SUPPLEMENTARY MATERIAL

Supplementary Material to "High-Dimensional Structure Learning of Ising Models: Local Separation Criterion"
(http://newport.eecs.uci.edu/anandkumar/).

1. Preliminaries and Tools.

Notation. For any two functions $f(p), g(p)$, $f(p) = O(g(p))$ if there exists a constant c such that $f(p) \leq cg(p)$ for all $p \geq p_0$ for a fixed $p_0 \in \mathbb{N}$. Similarly, $f(p) = \Omega(g(p))$ if there exists a constant c' such that $f(p) \geq c'g(p)$ for all $p \geq p_0$ for a fixed $p_0 \in \mathbb{N}$, and $f(p) = \Theta(g(p))$ if $f(p) = \Omega(g(p))$ and $f(p) = O(g(p))$. Also, $f(p) = o(g(p))$ when $f(p)/g(p) \to 0$ and $f(p) = \omega(g(p))$ when $f(p)/g(p) \to \infty$ as $p \to \infty$. We use the notation $f(p) = \widetilde{O}(g(p))$ if $f(p) \leq cg(p) \log p$, for some constant c and for all $p \geq p_0$. Similarly, we have $f(p) = \widetilde{\omega}(g(p))$, if $\frac{f(p)}{g(p)\log p} \to \infty$ and $f(p) = \widetilde{o}(g(p))$ if $\frac{f(p)\log p}{g(p)} \to 0$, as $p \to \infty$.

For a graph G, let $v(G)$ denote the vertex set of G. Let $\mathcal{N}(i)$ denote the neighbors of node i and $\mathcal{N}[i]$ denote the closed neighborhood, i.e., including node i as well. We let $\mathrm{Path}(i, j; G) = \mathrm{Path}_1(i, j; G)$ denote the subgraph spanning the corresponding shortest path and $d(i, j; G) := |\mathrm{Path}(i, j; G)|$ denote the graph distance or the shortest path distance between nodes i and j. Let the set of nodes at distance[1] exactly l from i in G be denoted as

$$B_l(i; G) := \{k \in V : d(i, k; G) = l\}. \tag{1}$$

Let $\mathrm{Path}_l(i, j; G)$ denote[2] the l^{th} shortest path from i to j and $d_l(i, j; G)$ the corresponding length of the path. Let $N_l^{\mathrm{Path}}(i, j; G)$ denote the number of paths of length l from node i to node j in G without repeating any node in the intermediate steps.

Denote the correlation between any two variables X_i and X_j , $i, j \in V_p$ as

$$C(i, j) := \mathbb{E}[X_i X_j]. \tag{2}$$

Given n samples x_i^n, x_j^n drawn i.i.d. from X_i, X_j, let $\widehat{C}(i, j; x_i^n, x_j^n)$ denote the empirical correlation between node i and j is defined as

$$\widehat{C}_{i,j}^n := \widehat{C}(i, j; x_i^n, x_j^n) := \frac{1}{n} \sum_{k=1}^{n} x_{i,k} x_{j,k}. \tag{3}$$

For any distributions P, Q on a finite alphabet $\mathcal{X}$, recall that $\nu(P, Q)$ denotes the total variation distance, given by

$$\nu(P, Q) := \frac{1}{2}\|P - Q\|_1 = \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)|. \tag{4}$$

1.1. *Analysis of Ising Models on Trees.*

We first derive simple expressions for Ising models Markov on trees. This will be later used upon reduction of general models to tree models via self-avoiding walk-tree construction. We first note the correlation between any two node pairs on a tree model.

[1] We follow the convention that if l is not an integer, the distance is $\lfloor l \rfloor$.
[2] We abbreviate $\mathrm{Path}_1(i, j; G)$ as $\mathrm{Path}(i, j; G)$ and $d_1(i, j; G)$ as $d(i, j; G)$.

