Submitted to the Annals of Statistics
arXiv: math.ST/1011.0129

HIGH DIMENSIONAL STRUCTURE LEARNING OF ISING MODELS ON SPARSE RANDOM GRAPHS[*]

By Animashree Anandkumar[†,¶], Vincent Tan[‡,§,‖] and Alan Willsky[§,‖]

University of California Irvine[¶] *and Massachusetts Institute of Technology*[‖]

We consider the problem of learning the structure of ferromagnetic Ising models Markov on sparse Erdős-Rényi random graphs $G_n \sim \mathcal{G}(n, c/n)$ with $c \in \mathbb{R}^+$. We propose simple local algorithms and analyze their performance in the regime of correlation decay, i.e., when $c \tanh J_{\max} < 1$, where $J_{\max}$ is the maximum potential in the model. We prove that an algorithm based on a set of conditional mutual information tests is consistent for structure learning throughout the regime of correlation decay. This algorithm requires the number of samples to scale as $\omega(\log n)$, and has a computational complexity of $O(n^4)$. A simpler algorithm based on correlation thresholding outputs a graph with a constant edit distance to the original graph when there is correlation decay and sufficient homogeneity, and the number of samples required is $m = \Omega(\log n)$. Under a more stringent condition, given by $2 \tanh^2 J_{\max} < \tanh J_{\min}$, where $J_{\min}$ is the minimum potential, correlation thresholding is consistent for structure estimation. We finally prove a lower bound that $m = \Omega(c \log n)$ samples are also needed for consistent reconstruction of random graphs by any algorithm with positive probability, for any $c \leq 0.5n$. Thus, we establish that consistent structure estimation is possible with almost order-optimal sample complexity throughout the regime of correlation decay.

1. Introduction.

Many naturally occurring large networks such as online social networks are well modeled by random graphs [32]. The connectivity of these networks is usually unknown and has to be estimated using noisy and incomplete observations. The challenges in structure estimation of these networks are compounded by the fact that these networks are typically in the high-dimensional regime, where the size of the network is much

[*] Part of this material was presented at the Neural Information Processing Workshop 2010 on Robust Statistical Learning.

[†] Supported by the setup funds at UCI.

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

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

AMS 2000 subject classifications: Primary 62H12; secondary 05C80

Keywords and phrases: Ising models, Sparse random graphs, High-dimensional structure learning, Correlation decay

larger than the number of observations. It is thus imperative to design efficient algorithms for learning such large random structures with low sample complexity, i.e., those that can produce consistent estimates when the graph size grows much faster than the number of samples obtained for learning.

The interactions among the nodes in the network can be modeled via a graphical model, also known as a Markov random field [22], where the probability distribution of the observations at the nodes factorizes according to the given graph. In this paper, we consider Ising models Markov on sparse Erdős-Rényi random graphs [3]. Ising models are arguably the simplest and also the most extensively studied class of discrete graphical models and among random graphs, the same holds for Erdős-Rényi random graphs. It is well known that consistent structure learning of tree graphical models is tractable [9] while structure learning of general models is NP-hard [19]. Given that sparse random graphs are "locally tree-like" [3] (meaning that a typical node is not part of a short cycle), our work explores if consistent learning is tractable for random graphs as well.

The questions addressed in this paper are: Given i.i.d. observations from the Ising model on sparse random graphs, are there simple, efficient and preferably local algorithms which can recover the underlying graph structure? If so, under what conditions on the model can we guarantee these algorithms to be consistent? What is the sample complexity of these algorithms? Is there a lower bound or a converse result on the sample complexity for reconstructing random graphs by *any* algorithm? The answers to these questions provide insights to the complex interplay between the underlying graph structure and parameters, and the resulting learning performance in large-scale models.

While many new algorithms have been recently proposed for structure learning (see Section 1.2 for a review), most of these algorithms require the graph to have a bounded maximum degree (independent of the size of the graph) for consistent learning and polynomial-time computational complexity. Hence, these algorithms are not applicable for random graphs which are sparse on average but contain nodes with large degrees. To the best of our knowledge, this paper presents the first class of algorithms that guarantee consistent reconstruction of random graphs with growing maximum degree and logarithmic sample complexity in the so-called regime of *correlation decay* [27]. The notion of correlation decay has been extensively studied by the physics, computer science and statistics community, and is related to many phenomena such as fast mixing of Markov chain Monte Carlo (MCMC) algorithms [23] and non-reconstructability of a variable at a node, given variables "far away" from it [27]. This work establishes that correlation decay is sufficient to ensure consistent structure learning in ferromagnetic (attractive) Ising models on random graphs.

1.1. *Summary of Contributions.* Our main contributions in this work are threefold. We propose two local algorithms for structure learning of ferromagnetic Ising models and establish sufficient conditions for consistency and the required sample complexity. The algorithm based on a set of conditional mutual information tests is shown to be consistent throughout the regime of correlation decay. A simpler algorithm based on correlation thresholding is shown to be consistent under more stringent conditions. Third, we prove a lower bound on sample complexity required by any learning algorithm to succeed.

We establish that consistent structure learning is achieved by a set of conditional mutual information tests in the regime of correlation decay. This algorithm requires the number of samples scaling as $m = \omega(\log n)$, for a n node graph, and it has a computational complexity of $O(n^4)$. This algorithm is completely localized and independently tests whether a node pair forms an edge in the graph. Moreover, the algorithm is parameter-free, meaning that it does not require the knowledge of any parameters such as the minimum/maximum potentials of the Ising model or the average degree of the random graph.

We also analyze the performance of a simpler algorithm based on thresholding of pairwise correlations. Under correlation decay, we prove that this algorithm is consistent under an additional condition of homogeneity, given by $2\tanh^2 J_{\max} < \tanh J_{\min}$, where $J_{\min}, J_{\max}$ are minimum and maximum potentials of the Ising model. When this condition is not satisfied, but there is still correlation decay and sufficient homogeneity, we prove that the edit distance between the reconstructed and the original graphs is at most a constant. The number of samples required is $m = \Omega(\log n)$, and this matches the lower bound on sample complexity required by any algorithm (up to constant factors).

The third main contribution is a lower bound on the sample complexity required for consistent reconstruction with positive probability[1] by any algorithm. We prove that $m = \Omega(c \log n)$ number of samples is required by any algorithm to ensure consistent learning of Erdős-Rényi random graphs, where c is the average degree and n is the number of nodes. The proof uses an information-theoretic covering argument, on lines of [5], and is valid for

[1] In information-theoretic terms, our result is a strong converse [10], meaning we establish that when $m < \epsilon c \log n$ for some $\epsilon > 0$, the probability of inconsistent reconstruction goes to one instead of merely being bounded away from zero.

any $c < 0.5n$ (and hence, is not limited to the regime of sparse graphs).

Thus, our work establishes that consistent structure learning is tractable throughout the regime of correlation decay for ferromagnetic Ising models on sparse random graphs. Our results are based on novel bounds on correlations and conditional mutual information quantities for models under correlation decay on sparse random graphs. These bounds require careful analysis of cycles and paths in random graphs, and go beyond the locally tree-like property at a typical node in the graph. This is because the bounds are for the worst-case correlation and mutual information quantities, which are indeed affected by the presence of cycles in the graph. Moreover, to obtain the stated sample complexity result for conditional mutual information thresholding test requires careful design of the threshold, based both on the number of samples and the number of nodes.

While our work establishes feasibility of structure learning under correlation decay, a converse result on learning beyond the correlation decay regime is currently elusive. It remains an open question on whether polynomial-time algorithms exist for structure learning of graphs with growing maximum degree beyond the regime of correlation decay.

1.2. *Related Work.* This paper is situated at the intersection of two main research themes: (i) learning of graphical models and related analysis on consistency, sample and computational complexities and (ii) correlation decay and related analysis of Ising models on random graphs. We describe the related literature of these research themes in this section.

The problem of structure estimation of a general graphical model [19] is NP-hard[2]. However, for tree graphical models, it can be performed in polynomial time and the classical Chow-Liu algorithm [9] provides an efficient implementation for maximum-likelihood structure estimation. The authors show that it is possible to reduce the learning problem to a maximum-weight spanning tree problem where the edge weights are the empirical mutual information quantities that can be computed from the data. Error-exponent analysis of the Chow-Liu algorithm was performed in [37, 39] and extensions to general forest models [24, 38] and trees with latent (hidden) variables [8] have also been studied recently.

Given that structure learning can be solved efficiently for tree models, a natural extension is to consider learning the structures of *junction trees.* Junction trees are formed by triangulating a given graph, and its nodes correspond to the maximal cliques of the triangulated graph [40]. The *treewidth*

[2] It is also known that the complexity of distinguishing general graphical models with hidden nodes is NP-hard [2].

of a graph is one less than the minimum possible size of the maximum clique in the triangulated graph over all possible triangulations. Efficient algorithms have been previously proposed for learning junction trees with bounded treewidth (e.g., [7]). However, these algorithms entail exponential complexity in the tree width, and hence, are not applicable to our problem since the treewidth of random graphs grows with the number of nodes[3].

Learning sparse graphical model structures have been extensively studied in literature before. In [5], a consistent algorithm for structure learning of general[4] Markov random fields with bounded-degree graphs is proposed. For a graph with degree bound Δ, the neighborhood selection at each node is based on a series of conditional-independence tests on subsets of cardinality at most Δ and the computational complexity of structure learning is $O(n^{\Delta+2}\log n\epsilon^{-2}\delta^{-4})$, where δ, ϵ depend on the model parameters. If the model satisfies correlation decay[5], it is shown that the running time can be reduced to $O(n^2 \log n)$ for any fixed Δ. In [31], the authors suggested an alternative greedy algorithm, based on conditional entropy, for graphs with large girth and bounded degree under correlation decay. However, the algorithms in [5, 31] are not efficient on random graphs since the maximum degree is not bounded[6], and the girth is not small (although a typical node is not part of a short cycle). Moreover, unlike the work here, the algorithms proposed in [5, 31] are not parameter-free, meaning they require knowledge of the parameters of the model (such as the degree bound, strength of local interactions) for structure learning. Structure learning algorithms for homogeneous Ising models on lattices are proposed in [11], but the algorithm is not computationally efficient.

Recent works explore the applicability of convex relaxation methods for structure learning. In [34], each node independently performs a local neighborhood selection based on ℓ_1 logistic regression. This method can efficiently learn Ising models on graphs with slowly growing maximum degree. However, the incoherence conditions required for consistency guarantees are not straightforward to verify for random graphical models, and it is not clear whether they coincide with the regime of correlation decay. Moreover, the number of samples required for ℓ_1 logistic regression is $m = \Omega(\Delta^3 \log n)$, where Δ is the maximum degree. In contrast, our proposed algorithm requires $m = \omega(\log n)$ samples. This ℓ_1-based method is extended in [21] to

[3] For a random graph $G_n \sim \mathcal{G}(n, c/n)$ in the super-critical regime ($c > 1$), the tree-width is greater than n^ϵ, for some $\epsilon > 0$ [20].

[4] The work in [5] allows for general Markov random fields with higher- order potentials.

[5] For Ising models on graphs with degree bound Δ, the condition $\Delta \tanh J_{\max} < 1$ implies correlation decay, where $J_{\max}$ is the maximum potential [42].

[6] In fact, the maximum degree is $\Theta(\frac{\log n}{\log\log n})$ for a.e. G_n $\mathcal{G}(n, c/n)$ [3, Ex. 3.6].

time-varying models, and in [18] to joint structure and parameter estimations. Analogously, [26] propose Lasso-like algorithms for neighborhood selection and [35] propose ℓ_1-penalized likelihood method for Gaussian graphical models. Recently, in [6], convex relaxation methods are used for estimating Gaussian graphical models with hidden nodes, and sufficient conditions for correct recovery are provided. A major disadvantage in using convex-relaxation methods is that the incoherence conditions required for consistent estimation are hard to interpret. In contrast, we provide consistency guarantees under a transparent condition of correlation decay. Moreover, convex relaxation methods require full-order statistics of data, while our proposed algorithms require only pairwise and fourth-order statistics respectively.

Converse results on structure learning provide a lower bound on sample complexity for structure learning and have been explored before in [28, 36, 41]. But these works consider graphs drawn uniformly from the class of bounded degree graphs. For this scenario, it is shown that $m = \Omega(\Delta^k \log n)$ samples are required for consistent structure estimation, in an n-node graph with maximum degree Δ, for some $k \in \mathbb{N}$. In contrast, our converse result is based on the *average degree* of the random graph instead of the maximum degree.

This work establishes tractable structure learning in the presence of correlation decay. Correlation decay[7] refers to the property in large models, where there are no long-range correlations [15, 27]. This regime is also known as the *uniqueness regime* since under such an assumption, the variable at a node is asymptotically independent of a growing boundary. Correlation decay and related analysis of Ising models have been carried out extensively in the literature, e.g., on trees in [25, 33] and the references therein, and on random graphs in [13, 16, 29, 30]. We leverage on these results but also use new techniques to provide novel bounds on the correlations and the conditional mutual information quantities on the edges and non-edges of the random Ising model.

The motivation for this work comes from the work in [1], where the relationship between structure learning and correlation decay is discussed. It is shown that some well-known local learning algorithms, discussed above, fail when the long-range correlations are sufficiently large. The analysis in [1] is carried out on homogeneous Ising models on bounded-degree graphs unlike heterogeneous Ising models on random graphs considered here. Moreover, the regime of failure of local algorithms does not coincide with the phase transition (from the regime of correlation decay). On the other hand, we

[7] Technically, correlation decay can be defined in multiple ways [27, p. 520] and the notion we use is the uniqueness or the extremality condition.

establish feasibility of structure learning within the regime of correlation decay, and the question of whether learning is feasible beyond this regime remains open.

2. Problem Formulation.

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

2.1. *Notation.*

We introduce basic notions in graph theory [3] and in information theory [10]. Let $G_n = (V_n, E_n)$ be a (labeled) undirected graph where $V_n = \{1, 2, \ldots, n\}$ is the vertex set and $E_n \subset \binom{V_n}{2}$ is the edge set. Let $G_n \sim \mathcal{G}(n, c/n), c \in \mathbb{R}^+$ denote a realization of the Erdős-Rényi random graph. Here, the appearance probability of each edge is $\frac{c}{n}$, and thus the average node degree is $c/2$. Let $\mathcal{Q}$ be a graph property (such as being connected). We say that the property $\mathcal{Q}$ for a sequence of random graphs $\{G_n\}_{n\in\mathbb{N}}$ holds asymptotically almost surely (a.a.s.) if, $\mathbb{P}(G_n \text{ satisfies } \mathcal{Q}) \to 1$ as $n \to \infty$. Here $\mathbb{P}$ is the probability measure associated to the ensemble of random graphs. Alternatively, we say that almost every (a.e.) graph G_n satisfies property $\mathcal{Q}$.

Denote the pairwise correlation between variables X_i and X_j , $i, j \in V_n$ as

$$C(i, j) := \mathbb{E}[X_i X_j] \tag{1}$$

Given m samples x_i^m, x_j^m drawn i.i.d. from X_i, X_j, the *empirical correlation* between node i and j defined as

$$\widehat{C}_{i,j}^m := \widehat{C}(i, j; x_i^m, x_j^m) := \frac{1}{m}\sum_{k=1}^{m} x_{i,k}x_{j,k}, \tag{2}$$

and we first consider a graph estimator that only requires empirical correlations $\{\widehat{C}(i, j; x_i^m, x_j^m)\}_{i,j\in V_n}$ as inputs.

For two probability mass functions P and Q defined a common countable sample space $\mathcal{X}$, 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 random variables (X, Y) taking values in the countable set $\mathcal{X} \times \mathcal{Y}$ and distributed as $P = P_{X,Y}$, the *mutual information* is defined as

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

The mutual information, an information-theoretic quantity, roughly measures the amount of dependence between the two random variables X and Y. It is also well-known that $I(X;Y) = 0$ if and only if X is independent of Y. On 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

$$I(X;Y|Z) := \sum_{x\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}} P(x,y,z)\log\frac{P(x,y|z)}{P(x|z)P(y|z)}. \tag{4}$$

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

$$\widehat{P}^m(x,y;x^m,y^m) := \frac{1}{m}\sum_{i=1}^m \mathbb{I}\{(x,y) = (x_i,y_i)\}. \tag{5}$$

The *empirical mutual information* is given by

$$\widehat{I}^m(X;Y) := \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \widehat{P}^m(x,y)\log\frac{\widehat{P}^m(x,y)}{\widehat{P}^m(x)\widehat{P}^m(y)}. \tag{6}$$

The *empirical conditional mutual information* is defined analogously given (x^m, y^m, z^m). Our second algorithm for graph estimation will be based on empirical conditional mutual information.

