The Annals of Statistics
2012, Vol. 40, No. 3, 1346–1375
DOI: 10.1214/12-AOS1009

HIGH-DIMENSIONAL STRUCTURE ESTIMATION IN ISING MODELS: LOCAL SEPARATION CRITERION[1]

BY ANIMASHREE ANANDKUMAR[2], VINCENT Y. F. TAN[3,4], FURONG HUANG[2] AND ALAN S. WILLSKY[4]

University of California Irvine, Institute for Infocomm Research and National University of Singapore, University of California Irvine and Massachusetts Institute of Technology

We consider the problem of high-dimensional Ising (graphical) model selection. We propose a simple algorithm for structure estimation based on the thresholding of the empirical conditional variation distances. We introduce a novel criterion for tractable graph families, where this method is efficient, based on the presence of sparse local separators between node pairs in the underlying graph. For such graphs, the proposed algorithm has a sample complexity of $n = \Omega(J_{\min}^{-2} \log p)$, where p is the number of variables, and $J_{\min}$ is the minimum (absolute) edge potential in the model. We also establish nonasymptotic necessary and sufficient conditions for structure estimation.

1. Introduction. The use of probabilistic graphical models allows for succinct representation of high-dimensional distributions, where the conditional-independence relationships among the variables are represented by a graph. Such models have found many applications in a variety of areas, including computer vision [14], bio-informatics [21], financial modeling [15] and social networks [25]. For instance, graphical models are employed for contextual object recognition to improve detection performance based on object co-occurrences [14] and for modeling opinion formation and technology adoption in social networks [25, 30].

A major challenge involving graphical models is structure estimation, given samples drawn from the model. It is known that such a learning task is NP-hard [7, 27]. This challenge is compounded in the *high-dimensional* regime, where the number of available observations is typically much smaller than the number of dimensions (or variables). It is thus imperative to design efficient algorithms for structure estimation of graphical models with low sample complexity.

In their seminal work, Chow and Liu presented an efficient algorithm for structure estimation of tree-structured graphical models based on a maximum weight

[1]An abridged version of this paper appears in Proc. of NIPS 2011.

[2]Supported by the setup funds at UCI and the AFOSR Award FA9550-10-1-0310.

[3]Supported in part by A*STAR, Singapore.

[4]Supported in part by AFOSR under Grant FA9550-08-1-1080.

MSC2010 subject classifications. Primary 62H12; secondary 05C80.

Key words and phrases. Ising models, graphical model selection, local-separation property.

spanning tree algorithm [16]. Since then, various algorithms have been proposed for structure estimation of sparse graphical models. They can be broadly classified into two categories: combinatorial algorithms [10, 39] and those based on convex relaxation [11, 37, 41, 42]. The former approach is typically based on certain local tests on small groups of data, and then combining them to output a graph structure, while the latter approach involves solving a penalized convex optimization problem. See Section 1.2 for a detailed discussion of these approaches.

In this paper, we propose a novel local algorithm and analyze its performance for structure estimation of Ising models, which are pairwise binary graphical models. Our proposed algorithm circumvents one of the primary limitations of existing local algorithms [10, 39] for consistent estimation in high-dimensions—that the graphs have a bounded degree as the number of nodes p tends to infinity. We give a precise characterization of the class of graphs which can be consistently recovered by our algorithm with low computational and sample complexities. We demonstrate that a fundamental property shared by these graphs is that they have *sparse local vertex separators* between any two nonneighbors in the graph. A wide variety of graphs satisfy this property. These include large girth graphs, the Erdős–Rényi random graphs[5] [8] and the power-law graphs [18], as well as graphs with short cycles such as the small-world graphs [51] and other hybrid graphs [18, Chapter 12].

Our results are applicable in the realms of social networks, bio-informatics, computer vision and so on. Here, we elaborate on its relevance to social networks. The aforementioned graphs (i.e., the power-law and the small-world graphs) have been employed extensively for modeling the topologies of social networks [2, 40]. More recently, Ising models on such topologies have been employed for modeling various phenomena in social networks [48], such as opinion formation [23, 25, 34] and technology adoption [30]. A concrete example is the use of an Ising model for the U.S. senate voting network [52]. The nodes of the graph represent the senators, and the data are the voting decisions made by the senators. Estimating the graph reveals interesting relationships between the senators and the effect of political affiliations on their decisions. Similarly, in many other scenarios (e.g., online social networks), we have access to a sequence of measurements at the nodes of the network. For instance, we may gather the opinions of different users or measure the popularity of new technologies. As a first-order approximation, we can regard such a sequence of measurements as being independent and identically distributed (i.i.d.) samples drawn from an Ising model. Our findings imply that the topology of such social-network models can be efficiently estimated under some mild and transparent conditions.

[5]The Erdős–Rényi graphs have sparse local vertex separators asymptotically almost surely (a.a.s.) with respect to the random graph measure. Indeed, whenever we mention ensembles of random graphs in the sequel, our statements are taken to hold a.a.s.

1.1. *Summary of results.* Our main contributions in this work are threefold. We propose a simple algorithm for structure estimation of Ising models. The algorithm is based on approximate conditional independence testing based on conditional variation distances. Second, we derive sample complexity results for consistent structure estimation in high dimensions. Third, we prove novel lower bounds on the sample complexity required for any learning algorithm to be consistent for model selection.

We propose an algorithm for structure estimation, termed as conditional variation distance thresholding (CVDT), which tests if two nodes are neighbors by searching for a node set which (approximately) separates them in the underlying Markov graph. It first computes the minimum empirical conditional variation distance in (14) of a given node pair over conditioning sets of bounded cardinality η. Second, if the minimum exceeds a given threshold (depending on the number of samples n and the number of nodes p), the node pair is declared as an edge. This test has a computational complexity of $O(p^{\eta+2})$. Thus, the computational complexity is low if η is small. Further, it requires only low-order statistics (up to order $\eta + 2$). We establish that the parameter η is a bound on the size of local vertex-separators between any two nonneighbors in the graph, and is small for many common graph families, introduced before.

We establish that under a set of mild and transparent assumptions, structure learning is consistent in high dimensions for CVDT when the number of samples scales as $n = \Omega(J_{\min}^{-2} \log p)$, for a p-node graph, where $J_{\min}$ is the minimum (absolute) edge-potential of the Ising model. We relate the conditions for successful graph recovery to certain phase transitions in the Ising model. We also derive (nonasymptotic) PAC guarantees for CVDT and provide explicit results for specific graph families.

We derive a lower bound (necessary condition) on the sample complexity required for consistent structure learning with positive probability by any algorithm. We prove that $n = \Omega(c \log p)$ number of samples is required by any algorithm to ensure consistent learning of Erdős–Rényi random graphs, where c is the average degree, and p is the number of nodes. We also present a nonasymptotic necessary condition which employs information-theoretic techniques such as Fano's inequality and typicality. We also provide results for other graph families such as the girth-constrained graphs and augmented graphs.

Our results have several ramifications: we characterize the trade-off between various graph parameters, such as the maximum degree, threshold for local path length and the strength of edge potentials for efficient and consistent structure estimation. For instance, we establish a natural relationship between maximum degree and girth of a graph for consistent estimation: graphs with large degrees can be consistently estimated by our algorithm when they also have large girths. Indeed, in the extreme case of trees which have infinite girth, they can be consistently estimated with no constraint on the node degrees, corroborating the initial observation

by Chow and Liu [16]. We also derive stronger guarantees for many random-graph families. For instance, for the Erdős–Rényi random graph family and the small-world family (which is the union of a d-dimensional grid and an Erdős–Rényi random graph), the minimum sample complexity scales as $n = \Omega(c^2 \log p)$, where c is the average degree of the Erdős–Rényi random graph. Thus, when the average degree is bounded [$c = O(1)$], the sample complexity of our algorithm scales as $n = \Omega(\log p)$. Recall that the sample complexity of learning tree models is $\Omega(\log p)$ [47]. Thus, we establish that the complexity of learning sparse random graphs using the proposed algorithm is akin to learning tree models in certain parameter regimes.

Our sufficient conditions for consistent structure estimation impose transparent constraints on the graph structure and the parameters. The structural property is related to the presence of sparse local vertex separators between nonadjacent node pairs in the graph. The conditions on the parameters require that the edge potentials of the Ising model be below a certain threshold, which we explicitly characterize. In fact, we establish that below this threshold, the effect of long-range paths in the model decays and that graph estimation is feasible via local conditioning, as prescribed by our algorithm. Similar notions have been previously established in other contexts, for example, to establish polynomial mixing time for Gibbs sampling of the Ising model [32]. We compare these different criteria and show that we can guarantee consistent learning in high dimensions under weaker conditions than those required for polynomial mixing of Gibbs sampling. Ours is the first work (to the best of the authors' knowledge) to establish such explicit connections between structure estimation and the statistical physics properties (i.e., phase transitions) of Ising models. Establishing these results requires the development and use of tools (e.g., self-avoiding walk trees), not previously employed for learning problems.

1.2. *Related work.* The problem of structure estimation of a general graphical model [7, 27] is NP-hard. However, for tree-structured graphical models, the maximum-likelihood (ML) estimation can be implemented efficiently via the Chow–Liu algorithm [16] since ML estimation reduces to a maximum-weight spanning tree problem where the edge weights are the empirical mutual information quantities, computed from samples. It can be established that the sample complexity for the Chow–Liu algorithm scales as $n = \Omega(\log p)$, where p is the number of variables [47]. Error-exponent analysis of the Chow–Liu algorithm was performed in [45, 46], and extensions to general acyclic models [33, 47] and trees with latent (or hidden) variables [15] have also been studied recently.

Given the feasibility of structure learning of tree models, a natural extension is to consider learning the structures of *junction trees*.[6] Efficient algorithms

[6]Junction trees are formed by triangulating a given graph, and its nodes correspond to the maximal cliques of the triangulated graph [49]. The *treewidth* of a graph is one less than the minimum possible size of the maximum clique in the triangulated graph over all possible triangulations.

have been previously proposed for learning junction trees with bounded treewidth (e.g., [12]). However, the complexity of these algorithms is exponential in the tree width, and hence are not practical when the graphs have unbounded treewidth.[7]

There are mainly two classes of algorithms for graphical model selection: local-search based approaches [10, 39] and those based on convex optimization [11, 37, 41, 42]. The latter approach typically incorporates an ℓ_1 penalty term to encourage sparsity in the graph structure. In [41], structure estimation of Ising models is considered where neighborhood selection for each node is performed, based on ℓ_1-penalized logistic regression. It was shown that this algorithm has a sample complexity of $n = \Omega(\Delta^3 \log p)$ under a set of so-called "incoherence" conditions. However, the incoherence conditions are not easy to interpret and NP-hard to verify in general models [6]. For more detailed comparison, see Section 3.5.

