured images using Markov random fields," *CVGIP: Graphical Models and Image Processing*, vol. 54, no. 4, pp. 308–328, 1992.

[4] X. Chen, F. J. Slack, and H. Zhao, "Joint analysis of expression profiles from multiple cancers improves the identification of microRNA–gene interactions," *Bioinformatics*, vol. 29, no. 17, pp. 2137–2145, 2013.

[5] J. Fang, L. S. Dongdong, S. Charles, Z. Xu, V. D. Calhoun, and Y.-P. Wang, "Joint sparse canonical correlation analysis for detecting differential imaging genetics modules," *Bioinformatics*, vol. 32, no. 15, pp. 3480–3488, 2016.

[6] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West, "Sparse graphical models for exploring gene expression data," *Journal of Multivariate Analysis*, vol. 90, no. 1, pp. 196–212, 2004.

[7] Y. Jacob, L. Denoyer, and P. Gallinari, "Learning latent representations of nodes for classifying in heterogeneous social networks," in *Proc. ACM international Conference on Web Search and Data Mining*, New York, Feb. 2014, pp. 373–382.

[8] K. Dvijotham, M. Chertkov, P. V. Hentenryck, M. Vuffray, and S. Misra, "Graphical models for optimal power flow," *Constraints*, vol. 22, no. 1, pp. 24–49, 2017.

[9] M. Rabbat, R. Nowak, and M. Coates, "Network tomography and the identification of shared infrastructure," in *Proc. Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, Nov. 2002, pp. 34–38.

[10] J. Guo, J. Cheng, E. Levina, G. Michailidis, and J. Zhu, "Estimating heterogeneous graphical models for discrete data with an application to roll call voting," *The Annals of Applied Statistics*, vol. 9, no. 2, pp. 821 – 848, Jun. 2015.

[11] N. P. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions." *IEEE Trans. Information Theory*, vol. 58, no. 7, pp. 4117–4134, May 2012.

[12] R. Tandon, K. Shanmugam, P. K. Ravikumar, and A. G. Dimakis, "On the information-theoretic limits of learning Ising models," in *Proc. Advances in Neural Information Processing Systems*, Montreal, Canada, Dec. 2014, pp. 2303–2311.

[13] J. Scarlett and V. Cevher, "On the difficulty of selecting Ising models with approximate recovery," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 4, pp. 625–638, Dec. 2016.

[14] D. Vats and J. M. Moura, "Necessary conditions for consistent set-based graphical model selection," in *Proc. IEEE International Symposium on Information Theory*, Saint-Petersburg, Russia, Jul. 2011, pp. 303–307.

[15] A. K. Das, P. Netrapalli, S. Sanghavi, and S. Vishwanath, "Learning Markov graphs up to edit distance," in *Proc. IEEE International Symposium on Information Theory*, Cambridge, MA, Jul. 2012, pp. 2731–2735.

[16] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic bounds on model selection for Gaussian Markov random fields," in *Proc. IEEE International Symposium on Information Theory*, Austin, TX, Jun. 2010.

[17] R. Tandon and P. Ravikumar, "On the difficulty of learning power law graphical models," in *Proc. IEEE International Symposium on Information Theory*, Istanbul, Turkey, Jul. 2013, pp. 2493–2497.

[18] P. Danaher, P. Wang, and D. M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 2, pp. 373–397, Mar. 2014.

[19] S. Yang, Z. Lu, X. Shen, P. Wonka, and J. Ye, "Fused multiple graphical lasso," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 916–943, 2015.

[20] K. Mohan, P. London, M. Fazel, D. Witten, and S.-I. Lee, "Node-based learning of multiple Gaussian graphical models," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 445–488, 2014.

[21] C. B. Peterson, F. C. Stingo, and M. Vannucci, "Bayesian inference of multiple Gaussian graphical models," *Journal of the American Statistical Association*, vol. 110, no. 509, pp. 159–174, 2015.

[22] H. Qiu, F. Han, H. Liu, and B. Caffo, "Joint estimation of multiple graphical models from high-dimensional time series," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 78, no. 2, pp. 487–504, 2016.