2.2. *Ising Models.* A *graphical model* is a family of multivariate distributions which are Markov in accordance to a particular undirected graph [22]. 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] $G = (V, E)$ $E \subset \binom{V}{2}$ capture the set of conditional independence relations among the random variables. We say that a vector of random variables $\mathbf{X} := (X_1, \ldots, X_n)$ with distribution P is Markov on the graph G if the local Markov property

$$P(x_i|x_{\mathcal{N}(i)}) = P(x_i|x_{V\setminus i}) \tag{7}$$

holds for all nodes $i \in V$. More generally, we say that a multivariate distribution P satisfies the global Markov property, if for all disjoint sets $A, B \subset V$, we have

$$P(\mathbf{x}_A, \mathbf{x}_B|\mathbf{x}_S) = P(\mathbf{x}_A|\mathbf{x}_S)P(\mathbf{x}_B|\mathbf{x}_S). \tag{8}$$

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

where set S is the *separator*[9] between A and B. The local and global Markov properties are equivalent under the *positivity* condition, given by $P(\mathbf{x}) > 0$, for all $\mathbf{x} \in \mathcal{X}^n$ [22].

The Hammersley-Clifford theorem [4] states that under the positivity condition, a distribution P satisfies the Markov property according to a graph G iff. it factorizes according to the cliques of G, i.e.,

$$P(\mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{c \in \mathcal{C}} \Psi_c(\mathbf{x}_c)\right), \tag{9}$$

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,

$$P(\mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{e \in E} \Psi_e(\mathbf{x}_e)\right). \tag{10}$$

One of the most well-studied pairwise models is the Ising model. Here, each random variable X_i takes values in the set $\mathcal{X} = \{-1, +1\}$. When the underlying graph G_n is random, we let $P_{\mathbf{X}_n|G_n}(\cdot|G_n)$ denote the Ising model probability mass function associated with random vector $\mathbf{X} = (X_1, \ldots, X_n)$ conditioned on a particular graph realization G_n. The distribution[10] is given by

$$P_{\mathbf{X}|G_n}(\mathbf{x}|G_n; \mathbf{J}) = \frac{1}{Z(\mathbf{J})} \exp\left(\sum_{(i,j) \in G_n} J_{i,j} x_i x_j\right), \tag{11}$$

where $J_{i,j}$ is known as the *potential* for node pair (i,j), and $J_{i,j} = 0$ for $(i,j) \notin G_n$. We also assume that there exists $J_{\min}, J_{\max} \in \mathbb{R}$ and independent of n such that the potentials on the edges are uniformly bounded, i.e.,

$$J_{i,j} \in [J_{\min}, J_{\max}], \quad \forall\, (i,j) \in G_n,\ n \in \mathbb{N}. \tag{12}$$

[9] A set $S \subset V$ is a separator for sets A and B if the removal of nodes in S separates A and B into distinct components.

[10] Note that the marginal node distribution is uniform in (11), and hence, this model is also known as the symmetric Ising model. A generalization of this model where the marginal node distributions are non-uniform has additional parameters known as the external magnetic fields [27].

2.3. *Consistent Graph Reconstruction.* Conditioned on a graph G_n, let $\mathbf{x}^m := \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ be m i.i.d. samples drawn from $P_{\mathbf{X}|G_n}(\cdot|G_n; \mathbf{J})$ and let $x_i^m := \{x_{i,1}, \dots, x_{i,m}\}$ denote m i.i.d samples from $P_{X_i|G_n}(\cdot|G_n; \mathbf{J})$ the marginal distribution at node i. Given $\mathbf{x}^m$, a *graph estimator* is a (deterministic or random) function[11] $\widehat{G}_n^m : (\mathcal{X}^n)^m \to \mathcal{G}_n$, where $\mathcal{G}_n$ is the set of all possible graphs on n (labeled) nodes. We formally define consistency and sample complexity.

Definition 1 (Structural Consistency and Sample Complexity). *An estimator $\widehat{G}_n^m$ is said to be consistent in high-dimensions with sample complexity $O(f(n))$ if*

$$\lim_{\substack{m,n\to\infty\\ m=O(f(n))}} \mathbb{P}[\widehat{G}_n^m \neq G] = 0.$$

3. Method and Guarantees. We propose two algorithms for reconstructing random graphs. We first present the simple correlation thresholding algorithm and then discuss the algorithm based on conditional mutual information tests. We provide performance guarantees for both these algorithms and in particular, show that the conditional mutual information test leads to consistent structure learning throughout the regime of correlation decay.

3.1. *Assumptions.* The following assumptions are made under which structure learning is carried out.

(A1) **High Dimensionality:** We consider the asymptotic setting where both the number of variables (nodes) n and the number of samples m go to infinity. Our proposed algorithms have low sample complexity meaning that the number of variables is allowed to be large compared to the number of samples $n \gg m$. The precise sample complexity is specified later. We require the number of nodes $n \to \infty$ to exploit the locally tree-like property of random graphs.

(A2) **Random Graph:** As stated, earlier, we consider Ising models which are Markov on a realization of the random graph $G_n \sim \mathcal{G}(n, \frac{c}{n})$, for some constant $c > 0$.

(A3) **Non-singularity:** The minimum potential on the edges of the Ising model satisfies

$$J_{i,j} \geq J_{\min} > 0, \quad \forall\,(i,j) \in G_n,\, n \in \mathbb{N}.$$

In other words, the Ising model is strictly *ferromagnetic*.

[11] We will sometimes denote the estimator as $\widehat{G}_n$ (instead of $\widehat{G}_n^m$) for simplicity.

(A4) **Regime of Correlation Decay:** The average node degree c of the graph G_n and the maximum potential $J_{\max}$ satisfy

$$\alpha := c \tanh J_{\max} < 1 \tag{13}$$

The models satisfying the condition in (13) belong to the so-called *uniqueness regime* and have a decay of long-range point-to-set correlations [15].

Assumption (A1) is concerned with the sample complexity. The sample size needs to be sufficiently large with respect to the number of variables in the model, and in fact, we show in Theorem 3 that $m = \Omega(c \log n)$ is necessary for any algorithm for consistent structure reconstruction. Assumption (A3) is required so that there are no "weak" edges with low correlations. Such assumptions have been previously made for structure learning in other contexts, e.g., [26, 34]. The ferromagnetic assumption in (A3) and correlation decay assumption in (A4) allow us to obtain bounds on correlations and mutual information quantities.

3.2. *Algorithms for Graph Reconstruction.*

3.2.1. *Correlation Thresholding Algorithm.* We first provide results for the simple correlation thresholding algorithm, provided in Algorithm 1. We denote $\mathsf{CT}(\{\widehat{C}^m_{i,j}\}; \delta)$ as the output edge set due to correlation thresholding, based on the set of empirical correlations $\{\widehat{C}^m_{i,j}\}_{i,j \in V}$, formed from m samples $\mathbf{x}^m$, and a threshold $\delta > 0$.

The threshold δ is chosen as follows: assuming that the minimum and maximum potentials $J_{\min}$ and $J_{\max}$ are known[12] and that[13] $\tanh J_{\min} > \tanh^2 J_{\max}$, define the constants $\zeta(J_{\min}, J_{\max})$, $\zeta_2(J_{\min}, J_{\max})$, and $\eta(J_{\min}, J_{\max})$ as

$$\zeta(J_{\min}, J_{\max}) := \min\{k \geq 2 : 2 \tanh^k J_{\max} < \tanh J_{\min}\}, \tag{14}$$

$$\zeta_2(J_{\min}, J_{\max}) := \min\{k \geq 2 : \tanh^k J_{\max} < \tanh J_{\min} - \tanh^2 J_{\max}\}, \tag{15}$$

$$\eta(J_{\min}, J_{\max}) := \max(2 \tanh^{\zeta} J_{\max}, \tanh^2 J_{\max} + \tanh^{\zeta_2} J_{\max}). \tag{16}$$

Note that ζ, ζ_2, η are positive and bounded when $J_{\min}$ and $J_{\max}$ are also positive and bounded. For the CT algorithm, we choose the threshold $\delta > 0$

[12] Our next algorithm CMIT, removes the assumption that these parameters are known and does not require the knowledge of *any* model parameters.

[13] When $\tanh J_{\min} \leq \tanh^2 J_{\max}$, we can redefine δ as $\frac{1}{2}(2 \tanh^{\zeta} J_{\max} + \tanh J_{\min})$, but results in worse edit-distance guarantees than in Theorem 1 for correlation thresholding.

Algorithm 1 Algorithm $\mathsf{CT}(\{\widehat{C}^m_{i,j}\};\delta)$ for structure learning of G_n from empirical correlations $\{\widehat{C}^m_{i,j}\}_{i,j\in V}$ using m samples and threshold $\delta(J_{\min}, J_{\max})$ according to (17).

```
Correlation Thresholding: For each $i, j \in V$, if $\widehat{C}^m_{i,j} > \delta$, add $(i, j)$ to $\widehat{G}^m_n$.
Output: $\widehat{G}^m_n$.
```

according to

$$\delta(J_{\min}, J_{\max}) = \frac{\tanh J_{\min} + \eta(J_{\min}, J_{\max})}{2}. \tag{17}$$

3.2.2. *Guarantees for Correlation Thresholding.* Assuming (A1) – (A4), and with the number of samples m satisfying

$$m = \Omega(\log n), \tag{18}$$

where n is the number of variables in the model, we have the following sufficiency result for structure reconstruction.

Theorem 1 (Structural consistency of CT). *For structure learning of the Ising model on the random graph $G_n = (V_n, E_n) \sim \mathcal{G}(n, \frac{c}{n})$ and threshold $\delta(J_{\min}, J_{\max})$ chosen according to (17),*

1. *If $\zeta(J_{\min}, J_{\max}) = 2$, then the CT algorithm is consistent for structure reconstruction for a.e. G_n, i.e.,*

$$\lim_{\substack{m,n\to\infty\\ m=\Omega(\log n)}} \mathbb{P}_{\mathbf{X}^m_n|G_n}\left[\mathsf{CT}(\{\widehat{C}^m_{i,j}\};\delta) \neq G_n\right] = 0 \tag{19}$$

2. *If $\zeta(J_{\min}, J_{\max}) > 2$ and $\tanh J_{\min} > \tanh^2 J_{\max}$, the edit distance between the true graph structure G_n and the estimated structure $\mathsf{CT}(\{\widehat{C}^m_{i,j}\};\delta)$ is finite for a.e. G_n, with*

$$\lim_{\substack{m,n\to\infty\\ m=\Omega(\log n)}} \mathbb{P}_{\mathbf{X}^m_n|G_n}\left[\left|\mathsf{CT}(\{\widehat{C}^m_{i,j}\};\delta) \setminus G_n\right| > \omega(1)\right] = 0, \tag{20}$$

$$\lim_{\substack{m,n\to\infty\\ m=\Omega(\log n)}} \mathbb{P}_{\mathbf{X}^m_n|G_n}\left[\left|G_n \setminus \mathsf{CT}(\{\widehat{C}^m_{i,j}\};\delta)\right| \neq 0\right] = 0 \tag{21}$$

Hence, asymptotically, CT recovers a supergraph of the original graph G_n with constant number of additional edges. The proof of the theorem is given in Section 5.2.

Remarks:

1. Correlation thresholding is consistent when $2\tanh^2 J_{\max} < \tanh J_{\min}$ which means that the correlations are sufficiently weak, and the potentials $J_{i,j}$ on different edges are nearly homogeneous and sufficiently weak.
2. When the above condition is not satisfied, but $\tanh J_{\min} > \tanh^2 J_{\max}$ and in the correlation-decay regime, correlation-thresholding results in a constant edit distance with respect to the original graph. The condition $\tanh J_{\min} > \tanh^2 J_{\max}$ is a natural condition to impose, since under this condition, correlation thresholding produces a consistent structure estimate on a tree model.
3. Correlation thresholding requires only pairwise statistics and not any higher order statistics, and has a low computational complexity of $O(n^2)$. Moreover, the algorithm has logarithmic sample complexity, as in (18). We later provide a matching lower bound for random graph reconstruction in Section 3.3.
4. Note that the homogeneity condition for consistency of correlation thresholding, given by $2\tanh^2 J_{\max} < \tanh J_{\min}$, is less stringent than that needed for general degree-bounded graphs [1], given by $\tanh J_{\max} < \frac{1}{2\Delta}$. Since the maximum degree $\Delta = \Theta(\frac{\log n}{\log \log n})$ for a.e. $G_n \sim \mathcal{G}(n, c/n)$ [3, Ex. 3.6], the condition in [1] thus leads to a decaying condition on the maximum potential, which is very restrictive.

Performance of Chow-Liu Structure Learning Algorithm

We now explore the performance of Chow-Liu algorithm for learning structure of Ising models on random graphs. Since the Chow-Liu algorithm outputs a tree, it cannot recover all the edges of the random graph. We are thus interested in analyzing the number of wrong edges produced by the Chow-Liu algorithm. This analysis is relevant since Chow-Liu algorithm is simple to implement, without requiring the knowledge of any model parameters.

Recall that the Chow-Liu algorithm is the maximum likelihood algorithm for tree structure learning and outputs the maximum weighted spanning tree, where the edge weights are the empirical mutual information quantities. This algorithm was recently extended to consistently learning forest distributions [38] by pruning out weak edges in the learnt tree based on a threshold τ_m, which depends on the number of samples m. A convenient choice for τ_m is shown to be $m^{-\beta}$ for some $\beta \in (0, 1)$. Note that the above threshold does not require the knowledge of model parameters (such as $J_{\min}, J_{\max}$ or c) and thus this algorithm is parameter-free. We consider a variant of this algorithm, where the weights are empirical correlations[14], which is also consistent for tree models. After building such a tree, "weak" edges are pruned out using the threshold τ_m to produce a forest. Let the resulting output be denoted as $\mathsf{CLThres}(\{\widehat{C}^m_{i,j}\}_{i,j\in V};\tau_m)$ [38]. The procedure is summarized in Algorithm 2.

For a general graph, no correspondence can be made between the output structure of CLThres and the underlying graph of the Ising model. However, for Ising models on random graphs, we obtain an interesting result that the forest structure is asymptotically contained in the original graph under certain conditions. Hence there are no spurious edges in the output tree asymptotically.

On lines of Theorem 1, assuming (A1) – (A4) and the number of samples m exceeding (18), we have the following result.

COROLLARY 1 (Guarantees for CLThres). *For Ising model on a random graph $G_n = (V_n, E_n) \sim \mathcal{G}(n, \frac{c}{n})$ and threshold τ_m chosen according to [38],*

1. *If $\zeta(J_{\min}, J_{\max}) = 2$, then the* CLThres *algorithm recovers a subgraph for a.e. G_n, i.e.,*

$$\lim_{\substack{m,n\to\infty\\ m=\Omega(\log n)}} \mathbb{P}_{\mathbf{X}^m_n|G_n}\left[\mathsf{CLThres}(\{\widehat{C}^m_{i,j}\};\tau_m) \subset G_n\right] = 1 \tag{22}$$

2. *If $\zeta(J_{\min}, J_{\max}) > 2$,* CLThres *outputs a forest satisfying*

$$\lim_{\substack{m,n\to\infty\\ m=\Omega(\log n)}} \mathbb{P}_{\mathbf{X}^m_n|G_n}\left[\left|\mathsf{CLThres}(\{\widehat{C}^m_{i,j}\};\tau_m) \setminus G_n\right| > \omega(1)\right] = 0. \tag{23}$$

Hence, simple algorithms based on empirical correlations have good performance for structure reconstruction in the regime of correlation decay for Ising models on random graphs.

The proof of the above result in (22) follows directly from the proof of Theorem 1. The proof of the result in (23) is in Section 5.2.4.

Remarks:

1. The CT algorithm in the previous section outputs a graph with a constant number of spurious edges, as in (20), under the additional condition that $\tanh^2 J_{\max} < \tanh J_{\min}$, while the Chow-Liu algorithm

[14] For symmetric Ising models on trees, it can be shown that the mutual information is a monotonic function in the correlation and hence, the two trees are equivalent. However, this is not true in general.

Algorithm 2 Algorithm $\mathsf{CLThres}(\{\widehat{C}^m_{i,j}\}_{i,j\in V};\tau_m)$ outputs a forest structure using empirical correlation coefficients $\{\widehat{C}^m_{i,j}\}_{i,j\in V}$ from m samples and threshold τ_m.

```
1. Construct the Chow-Liu tree $\widehat{T}$ or the maximum weight spanning tree using weights
   $\{\widehat{C}_{i,j}\}_{i,j\in V}$
2. Prune the Chow-Liu tree by retaining only those edges of $\widehat{T}$ whose empirical cor-
   relation $\widehat{C}_{i,j} > \tau_m$ to obtain forest $\widehat{F}$
                              Output: $\widehat{F}$.
```

removes this requirement to obtain (23). On the other hand, the CT algorithm recovers all the edges in the original graph asymptotically, as in (21), while the Chow-Liu algorithm does not have this guarantee.