In contrast to convex-relaxation approaches, the local-search based approach relies on a series of simple local tests for neighborhood selection at individual nodes. For instance, the work in [10] performs neighborhood selection at each node based on a series of conditional-independence tests. Abbeel et al. [1] propose an algorithm, similar in spirit to learning factor graphs with bounded degree. The authors in [44] and [13] consider conditional-independence tests for learning Bayesian networks. In [39], the authors suggest an alternative, greedy algorithm, based on minimizing conditional entropy, for graphs with large girth and bounded degree. However, these works [1, 10, 13, 39, 44] require the maximum degree in the graph to be bounded ($\Delta = O(1)$) which may be restrictive in practical scenarios. We consider graphical model selection on graphs where the maximum degree is allowed to grow with the number of nodes (albeit at a controlled rate). Moreover, we establish a natural trade-off between the maximum degree and other parameters of the graph (e.g., girth) required for consistent structure estimation.

Necessary conditions on structure learning provide lower bounds on the sample complexity for structure learning and have been studied in [38, 43, 50]. However, a standard assumption that these works make is that the underlying set of graphs is uniformly distributed with bounded degree. For this scenario, it is shown that $n = \Omega(\Delta^k \log p)$ samples are required for consistent structure estimation, for a graph with p nodes and maximum degree Δ, for some $k \in \mathbb{N}$, say $k = 3$ or 4. In contrast, our converse result is stated in terms of the *average degree*, instead of the maximum degree.

2. System model. In this section, we define the relevant notation to be used in the rest of the paper.

[7]For instance, it is known that for a Erdős–Rényi random graph $G_p \sim \mathcal{G}(p, c/p)$ when $(c > 1)$, the tree-width is greater than p^{ε}, for some $\varepsilon > 0$ [29].

2.1. *Notation.* We introduce some basic notions. Let $\|\cdot\|_1$ denote the ℓ_1 norm. For any two discrete distributions P, Q on the same alphabet $\mathcal{X}$, the total variation distance is given by

$$v(P, Q) := \frac{1}{2}\|P - Q\|_1 = \frac{1}{2}\sum_{x\in\mathcal{X}} |P(x) - Q(x)|, \tag{1}$$

and the Kullback–Leibler distance (or relative entropy) is given by

$$D(P\|Q) := \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)}.$$

Given a pair of discrete random variables (X, Y) taking values on the set $\mathcal{X}\times\mathcal{Y}$ and distributed as $P = P_{X,Y}$, the *mutual information* is defined as

$$I(X; Y) := D(P(x, y)\|P(x)P(y)) = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} P(x, y)\log\frac{P(x, y)}{P(x)P(y)}. \tag{2}$$

Along similar lines, the *conditional mutual information* of X and Y given another random variable Z, taking values on a countable set $\mathcal{Z}$, is defined as

$$I(X; Y|Z) := \sum_{x\in\mathcal{X}, y\in\mathcal{Y}, z\in\mathcal{Z}} P(x, y, z)\log\frac{P(x, y|z)}{P(x|z)P(y|z)}. \tag{3}$$

It is also well known that $I(X; Y|Z) = 0$ if and only if X and Y are independent given Z, that is, $P(x, y|z) = P(x|z)P(y|z)$.

Given n samples drawn i.i.d. from $P(x, y)$, denoted by $(x^n, y^n) = \{(x_i, y_i)\}_{i=1}^n$, the (joint) *empirical distribution* or the (joint) *type* is defined as

$$\widehat{P}^n(x, y; x^n, y^n) := \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{(x, y) = (x_i, y_i)\}. \tag{4}$$

We loosely use the term *empirical distance* to refer to distances between empirical distributions. For instance, the empirical variation distance is given by

$$v(\widehat{P}^n, \widehat{Q}^n) := \frac{1}{2}\sum_{x\in\mathcal{X}} |\widehat{P}^n(x) - \widehat{Q}^n(x)|. \tag{5}$$

Our algorithm for graph estimation will be based on empirical variation distance between conditional distributions. We employ such empirical estimates for testing conditional independencies between specific distributions.

2.2. *Ising models.* A *graphical model* is a family of multivariate distributions which are Markov in accordance to a particular undirected graph [31]. Each node in the graph $i \in V$ is associated to a random variable X_i, taking value in a set

$\mathcal{X}$. The set of edges[8] $E \subset \binom{V}{2}$ captures the set of conditional-independence relationships among the random variables. We say that a vector of random variables $\mathbf{X} := (X_1, \ldots, X_p)$ with a joint probability mass function (p.m.f.) P is Markov on the graph G if the *local Markov property*

$$P(x_i | x_{\mathcal{N}(i)}) = P(x_i | x_{V \setminus i}) \tag{6}$$

holds for all nodes $i \in V$. More generally, we say that P satisfies the *global Markov property*, if for all disjoint sets $A, B \subset V$ such that $A \cap \mathcal{N}(B) = \mathcal{N}(A) \cap B = \varnothing$, we have

$$P(\mathbf{x}_A, \mathbf{x}_B | \mathbf{x}_{\mathcal{S}(A,B;G)}) = P(\mathbf{x}_A | \mathbf{x}_{\mathcal{S}(A,B;G)}) P(\mathbf{x}_B | \mathbf{x}_{\mathcal{S}(A,B;G)}). \tag{7}$$

where the set $\mathcal{S}(A, B; G)$ is a *node separator*[9] between A and B, and $\mathcal{N}(A)$ denotes the neighborhood of A in G. The local and global Markov properties are equivalent under the *positivity* condition, given by $P(\mathbf{x}) > 0$, for all $\mathbf{x} \in \mathcal{X}^p$ [31].

The Hammersley–Clifford theorem [9] states that under the positivity condition, a distribution P satisfies the Markov property according to a graph G if and only if it factorizes according to the cliques of G, that is,

$$P(\mathbf{x}) = \frac{1}{Z} \exp\Big(\sum_{c \in \mathcal{C}} \Psi_c(\mathbf{x}_c) \Big), \tag{8}$$

where $\mathcal{C}$ is the set of cliques of G, and $\mathbf{x}_c$ is the set of random variables on clique c. The quantity Z is known as the *partition function* and serves to normalize the probability distribution. The functions Ψ_c are known as *potential* functions. An important class of graphical models is the class of pairwise models, which factorize according to the edges of the graph,

$$P(\mathbf{x}) = \frac{1}{Z} \exp\Big(\sum_{e \in E} \Psi_e(\mathbf{x}_e) \Big). \tag{9}$$

One of the most well-studied pairwise models is the Ising model. Here, each random variable X_i takes values in the set $\mathcal{X} = \{-1, +1\}$ and the probability mass function (p.m.f.) is given by

$$P(\mathbf{x}) = \frac{1}{Z} \exp\Big[\frac{1}{2} \mathbf{x}^T \mathbf{J}_G \mathbf{x} + \mathbf{h}^T \mathbf{x} \Big], \qquad \mathbf{x} \in \{-1, 1\}^p, \tag{10}$$

where $\mathbf{J}_G$ is known as the potential matrix, and $\mathbf{h}$ as the potential vector. By convention, $J(i, i) = 0$ for all $i \in V$. The sparsity pattern of $\mathbf{J}_G$ corresponds to that of the graph G, that is, $J_{i,j} = 0$ for $(i, j) \notin G$. A model is said to be *attractive* or

[8] We use the notation E and G interchangeably to denote the set of edges.

[9] A set $\mathcal{S}(A, B; G) \subset V$ is a separator of sets A and B if the removal of nodes in $\mathcal{S}(A, B; G)$ separates A and B into distinct components.

ferromagnetic if $J_{i,j} \geq 0$ and $h_i \geq 0$, for all $i, j \in V$. An Ising model is said to be *symmetric* if $\mathbf{h} = \mathbf{0}$.

We assume that there exists $J_{\min}, J_{\max} \in \mathbb{R}$ such that the absolute values of the edge potentials are uniformly bounded, that is,

$$|J_{i,j}| \in [J_{\min}, J_{\max}] \qquad \forall (i, j) \in G. \tag{11}$$

We can provide guarantees on structure recovery, subject to conditions on $J_{\min}$ and $J_{\max}$. We assume that the node potentials h_i are uniformly bounded away from $\pm\infty$.

Given an Ising model, nodes $i, j \in V$ and a subset $S \subset V \setminus \{i, j\}$, we define *conditional variation distance* as

$$\nu_{i|j;S} := \min_{\mathbf{x}_S \in \{\pm 1\}^{|S|}} \nu\big(P(X_i|X_j = +, \mathbf{X}_S = \mathbf{x}_S), P(X_i|X_j = -, \mathbf{X}_S = \mathbf{x}_S)\big) \tag{12}$$

$$= \min_{\mathbf{x}_S \in \{\pm 1\}^{|S|}} \frac{1}{2} \sum_{x_i = \pm 1} |P(X_i = x_i|X_j = +, \mathbf{X}_S = \mathbf{x}_S) - P(X_i = x_i|X_j = -, \mathbf{X}_S = \mathbf{x}_S)|. \tag{13}$$

The empirical conditional variation distance $\widehat{\nu}_{i|j;S}$ is defined by replacing the actual distributions with their empirical versions

$$\widehat{\nu}^n_{i,j;S} := \min_{\mathbf{x}_S \in \{\pm 1\}^{|S|}} \nu\big(\widehat{P}^n(X_i|X_j = +, \mathbf{X}_S = \mathbf{x}_S), \widehat{P}^n(X_i|X_j = -, \mathbf{X}_S = \mathbf{x}_S)\big). \tag{14}$$

Our algorithm will be based on empirical conditional variation distances. This is because the conditional variation distances[10] can be used as a test for conditional independence

$$\{X_i \perp\!\!\!\perp X_j|\mathbf{X}_S\} \equiv \{\nu_{i|j;S} = 0\} \qquad \forall i, j \in V, S \subset V \setminus \{i, j\}. \tag{15}$$

2.3. *Tractable graph families.* We consider the class of Ising models Markov on a graph G_p belonging to some ensemble $\mathcal{G}(p)$ of graphs with p nodes. We consider the high-dimensional regime, where both p and the number of samples n grow simultaneously; typically, the growth of p is much faster than that of n. We emphasize that in our formulation, the graph ensemble $\mathcal{G}(p)$ can either be deterministic or random—in the latter, we also specify a probability measure over the set of graphs in $\mathcal{G}(p)$. In the setting where $\mathcal{G}(p)$ is a random-graph ensemble, let $P_{\mathbf{X},G}$ denote the joint probability distribution of the variables $\mathbf{X}$ and the graph $G \sim \mathcal{G}(p)$, and let $P_{\mathbf{X}|G}$ denote the conditional distribution of the variables given a graph G. Let P_G denote the probability distribution of graph G drawn from a random ensemble $\mathcal{G}(p)$. In this setting, we use the term *almost every* (a.e.) graph G satisfies a certain property $\mathcal{Q}$ if

$$\lim_{p \to \infty} P_G[G \text{ satisfies } \mathcal{Q}] = 1.$$

[10]Note that the conditional variation distances are in general asymmetric, that is, $\nu_{i|j;S} \neq \nu_{j|i;S}$.