FACT 1 (Markov Property for Correlations on a Tree). *For a symmetric Ising model* ($\mathbf{h} = \mathbf{0}$) *Markov on a tree* T*, the correlation is given by*

$$C(i,j;T) = \prod_{(k,l)\in \mathrm{Path}(i,j;T)} C(k,l;T), \quad \forall i,j \in V, \tag{5}$$

and the correlation between any two neighbors is,

$$C(i,j;T) = \tanh(J_{i,j}), \quad \forall\, (i,j) \in T. \tag{6}$$

Proof: Eqn. (5) is obtained by successive conditioning on the intermediate nodes in the path between i and j in the tree T. Eqn. (6) is a consequence of the form of the symmetric Ising model. □

Given an Ising model P Markov on G, define a corresponding model $\widetilde{P}$ obtained by setting all the node potentials h_i to zero and all the edge potentials $J_{i,j}$ to their corresponding absolute values $|J_{i,j}|$. We term $\widetilde{P}$ as the corresponding symmetric attractive model for P. We make the following observation.

PROPOSITION 1 (Dominance by Symmetric Attractive Model on Trees). *For an Ising model* P *Markov on a tree* T *and for* $\widetilde{P}$ *its corresponding symmetric attractive model, we have*

$$\|P[X_i|X_j = +;T] - P[X_i|X_j = -;T]\|_1 \leq \|\widetilde{P}[X_i|X_j = +;T] - \widetilde{P}[X_i|X_j = -;T]\|_1 \tag{7}$$

Proof: The proof is along the lines of [3, Lemma 4.1], but we make the simple observation that it also holds when the model P is not necessarily attractive (or ferromagnetic).

We first note that it suffices to show (7) for the special case when P is a Markov chain on $k+2$ variables, for some $k \in \mathbb{N}$, i.e., the tree T is a path graph $T = i, 1, \ldots, k, j$ with i and j as endpoints. This is because we can reduce the conditional probability $P[X_i|X_j;T]$ on any tree T to a corresponding conditional probability on the path from i to j by suitably modifying the node potentials. See [3, Lemma 4.1] for details.

We now show that (7) holds when the tree is a path $T_k := i, 1, \ldots, k, j$, for all $k \in \mathbb{N}$, by doing an induction on k. For $k = 1$ (path of length two), we have[3]

$$\begin{aligned}
&\|P[X_i|X_j = +;T_1] - P[X_i|X_j = -;T_1]\|_1 \\
&= \left| \frac{e^{J_{i,1}+h_i} - e^{-J_{i,1}-h_i}}{e^{J_{i,1}+h_i} + e^{-J_{i,1}-h_i}} - \frac{e^{-J_{i,1}+h_i} - e^{J_{i,1}-h_i}}{e^{-J_{i,1}+h_i} + e^{J_{i,1}-h_i}} \right| \\
&= |\tanh(J_{i,1} + h_i) + \tanh(J_{i,1} - h_i)| \\
&= (\tanh(|J_{i,1}| + h_i) + \tanh(|J_{i,1}| - h_i)) \qquad (8) \\
&\leq \|\widetilde{P}[X_i|X_j = +;T] - \widetilde{P}[X_i|X_j = -;T]\|_1.
\end{aligned}$$

The expression in (8) has a unique maximum when $h_i = 0$ and thus, the subsequent inequality. The induction step on k now proceeds as in [3, Lemma 4.1], and we have the result. □

[3]Note the simple fact that $\|P[X_i|X_j = +] - P[X_i|X_j = -]\|_1 = |\mathbb{E}[X_i|X_j = +] - \mathbb{E}[X_i|X_j = -]|$. The result in [3, Lemma 4.1] is expressed in terms of expectations.

1.2. *Self-Avoiding Walk Tree Construction.* We now review the notion of a self-avoiding walk (SAW) tree for graphical models with binary variables, first introduced in [16]. Given an Ising model Markov on a general graph G and a particular node $i \in V$, the corresponding SAW tree rooted at i is denoted by $T_{\text{saw}}(i;G)$. It is essentially the tree of self-avoiding walks originating from node i, except that whenever a cycle in G is closed by the walk, a terminal node is included in $T_{\text{saw}}(i;G)$ and is fixed to be either $+1$ or -1; the actual value is determined by the direction in which the cycle is traversed by the walk (for instance, by convention, we can fix terminal nodes upon clockwise traversal of cycles as $+1$). Let A denote the set of all terminal nodes in $T_{\text{saw}}(i;G)$ and $\mathbf{x}_A$, the corresponding fixed configuration. In effect, $T_{\text{saw}}(i;G)$ involves conditioning with respect to the terminal nodes A. See Fig. 1 for an illustration.