2. The Chow-Liu algorithm is parameter-free and does not require the knowledge of $J_{\min}$ and $J_{\max}$, while the CT algorithm requires these parameters for the design of the threshold.

3.2.3. *Conditional Mutual Information Thresholding.* The simple correlation thresholding algorithm CT proposed in the previous section is consistent for graph reconstruction under an additional condition $(2\tanh^2 J_{\max} < \tanh J_{\min})$. We now propose an algorithm, termed as conditional mutual information thresholding (CMIT) which is proven to be consistent for graph reconstruction throughout the regime of correlation decay.

The procedure for CMIT is provided in Algorithm 3. Denote $\mathsf{CMIT}(\mathbf{x}^m;\xi_{n,m})$ as the output edge set from CMIT given m i.i.d. samples $\mathbf{x}^m$ and threshold $\xi_{n,m}$. The conditional mutual information test in the CMIT algorithm computes the empirical conditional mutual information for each node pair $(i,j)\in V^2$ and finds the conditioning set which achieves the minimum over all sets of cardinality 2. If the minimum exceeds the threshold $\xi_{n,m}$, the node pair is declared an edge.

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

$$\begin{aligned} &\xi_{n,m} = o_n(1),\ \xi_{n,m} = \omega_n(n^{-\kappa}),\ \forall \kappa > 0, \\ &\xi_{n,m} = \Omega(\frac{\log n}{m}). \qquad \text{(24a)} \end{aligned}$$

For example, when $m = \Omega(g_n \log n)$, for some sequence $g_n = \omega(1)$, we can choose $\xi_{n,m} = \frac{1}{\min(g_n,\log n)}$.

Algorithm 3 Algorithm $\mathsf{CMIT}(\mathbf{x}^m; \xi_{n,m})$ for structure learning from $\mathbf{x}^m$ samples.

```
For each $(i, j) \in \widehat{G}_n^m$, if
    $\min_{\substack{S \subset V \setminus \{i,j\} \\ |S| \leq 2}} \widehat{I}(X_i; X_j | \mathbf{X}_S) > \xi_{n,m}$,
then add $(i, j)$ to $\widehat{G}_n^m$.
Output: $\widehat{G}_n^m$.
```

Note that there is dependence on both m and n, since we need to regularize for sample size as well as the size of the graph. In other words, with finite number of samples m, the empirical conditional mutual information quantities are noisy and the threshold $\xi_{n,m}$ takes this into account via its inverse dependence on m. Similarly, as the graph size n increases, the empirical conditional mutual information decays at a ceratin rate. Hence, the threshold $\xi_{n,m}$ also depends on the graph size n. Moreover, note that for all the conditions in (24a) to be satisfied, the number of samples m should scale at least at a certain rate with respect to n, as discussed below.

3.2.4. *Consistency of Conditional Mutual Information Thresholding.* Assuming (A1) – (A4), and with the number of samples m satisfying

$$m = \omega(\log n), \tag{25}$$

where n is the number of variables in the model, we have the following sufficient condition for asymptotic graph structure recovery.

Theorem 2 (Structural consistency of CMIT). *For structure learning of the Ising model on the random graph $G_n = (V_n, E_n) \sim \mathcal{G}(n, \frac{c}{n})$, CMIT is consistent for a.e. graph G_n:*

$$\lim_{\substack{m,n \to \infty \\ m = \omega(\log n)}} \mathbb{P}_{\mathbf{X}_n^m | G_n} \left[\mathsf{CMIT} \left(\{\mathbf{x}^m\}; \xi_{n,m} \right) \neq G_n \right] = 0 \tag{26}$$

The proof of this theorem is provided in Section 5.3.

Remarks:

1. The CMIT algorithm is thus structurally consistent and has low sample complexity throughout the correlation-decay regime. Unlike the CT algorithm, the CMIT requires higher order statistics since it involves the computation of (empirical) conditional mutual information. It is not clear if consistent graph reconstruction can be accomplished throughout the correlation-decay regime using solely pairwise statistics.

2. The CMIT algorithm has sample complexity as in (25) which is slightly worse than logarithmic complexity in (18) required for the CT algorithm. The computational complexity is $O(n^4)$.
3. Although the Ising model assumed here is symmetric (zero external magnetic fields), the results for CMIT are identical for general Ising models with bounded external magnetic fields under (13). Similarly, the result also holds when the potentials $J_{i,j}$ are not necessarily positive, but when $0 < J_{\min} \leq |J_{i,j}| \leq J_{\max}$ and (13) are satisfied.
4. The algorithms proposed here require correlation decay (assumption A4). It is an open question on whether consistent reconstruction of random graphs is feasible with polynomial computational complexity and low sample complexity when there is no correlation decay.

3.3. *Lower Bound on Sample Complexity.* We have so far proposed algorithms and provided performance guarantees for random graph reconstruction. We now provide a lower bound on sample complexity for random graph reconstruction by any algorithm. Recall that n is the number of nodes in the model and m is the number of samples. In the following result, c is allowed to depend on n and is thus more general than the previous results.

Theorem 3 (Lower bound on sample complexity). *Assume that $c \leq 0.5n$ and $G_n \sim \mathcal{G}(n, c/n)$. Then if $m \leq \epsilon c \log n$ for sufficiently small $\epsilon > 0$, we have*

$$\lim_{n\to\infty} \mathbb{P}_{\mathbf{X}_n^m|G_n}[\widehat{G}_n^m(\mathbf{X}_n^m) \neq G_n] = 1 \tag{27}$$

for any estimator $\widehat{G}_n$.

The proof of this theorem can be found in Section 5.5, and is on lines of [5, Thm. 1].

Remarks:

1. Thus, $m = \Omega(c \log n)$ number of samples are *necessary* for asymptotic 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, our result is a strong converse, as in [28], 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, a weak converse based on the Fano's inequality [10] is too loose to be applicable to our problem.

2. In [36], it is shown that for graphs uniformly drawn from the class of graphs with maximum degree Δ, when $m < \epsilon \Delta^k \log n$ 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 non-uniform for the class of random graphs.
3. The result is not dependent on the potentials $J_{i,j}$ (for example, the model need not be ferromagnetic). In fact, the result is not restricted to Ising models, and it holds for *any* pairwise discrete Markov random field (i.e., $\mathcal{X}$ is a finite set). The result also does not require locally-tree like property and is valid for dense graphs as well, i.e., c can be any function of n satisfying $c < 0.5n$.

3.4. *Discussion and Proof Steps.*

3.4.1. *Correlation Thresholding.* The correlation thresholding algorithm CT is based on the intuition that edges tend to have higher correlations than those between non-adjacent node pairs. Note that this property is true in (nearly) homogeneous Ising models on trees but need not hold for general graphical models. The key aspect is to establish correlation bounds by using the properties of paths and cycles in random graphs. The main steps in the proof of Theorem 1 are as follows:

1. We first summarize some preliminary results for random graphs in Section 5.1.1, and then review the results for ferromagnetic Ising models in Section 5.1.2.
2. We derive bounds on correlation between any node pair under a ferromagnetic Ising model Markov on a general graph in Section 5.2.1. This is accomplished via the construction of the *self-avoiding walk tree* (SAW) [42]. The bound involves the shortest-path correlation, the number of paths between the two nodes within any chosen distance, and the size of the boundary at that distance. The distance needs to be chosen appropriately (depending on the graph) to obtain a tight bound.
3. We specialize the correlation bounds for random graphs in Section 5.2.2 using the properties of paths and cycles in random graphs. In particular, we choose an appropriate distance for the boundary to obtain a tight bound on correlation. We crucially use the property that in a.e. random graph, there are no overlapping cycles of length less than $\epsilon \log n$, for a sufficiently small constant $\epsilon > 0$ (the so-called "short" distances). Equivalently, there are at most two short paths between any two nodes in the random graph. The boundary is chosen such

that only these short paths are counted towards the upper bound on correlation.

4. Using the bounds on correlation, we analyze the error events for the correlation-thresholding algorithm, first under exact statistics, and then under sample statistics in Section 5.2.3. For the former analysis, we characterize the node pairs whose correlations exceed the threshold δ of the correlation-thresholding algorithm in (17). For the sample-based analysis, we employ the standard concentration inequalities [14].

3.4.2. *Conditional Mutual Information Thresholding.* The conditional mutual information test is based on the following property: for a fixed graphical model $P(\mathbf{x})$ Markov on graph $G = (V, E)$, given any two non-adjacent nodes, we have

$$I(X_u; X_v | \mathbf{X}_S) = 0, \quad \forall\, (u, v) \notin E, \tag{28}$$

where $S \subset V$ is the separator set with respect to u and v on G. On the other hand, for adjacent nodes[15] we have

$$I(X_u; X_v | \mathbf{X}_S) > 0, \quad \forall\, (u, v) \in E, \forall\, S \subset V, \tag{29}$$

for the Ising model under consideration with $0 < J_{\min}, J_{\max} < \infty$, as assumed earlier.

The mutual information relationships in (28) and (29) can be used for graph reconstruction as follows: for any node pair u, v, find the minimum mutual information by conditioning on all possible subsets. This is efficient only if the separator sets in (28) are bounded (with known bounds), otherwise this requires exponential time in the number of nodes. Moreover, when only samples are available, computing empirical mutual information, even for a specific conditioning set, requires exponential computational and sample complexities in the size of the conditioning set [7]. Hence, the criteria in (28) and (29) cannot be used directly used for efficient structure estimation.

Instead, we show that for Ising models on random graphs under correlation decay, there exist a sparse approximate separator for any non-adjacent node pair. Such an approximate separator is a subset of the exact separator, consisting only of nodes on short paths (the precise definition is provided in Definition 2). Under the correlation-decay condition in (13), we show that the approximate-separator set suffices to distinguish edges from non-adjacent node pairs for an Ising model on a random graph. The main steps in the proof of Theorem 2 are as follows:

[15] Eqn. (29) is not true for general models and we prove that it holds for Ising models under the given assumptions.

1. Conditional mutual information between any two nodes on general graphs are related to total variation distance on the self-avoiding walk tree [42] via information-theoretic inequalities. Bounds on the self-avoiding walk tree are then derived on similar lines, as in proof of Theorem 1.
2. Asymptotically almost-sure (a.a.s.) bounds on the size of approximate-separator sets in random graphs are derived. This uses the fact that there are no overlapping short cycles in random graphs. Note that this is the reason we constrain the conditional set for mutual information in CMIT algorithm to be of cardinality 2.
3. Concentration bounds for the empirical conditional mutual information quantities are derived using the method of types [10, Chapter 12]. Combining the above results, a bound on the probability of the error for structure reconstruction is obtained.

3.4.3. *Lower Bound on Random Graph Reconstruction.* The proof for the converse uses an information-theoretic covering argument, as in [5, Thm. 1] and [38, Thm. 7].

1. The error probability for structure estimation is lower bounded by the likelihood of graphs occurring outside the range of the estimator. Note that the size of the range of any estimator is at most $|\mathcal{X}|^{nm}$, where n is the size of the graph and m is the number of samples and $|\mathcal{X}| = 2$ for Ising models.
2. We find the covering of the range of the estimator by the set of random graphs with large likelihoods, and using standard bounds on the tails of binomial distributions, we obtain the desired result.

4. Conclusion.

In this work, we considered structure estimation of Ising models Markov on sparse random graphs. We proved that consistent structure estimation is tractable with almost order-optimal sample complexity throughout the regime of correlation decay. There are many other open questions at the intersection of statistical learning theory and statistical physics. Perhaps the most important one is the feasibility of consistent structure learning in growing graphical models beyond the regime of correlation decay.

Acknowledgement. The authors thank Sujay Sanghavi (U.T. Austin) for suggesting the problem, and for extensive discussions and comments, Jorge Silva (University of Chile) for help on bounds on empirical mutual information, Sebastien Roch (UCLA) and Rui Wu (UIUC) for pointing out errors in Theorem 1, Elchanan Mossel (Berkeley) for pointing out additional references and suggesting future directions, and Béla Bollobás (Cambridge), Paul Balister (Univ. of Memphis), and Oliver Riordan (Oxford) for extensive discussions on random graphs.

5. Detailed Proofs.

Notation. We let $\mathrm{path}(i,j;G_n) = \mathrm{path}_1(i,j;G_n)$ denote the subgraph spanning the corresponding shortest path and $d(i,j;G_n) := |\,\mathrm{path}(i,j;G_n)|$ denote the graph distance or the shortest path distance between nodes i and j. Let the set of nodes at distance[16] exactly l from i in G_n be denoted as

$$B_l(i;G_n) := \{k \in V_n : d(i,k;G_n) = l\}. \tag{30}$$

Let $\mathrm{path}_l(i,j;G_n)$ denote[17] the l^{th} shortest path from i to j and $d_l(i,j;G_n)$ the corresponding length of the path. Let $N_l^{\mathrm{path}}(i,j;G_n)$ denote the number of paths of length l from node i to node j in G_n without repeating any node in the intermediate steps. Hence, $N_l^{\mathrm{cycles}}(i;G) := N_l^{\mathrm{path}}(i,i;G)$ denotes the number of cycles passing through node i of length l and let $N_l^{\mathrm{cycles}} := \sum_{i\in V} N_l^{\mathrm{cycles}}(i)$ denote the total number of cycles, and let $\mathcal{C}_l$ denote the (unlabeled) graph which is a cycle of length l. Let $\mathrm{Diam}(G_n)$ denote the graph diameter. We say that two subgraphs $H_1, H_2 \subset G$ *overlap* if the vertices spanned by H_1 and H_2 satisfy $v(H_1)\cap v(H_2) \neq \emptyset$, otherwise the two subgraphs are disjoint.

5.1. *Preliminaries.*

5.1.1. *Properties of Random Graphs.* We first study the properties of random graphs. Let $H_{k_1,k_2,l}$ denote a graph which is the union of two cycles each of length $k_1, k_2 \geq 3$ separated by a path of length at most l. That is, if $\mathcal{C}(k_1)$ and $\mathcal{C}(k_2)$ are the two cycles then,

$$\min_{a\in\mathcal{C}(k_1),b\in\mathcal{C}(k_2)} d(a,b;H) = l. \tag{31}$$

The graph $H_{k_1,k_2,l}$ is also known as a *generalized cycle*. Let $N_{H_{k_1,k_2,l}}$ denote the number of $H_{k_1,k_2,l}$ subgraphs in G_n.

Lemma 1 (Cycles and Paths in Random Graphs). *In $G_n \sim \mathcal{G}(n, \frac{c}{n})$,*

[16] We follow the convention that if l is not an integer, the distance is $\lfloor l \rfloor$.

[17] We abbreviate $\mathrm{path}_1(i,j;G_n)$ as $\mathrm{path}(i,j;G_n)$ and $d_1(i,j;G_n)$ as $d(i,j;G_n)$.

1. *For uniformly chosen nodes i, j such that $d(i, j; G_n) = l$ for r any $1 \leq l \leq n-1$ and any $k \geq l$, we have*
(32)
$$\mathbb{E}_{\mathcal{G}_n}[N_k^{\text{path}}(i,j;G_n)|d(i,j;G_n) = l] = \mathbb{I}\{k = l\} + O(n^{-2}c^{k+l}(k+l)).$$
The expectation in (32) is taken with respect to the Erdős-Rényi random graph.
2. *The number of cycles of lengths $3, 4, \dots, k$ are asymptotically independent Poisson random variables with mean values $\frac{c^3}{6}, \dots, \frac{c^k}{2k}$ respectively for $k = o(\log n)$. Moreover, there are no subgraphs $H_{k_1,k_2,l}$ in a.e. graph G_n when $k_1, k_2, l < \frac{\log n}{3 \log c}$, i.e.,*
$$N_{H_{k_1,k_2,l}} = o(1), \quad a.a.s., \; \forall\, k_1, k_2, l < \frac{\log n}{3 \log c}. \tag{33}$$

Remarks: It is thus unlikely that for a node pair at short distances to have other short paths. Moreover, short-length cycles are unlikely to overlap asymptotically.

Proof: By a counting argument,

$$\begin{aligned}\mathbb{E}_{\mathcal{G}_n}[N_k^{\text{path}}(i,j;G_n)|d(i,j;G_n) = l] &= \mathbb{I}\{k = l\} + \mathbb{E}_{\mathcal{G}_n}[N_{k+l}^{\text{cycles}}(\{i,j\};G_n)] \\ &= \mathbb{I}\{k = l\} + \binom{n-2}{l+k-2}\frac{(l+k)!}{2(l+k)}\left(\frac{c}{n}\right)^{l+k} \\ &= \mathbb{I}\{k = l\} + O(n^{-2}c^{l+k}(l+k-1)),\end{aligned}$$

where $N_{k+l}^{\text{cycles}}(\{i,j\};G_n)$ denotes the number of cycles of length $k+l$ containing both i and j.