In other words, the property $\mathcal{Q}$ holds asymptotically almost surely[11] (a.a.s.) with respect to the random-graph ensemble $\mathcal{G}(p)$. Our conditions and theoretical guarantees will be based on this notion for random graph ensembles. Intuitively, this means that graphs that have a vanishing probability of occurrence as $p \to \infty$ are ignored.

We now characterize the ensemble of graphs amenable for consistent structure estimation under our formulation. To this end, we characterize the so-called *local separators* in graphs. See Figure 1 for an illustration. For $\gamma \in \mathbb{N}$, let $B_\gamma(i; G)$ denote the set of vertices within distance γ from i with respect to graph G. Let $F_{\gamma,i} := G(B_\gamma(i))$ denote the subgraph of G spanned by $B_\gamma(i; G)$, but in addition, we retain the nodes not in $B_\gamma(i)$ (and remove the corresponding edges).

DEFINITION 1 (γ-Local separator). Given a graph G, a *γ-local separator* $S_\gamma(i, j)$ between i and j, for $(i, j) \notin G$, is a *minimal* vertex separator[12] with respect to the subgraph $F_{\gamma,i}$. In addition, the parameter γ is referred to as the *path threshold* for local separation.

In other words, the γ-local separator $S_\gamma(i, j)$ separates nodes i and j with respect to paths in G of length at most γ. We now characterize the ensemble of graphs based on the size of local separators.

DEFINITION 2 ((η, γ)-Local separation property). An ensemble of graphs $\mathcal{G}(p; \eta, \gamma)$ satisfies (η, γ)-local separation property if for a.e. $G_p \in \mathcal{G}(p; \eta, \gamma)$,

$$\max_{(i,j)\notin G_p} |S_\gamma(i, j)| \leq \eta. \tag{16}$$

In Section 3, we propose an efficient algorithm for graphical model selection when the underlying graph belongs to a graph ensemble $\mathcal{G}(p; \eta, \gamma)$ with sparse local separators [i.e., small η, for η defined in (16)]. We will see that the computational complexity of our proposed algorithm scales as $O(p^{\eta+2})$. In Section 3.3, we provide examples of many graph families satisfying (16), which include the random regular graphs, Erdős–Rényi random graphs and small-world graphs.

REMARK. The criterion of local separation for tractable learning is novel to the best of our knowledge. The complexity of a graphical model is usually expressed in terms of its *tree-width* [49]. We note that the criterion of sparse local separation is weaker than the tree-width; that is, $\eta \leq t$, where t is the tree-width of the graph. In fact, our criterion is also weaker than the criterion of bounded local tree-width, introduced in [22].

[11] Note that the term a.a.s. does not apply to deterministic graph ensembles $\mathcal{G}(p)$ where no randomness is assumed, and in this setting, we assume that the property $\mathcal{Q}$ holds for every graph in the ensemble.

[12] A minimal separator is a separator of smallest cardinality.

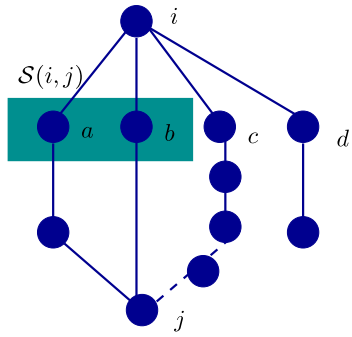


FIG. 1. *Illustration of l-local separator set $\mathcal{S}(i,j;G,l)$ for the graph shown above with $l=4$. Note that $\mathcal{N}(i)=\{a,b,c,d\}$ is the neighborhood of i and the l-local separator set $\mathcal{S}(i,j;G,l)=\{a,b\}\subset\mathcal{N}(i;G)$. This is because the path along c connecting i and j has a length greater than l and hence node $c\notin\mathcal{S}(i,j;G,l)$.*

3. Method and guarantees.

3.1. *Assumptions.*

(A1) *Sample complexity*: We consider the asymptotic setting where both the number of variables (nodes) p and the number of i.i.d. samples n go to infinity. The required sample complexity is

$$n=\Omega(J_{\min}^{-2}\log p). \tag{17}$$

We require that the number of nodes $p\to\infty$ to exploit the local-separation properties of the class of graphs under consideration.

(A2) *Bounded edge potentials*: The Ising model Markov on a.e. $G_p\sim\mathcal{G}(p)$ has the maximum absolute potential below a threshold J^*. More precisely,

$$\alpha:=\frac{\tanh J_{\max}}{\tanh J^*}<1, \tag{18}$$

where the threshold J^* depends on the specific graph ensemble $\mathcal{G}(p)$. See Section 8.1 in the supplementary material [4] for an explicit characterization of J^* for specific ensembles.

(A3) *Local-separation property*: We consider the ensemble of graphs $\mathcal{G}(p)$ such that almost every graph G drawn from $\mathcal{G}(p)$ satisfies the local-separation property (η,γ), according to Definition 2, for some $\eta=O(1)$ and $\gamma\in\mathbb{N}$ such that[13]

$$J_{\min}\alpha^{-\gamma}=\widetilde{\omega}(1), \tag{19}$$

where we say that a function $f(p)=\widetilde{\omega}(g(p))$, if $\frac{f(p)}{g(p)\log p}\to\infty$ as $p\to\infty$.

[13]The condition in (19) involving $\widetilde{\omega}(1)$ is required for random graph ensembles such as Erdős–Rényi random graphs. It can be weakened as $J_{\min}\alpha^{-\gamma}=\omega(1)$ for degree-bounded ensembles $\mathcal{G}_{\mathrm{Deg}}(\Delta)$.

(A4) *Generic edge-potentials*: The edge potentials $\{J_{i,j}, (i,j) \in G\}$ of the Ising model are assumed to be generically drawn from $[-J_{\max}, -J_{\min}] \cup [J_{\min}, J_{\max}]$; that is, our results hold except for a set of Lebesgue measure zero. We also characterize specific classes of models where this assumption can be removed, and we allow for any choice of edge potentials. See Section 8.3 in the supplementary material [4] for details.

Assumption (A1) provides on the bound on the sample complexity. Assumption (A2) limits the maximum edge potential $J_{\max}$ of the model. Assumption (A3) relates the path threshold γ with the minimum edge potential $J_{\min}$ in the model. For instance, if $J_{\min} = \Theta(1)$ and $\gamma = O(\log\log p)$, we require that $\alpha := \frac{\tanh J_{\max}}{\tanh J^*} = 1 - \Theta(1) < 1$.

Condition (A4) guarantees the success of our method for generic edge potentials. Note that if the neighbors are marginally independent, then our method fails, and thus, we cannot expect our method to succeed for all edge potentials. Condition (A4) can be removed if we limit to attractive models (see Section 8.3.1 in the supplementary material [4]), or if we allow for nonattractive models, but limit to graphs with bounded local paths (see Section 8.3.3 in the supplementary material [4]). For general models, we guarantee success of our methods for generic potentials; that is, we establish that the set of edge potentials where our method fails has Lebesgue measure zero. Similar assumptions have been previously employed; for example, in [26] where learning directed models is considered, it is assumed that the graphical model is faithful with respect to the underlying graph.

3.2. *Conditional variation distance thresholding*. We now propose an algorithm, termed as conditional variation distance thresholding (CVDT) which is proven to be consistent for graph reconstruction under the above assumptions. The procedure for CVDT is provided in Algorithm 1. Denote $\mathsf{CVDT}(\mathbf{x}^n; \xi_{n,p})$ as the output edge set from CVDT given n i.i.d. samples $\mathbf{x}^n$ and threshold $\xi_{n,p}$. The conditional variation distance test in the CVDT algorithm computes the empirical conditional variation distance in (14) for each node pair $(i,j) \in V^2$ and finds the conditioning set which achieves the minimum over all sets of cardinality η. If the minimum exceeds the threshold $\xi_{n,p}$, the node pair is declared an edge.

The threshold $\xi_{n,p}$ needs to separate the edges and the nonedges in the Ising model. It is chosen as a function of both number of nodes p and number of samples n and needs to satisfy the following conditions:

$$\xi_{n,p} = O(J_{\min}), \qquad \xi_{n,p} = \widetilde{\omega}(\alpha^{\gamma}), \qquad \xi_{n,p} = \Omega\left(\sqrt{\frac{\log p}{n}}\right). \tag{20}$$

For example, when $J_{\min} = \Omega(1)$, $\alpha < 1$, $\gamma = \Omega(\log p)$, $n = \Omega(g_p \log p)$, for some sequence $g_p = \omega(1)$, we can choose $\xi_{n,p} = \frac{1}{\min(g_p, \log p)}$.

Algorithm 1 Algorithm $\mathsf{CVDT}(\mathbf{x}^n; \xi_{n,p}, \eta)$ for structure learning from $\mathbf{x}^n$ samples based on empirical conditional variation distances. See (14).

Initialize $\widehat{G}_p^n = (V, \varnothing)$.
For each $i, j \in V$, if

$$\min_{\substack{S \subset V \setminus \{i,j\} \\ |S| \leq \eta}} \widehat{\mathcal{V}}_{i|j;S} > \xi_{n,p}, \tag{21}$$

then add (i, j) to $\widehat{G}_p^n$.
Output: $\widehat{G}_p^n$.

Note that there is dependence on both n and p, since we need to regularize for sample size, as well as for the size of the graph. In other words, with finite number of samples n, the empirical conditional variation distances are noisy, and the threshold $\xi_{n,p}$ takes this into account via its inverse dependence on n. Similarly, as the graph size p increases, we establish that the true conditional variation distance decays at a certain rate under assumption (A2). Hence the threshold $\xi_{n,p}$ also depends on the graph size p. Moreover, note that for all the conditions in (20) to be satisfied, the number of samples n should scale at least at a certain rate with respect to p, as given by (17).

3.2.1. *Structural consistency of* CVDT. Assuming (A1)–(A4), we have the following result on asymptotic graph structure recovery.

THEOREM 1 (Structural consistency of CVDT). *The algorithm* CVDT *is consistent for structure recovery of Ising models Markov on a.e. graph* $G_p \sim \mathcal{G}(p; \eta, \gamma)$:

$$\lim_{\substack{n,p \to \infty \\ n = \Omega(J_{\min}^{-2} \log p)}} P[\mathsf{CVDT}(\{\mathbf{x}^n\}; \xi_{n,p}, \eta) \neq G_p] = 0. \tag{22}$$

The proof of this theorem is provided in Section 8 in the supplementary material [4].

REMARKS.

(1) *Consistency guarantee*: The CVDT algorithm consistently recovers the structure of the graphical models, with probability tending to one, where the probability measure is with respect to both the graph and the samples. We extend our results and provide finite sample guarantees for specific graph families in Section 3.2.2. Moreover, if we require a *parameter-free* threshold, that is, we do not know the exact value of $J_{\min}$ but only its scaling with p, then we need to choose

$\xi_{n,p} = o(J_{\min})$ rather than $\xi_{n,p} = O(J_{\min})$. In this case, the sample complexity scales as $n = \omega(J_{\min}^{-2} \log p)$.