We now recap a powerful result of [16] that $T_{\text{saw}}(i;G)$ preserves the marginal and conditional distributions of node i with respect to the original graph G. Recall that $N_l^{\text{Path}}(i,Q;G) = \sum_{q\in Q} N_l^{\text{Path}}(i,q;G)$ denotes the number of paths of length l from i to a set $Q \subset V$ in G, $d(i,Q;G) = \min_{q\in Q} d(i,q;G)$ denotes the graph distance, and $\mathcal{S}(i,Q;G) = \cup_{q\in Q}\mathcal{S}(i,q;G)$ denotes a vertex separator between i and Q in G. Let

$$\mathcal{U}(j;T_{\text{saw}}(i;G)) = \{j_1,\ldots,j_{|\mathcal{U}(j;T_{\text{saw}}(i;G))|}\} \subset v(T_{\text{saw}}(i;G)) \tag{9}$$

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

THEOREM 1 (Properties of $T_{\text{saw}}(i;G)$). *The following properties hold for the self-avoiding walk tree $T_{\text{saw}}(i;G)$*

1. *The marginal and conditional distributions of node i are preserved*

$$P(x_i;G) = P(x_i|\mathbf{x}_A;T_{\text{saw}}(i;G)) \tag{10}$$
$$P(x_i|\mathbf{x}_Q;G) = P(x_i|\mathbf{x}_{\mathcal{U}(Q)},\mathbf{x}_A;T_{\text{saw}}(i;G)), \tag{11}$$

   for a fixed configuration $\mathbf{x}_A$ on the set of terminal nodes A, and for any set $Q \subset V\setminus\{i\}$.
2. *The paths in G from node i to any set V are preserved in $T_{\text{saw}}(i;G)$:*

$$N_l^{\text{Path}}(i,Q;G) = N_l^{\text{Path}}(i,\mathcal{U}(Q);T_{\text{saw}}(i;G)), \quad \forall l\in\mathbb{N},\ Q\subset V\setminus\{i\}. \tag{12}$$

3. *The graph distances from node i in G and $T_{\text{saw}}(i;G)$ are equal:*

$$d(i,Q;G) = d(i,\mathcal{U}(Q);T_{\text{saw}}(i;G)), \quad \forall Q\subset V\setminus\{i\}. \tag{13}$$

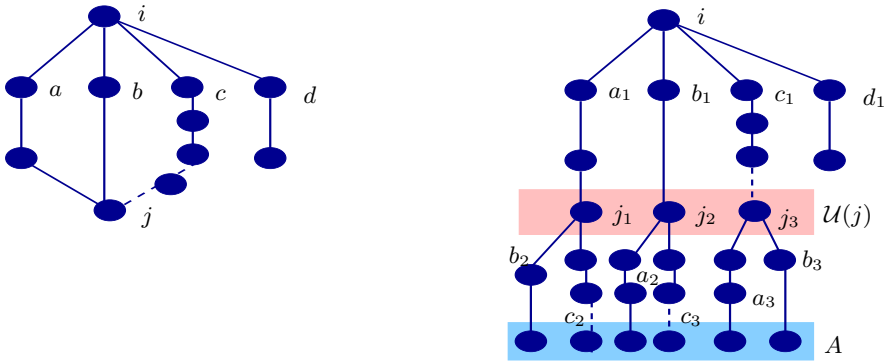
4. *The cardinality of the vertex separators are preserved:*

$$|\mathcal{S}(i,Q;G)| = |\mathcal{S}(i,\mathcal{U}(Q);T_{\text{saw}}(i;G))|, \quad \forall Q\subset V\setminus\{i\}. \tag{14}$$