The second result is from [3, Cor. 4.9], but we repeat it here specific to cycles. It suffices to show that for any increasing sequence of numbers $1 \leq l_1 \leq l_2 \ldots \leq l_r \leq k$,

$$\mathbb{E}\left[\prod_{i=1}^{r} N_{l_i}^{\text{cycles}}\right] = \prod_{i=1}^{r} O\left(\frac{c^{l_i}}{2l_i}\right), \quad \forall r \in \mathbb{N}. \tag{34}$$

Let $\mathbb{E}'[\prod_{i=1}^r N_{l_i}^{\text{cycles}}]$ denote the expectation by considering only cycles which span disjoint vertex sets. We have

$$\mathbb{E}'\left[\prod_{i=1}^{r} N_{l_i}^{\text{cycles}}\right] = O\left(\prod_{a=0}^{r-1}\binom{n-\sum_{i=1}^{a} l_i}{l_{a+1}}\frac{l_i!}{2l_i}\left(\frac{c}{n}\right)^{l_i}\right) = O\left(\prod_{i=1}^{r}\frac{c^{l_i}}{2l_i}\right). \tag{35}$$

Therefore, it suffices to show that $\mathbb{E}''[\prod_{i=1}^r N_{l_i}^{\text{cycles}}] := \mathbb{E}[\prod_{i=1}^r N_{l_i}^{\text{cycles}}] - \mathbb{E}'[\prod_{i=1}^r N_{l_i}^{\text{cycles}}] = o(1)$. Let $F_1, \ldots, F_r$ be r cycles of lengths $l_1, \ldots, l_r$ with $|v(\cup_{i=1}^r F_i)| = s < \sum_i l_i$ and $s > l_r$. We now claim that for such overlapping cycles, the edges satisfy

$$\left| \bigcup_{i=1}^r F_i \right| \geq s + 1. \tag{36}$$

The above statement is proven as follows: let t_j be the number of vertices in F_j not belonging to $\cup_{i=1}^{j-1} F_i$ (and setting $t_1 = l_1$). Hence, the number of edges satisfies

$$\left| \bigcup_{i=1}^j F_i \right| \geq \left| \bigcup_{i=1}^{j-1} F_i \right| + t_j + 1, \tag{37}$$

and noting that $\sum_{i=1}^r t_i = s$. Hence, the expected number of overlapping cycles is

$$\mathbb{E}'' \left[\prod_{i=1}^r N_{l_i}^{\text{cycles}} \right] \leq \binom{n}{s} s! \left(\frac{c}{n} \right)^{s+1} = O(n^{-1} c^{s+1}). \tag{38}$$

The above term is $o(1)$ when $kr < \frac{\log n}{\log c}$ for all $r \in \mathbb{N}$, which is true when $k = o(\log n)$.

We now prove (33) along the same lines as the previous proof. Let $|v(H_{k_1,k_2,l})| = s$ with $\max(k_1, k_2) \leq s \leq k_1 + k_2 + l$. Note that the number of edges $|H_{k_1,k_2,l}| \geq s + 1$. Hence,

$$\begin{aligned} \mathbb{E}[N_{H_{k_1,k_2,l}}] &\leq \binom{n}{s} s! \left(\frac{c}{n} \right)^{s+1} \\ &= O(n^{-1} c^{s+1}), \end{aligned}$$

which is $o(1)$ when $c^s = o(n)$ implying that $\max(k_1, k_2, l) = o(\frac{\log n}{3 \log c})$.

□

We now recall the result of [30, Lemma 2.6] which provides a bound on $|B_l(i; G_n)|$, the size of l-hop neighborhood of i for a random graph G_n.

LEMMA 2 (Self Avoiding Walk Tree of Random Graph [30]). *For $1 \leq l \leq a \log n$, where $0 < a < \frac{1}{2 \log c}$, we have*

$$\max_{i \in G_n} |B_l(i)| = O(c^l \log n), \quad a.a.s,\ G_n \sim \mathcal{G}(n, \frac{c}{n}). \tag{39}$$

5.1.2. *Preliminaries on Ferromagnetic Ising Models.* We require Griffiths second inequality [17] for our analysis.

Theorem 4 (Griffiths Second Inequality [17]). *For two ferromagnetic Ising models Markov on same graph $G = (V, E)$ with potentials $0 < J_{i,j} \leq J'_{i,j}$ for all $(i,j) \in E$, we have*

$$\mathbb{E}\left[\prod_{i \in U} X_i; G, \mathbf{J}\right] \leq \mathbb{E}\left[\prod_{i \in U} X_i; G, \mathbf{J}'\right], \quad \forall U \subset V. \tag{40}$$

In particular, this means that if the potentials of a model are increased, then the correlations $\mathbb{E}[X_i X_j; G, \mathbf{J}]$ are also increased.

We also note that the correlation between any two node pairs has a simple expression in symmetric (i.e., zero-field) Ising models.

Fact 1 (Correlation for Ising Models). *For the Ising model in (11), the correlation in (1) simplifies to*

$$C(i,j; G, \mathbf{J}) := \frac{1}{2}\left(\mathbb{E}[X_i | X_j = +; G, \mathbf{J}] - \mathbb{E}[X_i | X_j = -; G, \mathbf{J}]\right). \tag{41}$$

Proof: The mean value at any node $\mathbb{E}[X_i; G]$ is zero for the Ising model in (11) since $P_{\mathbf{X}}(\mathbf{x}; G) = P_{\mathbf{X}}(-\mathbf{x}; G)$ for all $\mathbf{x} \in \mathcal{X}^n$. Hence, the simplification in (41). □

Note that (41) can be interpreted as the total variation distance between the distributions at node i due change of configuration at node j.

We first study Ising models Markov on a tree. The following properties will be used for the analysis of Ising models on random graphs.

Fact 2 (Markov Property for Correlations on a Tree). *For an Ising model on a tree T, the correlation is given by*

$$C(i,j; T, \mathbf{J}) = \prod_{(k,l) \in \text{path}(i,j;T)} C(k,l; T), \quad \forall i,j \in V, \tag{42}$$

and the correlation between any two neighbors is,

$$C(i,j; T, \mathbf{J}) = \tanh(J_{i,j}), \quad \forall\, (i,j) \in T. \tag{43}$$

Proof: Eqn. (42) is obtained by successive conditioning on the intermediate nodes in the path between i and j in the tree T. Eqn. (43) is a consequence of the form of the Ising model in (11). □

5.2. *Proof of Theorem 1.* Let

$$C_{\mathrm{SP}}(i,j;G_n,\mathbf{J}) := C(i,j;\mathrm{path}(i,j;G_n),\mathbf{J}) = \prod_{(k,l)\in \mathrm{path}(i,j;G_n)} \tanh J_{k,l}. \tag{44}$$

Hence, $C_{\mathrm{SP}}(i,j;G_n,\mathbf{J})$ denotes the correlation between i and j, when the graph is the shortest path between i and j in the graph G_n and hence, we term this as the “shortest-path” correlation. Note that this also turns out to be the true correlation when there is no other path from i and j, i.e., i and j are not part of any cycle. We now provide bounds on correlations between any two nodes in a general graph based on the shortest-path correlations. This is achieved via the self-avoiding walk tree construction.

5.2.1. *Correlation on Graphs via Self-Avoiding Walk Trees.* We first introduce some notation.

Notation for Self-Avoiding Walk Trees: Recall that $N_l^{\mathrm{path}}(i,j;G)$ denotes the number of paths of length l from i to j in G and $d(i,j;G)$ denotes graph distance between i and j. Let $T_{\mathrm{saw}}^{(i,G)}$ denote the self-avoiding walk tree [42] rooted at node i. See Fig. 1 for an illustration. Let

$$\mathcal{U}(j;T_{\mathrm{saw}}^{(i,G)}) = \{j_1,\ldots,j_{|\mathcal{U}(j;T_{\mathrm{saw}}^{(i,G)})|}\} \subset v(T_{\mathrm{saw}}^{(i,G)}) \tag{45}$$

denote the set of copies of node j in the self-avoiding walk tree $T_{\mathrm{saw}}^{(i,G)}$. The definition is extended to sets $A\subset V$ as $\mathcal{U}(A;T_{\mathrm{saw}}^{(i,G)}) := \cup_{a\in A}\mathcal{U}(a;T_{\mathrm{saw}}^{(i,G)})$.

The set of nodes in the self-avoiding walk tree that are at distance l from i is denoted as

$$B_l(i;T_{\mathrm{saw}}^{(i,G)}) := \left\{k\in T_{\mathrm{saw}}^{(i,G)} : d(i,k;T_{\mathrm{saw}}^{(i,G)}) = l\right\}. \tag{46}$$

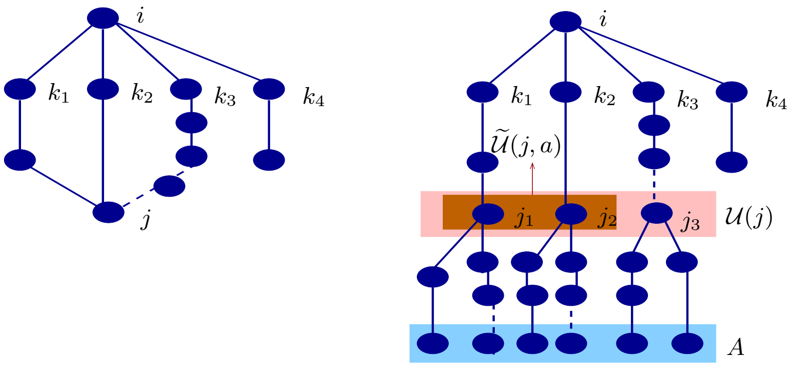
Define the subset of copies of any node j in the self-avoiding walk tree of distance less than a as

$$\widetilde{\mathcal{U}}(j,a;T_{\mathrm{saw}}^{(i,G)}) := \{k\in \mathcal{U}(j;T_{\mathrm{saw}}^{(i,G)}) : d(i,j_k;T_{\mathrm{saw}}^{(i,G)}) \le a\}. \tag{47}$$

We now have the result on bounds on the pairwise correlations.

Lemma 3 (Bounds on Correlation). *For an Ising model on a graph $G=(V,E)$, the correlation between any two nodes $i,j\in V$ satisfies*

$$\begin{aligned}0 \le{}& C(i,j;G,\mathbf{J}) - C_{\mathrm{SP}}(i,j;G,\mathbf{J})\\ \le{}& \inf_{\substack{a\in\mathbb{N}\\ a\ge d(i,j)}}\left\{\sum_{l=d(i,j)}^{a}\left((N_l^{\mathrm{path}}(i,j;G) - \mathbb{I}\{l=d(i,j)\})(\tanh J_{\max})^l\right)\right.\\ &\left. +|B_a(i;T_{\mathrm{saw}}^{(i,G)})|(\tanh J_{\max})^a\right\}.\end{aligned} \tag{48}$$



(a) A graph

(b) Its Self-Avoiding Walk Tree

FIG 1. *The figure on the right is self-avoiding walk tree $T_{\rm saw}^{(i,G)}$ rooted at node i for the graph shown in the left. The dotted line represents a long path with a large number of nodes. The copies of node j in the self-avoiding walk tree is $\mathcal{U}(j;T_{\rm saw}^{(i,G)})=\{j_1,j_2,j_3\}$. The subset of copies of the node j in $T_{\rm saw}^{(i,G)}$ of distance less than $a=4$ is $\widetilde{\mathcal{U}}(j,a;T_{\rm saw}^{(i,G)})=\{j_1,j_2\}$, defined in (47). Set A is the set of terminal nodes in $T_{\rm saw}^{(i,G)}$.*

Proof: Consider the self-avoiding walk tree $T_{\rm saw}^{(i,G)}$ rooted at node i. There exists a set of terminal nodes $A\subset v(T_{\rm saw}^{(i,G)})$ and a configuration $\mathbf{x}_A$ such that [42]

$$\begin{aligned}C(i,j;G):=&\frac{1}{2}(\mathbb{E}[X_i|X_j=+;G]-\mathbb{E}[X_i|X_j=-;G])\\ =&\frac{1}{2}(\mathbb{E}[X_i|\mathbf{X}_{\mathcal{U}(j)}=+,\mathbf{X}_A=\mathbf{x}_A;T_{\rm saw}^{(i,G)}]\\ &-\mathbb{E}[X_i|\mathbf{X}_{\mathcal{U}(j)}=-,\mathbf{X}_A=\mathbf{x}_A;T_{\rm saw}^{(i,G)}]).\end{aligned}\tag{49}$$

Upper Bound: It is shown in [30, Lemma 2.8] that the right-hand size in (49) is upper bounded by removing the conditioning on $\mathbf{X}_A$ and hence,

$$C(i,j;G)\le\frac{1}{2}(\mathbb{E}[X_i|\mathbf{X}_{\mathcal{U}(j)}=+;T_{\rm saw}^{(i,G)}]-\mathbb{E}[X_i|\mathbf{X}_{\mathcal{U}(j)}=-;T_{\rm saw}^{(i,G)}]),\tag{50}$$

$$\begin{aligned}\le&\frac{1}{2}(\mathbb{E}[X_i|\mathbf{X}_{\mathcal{U}(j)}=+,\mathbf{X}_{B_a(i)}=+;T_{\rm saw}^{(i,G)}]\\ &-\mathbb{E}[X_i|\mathbf{X}_{\mathcal{U}(j)}=-,\mathbf{X}_{B_a(i)}=-;T_{\rm saw}^{(i,G)}]),\end{aligned}\tag{51}$$

for all $a\in\mathbb{N}$ where $B_a(i):=B_a(i;T_{\rm saw}^{(i,G)})$ is the set of nodes at distance a from i in $T_{\rm saw}^{(i,G)}$ defined in (46). The result in (51) uses the fact that for a ferromagnetic model, conditioning on a larger set of nodes increases[18] the variation distance between the conditional distributions at any node $i \in V$. Thus, we have for every $a \in \mathbb{N}$,

$$\begin{aligned}
C(i,j;G) \leq & \frac{1}{2}(\mathbb{E}[X_i | \mathbf{X}_{\widetilde{\mathcal{U}}(j,a)} = +, \mathbf{X}_{B_a(i)} = +; T_{\text{saw}}^{(i,G)}] \\
& - \mathbb{E}[X_i | \mathbf{X}_{\widetilde{\mathcal{U}}(j,a)} = -, \mathbf{X}_{B_a(i)} = -; T_{\text{saw}}^{(i,G)}]), \\
\overset{(a)}{\leq} & \prod_{\substack{(u,v) \in \text{path}(i,j_1; T_{\text{saw}}^{(i,G)}): \\ d(i,j_1; T_{\text{saw}}^{(i,G)}) = d(i,j;G)}} (\tanh J_{u,v})^{d(u,v;T_{\text{saw}}^{(i,G)})} \\
& + \sum_{\substack{k \in \widetilde{\mathcal{U}}(j,a), \\ d(i,j_k; T_{\text{saw}}^{(i,G)}) > d(i,j;G)}} (\tanh J_{\max})^{d(i,j_k;T_{\text{saw}}^{(i,G)})} \\
& + |B_a(i; T_{\text{saw}}^{(i,G)})| (\tanh J_{\max})^a.
\end{aligned}$$

Inequality (a) is obtained by telescoping from all + to all − configuration on nodes in $\widetilde{\mathcal{U}}(j,a)$ by going through configurations that only differ at a single vertex. See [30, Lemma 2.8] for the details. The final result can be obtained from the fact that correlation between two nodes in $T_{\text{saw}}^{(i,G)}$ separated by a path of length l is bounded above by $(\tanh J_{\max})^l$ and that the length of a path in $T_{\text{saw}}^{(i,G)}$ from i to a copy of j is equal to length of some path from i to j in G (without visiting any node more than once).

Lower Bound: Consider a new Ising model Markov on G with the potential matrix $\mathbf{J}'$ as

$$J'_{i,j} = \begin{cases} J_{i,j} & \text{if } (i,j) \in \text{path}(i,j;G), \\ 0 & \text{o.w.} \end{cases}$$

From the Griffiths inequality in (40), we have

$$C(i,j;G,\mathbf{J}) \geq C(i,j;G,\mathbf{J}') = C_{\text{SP}}(i,j;G,\mathbf{J}).$$

This completes the proof. □

Hence, the bounds on correlation between a node pair can be expressed in terms of the shortest-path correlation, the number of paths between the two nodes within a chosen boundary and the number of nodes on the boundary. Depending on the graph, the boundary can be appropriately chosen to obtain a tight bound on the correlation.