(2) *Other tests for conditional independence*: We consider a test based on variation distances. Alternatively other distance measures can be employed. For instance, it can be proven that the Hellinger distance and the Kullback–Leibler distance have similar sample complexity results, while a test based on mutual information has a worse sample complexity of $\Omega(J_{\min}^{-4} \log p)$ under the assumptions (A1)–(A4). We term the test based on mutual information as CMIT and compare its experimental performance with CVDT in Section 5.

(3) *Extension to other models*: The CVDT algorithm can be extended to general discrete models by considering pairwise variation distance between different configurations. For instance, we can set

$$\nu_{i|j;S} := \sum_{\substack{\lambda_1 \neq \lambda_2 \\ \lambda_1, \lambda_2 \in \mathcal{X}}} \min_{\mathbf{x}_S \in \mathcal{X}^{|S|}} \nu\big(P(X_i|X_j = \lambda_1, \mathbf{X}_S = \mathbf{x}_S), P(X_i|X_j = \lambda_2, \mathbf{X}_S = \mathbf{x}_S)\big). \tag{23}$$

In [3], we derive analogous conditions for Gaussian graphical models. Our approach is also applicable to models with higher order potentials since it does not depend on the pairwise nature of Ising models. The conditions for recovery are based on the notion of *conditional uniqueness* and can be imposed on any model. Indeed the regime of parameters where conditional uniqueness holds depends on the model and is harder to characterize for more complex models. Notice that our algorithm requires only low-order statistics [up to $O(\eta+2)$] for any class of graphical models which is relevant when we are dealing with models with higher order potentials.

PROOF OUTLINE. We first analyze the scenario when exact statistics are available. (i) We establish that for any two nonneighbors $(i, j) \notin G$, the conditional variation distance in (21) (based on exact statistics) does not exceed the threshold $\xi_{n,p}$. (ii) Similarly, we also establish that the conditional variation distance in (21) exceeds the threshold $\xi_{n,p}$ for all neighbors $(i, j) \in G$. (iii) We then extend these results to empirical versions using concentration bounds. □

3.2.2. *PAC Guarantees for* CVDT. We now provide stronger results for CVDT method in terms of the probably approximately correct (PAC) model of learning [28]. This provides additional insight into the task of graph estimation. Given an Ising model P on graph G_p, recall the definition of conditional variation distance

$$\nu_{i|j;S} := \min_{\mathbf{x}_S \in \{-1,+1\}^{|S|}} \nu\big(P(X_i|X_j = +, \mathbf{X}_S = \mathbf{x}_S), P(X_i|X_j = -, \mathbf{X}_S = \mathbf{x}_S)\big).$$

Given a graph G_p and $\lambda, \eta > 0$, define

$$G'_p(V;\lambda) := \Big\{(i,j) \in G_p : \min_{\substack{|S|\le\eta \\ S\subset V\setminus\{i,j\}}} \nu_{i|j;S} > \lambda\Big\}, \tag{24}$$

$$\nu_{\max}(p;\eta) := \max_{(i,j)\notin G_p} \min_{\substack{|S|\le\eta \\ S\subset V\setminus\{i,j\}}} \nu_{i|j;S}. \tag{25}$$

For any $\delta > 0$, choose the threshold $\xi_{n,p}$ as

$$\xi_{n,p}(\delta) = \nu_{\max}(p;\eta) + \delta. \tag{26}$$

Define

$$P_{\min} := \min_{\substack{S\subset V, |S|\le\eta+1 \\ \mathbf{x}=\{\pm1\}^{|S|}}} P(\mathbf{X}_S = \mathbf{x}_S). \tag{27}$$

THEOREM 2 (PAC guarantees for CVDT). *Given an Ising model Markov on graph G and threshold $\xi_{n,p}(\delta)$ according to* (26), $\mathsf{CVDT}(\{\mathbf{x}^n\}; \xi_{n,p}(\delta), \eta)$ *recovers* $G'_p(V; \nu_{\max}(p;\eta) + 2\delta)$ *for any* $\delta > 0$, *defined in* (24), *with probability at least* $1-\varepsilon$, *when the number of samples is*

$$n > \frac{2(\delta+2)^2}{\delta^2 P_{\min}^2}\left[\log\left(\frac{1}{\varepsilon}\right) + (\eta+2)\log p + (\eta+4)\log 2\right], \tag{28}$$

and the computational complexity scales as $O(p^{\eta+2})$.

PROOF. The proof is provided in Section 9 in the supplementary material [4]. □

Thus, the above result characterizes the relationship between the separation between edges and nonedges (in terms of conditional variation distances) and the number of samples required to distinguish them. A critical parameter in the above result is $\nu_{\max}(p;\eta)$, the maximum conditional variation distance between nonneighbors. We now provide nonasymptotic bounds on $\nu_{\max}(p;\eta)$ for specific graph families satisfying the (η,γ)-local separation condition. A detailed description of the graph families considered below is provided in Section 3.3. On lines of assumption (A2) in Section 3.1, define

$$\alpha := \frac{\tanh J_{\max}}{\tanh J^*}. \tag{29}$$

As we noted earlier, the threshold J^* depends on the graph family. We characterize both J^* and $\nu_{\max}(p;\eta)$ for various graph families below.

LEMMA 1 [Nonasymptotic bounds on $\nu_{\max}(p;\eta)$ for graph families]. *The following statements hold for* α *in* (29):

(1) *For the degree-bounded ensemble* $\mathcal{G}_{\mathrm{Deg}}(p;\Delta)$,

$$J^*_{\mathrm{Deg}} = \infty, \qquad \nu_{\max}(p;\Delta) = 0. \tag{30}$$

(2) *For the girth-bounded ensemble* $\mathcal{G}_{\mathrm{Girth}}(p;g,\Delta)$,

$$J^*_{\mathrm{Girth}} = \operatorname{atanh}\left(\frac{1}{\Delta}\right), \qquad \nu_{\max}(p;1) \le \alpha^{g/2}, \tag{31}$$

where Δ *is the maximum degree and* g *is the girth.*

(3) *For the ensemble of* Δ*-random regular graphs* $\mathcal{G}_{\mathrm{Reg}}(p;\Delta)$,

$$J^*_{\mathrm{Reg}} = \operatorname{atanh}\left(\frac{1}{\Delta}\right). \tag{32}$$

Choose any $l \in \mathbb{N}$ *such that* $l < 0.25(0.25p\Delta + 0.5 - \Delta^2)$. *Then, with probability at least* $1 - \Delta^{16l-2}(p\Delta - 4\Delta^2 - 16l)^{-(8l-1)}$,

$$\nu_{\max}(p;2) \le \alpha^l, \tag{33}$$

where Δ *is the degree.*

(4) *For the Erdős–Rényi ensemble* $\mathcal{G}_{\mathrm{ER}}(p, c/p)$,

$$J^*_{\mathrm{ER}} = \operatorname{atanh}\left(\frac{1}{c}\right). \tag{34}$$

Choose any $l \in \mathbb{N}$ *such that* $l < \frac{\log p}{4\log c}$. *When* $c > 1$, *then with probability at least* $1 - le^{\sqrt{125}}p^{-2.5} - l!c^{4l+1}p^{-1}$,

$$\nu_{\max}(p;2) \le 2l^3\alpha^l \log p, \tag{35}$$

where c *is the average degree.*

(5) *For the small-world graph ensemble* $\mathcal{G}_{\mathrm{Watts}}(p,d,c/p)$, *similar results apply.*

$$J^*_{\mathrm{Watts}} = \operatorname{atanh}\left(\frac{1}{c}\right), \tag{36}$$

Choose any $l \in \mathbb{N}$ *such that* $l < \frac{\log p}{4\log c}$. *When* $c > 1$, *with probability at least* $1 - le^{\sqrt{125}}p^{-2.5} - l!c^{4l-1}p^{-1}$,

$$\nu_{\max}(p;d+2) \le 4l^3\alpha^l \log p, \tag{37}$$

where c *is the average degree of the Erdős–Rényi subgraph.*

PROOF. See Corollaries 1 and 2 in Section 8.1 in the supplementary material [4]. □

Thus, we note that the conditional variation distance is small for nonneighbors when the maximum edge potential $J_{\max}$ is suitably bounded. Combining the results above on $\nu_{\max}(p;\eta)$ and the PAC guarantees in Theorem 2, we note that a majority of edges in the Ising model can be learned efficiently under a logarithmic sample complexity.

3.3. *Examples of tractable graph families.* We now show that the local-separation property in Definition 2 and the assumptions in Section 3.1 hold for a rich class of graphs.

EXAMPLE 1 (Bounded-degree). Any (deterministic or random) ensemble of degree-bounded graphs $\mathcal{G}_{\mathrm{Deg}}(p, \Delta)$ satisfies (η, γ)-local separation property with $\eta = \Delta$ and arbitrary $\gamma \in \mathbb{N}$. This is because for any node $i \in V$, its neighborhood $\mathcal{N}(i)$ exactly separates it from nonneighbors. Since there is exact separation, we can establish that the threshold in (18) is infinite ($J^*_{\mathrm{Deg}} = \infty$); that is, there is no constraint on the maximum edge potential $J_{\max}$. However, the computational complexity of our proposed algorithm scales as $O(p^{\Delta+2})$; see also [10]. Thus, when Δ is large, our proposed algorithm, as well as the algorithm in [10], are computationally intensive. Our goal in this paper is to relax the bounded-degree assumption and to consider sequences of ensembles of graph $\mathcal{G}(p)$ whose maximum degrees may grow with the number of nodes p. To this end, we discuss other structural constraints which can lead to graphs with sparse local separators.

EXAMPLE 2 (Bounded local paths). Another sufficient condition[14] for the (η, γ)-local separation property in Definition 2 to hold is that there are at most η paths of length at most γ in G between any two nodes [henceforth, termed as the (η, γ)-*local paths property*]. In other words, there are at most $\eta - 1$ number of overlapping[15] cycles of length smaller than 2γ. We denote this ensemble of graphs as $\mathcal{G}_{\mathrm{LP}}(p; \eta, \gamma)$.

In particular, a special case of the local-paths property described above is the so-called girth property. The *girth* of a graph is the length of the shortest cycle. Thus, a graph with girth g satisfies (η, γ)-local separation property with $\eta = 1$ and $\gamma = g/2$. Let $\mathcal{G}_{\mathrm{Girth}}(p; g)$ denote the ensemble of graphs with girth at most g. There are many graph constructions which lead to large girth. For example, the bipartite Ramanujan graph [17], page 107 and the random Cayley graphs [24] have large girths. Recently, efficient algorithms have been proposed to generate large girth graphs efficiently [5].

The girth condition can be weakened to allow for a small number of short cycles, while not allowing for typical node neighborhoods to contain short cycles. Such graphs are termed as *locally tree-like*. For instance, the ensemble of Erdős–Rényi graphs $\mathcal{G}_{\mathrm{ER}}(p, c/p)$, where an edge between any node pair appears with a

[14]For any graph satisfying (η, γ)-local separation property, the number of vertex-disjoint paths of length at most γ between any two nonneighbors is bounded above by η, by appealing to Menger's theorem for bounded path lengths [35]. However, the property of local paths that we describe above is a stronger notion than having sparse local separators, and we consider all distinct paths of length at most γ and not just vertex disjoint paths in the formulation.

[15]Two cycles are said to overlap if they have common vertices.

probability c/p, independent of other node pairs, is locally tree-like. The parameter c may grow with p, albeit at a controlled rate for tractable structure learning, made precise later. In Section 11 in the supplementary material [4], we establish that there are at most two paths of length smaller than $\gamma < \frac{\log p}{4 \log c}$ between any two nodes in Erdős–Rényi graphs a.a.s., or equivalently, there are no overlapping cycles of length smaller than 2γ a.a.s. Similar observations apply for the more general *scale-free* or *power-law* graphs [18, 20], and we derive the precise relationships in Section 11 in the supplementary material [4]. Along similar lines, the ensemble of Δ-random regular graphs, denoted by $\mathcal{G}_{\mathrm{Reg}}(p, \Delta)$, which is the uniform ensemble of regular graphs with degree Δ has no overlapping cycles of length at most $\Theta(\log_{\Delta-1} p)$ a.a.s. [36], Lemma 1.

We now discuss the conditions under which a general local-paths graph ensemble $\mathcal{G}_{\mathrm{LP}}(p; \eta, \gamma)$ satisfies assumption[16] (A3) in Section 3.1, required for our graph estimation algorithm CVDT to succeed. Denote the maximum degree for the $\mathcal{G}_{\mathrm{LP}}(p; \eta, \gamma)$ ensemble as Δ (possibly growing with p). Note that we can now implement the CVDT algorithm with parameter η. In Section 8.1 in the supplementary material [4], we establish that the threshold J^* in (18) is given by $J^*_{\mathrm{LP}} = \Theta(1/\Delta)$. When the minimum edge potential $J_{\min}$ achieves the bound, that is, $J_{\min} = \Theta(1/\Delta)$, the assumption (A3) simplifies as

$$\Delta \alpha^{\gamma} = o(1). \tag{38}$$

Note that $\alpha < 1$ under (A2). We obtain a natural trade-off between the maximum degree Δ and the path threshold γ.

When $\Delta = O(1)$, we can allow the path threshold in (38) to scale as $\gamma = O(\log \log p)$. This implies that graphs with fairly small path threshold γ can be incorporated under our framework. In particular, this includes the class of girth-bounded graph with fairly small girth [i.e., the girth g scaling as $O(\log \log p)$].

We can also incorporate graph families with growing maximum degrees in (38). For instance, when $\Delta = O(\mathrm{poly} \log p)$, we require the path threshold to scale as $\gamma = O(\log p)$. In particular, the Δ-random-regular ensemble satisfies (38) when $\Delta = O(\mathrm{poly} \log p)$.

Thus, (38) represents a natural trade-off between node degrees and path threshold for consistent structure estimation; graphs with large degrees can be learned efficiently if their path thresholds are large. Indeed, in the extreme case of trees which have infinite threshold (since they have infinite girth), in accordance with (38), there is no constraint on node degrees for successful recovery, and recall that the Chow–Liu algorithm [16] is an efficient method for model selection on tree distributions.

Moreover, the constraint in (38) can be weakened for random graph ensembles by replacing the maximum degree with the average degree. Recall that in

[16] In fact, a weaker version of (A3) as $J_{\min} \alpha^{-\gamma} = \omega(1)$ suffices for degree-bounded ensembles $\mathcal{G}_{\mathrm{Deg}}(\Delta)$.

the Erdős–Rényi ensemble $\mathcal{G}_{\mathrm{ER}}(p, c/p)$, an edge between any two nodes occurs with probability c/p and that this ensemble satisfies the (η, γ) property with path threshold $\gamma = O(\frac{\log p}{\log c})$ and $\eta = 2$. In Section 8.1 in the supplementary material [4], we establish that the threshold in (18) is given by $J^*_{\mathrm{ER}} = \Theta(1/c)$. Comparing with the threshold for Δ-degree bounded graphs $J^* = \Theta(1/\Delta)$ discussed above, we see that we can obtain better bounds for random-graph ensembles.

When the minimum edge potentials achieves the threshold ($J_{\min} = \Theta(1/c)$), the requirement in assumption (A3) in Section 3.1 simplifies to

$$c\alpha^{\gamma} = \widetilde{o}(1), \tag{39}$$

which is true when $c = O(\mathrm{poly}\log p)$. Thus, we can guarantee consistent structure estimation for the Erdős–Rényi ensemble when the average degree scales as $c = O(\mathrm{poly}\log p)$. This regime is typically known as the "sparse" regime and is relevant, since in practice, our goal is to fit the measurements to a sparse graphical model.

EXAMPLE 3 (Small-world graphs). The previous two examples showed that local separation holds under two different conditions: bounded maximum degree and bounded number of local paths. The former class of graphs can have short cycles, but the maximum degree needs to be constant, while the latter class of graphs can have a large maximum degree but the number of overlapping short cycles needs to be small. We now provide instances which incorporate both these features, large degrees and short cycles, and yet satisfy the local separation property.

The class of hybrid graphs or augmented graphs ([18], Chapter 12) consists of graphs which are the union of two graphs: a "local" graph, having short cycles, and a "global" graph, having small average distances. Since the hybrid graph is the union of these local and global graphs, it simultaneously has large degrees and short cycles. The simplest model $\mathcal{G}_{\mathrm{Watts}}(p, d, c/p)$, first studied by Watts and Strogatz [51], consists of the union of a d-dimensional grid and an Erdős–Rényi random graph with parameter c. It is easily seen that a.e. graph $G \sim \mathcal{G}_{\mathrm{Watts}}(p, d, c/p)$ satisfies (η, γ)-local separation property in (16), with

$$\eta = d + 2, \qquad \gamma \leq \frac{\log p}{4 \log c}.$$

Similar observations apply for more general hybrid graphs studied in [18], Chapter 12.

In Section 8.1 in the supplementary material, we establish that the threshold in (18) for the small-world ensemble $\mathcal{G}_{\mathrm{Watts}}(p, d, c/p)$ is given by $J^*_{\mathrm{Watts}} = \Theta(1/c)$ and is independent of d, the degree of the grid graph. Comparing with the threshold J^*_{ER} for Erdős–Rényi ensemble $\mathcal{G}_{\mathrm{ER}}(p, c/p)$, we note that the two thresholds are identical. This further implies that (39) holds for the small-world graph ensemble as well.

3.4. *Explicit bounds on sample complexity of* CVDT. Recall that the sample complexity of the CVDT is required to scale as $n = \Omega(J_{\min}^{-2} \log p)$ for structural consistency in high dimensions. Thus, the sample complexity is small when the minimum edge potential $J_{\min}$ is large. On the other hand, $J_{\min}$ cannot be arbitrarily large due to assumption (A2) in Section 3.1, which entails that $J_{\min} < J^*$. The minimum sample complexity is thus attained when $J_{\min}$ achieves the threshold J^*.

We now provide explicit results for the minimum sample complexity for various graph ensembles, based on the threshold J^*. Recall that in Section 3.3, we discussed that for the graph ensemble $\mathcal{G}_{\mathrm{LP}}(p, \eta, \gamma, \Delta)$ satisfying the (η, γ)-local paths property and having maximum degree Δ, the threshold is $J^*_{\mathrm{LP}} = 1/\Delta$. Thus, the minimum sample complexity for this graph ensemble is $n = \Omega(\Delta^2 \log p)$, that is, when $J_{\min} = \Theta(1/\Delta)$.

For the Erdős–Rényi random graph ensemble $\mathcal{G}_{\mathrm{ER}}(p, c/p)$ and the small-world graph ensemble $\mathcal{G}_{\mathrm{Watts}}(p, d, c/p)$, recall that the thresholds are given by $J^*_{\mathrm{ER}} = J^*_{\mathrm{Watts}} = 1/c$, where c is the mean degree of the Erdős–Rényi graph. Thus, the minimum sample complexity can be improved to $n = \Omega(c^2 \log p)$, by setting $J_{\min} = \Theta(1/c)$. This implies that when the Erdős–Rényi random graphs and small-world graphs have a bounded average degree [$c = O(1)$], the minimum sample complexity is $n = \Omega(\log p)$. Recall that the sample complexity of learning tree models is $\Omega(\log p)$ [47]. Thus, we observe that the complexity of learning sparse Erdős–Rényi random graphs and small-world graphs using our algorithm CVDT is akin to learning tree structures in certain parameter regimes.

3.5. *Comparison with previous results*. We now compare the performance of our algorithm CVDT with ℓ_1-penalized logistic regression proposed in [41]. We first compare the computational complexities. The method in [41] has a computational complexity of $O(p^4)$ for any input (assuming $p > n$). On the other hand, the complexity of our method depends on the graph family under consideration. It can be as low as $O(p^3)$ for girth-bounded ensembles, $O(p^4)$ for random graph families and as high as $O(p^\Delta)$ for degree-bounded ensembles (without any additional characterization of the local separation property). Clearly our method is not efficient for general degree-bounded ensembles since it is tailored to exploit the sparse local-separation property in the underlying graph.

We now compare the sample complexities under the two methods. It was established that the method in [41] has a minimum sample complexity of $n = \Omega(\Delta^3 \log p)$ for a degree-bounded ensemble $\mathcal{G}_{\mathrm{Deg}}(p, \Delta)$ satisfying certain "incoherence" conditions. The sample complexity of our CVDT algorithm is better at $n = \Omega(\Delta^2 \log p)$. Moreover, we can guarantee improved sample complexity of $n = \Omega(c^2 \log p)$ for Erdős–Rényi random graphs $\mathcal{G}_{\mathrm{ER}}(p, c/p)$ and small-world graphs $\mathcal{G}_{\mathrm{Watts}}(p, d, c/p)$ under the modified CVDT algorithm. Note that these random graph ensembles have maximum degrees (Δ) much larger than the average degrees (c), and thus, we can provide stronger sample complexity results. Moreover, our algorithm is local and requires only low-order statistics for any class of

graphical models of arbitrary order, while the method in [41] requires full-order statistics since it undertakes neighborhood selection through regularized logistic regression. This is relevant in practice, since our algorithm is better equipped to handle missing samples.