[18] Note that in (49), we condition on the same configuration of nodes, while in (50), we condition on nodes taking opposite signs.

5.2.2. *Ising Models on Random Graphs Under Correlation Decay.* From the previous result in Lemma 3, the upper bound on correlation can thus be expressed in terms of number of paths within an appropriately chosen boundary. In order to obtain tight bounds on correlation, the boundary must be chosen effectively, and this depends on the properties of the random graph, which were analyzed in Section 5.1.1. We now combine these results to obtain bounds on correlation for Ising models on random graphs.

Recall that a property $\mathcal{Q}$, for a sequence of random graphs $\{G_n\}_{n\in\mathbb{N}}$, holds asymptotically almost surely (a.a.s.) if, $\mathbb{P}(G_n \text{ satisfies } \mathcal{Q}) \to 1$ as $n \to \infty$. The notation $\text{path}(i,j;G_n)$ denotes the shortest path between nodes i and j in G_n and that $C_{\text{SP}}(i,j;G_n,\mathbf{J})$ in (44) denotes the "shortest path" correlations between i and j and $d_2(i,j;G_n)$ denotes the second shortest distance between i and j in G_n.

Theorem 5 (Bounds on Correlation in the Uniqueness Regime). *For an Ising model Markov on $G_n \sim \mathcal{G}(n,\frac{c}{n})$ in the uniqueness regime ($\alpha := c\tanh J_{\max} < 1$),*

1. *For two uniformly chosen nodes i and j, we have for a.e. G_n,*
$$C_{\text{SP}}(i,j;G_n,\mathbf{J}) \le C(i,j;G_n,\mathbf{J}) \le C_{\text{SP}}(i,j;G_n,\mathbf{J}) + O(n^{-2}\alpha^{2d(i,j;G_n)}). \tag{52}$$
2. *The maximum correlation between a node pair conditioned on their graph distance satisfies*
$$\max_{\substack{i,j\in V\\ d(i,j;G_n)=l}} C(i,j;G_n,\mathbf{J}) = O(\alpha^l \log n), \quad a.a.s. \tag{53}$$
3. *The above result can be strengthened when $l = o(\log n)$ as*
$$\max_{\substack{i,j\in V\\ d(i,j;G_n)=l}} C(i,j;G_n,\mathbf{J}) \le 2(\tanh J_{\max})^l + o(1), \quad a.a.s. \tag{54}$$
Moreover, if conditioned on graph distance l and second shortest distance $s \ge l$, we have for a.e. G_n
$$\max_{\substack{i,j\in V\\ d(i,j;G_n)=l,\\ d_2(i,j;G_n)=s}} C(i,j;G_n,\mathbf{J}) \le (\tanh J_{\max})^l + (\tanh J_{\max})^s + o(1). \tag{55}$$

Proof: Proof of statement (1): The lower bound in (52) is from Lemma 3. Conditioned on the fact that the minimum distance between i and j equals

l, i.e., $d(i,j;G_n) = l$, we have from (32) in Lemma 1,

$$\mathbb{E}\left[\sum_{k>l} N_k^{\text{path}}(i,j;G_n)(\tanh J_{\max})^k \middle| d(i,j;G_n) = l\right] = O(n^{-2}\alpha^{2l}).$$

The expectation above is taken with respect to the random graph. Hence, from the upper bound in (48) in Lemma 3, by taking $a \to \infty$,

$$\mathbb{E}[C(i,j;G_n,\mathbf{J})|d(i,j;G_n) = l] \leq C_{\text{SP}}(i,j;G_n;\mathbf{J}) + O(n^{-2}\alpha^{2l}).$$

From Markov's inequality, for any $\epsilon > 0$ and a positive sequence $\{\beta_n\}_{n\in\mathbb{N}}$

$$\begin{aligned}
&\mathbb{P}[C(i,j;G_n,\mathbf{J}) - C_{\text{SP}}(i,j;G_n;\mathbf{J}) > \epsilon\beta_n | d(i,j;G_n) = l] \\
&\leq \frac{1}{\epsilon\beta_n}\mathbb{E}[C(i,j;G_n,\mathbf{J}) - C_{\text{SP}}(i,j;G_n;\mathbf{J})|d(i,j;G_n) = l] \\
&= \frac{1}{\epsilon\beta_n}O(n^{-2}\alpha^{2l})
\end{aligned}$$

which is $o(1)$ when $\beta_n = \omega(n^{-2}\alpha^{2l})$. Hence, (52) holds.

Proof of statement (2): Eqn. (53) is obtained from (39), and using the fact that

$$\max_{\substack{i,j\in V\\ d(i,j;G_n)=l}} C(i,j;G,\mathbf{J}) \leq \max_{i\in V} |B_l(i;T_{\text{saw}}^{(i,G)})|(\tanh J_{\max})^l.$$

by setting $a = d(i,j;G_n)$ in (48) in Lemma 3.

Proof of statement (3): We now proceed to prove (54). From Lemma 3,

$$C(i,j;G,\mathbf{J}) \leq \inf_{a\geq d(i,j;G_n)} \Psi(a), \tag{56}$$

where

$$\Psi(a) := \sum_{k=d(i,j;G)}^{a} \left(N_k^{\text{path}}(i,j;G)(\tanh J_{\max})^k\right) + |B_a(i;T_{\text{saw}}^{(i,G)})|(\tanh J_{\max})^a.$$

Now the number of overlapping paths can be stated in terms of overlapping cycles. Recall that $H_{k_1,k_2,l}$ denotes a graph consisting of union of two cycles of lengths $k_1, k_2 \geq 3$ which are separated by a path of length at most $l \geq 0$, and $N_{H_{k_1,k_2,l}}$ denotes the number of subgraphs $H_{k_1,k_2,l}$ in G_n. Given that the minimum distance between i and j as $d(i,j;G_n) = l$, and that there is another path from i to j of length $l \leq s \leq a$, for $a > 0$, there exists a cycle of length between 3 and $l + s$ which overlaps with the shortest path, denoted by $\mathrm{path}(i, j; G_n)$. Hence, we can bound the number of paths in G_n other than $\mathrm{path}(i, j; G_n)$ from i to j as

(57)
$$\max_{\substack{i,j\in V,\\ d(i,j;G_n)=l}} \sum_{b=l}^{a} N_b^{\mathrm{path}}(i, j; G)(\tanh J_{\max})^b \leq (\tanh J_{\max})^l + \sup_{l\leq s\leq a} \Upsilon(s, a, l),$$

where

$$\Upsilon(s, a, l) := (\tanh J_{\max})^s + \sum_{t_1=3}^{l+s} \sum_{t_2=3}^{s+a} \sum_{t_3=2}^{a} N_{H_{t_1,t_2,t_3}} (\tanh J_{\max})^{t_3}$$

Let $a = o(\log n)$, say $a = O(\frac{\log n}{\log\log n})$, and by assumption $l = o(\log n)$. From Lemma 1, we have

$$\sum_{t_1=3}^{l+s} \sum_{t_2=3}^{s+a} \sum_{t_3=2}^{a} \mathbb{E}\left[N_{H_{t_1,t_2,t_3}} (\tanh J_{\max})^{t_3}\right] = o(1), \quad \forall\, l, s, a = o(\log n).$$

Hence, when $a = O(\frac{\log n}{\log\log n})$,

$$\sup_{l\leq s\leq a} \Upsilon(s, a, l) = \Upsilon(l, l, a) \overset{a.a.s}{=} (\tanh J_{\max})^l + o(1).$$

Substituting in (56), and using (39), we conclude that

$$|B_a(i; T_{\mathrm{saw}}^{(i,G)})| \overset{a.a.s}{=} O(\alpha^a \log n) \overset{a.a.s}{=} o(1),$$

when $a = O(\frac{\log n}{\log\log n})$. Hence,

$$C(i, j; G_n, \mathbf{J}) \leq 2(\tanh J_{\max})^l + o(1), \quad a.a.s,$$

thereby yielding (54). Eqn. (55) is obtained by following (57) but the value of s is constant. □

5.2.3. *Error Events for Correlation Thresholding.* We are now ready to prove Theorem 1. We first consider error events for the correlation-thresholding algorithm under exact statistics and then analyze the events under sample statistics. Define the events under exact statistics as

$$\mathcal{E}_1(i, j; G_n, \mathbf{J}) := \{C(i, j; G_n, \mathbf{J}) > \delta(J_{\min}, J_{\max})\}, \tag{58}$$
$$\mathcal{E}_2(i, j; G_n, \mathbf{J}) := \{C(i, j; G_n, \mathbf{J}) < \delta(J_{\min}, J_{\max})\} \tag{59}$$

Similarly, we let $\widehat{\mathcal{E}}_k^m$ for $k = 1, 2$ to be the events obtained by replacing the exact correlations $C(i, j; G_n, \mathbf{J})$ with empirical correlations $\widehat{C}(i, j; G_n, \mathbf{J})$, defined in (2), i.e.,

$$\widehat{\mathcal{E}}_1^m(i, j; G_n, \mathbf{J}) := \{\widehat{C}^m(i, j; G_n, \mathbf{J}) > \delta(J_{\min}, J_{\max})\}, \tag{60}$$

$$\widehat{\mathcal{E}}_2^m(i, j; G_n, \mathbf{J}) := \{\widehat{C}^m(i, j; G_n, \mathbf{J}) < \delta(J_{\min}, J_{\max})\} \tag{61}$$

Hence, the event that the estimated graph from CT is not equal to the true graph can be expressed as

$$\left\{G_n \neq \mathsf{CT}(\{\widehat{C}_{i,j}^m\}; \delta)\right\} = \left\{\bigcup_{(i,j)\notin G_n} \widehat{\mathcal{E}}_1^m(i, j)\right\} \cup \left\{\bigcup_{(i,j)\in G_n} \widehat{\mathcal{E}}_2^m(i, j)\right\}. \tag{62}$$

The *edit distance* between the output of CT and the true graph can be expressed as

$$\left|\mathsf{CT}(\{\widehat{C}_{i,j}^m\}; \delta) \triangle G_n\right| = \left|\mathsf{CT}(\{\widehat{C}_{i,j}^m\}; \delta) \setminus G_n\right| + \left|G_n \setminus \mathsf{CT}(\{\widehat{C}_{i,j}^m\}; \delta)\right|, \tag{63}$$

where

$$\left|\mathsf{CT}(\{\widehat{C}_{i,j}^m\}; \delta) \setminus G_n\right| = \sum_{(i,j)\notin G_n} \mathbb{I}(\widehat{\mathcal{E}}_1^m(i, j)), \tag{64}$$

$$\left|G_n \setminus \mathsf{CT}(\{\widehat{C}_{i,j}^m\}; \delta)\right| = \sum_{(i,j)\in G_n} \mathbb{I}(\widehat{\mathcal{E}}_2^m(i, j)) \tag{65}$$

Thus, we have

Correlation Thresholding with Exact Statistics. We first consider the graph $\mathsf{CT}(\{C_{i,j}\}; \delta)$ learned using exact correlations $C(i, j; G_n, \mathbf{J})$ and then consider the scenario where only samples are available.

For the threshold δ in (17), we have $\mathbb{I}[\mathcal{E}_2(i, j)] = 0$ for all $(i, j) \in G_n$, since $\delta < \tanh J_{\min}$ and $C(i, j; G_n) \geq \tanh J_{\min}$ for $(i, j) \in G_n$ from Theorem 5. Thus, with exact statistics, we only need to consider the event $\mathcal{E}_1(i, j)$ for $(i, j) \notin G_n$.

Recall that ζ is defined in (14) as

$$\zeta(J_{\min}, J_{\max}) := \min\{k \geq 2 : 2(\tanh J_{\max})^k < \tanh J_{\min}\}. \tag{66}$$

From Theorem 5, intuitively, we can see that any node pair i, j with shortest path $d(i, j; G_n) > \zeta$ will not be asymptotically (in the number of nodes) chosen as an edge by correlation thresholding under exact statistics. Similarly, recall that ζ_2 is defined in (15) as

$$\zeta_2(J_{\min}, J_{\max}) := \min\{k \geq 2 : (\tanh J_{\max})^k < \tanh J_{\min} - \tanh^2 J_{\max}\}. \tag{67}$$

From Theorem 5, intuitively, we can see that any node pair i, j with second shortest path $d_2(i,j;G_n) > \zeta_2$ will not be asymptotically (in the number of nodes) chosen as an edge by correlation thresholding under exact statistics. We now formally establish these results and obtain the resulting edit distance. We consider the two cases (i) $\zeta = 2$ and (ii) $\zeta > 2$ separately.

Case 1 ($\zeta = 2$)*:* The probability of not obtaining the correct graph can be upper bounded as

$$\begin{aligned}\mathbb{P}\left[\mathsf{CT}(\{C_{i,j}\};\delta) \neq G_n\right] =& \mathbb{P}\left[\bigcup_{i,j\in V, d(i,j;G_n)>1} \mathcal{E}_1(i,j;G_n,\mathbf{J})\right]\\ \leq& \mathbb{P}\left[\bigcup_{2\leq d(i,j;G_n)\leq h} \mathcal{E}_1(i,j;G_n,\mathbf{J})\right]\\ &+\mathbb{P}\left[\bigcup_{h\leq d(i,j;G_n)\leq \mathrm{Diam}(G_n)} \mathcal{E}_1(i,j;G_n,\mathbf{J})\right], \end{aligned} \tag{68}$$

where $h = o(\log n)$, say $h = O(\frac{\log n}{\log\log n})$. We now bound the first term in (68) using (54) in Theorem 5 as

$$\begin{aligned}&\mathbb{P}\left[\bigcup_{2\leq d(i,j;G_n)\leq h} \mathcal{E}_1(i,j;G_n,\mathbf{J})\right]\\ &\overset{(a)}{\leq} \mathbb{P}\left[\max_{2\leq d(i,j;G_n)\leq h}\{C(i,j;G_n,\mathbf{J}) > 2(\tanh J_{\max})^{\zeta}\}\right], &(69)\\ &\overset{(b)}{\leq} \mathbb{P}\left[\max_{2\leq d(i,j;G_n)\leq h}\{C(i,j;G_n,\mathbf{J}) > 2(\tanh J_{\max})^{d(i,j;G_n)}\}\right] \overset{(c)}{=} o(1), &(70)\end{aligned}$$

where inequality (a) follows from the choice of δ in (17) with $\delta > 2\tanh^{\zeta} J_{\max}$. Inequality (a) is obtained using $d(i,j;G_n) \geq \zeta = 2$ for non-neighbors i, j, and (c) is obtained from Theorem 5. The second term in (68) can be bounded using (53) in Theorem 5 as

$$\mathbb{P}\left[\max_{h\leq d(i,j;G_n)\leq \mathrm{Diam}(G_n)} C(i,j;G_n,\mathbf{J}) > \delta\right] = o(1). \tag{71}$$

Combining the results in (70) and (71), we conclude that $\mathbb{P}\left[\mathsf{CT}(\{C_{i,j}\};\delta) \neq G_n\right] = o(1)$.