The incoherence conditions required for the success of ℓ_1 penalized logistic regression in [41] are NP-hard to establish for general models since they involve the partition function of the model [6]. In contrast, our conditions are transparent and relate to the phase transitions in the model. It is an open question as to whether the incoherence conditions are implied by our assumptions or vice-versa for general models. It appears that our conditions are weaker than the incoherence conditions for random-graph models. For instance, for the Erdős–Rényi model $\mathcal{G}_{\mathrm{ER}}(p, c/p)$, we require that $J_{\max} = O(1/c)$, where c is the average degree, while a sufficient condition for incoherence is $J_{\max} = O(1/\Delta)$, where Δ is the maximum degree. Note that $\Delta = O(\log p \log c)$ a.a.s. for the Erdős–Rényi model. Similar observations also hold for the power-law and small-world graph ensembles. This implies that we can guarantee consistent structure estimation under weaker conditions (i.e., a wider range of parameters) and better sample complexity for the Erdős–Rényi, power-law and small-world models.

4. Necessary conditions for graph estimation. We have so far proposed algorithms and provided performance guarantees for graph estimation given samples from an Ising models. We now analyze necessary conditions for graph estimation.

4.1. *Erdős–Rényi random graphs*. Necessary conditions for graph estimation have been previously characterized for degree-bounded graph ensembles $\mathcal{G}_{\mathrm{Deg}}(p, \Delta)$ [43]. However, these conditions are too loose to be useful for the ensemble of Erdős–Rényi graphs $\mathcal{G}_{\mathrm{ER}}(p, c/p)$, where the average degree[17] (c) is much smaller than the maximum degree.

We now provide a lower bound on sample complexity for graph estimation of Erdős–Rényi graphs using any deterministic estimator. Recall that p is the number of nodes in the model, and n is the number of samples. In the following result, c is allowed to depend on p and is thus more general than the previous results.

THEOREM 3 (Necessary conditions for model selection). *Assume that* $c \leq 0.5p$ *and* $G_p \sim \mathcal{G}_{\mathrm{ER}}(p, c/p)$. *Then if* $n \leq \varepsilon c \log p$ *for sufficiently small* $\varepsilon > 0$, *we have*

$$\lim_{p\to\infty} P[\widehat{G}_p^n(\mathbf{X}_p^n) \neq G_p] = 1 \tag{40}$$

for any deterministic estimator $\widehat{G}_p$.

[17] The techniques in this section is applicable when the average sparsity parameter c of $\mathcal{G}_{\mathrm{ER}}(p, c/p)$ ensemble is a function of p and satisfies $c \leq p/2$.

Thus, when $n \leq \varepsilon c \log p$ for sufficiently small $\varepsilon > 0$, the probability of error for structure estimation tends to one, where the probability measure is with respect to both the Erdős–Rényi random graph and the samples. The proof of this theorem can be found in Section 10 in the supplementary material, and is along the lines of [10], Theorem 1.

The result in Theorem 3 provides an asymptotic necessary condition for structure learning and involves an additional auxiliary parameter ε. In the following result, we remove the requirement for the auxiliary parameter ε and provide a nonasymptotic necessary condition, but at the expense of having a weak (instead of a strong) converse.

THEOREM 4 (Nonasymptotic necessary conditions for model selection). *Assume that $G \sim \mathcal{G}_{\mathrm{ER}}(p, c/p)$, where c may depend on p. Let $P_e^{(p)} := P(\hat{G}_p \neq G_p)$ be the probability of error. If $P_e^{(p)} \to 0$, the number of samples n must satisfy*

$$n \geq \frac{1}{p \log_2 |\mathcal{X}|} \binom{p}{2} \mathcal{H}_b\left(\frac{c}{p}\right). \tag{41}$$

By expanding the binary entropy function $\mathcal{H}_b(\cdot)$, it is easy to see that the statement in (41) can be weakened to the more easily interpretable (albeit weaker) necessary condition

$$n \geq \frac{c \log_2 p}{2 \log_2 |\mathcal{X}|}. \tag{42}$$

The above result differs from Theorem 3 in two aspects: the bound in (41) does not involve any asymptotic notation and is a weak converse result (instead of a strong converse). The proof is provided in Section 10.3 in the supplementary material [4].

REMARKS.

(1) Thus, $n = \Omega(c \log p)$ number of samples are *necessary* for structure recovery. Hence, the larger the average degree, the higher is the required sample complexity. Intuitively this is because as c grows, the graph is denser, and hence we require more samples for learning. In information-theoretic terms, Theorem 3 is a strong converse [19], since we show that the error probability of structure learning tends to one (instead of being merely bounded away from zero). On the other hand, the result in Theorem 4 is a weak converse result.

(2) In [43], it is shown that for graphs uniformly drawn from the class of graphs with maximum degree Δ, when $n < \varepsilon \Delta^k \log p$ for some $k \in \mathbb{N}$, there exists a graph for which any estimator fails with probability at least 0.5. These results cannot be applied here since the probability mass function is nonuniform for the class of Erdős–Rényi random graphs.

(3) The result is not dependent on the Ising model assumption, and holds for *any* pairwise discrete Markov random field (i.e., $\mathcal{X}$ is a finite set).

We now provide an outline for the proof of Theorem 4. A naïve application of Fano's inequality for this problem does not yield any meaningful result since the set of all graphs (which can be realized by $\mathcal{G}_{\mathrm{ER}}$) is "too large." We employ another information-theoretic idea known as *typicality*. We identify a set of graphs with p nodes whose average degree is ε-close to c (which is the expected degree for $\mathcal{G}_{\mathrm{ER}}(p, c/p)$. The set of typical graphs has a small cardinality but high probability when p is large. The novelty of our proof lies in our use of both typicality as well as Fano's inequality to derive necessary conditions for structure learning. We can show that (i) the probability of the typical set tends to one as $p \to \infty$; (ii) the graphs in the typical set are almost uniformly distributed (the asymptotic equipartition property); (iii) the cardinality of the typical set is small relative to the set of all graphs. A detailed discussion of these techniques is given in [3].

4.2. *Other graph families.* We now provide necessary conditions for recovery of graphs belonging to various graph ensembles considered in this paper. We first recap the results of [10], Theorem 1, which is applicable for any uniform ensemble of graphs.

THEOREM 5 (Lower bound on sample complexity). *Assume that a graph G_p on p nodes is uniformly drawn from an ensemble $\mathcal{G}$. Given n i.i.d. samples from an Ising model Markov on G, we have*

$$P[\widehat{G}_p^n(\mathbf{X}_p^n) \neq G_p] \geq 1 - \frac{2^{np}}{|\mathcal{G}|} \tag{43}$$

for any deterministic estimator $\widehat{G}_p$.

We provide bounds on the number of graphs in specific graph families considered earlier in the paper which gives us necessary conditions for their recovery.

LEMMA 2 (Bounds on size of graph families). *The following bounds hold*:

(1) *For girth-bounded ensembles $\mathcal{G}_{\mathrm{Girth}}(p; g, \Delta_{\min}, \Delta_{\max}, k)$ with girth g, minimum degree $\Delta_{\min}$, maximum degree $\Delta_{\max}$ and number of edges k, we have*

$$p^k(p - g\Delta_{\max}^g)^k \leq |\mathcal{G}_{\mathrm{Girth}}(p; g, \Delta_{\min}, \Delta_{\max}, k)| \leq p^k(p - \Delta_{\min}^g)^k. \tag{44}$$

(2) *For local-path ensembles $\mathcal{G}_{\mathrm{LP}}(p; \eta, \gamma, \Delta_{\min}, \Delta_{\max}, k)$ having η paths of length less than $\gamma > 0$ between any two nodes, minimum degree $\Delta_{\min} > 0$, maximum degree $\Delta_{\max}$ and number of edges k,*

$$\begin{aligned} m_1 p^{k_1}(p - \gamma\Delta_{\max}^\gamma)^{k_1} \binom{\Delta_{\min}^\gamma}{2}^{\eta-1} &\leq |\mathcal{G}_{\mathrm{LP}}(p; \eta, \gamma, \Delta_{\min}, \Delta_{\max}, k)| \\ &\leq m_2 p^{k_2}(p - \Delta_{\min}^\gamma)^{k_2} \binom{\gamma\Delta_{\max}^\gamma}{2}^{\eta-1}, \end{aligned} \tag{45}$$

where $k_1 := k - m_2(\eta - 1)$, $k_2 := k - m_1(\eta - 1)$, $m_1 := \frac{p}{\gamma \Delta_{\max}^{\gamma}}$ *and* $m_2 := \frac{p}{\Delta_{\min}^{\gamma}}$.

(3) *For augmented ensembles* $\mathcal{G}_{\text{Aug}}(p; d, \eta, \gamma, \Delta_{\min}, \Delta_{\max}, k)$ *consisting of a local graph with (regular) degree* d *and a global graph* $\mathcal{G}_{\text{LP}}(p; \eta, \gamma, \Delta_{\min}, \Delta_{\max}, k)$, *we have*

$$\begin{aligned} & m_1 p^{k_1'} (p - \gamma \Delta_{\max}^{\gamma})^{k_1'} \binom{\Delta_{\min}^{\gamma}}{2}^{\eta-1} \binom{p-1}{d} \\ & \quad \le |\mathcal{G}_{\text{Aug}}(p; d, \eta, \gamma, \Delta_{\min}, \Delta_{\max}, k)| \\ & \quad \le m_2 p^{k_2'} (p - \Delta_{\min}^{\gamma})^{k_2'} \binom{\gamma \Delta_{\max}^{\gamma}}{2}^{\eta-1} \binom{p-1}{d}, \end{aligned} \tag{46}$$

where $k_1' := k_1 + 1 - \frac{pd}{2}$ *and* $k_2' := k_2 + 1 - \frac{pd}{2}$, *for* k_1, k_2, m_1, m_2 *defined previously.*

The proof of the above result is given in Section 10.2 in the supplementary material [4].

REMARKS. Using the above results on lower bounds on the number of graphs in a given family, in conjunction with Theorem 5, we can obtain necessary conditions for different graph families. For instance, for girth-constrained families, when the girth g and maximum degree $\Delta_{\max}$ scale as $O(\text{poly} \log p)$, we have that

$$n = \Omega\left[\frac{k}{p} \log p\right] \tag{47}$$

number of samples is necessary for structure estimation, where k is the number of edges. Similarly, for local path ensembles, when the path threshold γ and maximum degree $\Delta_{\max}$ scale as $O(\text{poly} \log p)$, the above bound in (47) changes only slightly, and we have

$$n = \Omega\left[\left(\frac{k}{p} - \frac{\eta - 1}{\Delta_{\min}^{\gamma}}\right) \log p\right]$$

as the necessary condition, by substituting for k_1, and noting that the other terms scale slower than $\log p$ under the above specified regime. Similarly, for augmented graphs, we have

$$n = \Omega\left[\left(\frac{k}{p} - \frac{\eta - 1}{\Delta_{\min}^{\gamma}} - \frac{d}{2}\right) \log p\right]$$

as the necessary condition. Thus, for a wide class of graphs, we can characterize necessary conditions for structure estimation.

5. Experiments. In this section experimental results are presented on synthetic data. We implement the proposed CVDT (based on conditional variation distances) and CMIT (based on conditional mutual information) methods under different thresholds, as well the ℓ_1 regularized logistic regression [41] under different regularization parameters.[18] The performance of the methods is compared using the notion of the edit distance between the estimated and the true graphs. We implement the proposed CVDT and CMIT methods in MATLAB and the ℓ_1 regularized logistic regression is evaluated using L1General package.[19] CONTEST[20] package is used to generate the synthetic graphs, and UGM[21] package is used for implementing Gibbs sampling from the Ising Model. The datasets, software code and results are available at http://newport.eecs.uci.edu/anandkumar.

5.1. *Data sets*. In order to evaluate the CVDT performance in terms of quantity of errors in recovering the graph structure, we generate samples from Ising model for three typical graphs, namely, a single cycle graph whose $\eta_{\text{cycle}} = 2$, Erdős–Rényi random graph $\mathcal{G}_{\text{ER}}(p, c/p)$ with average degree $c = 1$ and the Watts and Strogatz model $\mathcal{G}_{\text{WS}}(p, d, c/p)$ with degree of local graph $d = 2$ and average degree of the global graph $c = 1$. Graphs of size $p = 80$ and sample size $n \in \{10^2, 5 \times 10^2, 10^3, 5 \times 10^3, 10^4, 10^5\}$ are considered.

Based on the generated graph topologies, we generate the potential matrix $\mathbf{J}_G$ whose sparsity pattern corresponds to that of the graph G. By convention, diagonal elements $\mathbf{J}(i, i) = 0$ for all $i \in V$. We consider both attractive and general models. For attractive models, we consider the nonzero off-diagonal entries of $\mathbf{J}$ as uniformly distributed in $[0.1, 0.2]$. For the general model, we consider the nonzero off-diagonal entries of $\mathbf{J}$ as uniformly distributed in $[0.1, 0.2] \cup [-0.1, -0.2]$. Potential vector is set to $\mathbf{0}$ resulting in a symmetric Ising model. Gibbs sampling method is used to generate samples. The knowledge of the bound on local separators η is assumed to be available in our experiments. We employ normalized edit distances as the performance criterion. Since we know the ground truth for synthetic data, it is possible to evaluate this measure. The thresholds $\xi_{n,p}$ for CVDT/CMIT and the regularization parameter λ_n for the ℓ_1 regularized logistic regression are selected based on the best edit distances for each method.

5.2. *Experimental results*. Table 1 presents the experimental outcomes, and an explicit comparison of the three graph estimation methods is illustrated in Figure 2 for attractive models, and in Figure 3 for mixed models (with both positive and negative edge potentials). Similar trends are observed for both attractive and

[18] For the convex relaxation method in [41], the regularization parameter denotes the weight associated with the ℓ_1 term.

[19] L1General is available at http://www.di.ens.fr/~mschmidt/Software/L1General.html.

[20] CONTEST is at http://www.mathstat.strath.ac.uk/research/groups/numerical_analysis/contest.

[21] UGM is at http://www.di.ens.fr/~mschmidt/Software/UGM.html.

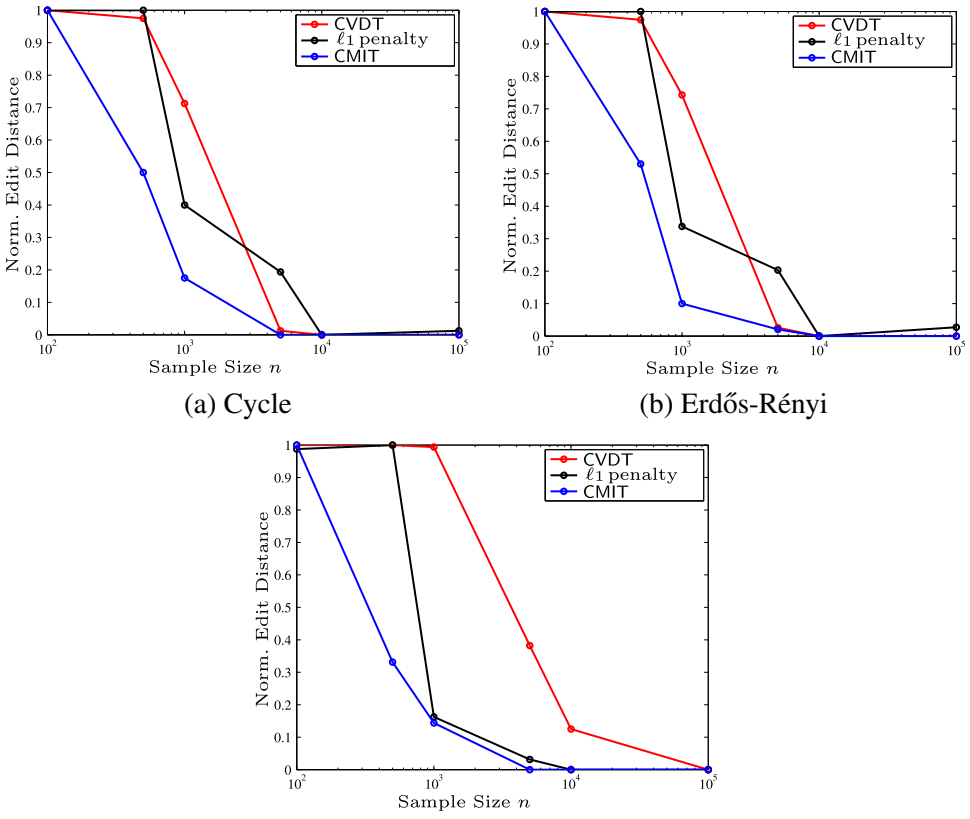
TABLE 1

Normalized edit distance under CVDT (*based on conditional variation distances*), CMIT (*based on conditional mutual information*) *and* ℓ_1 *penalized neighborhood selection on synthetic data from graphs listed above for attractive and mixed Ising models, where n denotes the number of samples*

Graph	n	**CVDT (attractive)**	**CMIT (attractive)**	ℓ_1 **penalty (attractive)**	**CVDT (mixed)**	**CMIT (mixed)**	ℓ_1 **penalty (mixed)**
Cycle	1×10^2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
ER	1×10^2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
WS	1×10^2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Cycle	5×10^2	1.0000	0.5000	1.0000	0.975	0.475	1.0000
ER	5×10^2	1.0000	0.5300	1.0000	0.9189	0.5946	1.0000
WS	5×10^2	1.0000	0.3313	1.0000	1.0000	0.3313	1.0000
Cycle	1×10^3	0.7125	0.1750	0.4000	0.7250	0.1500	0.3063
ER	1×10^3	0.7428	0.1020	0.3378	0.6757	0.1351	0.4342
WS	1×10^3	0.9937	0.1438	0.1625	0.9938	0.1438	0.4255
Cycle	5×10^3	0.0125	0.0000	0.1937	0.0125	0.0000	0.1500
ER	5×10^3	0.0000	0.0204	0.2031	0.0000	0.1053	0.0000
WS	5×10^3	0.3827	0.0000	0.0312	0.5688	0.0000	0.2671
Cycle	1×10^4	0.0000	0.0000	0.0000	0.3063	0.0000	0.0000
ER	1×10^4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
WS	1×10^4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

mixed models. We note that the edit distance decays as the number of samples increases, as expected. As long as there are enough number of samples (larger than 10,000), all the methods recover the graph structure accurately, that is, with zero error. In terms of the decaying rate of errors, the ℓ_1 logistic regression method has a faster rate than CVDT for the Watts–Strogatz graph in all regimes, while for the cycle graph and the Erdős–Rényi graph, the rates for CVDT and the ℓ_1 method are alternatively better depending on n. However, CMIT has the fastest rate of decay of edit distance for all the three graphs, although theoretically, CVDT has better sample complexity guarantees compared to CMIT; see Theorem 1 and related remarks. With regard to the running time, CVDT and CMIT are faster for the graphs under consideration, since there is one global threshold to be selected for finding all the edges, while for logistic regression, selection of the regularization parameter needs to be carried out for each neighborhood in the graph. This is especially expensive for large graphs.

6. Conclusion. In this paper, we adopted a novel and a unified paradigm for Ising model selection. We presented a simple local algorithm for structure estimation with low computational and sample complexities under a set of mild and



(a) Cycle (b) Erdős-Rényi

(c) Watts-Strogatz

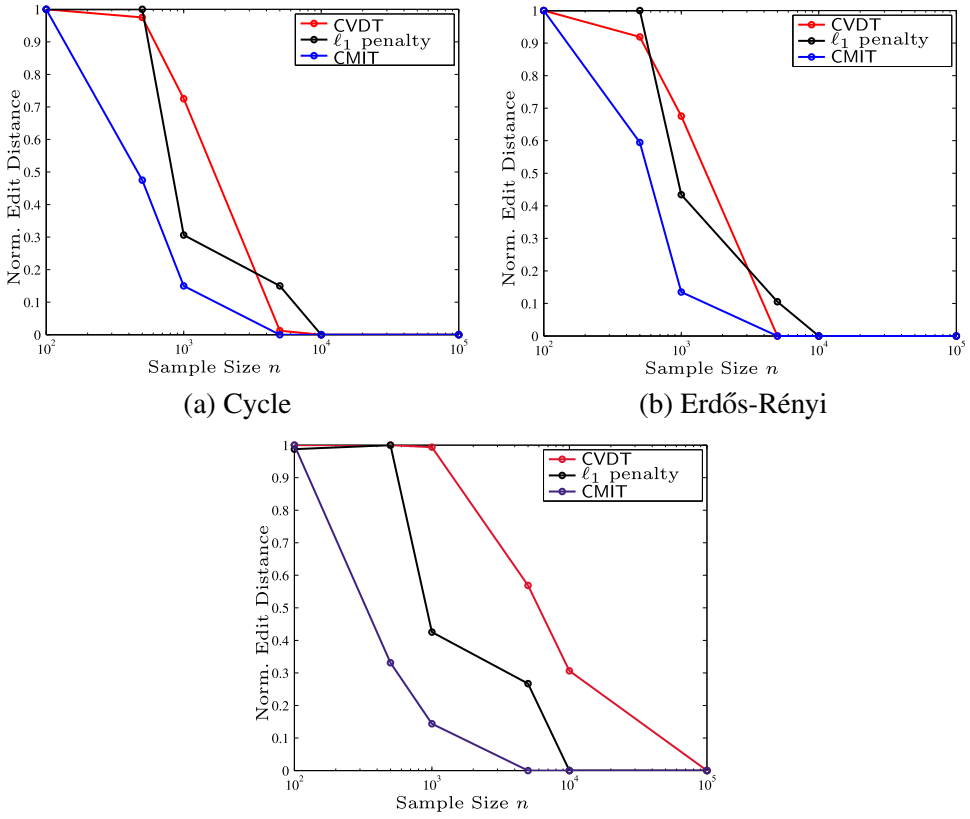
FIG. 2. CVDT, CMIT *and* ℓ_1 *penalized logistic regression on synthetic data from an attractive Ising model.*

transparent conditions. This algorithm succeeds on a wide range of graph ensembles such as the Erdős–Rényi ensemble, small-world networks etc. based on a local separation criterion.