Case 2 ($\zeta > 2$)*:* Recall that $d_2(i,j;G_n)$ is the length of the second shortest path between i and j. The edit distance satisfies

$$|\,\mathsf{CT}(\{C_{i,j}\};\delta)\setminus G_n\,| = \sum_{i,j\in V: d(i,j;G_n)>1} \mathbb{I}(\mathcal{E}_1(i,j;G_n,\mathbf{J}))$$

$$= \sum_{\substack{2\le d(i,j;G_n)\le\zeta,\\ 2\le d_2(i,j;G_n)\le\zeta_2}} \mathbb{I}(\mathcal{E}_1(i,j;G_n,\mathbf{J})) + \sum_{\substack{2\le d(i,j;G_n)\le\zeta,\\ d_2(i,j;G_n)>\zeta_2}} \mathbb{I}(\mathcal{E}_1(i,j;G_n,\mathbf{J}))$$

$$+ \sum_{\zeta+1\le d(i,j;G_n)\le h} \mathbb{I}(\mathcal{E}_1(i,j;G_n,\mathbf{J})) + \sum_{h\le d(i,j;G_n)\le \mathrm{Diam}(G_n)} \mathbb{I}(\mathcal{E}_1(i,j;G_n,\mathbf{J}))$$

From (70) and (71), the expected values of the third and the fourth terms above are $o(1)$. For the second term, we have

$$\mathbb{E}\left[\sum_{\substack{2\le d(i,j;G_n)\le\zeta,\\ d_2(i,j;G_n)\ge\zeta_2}} \mathbb{I}(\mathcal{E}_1(i,j;G_n,\mathbf{J}))\right]$$

$$\overset{(a)}{\le} \mathbb{E}\left[\sum_{\substack{2\le d(i,j;G_n)\le\zeta,\\ d_2(i,j;G_n)\ge\zeta_2}} \mathbb{I}\{C(i,j;G_n,\mathbf{J}) > (\tanh J_{\max})^2 + (\tanh J_{\max})^{\zeta_2}\}\right],$$

$$\overset{(b)}{\le} \mathbb{E}\left[\sum_{\substack{2\le d(i,j;G_n)\le\zeta,\\ d_2(i,j;G_n)\ge\zeta_2}} \mathbb{I}\{C(i,j;G_n,\mathbf{J}) > (\tanh J_{\max})^{d(i,j;G_n)} + (\tanh J_{\max})^{d_2(i,j;G_n)}\}\right]$$

$$\overset{(c)}{=} o(1),$$

where (a) follows from the choice of δ in (17) with $\delta > (\tanh J_{\max})^2 + (\tanh J_{\max})^{\zeta_2}$. Inequality (b) is using the fact that $d_2(i,j;G_n) \ge \zeta_2$ for node-pairs under consideration and the claim (c) is from (55) in Theorem 5. For the first term, we have

$$\sum_{\substack{2\le d(i,j;G_n)\le\zeta,\\ 2\le d_2(i,j;G_n)\le\zeta_2}} \mathbb{I}(\mathcal{E}_1(i,j;G_n,\mathbf{J})) \le \sum_{\substack{i,j\in v(\mathcal{C}_{l;G_n}):\\ l\le\zeta+\zeta_2}} \mathbb{I}(\mathcal{E}_1(i,j;G_n,\mathbf{J})), \tag{72}$$

where $\mathcal{C}_{l;G_n}$ is a cycle of length l. In words, the set under consideration above is the set of node pairs contained in a cycle of length less than $\zeta+\zeta_2$. Hence,

$$\begin{aligned}\mathbb{E}\left[\sum_{\substack{2\le d(i,j;G_n)\le\zeta,\\ 2\le d_2(i,j;G_n)\le\zeta_2}} \mathbb{I}(\mathcal{E}_1(i,j;G_n,\mathbf{J}))\right] &\le \sum_{l\le\zeta+\zeta_2}\mathbb{E}\left[l^2 N_l^{\text{cycles}}\right]\\ &= O((\zeta+\zeta_2)^2 c^{\zeta+\zeta_2}) = O(1),\end{aligned}$$

where $N_{\zeta+\zeta_2}^{\text{cycles}}$ is the number of cycles in G_n of length $\zeta+\zeta_2$. Hence, we have

$$\mathbb{E}\left[\sum_{i,j\in V:d(i,j;G_n)>1} \mathbb{I}(\mathcal{E}_1(i,j;G_n,\mathbf{J}))\right] = O(1).$$

Fix an increasing sequence $\{\beta_n\}_{n\in\mathbb{N}}$ with $\lim_{n\to\infty}\beta_n = \infty$. By Markov's inequality, we have

$$\mathbb{P}\left[\sum_{d(i,j;G_n)>1} \mathbb{I}(\mathcal{E}_1(i,j;G_n,\mathbf{J})) \ge \beta_n\right] = o(1) \tag{73}$$

Hence, the edit distance between the true graph and the estimated one from correlation thresholding satisfies

$$\mathbb{P}\left[|\mathsf{CT}(\{C_{i,j}\};\delta)\setminus G_n| \ge \beta_n\right] = o(1).$$

Correlation Thresholding with Samples. We now consider empirical correlations instead of exact statistics.

$$\begin{aligned}\mathbb{P}\left[\sum_{i,j\in V}\mathbb{I}(\widehat{\mathcal{E}}_1(i,j)) \ge \beta_n\right] &\le \mathbb{P}\left[\sum_{i,j\in V}\mathbb{I}(\widehat{\mathcal{E}}_1(i,j)) \ge \beta_n \middle| \sum_{(i,j\in V}\mathbb{I}(\mathcal{E}_1(i,j)) < \beta_n\right]\\ &+ \mathbb{P}\left[\sum_{i,j\in V}\mathbb{I}(\mathcal{E}_1(i,j)) \ge \beta_n\right]\end{aligned}$$

The second term is $o(1)$ from (73).

$$\mathbb{P}\left[\sum_{i,j\in V}\mathbb{I}(\widehat{\mathcal{E}}_1(i,j)) \ge \beta_n \middle| \sum_{i,j\in V}\mathbb{I}(\mathcal{E}_1(i,j)) < \beta_n\right] \le \sum_{i,j\in V}\mathbb{P}\left[\widehat{\mathcal{E}}_1(i,j)|(\mathcal{E}_1(i,j))^c\right]. \tag{74}$$

For $m > M \log n$, for sufficiently large $M > 0$, the above term decays to zero exponentially fast by the union bound and the Chernoff bound [14]. Similar analysis holds for the error event $\widehat{\mathcal{E}}_2^m$ in (59).

Hence, at most a constant number of events occur for the threshold δ as given in (17), if $m = \Omega(\log n)$.

5.2.4. *Proof of Corollary 1.* We now provide the proof for Corollary 1. Recall that we are interested in analyzing the number of spurious edges in the output by the Chow-Liu algorithm, given by $|\mathsf{CLThres}(\{\widehat{C}^m_{i,j}\}) \setminus G_n|$. The proof of (22) for the case $\zeta = 2$ in Corollary 1 follows directly from the proof above, and we prove the result in (23) for $\zeta > 2$.

From the cycle property of the max. weight spanning tree,

$$\left| \mathsf{CLThres}(\{\widehat{C}^m_{i,j}\}) \setminus G_n \right| \leq \sum_{(i,j)\notin G_n} \mathbb{I}(\widehat{\mathcal{A}}^m(i,j)),$$

where $\widehat{\mathcal{A}}^m(i,j)$ is defined as

$$\widehat{\mathcal{A}}^m(i,j) := \{\exists (k,l) \in \mathrm{path}(i,j;G_n) : \widehat{C}^m(i,j) > \widehat{C}^m(k,l)\}. \tag{75}$$

We have from union bound that

$$\mathbb{E}\left[| \mathsf{CLThres}(\{\widehat{C}^m_{i,j}\}) \setminus G_n | \right] \leq \sum_{(i,j)\notin G_n} d(i,j;G_n) \mathbb{P}[\widehat{C}^m(i,j) > \min_{(k,l)\in \mathrm{path}(i,j)} \widehat{C}^m(k,l)].$$

As before, first consider performance under exact statistics, i.e. , $|\mathsf{CLThres}(\{C_{i,j}\}) \setminus G_n|$. We have

$$\mathbb{P}[C(i,j) > \min_{(k,l)\in\mathrm{path}(i,j)} C(k,l)] \leq \mathbb{P}[(\tanh J_{\max})^{d_2(i,j)} > \min_{(k,l)\in\mathrm{path}(i,j)} C(k,l) - C_{\mathrm{SP}}],$$

from the upper bound on correlation in (55) in Theorem 5. Recall that C_{SP} is the correlation along the shortest path, given by (44). By definition,

$$C_{\mathrm{SP}} < \min_{(k,l)\in\mathrm{path}(i,j;G_n)} C(k,l).$$

Thus, we have

$$\begin{aligned}
&\mathbb{P}[(\tanh J_{\max})^{d_2(i,j)} > \min_{(k,l)\in\mathrm{path}(i,j)} C(k,l) - C_{\mathrm{SP}}] \\
&\leq \mathbb{P}[(\tanh J_{\max})^{d_2(i,j)} > \{\tanh J_{\min}(1 - (\tanh J_{\max})^{d(i,j)-1})\}], \\
&= \mathbb{P}[d_2(i,j;G_n) \leq K(J_{\min}, J_{\max}, d(i,j))],
\end{aligned}$$

for appropriately defined constant K. From Lemma 1, we have

$$\mathbb{P}[d_2(i,j;G_n) \leq K(J_{\min}, J_{\max}, d(i,j)) | d(i,j;G_n) \leq l] = O(n^{-2} c^{K(J_{\min}, J_{\max}, l)+l}).$$

By choosing $l = o(\log n)$, we obtain

$$\mathbb{E}\left[| \mathsf{CLThres}(\{C_{i,j}\}) \setminus G_n |\right] = O(1),$$

and using Markov's inequality, and Chernoff bound, we obtain the desired result.

5.3. *Proof of Theorem 2.* We now establish the consistency of CMIT algorithm. We first establish the presence of a sparse approximate separator for non-neighbor nodes which results in the decay of conditional mutual information. We then establish the exact rate of decay of conditional mutual information with respect to the graph size. We then consider empirical conditional mutual information, estimated from samples, and provide concentration results. Finally, we provide scaling laws for number of samples required to obtain learning consistency.

5.3.1. *Approximate-Separator Sets.* We provide a formal definition of approximate-separator sets.

Definition 2 (Approximate-separator Set). *The approximate-separator set $\mathcal{S}(i,j;G,\gamma_n)$ between non-adjacent nodes i and j with respect to a graph $G = (V,E)$ and a threshold $\gamma_n > 1$ is, $\forall\, (i,j) \notin G$,*

$$\mathcal{S}(i,j;G,\gamma_n) := (\mathcal{N}(i;G)) \bigcap \left(\bigcup_{\substack{l \in \mathbb{N}: \\ |\operatorname{path}_l(i,j;G)| \leq \gamma_n}} \operatorname{path}_l(i,j;G) \right). \tag{76}$$

The above definition for $\mathcal{S}(i,j;G,\gamma_n)$ can be equivalently[19] cast in terms of neighbors of j, $\mathcal{N}(j;G)$ instead of $\mathcal{N}(i;G)$. In words, the approximate-separator set between i and j is a subset of a true separator which only contains nodes along short paths (paths whose lengths are less than γ_n). See Fig. 2 for an illustration.

We show that $|\mathcal{S}(i,j;G_n,\gamma_n)| < 2 + o(1)$ for a.e. G_n for random graphs, when

$$\gamma_n < \frac{\log n}{(5+\epsilon)\log c}, \tag{77}$$

[19] Note that the approximate separator is empty when $d(i,j;G_n) \geq \gamma_n$, but under correlation decay, the mutual information $I(X_i;X_j)$ decays in this case without the need for any conditioning.

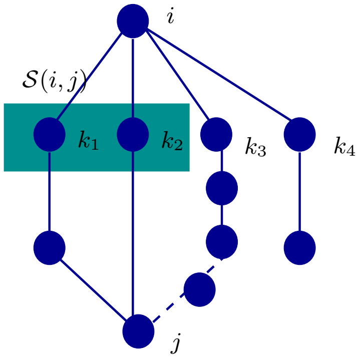


FIG 2. *Illustration of approximate-separator set $\mathcal{S}(i,j;G,\gamma_n)$ for the graph shown above with $\gamma_n = 4$. Note that $\mathcal{N}(i;G) = \{k_1,k_2,k_3,k_4\}$ is the neighborhood of i and the approximate separator set $\mathcal{S}(i,j;G,\gamma_n) = \{k_1,k_2\} \subset \mathcal{N}(i;G)$. This is because the path along k_3 connecting i and j has a length greater than γ_n and hence node $k_3 \notin \mathcal{S}(i,j;G,\gamma_n)$.*

for some $\epsilon > 0$. We now show that the size of the approximate-separator set is small in random graphs.

LEMMA 4 (Size of Approximate-separator Set). *For the set $\mathcal{S}(i,j;G_n,\gamma_n)$ defined in (76), we have a.a.s,*

$$|\mathcal{S}(i,j;G_n,\gamma_n)| \le 2 + o(1).$$

Proof: Consider the case when for a node pair i,j, the second shortest distance $d_2(i,j;G_n) > \gamma_n$. In this case, the result holds. Hence, conditioned on the event that $d_2(i,j;G_n) \le \gamma_n$, the size of the approximate separator is given by the number of overlapping cycles at node i as

$$|\mathcal{S}(i,j;G_n,\gamma_n)| \le 2 + \sum_{t_1=3}^{2\gamma_n} \sum_{t_2=3}^{2\gamma_n} \sum_{t_3=2}^{\gamma_n} N_{H_{t_1,t_2,t_3}}.$$

From Lemma 1, we have $\gamma_n^3 N_{H_{t_1,t_2,t_3}} = o(1)$ a.a.s. for γ_n chosen above. □

5.4. *Conditional Mutual Information for Ising Models.* We first derive an upper bound on mutual information between two nodes.

PROPOSITION 1. *For any symmetric Ising model, we have*

$$I(X_i;X_j) \le C^2(i,j), \quad \forall\, i,j \in V.$$

Proof: By symmetry, we can define

$$\begin{aligned}\lambda_+ &:= P(x_i = +|x_j = +) = P(x_i = -|x_j = -)\\ \lambda_- &:= P(x_i = +|x_j = -) = P(x_i = -|x_j = +).\end{aligned}$$

We have

$$\lambda_+ + \lambda_- = 1.$$

Reparameterize $\lambda_+ = 1/2 + \lambda$ and $\lambda_- = 1/2 - \lambda$. We can write mutual information as

$$\begin{aligned} I(X_i; X_j) &= (1/2 + \lambda)\log(1 + 2\lambda) + (1/2 - \lambda)\log(1 - 2\lambda),\\ &\overset{(a)}{\leq} (1/2 + \lambda)2\lambda - (1/2 - \lambda)2\lambda,\\ &= 4\lambda^2 = (\lambda_+ - \lambda_-)^2 = C^2(X_i, X_j) \end{aligned} \tag{78}$$

where inequality (a) is from the fact that $\log(1 + x) \leq x$. □

From Theorem 5, we immediately see that when $d(i, j; G_n) \geq \gamma_n$ for γ_n in (77), the mutual information decays with n, without the need for a conditioning set. Hence, we only need to bound the conditional mutual information for node pairs i, j with $d(i, j; G_n) < \gamma_n$. We establish a bound for such node pairs when the conditioning set is its approximate separator. Recall that $B_l(i; G)$ denotes the set of nodes at distance l from i on graph G.

Lemma 5 (Bounds on Conditional Mutual Information). *For any two nodes such that $(i, j) \notin G_n$, for $S := \mathcal{S}(i, j; G_n, \gamma_n)$, the approximate-separator defined in (76), we have*

$$I(X_i; X_j|\mathbf{X}_S) \leq \frac{1}{f_{\min}(S)} |B_{\gamma_n}(i; G_n)|^2 (\tanh J_{\max})^{2\gamma_n}, \tag{79}$$

where $f_{\min}(S) := \min_{x_i, \mathbf{x}_S} P(x_i|\mathbf{x}_S)$.

Remark: When S is a finite set, $f_{\min}(S)$ is positive. We use this bound to show that conditional mutual information between non-adjacent nodes decays for Ising model on a random graph in the uniqueness regime.

Proof: We have

$$\begin{aligned} I(X_i; X_j|\mathbf{X}_S) &:= \sum_{\mathbf{x}_S \in \mathcal{X}^{|S|}} P(\mathbf{x}_S) I(X_i; X_j|\mathbf{X}_s = \mathbf{x}_S),\\ &\overset{(a)}{\leq} \max_{\mathbf{x}_S} I(X_i; X_j|\mathbf{X}_s = \mathbf{x}_S), \end{aligned}$$

$$\overset{(b)}{\leq} \max_{x_j,\mathbf{x}_S} D(P(X_i|X_j = x_j, \mathbf{X}_S = \mathbf{x}_S)||P(X_i|\mathbf{X}_S = \mathbf{x}_S)), \tag{80}$$

where for both inequalities (a) and (b), we use the fact that the convex combination is bounded above by the maximum summand, and for inequality (b), we note that

$$I(X_i; X_j|\mathbf{X}_s = \mathbf{x}_S) := \sum_{x_j \in \mathcal{X}} P(x_j|\mathbf{x}_S) D(P(X_i|X_j = x_j, \mathbf{X}_S = \mathbf{x}_S)||P(X_i|\mathbf{X}_S = \mathbf{x}_S)).$$

For any two discrete distributions Q_1, Q_2, we have the upper bound

$$D(Q_1||Q_2) \leq \frac{1}{q_{\min}} ||Q_1 - Q_2||_2^2, \tag{81}$$

where $q_{\min} := \min_{x \in \mathcal{X}} Q_2(x)$ from [12, Lemma 6.3].

Applying the above to (80) and letting $f_{\min}(S) := \min_{x_i, \mathbf{x}_S} P(x_i|\mathbf{x}_S)$, we have

$$\begin{aligned} I(X_i; X_j|\mathbf{X}_S) &\leq \max_{x_j,\mathbf{x}_S} \frac{1}{f_{\min}(S)} ||P(X_i|X_j = x_j, \mathbf{X}_S = \mathbf{x}_S) - P(X_i|\mathbf{X}_S = \mathbf{x}_S)||_2^2, \\ &\overset{(a)}{\leq} \max_{\mathbf{x}_S} \frac{1}{f_{\min}(S)} ||P(X_i|X_j = +, \mathbf{X}_S = \mathbf{x}_S) - P(X_i|X_j = -, \mathbf{X}_S = \mathbf{x}_S)||_2^2, \\ &\overset{(b)}{\leq} \max_{\mathbf{x}_S} \frac{1}{f_{\min}(S)} (\mathbb{E}(X_i|X_j = +, \mathbf{X}_S = \mathbf{x}_S) - \mathbb{E}(X_i|X_j = -, \mathbf{X}_S = \mathbf{x}_S))^2 \end{aligned} \tag{82}$$

Inequality (a) is from the fact that conditioning on more nodes with opposite signs increases the l_2-norm. Inequality (b) comes from the fact that for ferromagnetic models $P(x_i = +|x_j = +, \mathbf{x}_S) \geq P(x_i = -|x_j = +, \mathbf{x}_S)$ for all $\mathbf{x}_S \in \mathcal{X}^{|S|}$. For an Ising model on a graph G, we appeal to the equivalence of the conditional distribution on the self-avoiding walk tree [42] i.e.,

$$P(x_i|x_j, \mathbf{x}_S; G) = P(x_i|\mathbf{x}_{\mathcal{U}(j)}, \mathbf{x}_{\mathcal{U}(S)}, \mathbf{x}_A; T_{\text{saw}}^{(i,G)}),$$

where $\mathcal{U}(S) = \mathcal{U}(S; T_{\text{saw}}^{(i,G)})$ is the set of copies of nodes in set S in $T_{\text{saw}}^{(i,G)}$, the self-avoiding walk tree rooted at i, and A is the set of terminal nodes in $T_{\text{saw}}^{(i,G)}$ with a given state $\mathbf{x}_A$.

Now, consider $S = \mathcal{S}(i, j; G_n, \gamma_n)$, the approximate-separator set between i and j defined in (76) and recall that $\widetilde{\mathcal{U}}(j; \gamma_n, T_{\text{saw}}^{(i,G)})$ denotes copies of j in $T_{\text{saw}}^{(i,G)}$ that are within distance γ_n, as defined in (47). The set $Y \subset \mathcal{U}(S; T_{\text{saw}}^{(i,G)})$ separates $\widetilde{\mathcal{U}}(j; \gamma_n)$ from i in $T_{\text{saw}}^{(i,G)}$. See Fig.3. Let $Z := \mathcal{U}(j; T_{\text{saw}}^{(i,G)}) \setminus$

$\widetilde{\mathcal{U}}(j;\gamma_n, T_{\text{saw}}^{(i,G)})$. Since the variables $\mathbf{X}_{\widetilde{\mathcal{U}}(j;\gamma_n)} - \mathbf{X}_Y - X_i$ form a Markov chain, we have

$$P(x_i|\mathbf{x}_{\mathcal{U}(j)}, \mathbf{x}_{\mathcal{U}(S)}, \mathbf{x}_A; T_{\text{saw}}^{(i,G)}) = P(x_i|\mathbf{x}_Z, \mathbf{x}_{\mathcal{U}(S)}, \mathbf{x}_A; T_{\text{saw}}^{(i,G)}). \tag{83}$$

Substituting this equivalence into (82), we have

$$\begin{aligned} I(X_i;X_j|\mathbf{X}_S) &\leq \max_{\mathbf{x}_S} f_{\min}^{-1}(S)(\mathbb{E}(X_i|\mathbf{X}_Z = +, \mathbf{x}_{\mathcal{U}(S)}, \mathbf{x}_A) \\ &\qquad - \mathbb{E}(X_i|\mathbf{X}_Z = -, \mathbf{x}_{\mathcal{U}(S)}, \mathbf{x}_A))^2 \\ &\overset{(a)}{\leq} \frac{1}{f_{\min}(S)}(\mathbb{E}(X_i|\mathbf{X}_Z = +) - \mathbb{E}(X_i|\mathbf{X}_Z = -))^2. \end{aligned} \tag{84}$$

where inequality (a) is obtained in [30, Lemma 2.8]. We abbreviate $B := B_{\gamma_n}(i; T_{\text{saw}}^{(i,G)})$ in the rest of the proof. Since the nodes in Z have at least distance $\lceil\gamma_n\rceil$ from i by definition, the variables $X_i - \mathbf{X}_B - \mathbf{X}_Z$ form a Markov chain. We can thus bound as (84) as,

$$\begin{aligned} I(X_i;X_j|\mathbf{X}_S) &\leq \frac{1}{f_{\min}(S)}(\mathbb{E}(X_i|\mathbf{X}_{B\cup Z} = +) - \mathbb{E}(X_i|\mathbf{X}_{B\cup Z} = -))^2 \\ &\leq \frac{1}{f_{\min}(S)}(\mathbb{E}(X_i|\mathbf{X}_B = +) - \mathbb{E}(X_i|\mathbf{X}_B = -))^2 \\ &\leq \frac{|B|^2(\tanh J_{\max})^{2\gamma_n}}{f_{\min}(S)}, \end{aligned}$$

which is obtained from telescoping from all + configuration on B to all − configuration by changing one sign at a time. □

The bounds on conditional mutual information for Ising models on random graphs $G_n \sim \mathcal{G}(n, \frac{c}{n})$ follow.

Lemma 6 (Bounds on Conditional Mutual Information in the Uniqueness Regime). *For an Ising model Markov on $G_n \sim \mathcal{G}(n, \frac{c}{n})$ in the uniqueness regime ($\alpha := c\tanh J_{\max} < 1$) if γ_n satisfies (77), then the conditional mutual information on non-edges satisfies*

$$\max_{(i,j)\notin G_n} I(X_i;X_j|\mathbf{X}_{\mathcal{S}(i,j;G_n,\gamma_n)}) = O(n^{-\kappa}), \tag{85}$$

for some $\kappa > 0$.

Proof: Using (79) and (39), we have for a.e. G_n,

$$\max_{(i,j)\notin G_n} I(X_i;X_j|\mathbf{X}_{\mathcal{S}(i,j;G_n,\gamma_n)}) \leq \alpha^{2\gamma_n}(\log n)^2,$$

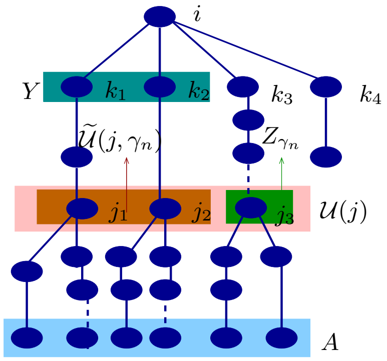


FIG 3. *Illustration of sets in (83) on $T_{\rm saw}^{(i,G)}$, which is the self-avoiding walk tree at node i corresponding to the graph in Fig.2. The nodes j_1, j_2 and j_3 are the copies of j in $T_{\rm saw}^{(i,G)}$ and similarly for node k. The set A is the set of terminal nodes in $T_{\rm saw}^{(i,G)}$. The set $\widetilde{\mathcal{U}}(j, \gamma_n; T_{\rm saw}^{(i,G)})$ contain the copies of j which are at a distance less than γ_n from i on $T_{\rm saw}^{(i,G)}$. The set Z_{γ_n} contains the rest of copies of j which are at distance at least γ_n from i. Set Y is the set of copies corresponding to nodes in $\mathcal{S}(i, j; G_n)$ which are on the paths from i to copies of j.*

and the result follows from the choice of γ_n in (77). □

We finally show that the conditional mutual information between neighbors is positive for a ferromagnetic Ising model, when the cardinality of the conditioning set is bounded. This implies that we can distinguish edges and non-edges through conditional mutual information thresholding.

LEMMA 7 (Positivity of Conditional Mutual Information between Neighbors). *For an Ising model Markov on a graph $G = (V, E)$ with minimum potential $J_{\min} > 0$, we have*

$$I(X_i; X_j | \mathbf{X}_S) > f(J_{\min}, J_{\max}, |S|), \quad \forall (i, j) \in G, S \subset V, i, j \notin S, \tag{86}$$

where f is a positive function that only depends on $J_{\min}, J_{\max}$ and $|S|$, i.e., the conditional mutual information is uniformly bounded away from zero if $J_{\min}$ is bounded away from zero and $J_{\max}, |S|$ are finite.

Proof: First we note that the conditional mutual information can be expressed as a KL-divergence

$$I(X_i; X_j | \mathbf{X}_S) = \sum_{\mathbf{x}_S} P(\mathbf{x}_S) D(P_{X_i, X_j | \mathbf{X}_S}(\cdot, \cdot | \mathbf{x}_S) \,||\, P_{X_i | \mathbf{X}_S}(\cdot | \mathbf{x}_S) P_{X_j | \mathbf{X}_S}(\cdot | \mathbf{x}_S)) \tag{87}$$

Now we make use of Pinsker's inequality [10, Ch. 11]

$$D(P\,||\,Q) \geq \frac{1}{2\log 2}\|P-Q\|_1^2$$

to lower bound each conditional divergence in (87) (indexed by $\mathbf{x}_S$:

$$\begin{aligned} & D(P_{X_i,X_j|\mathbf{x}_S}(\cdot,\cdot|\mathbf{x}_S)\,||\,P_{X_i|\mathbf{x}_S}(\cdot|\mathbf{x}_S)P_{X_j|\mathbf{x}_S}(\cdot|\mathbf{x}_S))^{1/2} \\ & \qquad \geq \frac{1}{\sqrt{2\log 2}}\|P_{X_i,X_j|\mathbf{X}_S=\mathbf{x}_S} - P_{X_i|\mathbf{X}_S=\mathbf{x}_S}P_{X_j|\mathbf{X}_S=\mathbf{x}_S}\|_1 \end{aligned} \tag{88}$$

We now find a lower bound for the ℓ_1 norm between the joint distribution and the product of the marginals in (88). Indeed, we have

$$\begin{aligned} & \|P_{X_i,X_j|\mathbf{X}_S=\mathbf{x}_S} - P_{X_i|\mathbf{X}_S=\mathbf{x}_S}P_{X_j|\mathbf{X}_S=\mathbf{x}_S}\|_1 \\ & = \sum_{x_i,x_j} |P(x_i,x_j|\mathbf{x}_S) - P(x_i|\mathbf{x}_S)P(x_j|\mathbf{x}_S)| \\ & \geq \min_{x_i,x_j} |P(x_i,x_j|\mathbf{x}_S) - P(x_i|\mathbf{x}_S)P(x_j|\mathbf{x}_S)| \\ & \geq \left(\min_{x_j} P(x_j|\mathbf{x}_S)\right)\left(\min_{x_i,x_j} |P(x_i|x_j,\mathbf{x}_S) - P(x_i|\mathbf{x}_S)|\right) \end{aligned} \tag{89}$$

Let us define $l(z) := 1/(1+\exp(-z))$ to be the logistic function. Using the locally tree-like property of random graphs, the first term in (89) can be lower bounded as

$$\min_{x_j} P(x_j|\mathbf{x}_S) \geq l(-M|S|J_{\max}), \tag{90}$$

for some constant M since there are at most two short paths between any two nodes in the random graph, meaning that the number of copies of nodes in S which are close to j does not "explode" in the self-avoiding walk tree. Similarly, the second term in (89) can be lower bounded as

$$\begin{aligned} & \min_{x_i,x_j} |P(x_i|x_j,\mathbf{x}_S) - P(x_i|\mathbf{x}_S)| \\ & \geq \min\left\{l(-M(|S|+1)J_{\max}) - l(-M|S|J_{\max}), l(-M(|S|+1)J_{\min}) - l(-M|S|J_{\min})\right\}. \end{aligned}$$

Putting (87), (88), and (90)

$$I(X_i;X_j|\mathbf{X}_S) \geq f(J_{\min},J_{\max},|S|), \tag{91}$$

for some positive function f, which only depends on $J_{\min}, J_{\max}$ and the cardinality of S. □

5.4.1. *Concentration of Empirical Mutual Information.* We have so far established bounds on conditional mutual information. We now provide concentration results for empirical mutual information estimated from samples.

LEMMA 8 (Concentration Results for Empirical Mutual Information). *For $\mathbf{X}$ drawn from an Ising model, for any nodes $i, j \in V$ and any set $S \subset V \setminus \{i, j\}$, we have the following bound for empirical conditional mutual information from m i.i.d. samples,*

$$\begin{aligned}&\mathbb{P}\left(|\widehat{I}^m(X_i; X_j|\mathbf{X}_S) - I(X_i; X_j|\mathbf{X}_S)| > \epsilon\right)\\ &\leq (m+1)^{2^{|S|+2}}\left[3\exp\left(-m\frac{\epsilon}{4}\right) + K\exp\left(-m\frac{\epsilon^2}{K'}\right)\right],\end{aligned} \tag{92}$$

where $K' := 64(\log 4)^3$ and $K > 0$ depends only on $J_{\min}, J_{\max}$ and $|S|$ and is finite and positive when $J_{\min} > 0$ and $J_{\max}, |S|$ are finite.

Proof: By adding and subtracting $\mathbb{E}_{\widehat{P}^m}[\log \frac{P(x_i,x_j|\mathbf{x}_S)}{P(x_i|\mathbf{x}_S)P(x_j|\mathbf{x}_S)}]$, the difference of mutual information quantities can be upper bounded as

$$\begin{aligned}&|\widehat{I}^m(X_i; X_j|\mathbf{X}_S) - I(X_i; X_j|\mathbf{X}_S)|\\ &\leq \left|\sum_{x_i,x_j,\mathbf{x}_S}(P(x_i, x_j|\mathbf{x}_S) - \widehat{P}^m(x_i, x_j|\mathbf{x}_S))\log\frac{P(x_i, x_j|\mathbf{x}_S)}{P(x_i|\mathbf{x}_S)P(x_j|\mathbf{x}_S)}\right|\\ &+ D(\widehat{P}^m(x_i|\mathbf{x}_S)||P(x_i|\mathbf{x}_S)) + D(\widehat{P}^m(x_j|\mathbf{x}_S)||P(x_j|\mathbf{x}_S))\\ &+ D(\widehat{P}^m(x_i, x_j|\mathbf{x}_S)||P(x_i, x_j|\mathbf{x}_S))\end{aligned} \tag{93}$$

From [38, Lemma 14],

$$\mathbb{P}(D(\widehat{P}^m(x_i|\mathbf{x}_S)||P(x_i|\mathbf{x}_S)) \geq \eta_1) \leq (m+1)^{2^{|S|+1}}\exp(-m\eta_1). \tag{94}$$

This enables us to bound the probability of the deviation of the empirical KL-divergence from zero in last three terms in (93). The first term in (93) can be further upper bounded as

$$\begin{aligned}&\left|\sum_{x_i,x_j}(P(x_i, x_j, \mathbf{x}_S) - \widehat{P}^m(x_i, x_j, \mathbf{x}_S))\log\frac{P(x_i, x_j|\mathbf{x}_S)}{P(x_i|\mathbf{x}_S)P(x_j|\mathbf{x}_S)}\right|\\ &\leq \max_{x_i,x_j,\mathbf{x}_S}\left|\log\frac{P(x_i, x_j|\mathbf{x}_S)}{P(x_i|\mathbf{x}_S)P(x_j|\mathbf{x}_S)}\right|\sum_{x_i,x_j,\mathbf{x}_S}|P(x_i, x_j, \mathbf{x}_S) - \widehat{P}^m(x_i, x_j, \mathbf{x}_S)|\end{aligned}$$

(95)

$$\leq K\|P - \widehat{P}^m\|_1,$$

where we upper bounded the absolute value of the information density $|\log(\frac{P(x_i,x_j|\mathbf{x}_S)}{P(x_i|\mathbf{x}_S)P(x_j|\mathbf{x}_S)})|$ as

$$K' \leq \max\left[\left|\log\frac{1+\exp(-2(|S|+1)J_{\min})}{1+\exp(-2|S|J_{\min})}\right|, \left|\log\frac{1+\exp(-2(|S|+1)J_{\max})}{1+\exp(-2|S|J_{\max})}\right|\right]$$

In (95), we also used the notation $\|P - \widehat{P}^m\|_1$ to denote the vector ℓ_1 norm between the vectorized versions of P and $\widehat{P}^m$. Now note from Pinsker's inequality [10, Ch. 11] that

$$\|P - \widehat{P}^m\|_1 \leq \sqrt{(\log 4)D(\widehat{P}^m \,||\, P)} \tag{96}$$

and thus the bound in (94) can be modified as

$$\mathbb{P}(\|\widehat{P} - P\|_1 \geq \eta_2) \leq (m+1)^{2^{|S|+2}} \exp\left(-m\frac{\eta_2^2}{\log 4}\right). \tag{97}$$

Putting (93), (94), (97) together, using the fact that $\mathbb{P}(\sum_{i=1}^k \mathcal{A}_i \geq \epsilon) \leq \sum_{i=1}^k \mathbb{P}(\mathcal{A}_i \geq \epsilon/k)$ and taking $\eta_1 = \epsilon/4$ and $\eta_2 = \epsilon/(4\log 4)$ yields[20] the result. □

5.4.2. *Error Events for CMIT.* For any $(i,j) \notin G_n$, define the event

$$\mathcal{F}_1(i,j;\{\mathbf{x}^m\},G_n) := \left\{\widehat{I}(X_i;X_j|\mathbf{X}_{\mathcal{S}(i,j;G_n;\gamma_n))}) > \xi_{n,m},\right\} \tag{98}$$

where $\xi_{n,m}$ is the threshold in (24a). Similarly for any edge $(i,j) \in G_n$, define the event that

(99)

$$\mathcal{F}_2(i,j;\{\mathbf{x}^m\},G_n) := \left\{\exists S \subset V : |S| \leq 2 + o(1), \widehat{I}(X_i;X_j|\mathbf{X}_S) < \xi_{n,m},\right\}.$$

The probability of error resulting from CMIT can thus be bounded by the two types of errors,

$$\mathbb{P}[\mathsf{CMIT}(\{\mathbf{x}^m\};\xi_{n,m}) \neq G_n] \leq \mathbb{P}\left[\bigcup_{(i,j)\in G_n} \mathcal{F}_2(i,j;\{\mathbf{x}^m\},G_n)\right]$$

[20] if $I(X_i;X_j|\mathbf{X}_S) = 0$, then $P(x_i,x_j|\mathbf{x}_S) = P(x_i|\mathbf{x}_S)P(x_j|\mathbf{x}_S)$ and the first term in (93) is exactly zero.