SUPPLEMENTARY MATERIAL

Supplement to "High-dimensional structure estimation in Ising models: Local separation criterion" (DOI: 10.1214/12-AOS1009SUPP; .pdf). Detailed analysis and proofs.

Acknowledgments. The authors thank Sujay Sanghavi (U.T. Austin), Elchanan Mossel (UC Berkeley), Martin Wainwright (UC Berkeley), Sebastien Roch (UCLA), Rui Wu (UIUC) and Divyanshu Vats (U. Minn.) for extensive comments, and Béla Bollobás (Cambridge) for discussions on random graphs.



(a) Cycle (b) Erdős-Rényi

(c) Watts-Strogatz

FIG. 3. CVDT, CMIT *and* ℓ_1 *penalized logistic regression on synthetic data from a mixed Ising model* (*with both positive and negative edge potentials*).

The authors thank the anonymous reviewers and the co-editor Peter Bühlmann (ETH) for valuable comments that significantly improved this manuscript.

REFERENCES

[1] ABBEEL, P., KOLLER, D. and NG, A. Y. (2006). Learning factor graphs in polynomial time and sample complexity. *J. Mach. Learn. Res.* **7** 1743–1788. MR2274423

[2] ALBERT, R. and BARABÁSI, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Modern Phys.* **74** 47–97. MR1895096

[3] ANANDKUMAR, A., TAN, V. Y. F., HUANG, F. and WILLSKY, A. S. (2011). High-dimensional Gaussian graphical model selection: Tractable graph families. Preprint. Available at arXiv:1107.1270.

[4] ANANDKUMAR, A., TAN, V. Y. F., HUANG, F. and WILLSKY, A. S. (2012). Supplement to "High-dimensional structure learning of Ising models: Local separation criterion." DOI:10.1214/12-AOS1009SUPP.

[5] BAYATI, M., MONTANARI, A. and SABERI, A. (2009). Generating random graphs with large girth. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms* 566–575. SIAM, Philadelphia, PA. MR2809261

[6] BENTO, J. and MONTANARI, A. (2009). Which graphical models are difficult to learn? In *Proc. of Neural Information Processing Systems* (*NIPS*).

[7] BOGDANOV, A., MOSSEL, E. and VADHAN, S. (2008). The complexity of distinguishing Markov random fields. In *Approximation, Randomization and Combinatorial Optimization. Lecture Notes in Comput. Sci.* **5171** 331–342. Springer, Berlin. MR2538798

[8] BOLLOBÁS, B. (1985). *Random Graphs*. Academic Press, London. MR0809996

[9] BRÉMAUD, P. (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues. Texts in Applied Mathematics* **31**. Springer, New York. MR1689633

[10] BRESLER, G., MOSSEL, E. and SLY, A. (2008). Reconstruction of Markov random fields from samples: Some observations and algorithms. In *Approximation, Randomization and Combinatorial Optimization. Lecture Notes in Computer Science* **5171** 343–356. Springer, Berlin. MR2538799

[11] CHANDRASEKARAN, V., PARRILO, P. A. and WILLSKY, A. S. (2010). Latent variable graphical model selection via convex optimization. *Ann. Statist.* To appear. Preprint. Available on ArXiv.

[12] CHECHETKA, A. and GUESTRIN, C. (2007). Efficient principled learning of thin junction trees. In *Advances in Neural Information Processing Systems* (*NIPS*).

[13] CHENG, J., GREINER, R., KELLY, J., BELL, D. and LIU, W. (2002). Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence* **137** 43–90. MR1906473

[14] CHOI, M. J., LIM, J. J., TORRALBA, A. and WILLSKY, A. S. (2010). Exploiting hierarchical context on a large database of object categories. In *IEEE Conf. on Computer Vision and Pattern Recognition* (*CVPR*).

[15] CHOI, M. J., TAN, V. Y. F., ANANDKUMAR, A. and WILLSKY, A. S. (2011). Learning latent tree graphical models. *J. Mach. Learn. Res.* **12** 1771–1812. MR2813153

[16] CHOW, C. and LIU, C. (1968). Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Tran. on Information Theory* **14** 462–467.

[17] CHUNG, F. R. K. (1997). *Spectral Graph Theory. CBMS Regional Conference Series in Mathematics* **92**. Published for the Conference Board of the Mathematical Sciences, Washington, DC. MR1421568

[18] CHUNG, F. R. K. and LU, L. (2006). *Complex Graphs and Network*. Amer. Math. Soc., Providence, RI.

[19] COVER, T. M. and THOMAS, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley, Hoboken, NJ. MR2239987

[20] DOMMERS, S., GIARDINÀ, C. and VAN DER HOFSTAD, R. (2010). Ising models on power-law random graphs. *J. Stat. Phys.* **141** 1–23.

[21] DURBIN, R., EDDY, S. R., KROGH, A. and MITCHISON, G. (1999). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge Univ. Press, Cambridge.

[22] EPPSTEIN, D. (2000). Diameter and treewidth in minor-closed graph families. *Algorithmica* **27** 275–291. MR1759751

[23] GALAM, S. (1997). Rational group decision making: A random field Ising model at $T = 0$. *Physica A: Statistical and Theoretical Physics* **238** 66–80.

[24] GAMBURD, A., HOORY, S., SHAHSHAHANI, M., SHALEV, A. and VIRÁG, B. (2009). On the girth of random Cayley graphs. *Random Structures Algorithms* **35** 100–117. MR2532876

[25] GRABOWSKI, A. and KOSINSKI, R. (2006). Ising-based model of opinion formation in a complex network of interpersonal interactions. *Physica A: Statistical Mechanics and Its Applications* **361** 651–664.

[26] KALISCH, M. and BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8** 613–636.
[27] KARGER, D. and SREBRO, N. (2001). Learning Markov networks: Maximum bounded tree-width graphs. In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms* (*Washington, DC*, 2001) 392–401. SIAM, Philadelphia, PA. MR1958431
[28] KEARNS, M. J. and VAZIRANI, U. V. (1994). *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, MA. MR1292868
[29] KLOKS, T. (1994). Only few graphs have bounded treewidth. *Springer Lecture Notes in Computer Science* **842** 51–60.
[30] LACIANA, C. E. and ROVERE, S. L. (2010). Ising-like agent-based technology diffusion model: Adoption patterns vs. seeding strategies. *Physica A: Statistical Mechanics and Its Applications* **390** 1139–1149.
[31] LAURITZEN, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. MR1419991
[32] LEVIN, D. A., PERES, Y. and WILMER, E. L. (2008). *Markov Chains and Mixing Times.* Amer. Math. Soc., Providence, RI.
[33] LIU, H., XU, M., GU, H., GUPTA, A., LAFFERTY, J. and WASSERMAN, L. (2011). Forest density estimation. *J. Mach. Learn. Res.* **12** 907–951. MR2786914
[34] LIU, S., YING, L. and SHAKKOTTAI, S. (2010). Influence maximization in social networks: An ising-model-based approach. In *Proc.* 48*th Annual Allerton Conference on Communication, Control, and Computing*.
[35] LOVÁSZ, L., NEUMANN LARA, V. and PLUMMER, M. (1978). Mengerian theorems for paths of bounded length. *Period. Math. Hungar.* **9** 269–276. MR0509677
[36] MCKAY, B. D., WORMALD, N. C. and WYSOCKA, B. (2004). Short cycles in random regular graphs. *Electron. J. Combin.* **11** Research Paper 66, 12 pp. (electronic). MR2097332
[37] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363
[38] MITLIAGKAS, I. and VISHWANATH, S. (2010). Strong information-theoretic limits for source/model recovery. In *Proc.* 48*th Annual Allerton Conference on Communication, Control and Computing*.
[39] NETRAPALLI, P., BANERJEE, S., SANGHAVI, S. and SHAKKOTTAI, S. (2010). Greedy learning of Markov network structure. In *Proc.* 48*th Annual Allerton Conference on Communication, Control and Computing*.
[40] NEWMAN, M. E. J., WATTS, D. J. and STROGATZ, S. H. (2002). Random graph models of social networks. *Proc. Natl. Acad. Sci. USA* **99** 2566–2572.
[41] RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343
[42] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766
[43] SANTHANAM, N. P. and WAINWRIGHT, M. J. (2008). Information-theoretic limits of high-dimensional model selection. In *International Symposium on Information Theory*.
[44] SPIRTES, P. and MEEK, C. (1995). Learning Bayesian networks with discrete variables from data. In *Proc. of Intl. Conf. on Knowledge Discovery and Data Mining* 294–299.
[45] TAN, V. Y. F., ANANDKUMAR, A., TONG, L. and WILLSKY, A. S. (2011). A large-deviation analysis of the maximum-likelihood learning of Markov tree structures. *IEEE Trans. Inform. Theory* **57** 1714–1735. MR2815845
[46] TAN, V. Y. F., ANANDKUMAR, A. and WILLSKY, A. S. (2010). Learning Gaussian tree models: Analysis of error exponents and extremal structures. *IEEE Trans. Signal Process.* **58** 2701–2714. MR2789417

[47] TAN, V. Y. F., ANANDKUMAR, A. and WILLSKY, A. S. (2011). Learning high-dimensional Markov forest distributions: Analysis of error rates. *J. Mach. Learn. Res.* **12** 1617–1653. MR2813149

[48] VEGA-REDONDO, F. (2007). *Complex Social Networks*. *Econometric Society Monographs* **44**. Cambridge Univ. Press, Cambridge. MR2361122

[49] WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* **1** 1–305.

[50] WANG, W., WAINWRIGHT, M. J. and RAMCHANDRAN, K. (2010). Information-theoretic bounds on model selection for Gaussian Markov random fields. In *IEEE International Symposium on Information Theory Proceedings* (*ISIT*).

[51] WATTS, D. J. and STROGATZ, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393** 440–442.

[52] GRAPHICAL MODEL OF SENATE VOTING. http://www.eecs.berkeley.edu/~elghaoui/StatNews/ex_senate.html.

A. ANANDKUMAR
F. HUANG
CENTER FOR PERVASIVE COMMUNICATIONS
& COMPUTING
ELECTRICAL ENGINEERING
& COMPUTER SCIENCE DEPARTMENT
4408 ENGINEERING HALL
IRVINE, CALIFORNIA 92697
USA
E-MAIL: a.anandkumar@uci.edu
furongh@uci.edu

V. Y. F. TAN
INSTITUTE FOR INFOCOMM RESEARCH
A*STAR SINGAPORE
AND
DEPARTMENT OF ELECTRICAL
AND COMPUTER ENGINEERING
NATIONAL UNIVERSITY OF SINGAPORE
SINGAPORE
E-MAIL: tanyfv@i2r.a-star.edu.sg

A. S. WILLSKY
LABORATORY OF INFORMATION
& DECISION SYSTEMS
STATA CENTER, 77 MASSACHUSETTS AVE.
CAMBRIDGE, MASSACHUSETTS 02139
USA
E-MAIL: willsky@mit.edu