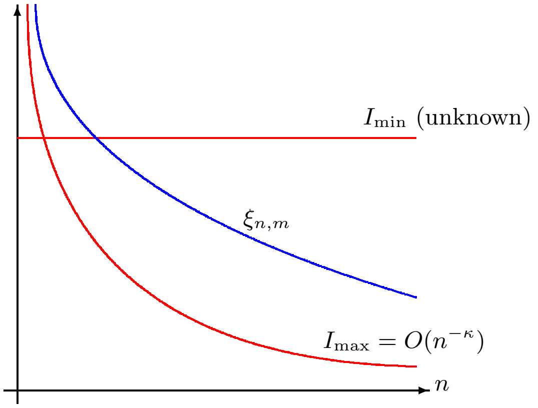


Fig 4. *The threshold $\xi_{m,n}$ in* CMIT *algorithm separates edges and non-edges with high probability. $I_{\min}$ and $I_{\max}$ are defined in* (102) *and* (104).

$$+\,\mathbb{P}\left[\bigcup_{(i,j)\notin G_n} \mathcal{F}_1(i,j;\{\mathbf{x}^m\},G_n)\right] \tag{100}$$

For the first term, applying union bound for both the terms and using the result (92) of Lemma 8,

$$\mathbb{P}\left[\bigcup_{(i,j)\in G_n} \mathcal{F}_2(i,j;\{\mathbf{x}^m\},G_n)\right] = O(m^\tau n \exp[-m(I_{\min} - \xi_{n,m})/4]) \tag{101}$$

for some constant $\tau > 0$ and where

$$I_{\min} := \inf_{\substack{(i,j)\in G_n\\ S\subset V, i,j\notin S\\ |S|\le 2+o(1)}} I(X_i;X_j|\mathbf{X}_S) > 0, \quad \forall\, n\in\mathbb{N}, \tag{102}$$

from Lemma 7. Since $\xi_{n,m} = o(1)$, (101) is $o(1)$ when $m = \omega(\log n)$. For the second term in (100),

$$\mathbb{P}\left[\bigcup_{(i,j)\notin G_n} \mathcal{F}_1(i,j;\{\mathbf{x}^m\},G_n)\right] = O(n^2 m^\tau \exp[-m(\xi_{n,m} - I_{\max}(n))/4]), \tag{103}$$

again for some $\tau > 0$ and where

$$I_{\max}(n) := \max_{\substack{(i,j)\notin G_n\\ \mathcal{S}(i,j;G_n)\subset S}} I(X_i;X_j|\mathbf{X}_S) = O(n^{-\kappa}), \tag{104}$$

for some $\kappa > 0$, from (85). For the choice of $\xi_{n,m}$ in (24a), (103) is $o(1)$. □

5.5. *Proof of Theorem 3.* This proof is inspired by [5, Thm. 1]. Fix any estimator $\widehat{G}_n$. Denote $\mathcal{R} := \widehat{G}_n((\mathcal{X}^n)^m)$ as the range of the estimator $\widehat{G}_n$. This is the set of all graphs that can be output by the estimator $\widehat{G}_n$. Then we have the sequence of lower bounds:

$$\begin{aligned}
\mathbb{P}_{\mathbf{X},\mathcal{G}_n}(\widehat{G}_n \neq G_n) &\overset{(a)}{=} \sum_{g\in\mathcal{R}^c} \mathbb{P}_{\mathbf{X}|G_n}(\widehat{G}_n \neq G_n | G_n = g)\mathbb{P}_{\mathcal{G}_n}(G_n = g) \\
&\quad + \sum_{g\in\mathcal{R}} \mathbb{P}_{\mathbf{X}|G_n}(\widehat{G}_n \neq G_n | G_n = g)\mathbb{P}_{\mathcal{G}_n}(G_n = g) \\
&\overset{(b)}{\geq} \sum_{g\in\mathcal{R}^c} \mathbb{P}_{\mathbf{X}|G_n}(\widehat{G}_n \neq G_n | G_n = g)\mathbb{P}_{\mathcal{G}_n}(G_n = g) \\
&\overset{(c)}{=} \sum_{g\in\mathcal{R}^c} \mathbb{P}_{\mathcal{G}_n}(G_n = g) \\
&\overset{(d)}{=} 1 - \sum_{g\in\mathcal{R}} \mathbb{P}_{\mathcal{G}_n}(G_n = g), \qquad (105)
\end{aligned}$$

where equality (a) comes from the fact that $\mathcal{G}_n = \mathcal{R} \cup \mathcal{R}^c$, inequality (b) lower bounds the sum by the term involving $\mathcal{R}^c$, inequality (c) is due to the fact that $\mathbb{P}_{\mathbf{X}|G_n}(\widehat{G}_n \neq G_n | G_n = g) = 1$ for all $g \in \mathcal{R}^c$ and finally inequality (d) is because $\sum_{g\in\mathcal{R}} \mathbb{P}_{\mathcal{G}_n}(G_n = g) + \sum_{g\in\mathcal{R}^c} \mathbb{P}_{\mathcal{G}_n}(G_n = g) = 1$.

Now we provide an asymptotic upper bound for the term

$$\Upsilon := \sum_{g\in\mathcal{R}} \mathbb{P}_{\mathcal{G}_n}(G_n = g).$$

To do so, first note that $|\mathcal{R}| \leq |\mathcal{X}^n|^m = 2^{nm}$. Furthermore, let $k_g \in \{1, \dots, \binom{n}{2}\}$ denote the number of edges in the graph $g \in \mathcal{G}_n$. Then,

$$\mathbb{P}_{\mathcal{G}_n}(G_n = g) = \left(\frac{c}{n}\right)^{k_g} \left(1 - \frac{c}{n}\right)^{\binom{n}{2} - k_g}. \qquad (106)$$

Eq. (106) says that if the probability of edge appearance $c/n < 1/2$ (which is the case of interest) then $\mathbb{P}(G_n = g)$ is maximized at $k_g = 0$. In fact, we have the general result that for graphs $g_1, g_2 \in \mathcal{G}_n$

$$k_{g_1} \leq k_{g_2} \quad \Rightarrow \quad \mathbb{P}_{\mathcal{G}_n}(G_n = g_1) \geq \mathbb{P}_{\mathcal{G}_n}(G_n = g_2). \qquad (107)$$

It is then straightforward to show that the natural number

$$z := \min\left\{ l \in \mathbb{N} : \sum_{k=1}^{l} \binom{\binom{n}{2}}{k} \geq 2^{nm} \right\} \qquad (108)$$

is of the order $nm/\log n$ (by solving for l in (108)). The quantity z defined in (108) is to be interpreted as the number of edges such that the sum of the number of graphs with no greater than z edges is at least 2^{nm}. Thus,

$$\begin{aligned}\Upsilon &\overset{(a)}{:=} \sum_{g\in\mathcal{R}} \mathbb{P}_{\mathcal{G}_n}(G_n = g)\\ &\overset{(b)}{\leq} \sum_{k=0}^{z} \binom{\binom{n}{2}}{k} \left(\frac{c}{n}\right)^k \left(1-\frac{c}{n}\right)^{\binom{n}{2}-k}\\ &\overset{(c)}{=} \sum_{k=0}^{O(nm/\log n)} \binom{\binom{n}{2}}{k} \left(\frac{c}{n}\right)^k \left(1-\frac{c}{n}\right)^{\binom{n}{2}-k}\\ &\overset{(d)}{\leq} \exp\left[-\frac{4}{nc}\Big(nc - O\big(\frac{nm}{\log n}\big)\Big)^2\right]\end{aligned}$$

where (a) follows from the definition of Υ, (b) follows from rewriting Υ in terms of z, the number of edges and by using (106), (c) follows from (108), and (d) follows from the fact that $\Pr(\mathrm{Bin}(N,q)\leq k)\leq \exp(-\frac{2}{Nq}(Nq-k)^2)$ for $k\leq Nq$ with the identifications $N=\binom{n}{2}$ and $q=c/n$. Finally, we observe from (d) that if $m = ac\log n$ for some $a>0$, then the term $\Upsilon\to 0$ as $n\to\infty$. Thus, referring back to (105) and noting the arbitrariness of $\widehat{G}_n$, we conclude that if $m\leq \epsilon c\log n$ for sufficiently small $\epsilon>0$, then $\mathbb{P}_{\mathbf{X},\mathcal{G}_n}(\widehat{G}_n\neq G_n)\to 1$. □

References.

[1] Bento, J. and Montanari, A. (2009). Which Graphical Models are Difficult to Learn? In *Proc. of Neural Information Processing Systems (NIPS).*

[2] Bogdanov, A., Mossel, E. and Vadhan, S. The Complexity of Distinguishing Markov Random Fields. *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques* 331–342.

[3] Bollobás, B. (1985). *Random Graphs.* Academic Press.

[4] Brémaud, P. (1999). *Markov Chains: Gibbs fields, Monte Carlo simulation, and queues.* Springer.

[5] Bresler, G., Mossel, E. and Sly, A. (2008). Reconstruction of Markov Random Fields from Samples: Some Observations and Algorithms. In *Intl. workshop APPROX Approximation, Randomization and Combinatorial Optimization* 343–356. Springer.

[6] Chandrasekaran, V., Parrilo, P. A. and Willsky, A. S. (2010). Latent Variable Graphical Model Selection via Convex Optimization. *Preprint. Available on ArXiv.*

[7] Chechetka, A. and Guestrin, C. (2007). Efficient Principled Learning of Thin Junction Trees. In *Advances in Neural Information Processing Systems (NIPS).*

[8] Choi, M. J., Tan, V., Anandkumar, A. and Willsky, A. (2011). Learning Latent Tree Graphical Models. *accepted to Journal of Machine Learning Research, available on Arxiv.*

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

[10] Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. John Wiley & Sons, Inc.

[11] Csiszár, I. and Talata, Z. (2006a). Consistent Estimation of the Basic Neighborhood of Markov Random Fields. *The Annals of Statistics* 123–145.

[12] Csiszár, I. and Talata, Z. (2006b). Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Tran. on Information Theory* **52** 1007.

[13] Dembo, A. and Montanari, A. (2008). Ising Models on Locally Tree-like Graphs. *Arxiv preprint arXiv:0804.4726*.

[14] Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications, 2nd ed.* Springer, NY.

[15] Georgii, H. O. (1988). *Gibbs Measures and Phase Transitions*. Walter de Gruyter.

[16] Gerschenfeld, A. and Montanari, A. (2007). Reconstruction for Models on Random Graphs. In *Annual Symposium on Foundations of Computer Science*.

[17] Griffiths, R. B. (1967). Correlations in Ising ferromagnets. III. *Communications in Mathematical Physics* **6** 121–127.

[18] Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2010). Joint Structure Estimation for Categorical Markov Networks. *Submitted. Available at http://www.stat.lsa.umich.edu/~elevina/*.

[19] Karger, D. and Srebro, N. (2001). Learning Markov networks: maximum bounded tree-width graphs. In *Symposium on Discrete Algorithms* 392–401.

[20] Kloks, T. (1994). Only few graphs have bounded treewidth. *Springer Lecture Notes in Computer Science* **842** 51-60.

[21] Kolar, M., Le Song, A. A. and Xing, E. P. (2008). Estimating Time-Varying Networks. In *Proc. of Intl. Conf. on Intelligent Systems for Molecular Biology*.

[22] Lauritzen, S. L. (1996). *Graphical models: Clarendon Press*. Clarendon Press.

[23] Levin, D. A., Peres, Y. and Wilmer, E. L. (2008). *Markov Chains and Mixing Times*. American Mathematical Society.

[24] Liu, H., Xu, M., Gu, H., Gupta, A., Lafferty, J. and Wasserman, L. (2010). Forest density estimation. *Arxiv preprint arXiv:1001.1557*.

[25] Martinelli, F., Sinclair, A. and Weitz, D. (2003). The Ising model on trees: Boundary conditions and mixing time. In *Proc. of Foundations of Computer Science* 628–639. IEEE.

[26] Meinshausen, N. and Buehlmann, P. (2006). High Dimensional Graphs and Variable Selection With the Lasso. *Annals of Statistics* **34** 1436–1462.

[27] Mezard, M. and Montanari, A. (2009). *Information, physics, and computation*. Oxford University Press, USA.

[28] Mitliagkas, I. and Vishwanath, S. (2010). Strong Information-Theoretic Limits for Source/Model Recovery . In *Proc. of Allerton Conf. on Communication, Control and Computing*.

[29] Montanari, A., Mossel, E. and Sly, A. (2009). The weak limit of Ising models on locally tree-like graphs. *Arxiv*.

[30] Mossel, E. and Sly, A. (2009). Rapid mixing of Gibbs sampling on graphs that are sparse on average. *Random Structures and Algorithms* **35** 250–270.

[31] Netrapalli, P., Banerjee, S., Sanghavi, S. and Shakkottai, S. (2010). Greedy Learning of Markov Network Structure . In *Proc. of Allerton Conf. on Communication, Control and Computing*.

[32] 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.

[33] Peres, Y. (1999). Probability on Trees: An Introductory Climb. *Springer Lecture*

notes in Math 1717 193–280.

[34] Ravikumar, P., Wainwright, M. J. and Lafferty, J. (2008). High-dimensional Ising Model Selection Using l1-Regularized Logistic Regression. *Annals of Statistics*.

[35] Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2008). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Arxiv preprint arXiv:0811.3628*.

[36] Santhanam, N. P. and Wainwright, M. J. (2008). Information-theoretic Limits of High-dimensional Model Selection. In *International Symposium on Information Theory*.

[37] Tan, V., Anandkumar, A. and Willsky, A. (2010). Learning Gaussian Tree Models: Analysis of Error Exponents and Extremal Structures. *IEEE Tran. on Signal Processing* **58** 2701-2714.

[38] Tan, V., Anandkumar, A. and Willsky, A. (2011). Learning Markov Forest Models: Analysis of Error Rates. *accepted to Journal of Machine Learning Research, available on Arxiv*.

[39] Tan, V., Anandkumar, A., Tong, L. and Willsky, A. (2011). A Large-Deviation Analysis for the Maximum Likelihood Learning of Tree Structures. *IEEE Tran. on Information Theory*.

[40] Wainwright, M. J. and Jordan, M. I. (2008). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning* **1** 1–305.

[41] 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)*.

[42] Weitz, D. (2006). Counting independent sets up to the tree threshold. In *Proc. of ACM symp. on Theory of computing* 140–149.

Center for Pervasive Communications & Computing,
Electrical Engineering & Computer Science Dept.,
4408 Engineering Hall, Irvine, CA, USA 92697.
E-mail: a.anandkumar@uci.edu

Laboratory of Information & Decision Systems,
Stata Center, 77 Massachusetts Ave.,
Cambridge, MA, USA 02139.
E-mail: vtan@mit.edu; willsky@mit.edu