5. *The maximum degrees in G and $T_{\text{saw}}(i;G)$ are equal.*

Proof: Property (1) is proven in [16]. It involves a recursive expression for marginal and conditional distributions of node i. Property (2) holds by definition since $T_{\text{saw}}(i;G)$ is constructed by self-avoiding walks from node i. Properties (3), (4) and (5) depend only on the paths in the graph and are thus preserved. □

Thus, for any graph G, we have a tree-representation $T_{\text{saw}}(i;G)$ which preserves many properties with respect to node i. However, in general, the tree $T_{\text{saw}}(i;G)$ can have exponential number of



(a) A graph

(b) Its Self-Avoiding Walk Tree

FIG 1. *The figure on the right is self-avoiding walk tree $T_{\text{saw}}(i;G)$ rooted at node i for the graph shown in the left. The set $\mathcal{U}(j)$ is the set of copies of node j and the set A is the set of terminal nodes in $T_{\text{saw}}(i;G)$.*

nodes (compared to G) and thus, we cannot use $T_{\text{saw}}(i;G)$ directly. This is also true for the class of graphs considered in this paper. However, the bound on maximum edge potentials and conditioning on local separators allows us to limit the neighborhoods under consideration.

We note the following property of graphs with local-paths property. Recall that a graph ensemble $\mathcal{G}_{\text{LP}}(p;\eta,\gamma)$ satisfies (η,γ)-local paths property if there are at most η paths of length less than γ.

LEMMA 1 (Neighborhood Size of $T_{\text{saw}}(i;G)$ for Graphs with Local-Paths Property). *For a.e. $G \sim \mathcal{G}_{\text{LP}}(p;\eta,\gamma)$ satisfying the (η,γ)-local paths property as per Definition 2 in the main paper [2], we have*

$$|B_l(i;T_{\text{saw}}(i;G))| \leq \eta |B_l(i;G)|, \quad \forall l \leq \gamma. \tag{15}$$

Proof: Recall that a.e. $G \sim \mathcal{G}_{\text{LP}}(p;\eta,\gamma)$ has at most η paths of length smaller than γ between any two nodes. This implies that there are at most η copies of any node $j \neq i$ in $T_{\text{saw}}(i;G)$ and at most η number of terminal nodes A, which are at distance at most γ from i in $T_{\text{saw}}(i;G)$ using Property (2) in Theorem 1. Thus (15) holds. □

2. Conditional Variation Distance Test.

2.1. *Conditional Uniqueness Regime.* We now characterize a sufficient condition for structure estimation of Ising models and term it as the *conditional uniqueness regime*. In Section 2, we will see that Definition 1 leads to structural consistency of the proposed CVDT algorithm. We use the term "conditional uniqueness regime", since it is similar to the so-called uniqueness regime[4], but involves the conditional distributions instead of marginal distribution. Our condition stated below, is in fact, a weaker condition than the usual notion of the uniqueness regime.

Notations: Given a graph $G = (V,E)$ and a graphical model P Markov on G, and any subset $A \subset V$, let $P[X_A;G]$ denote the marginal distribution[5] of variables in A. Recall that $d(i,j;G)$ denotes the

[4]Roughly, the uniqueness condition states that asymptotically, as the number of variables $p \to \infty$, any marginal distribution of variables in a local neighborhood of the graph is asymptotically independent of faraway variables. Refer to [10, 13] for details

[5]In the sequel, we abuse notation by using $P[X_i;G]$ to refer to the vector of length $|\mathcal{X}|$ containing the values of the pmf $P_{X_i;G}$.

graph distance, $B_l(i;G)$ denotes the set of nodes within graph distance l from node i and $\partial B_l(i)$ denotes the boundary nodes, i.e., nodes exactly at l from node i.

DEFINITION 1 (Conditional Uniqueness Regime). *A discrete graphical model P Markov on graph $G \sim \mathcal{G}(p)$ is in the conditional uniqueness regime if there exists $\alpha \in (0,1)$ such that for a.e. G and all $l \in \mathbb{N}$ such that*[6]*,*

$$\max_{\substack{(i,j)\notin V \\ \mathbf{x}_{S_l}\in\mathcal{X}^{|S_l|}}} \|P[X_i|X_j=+,\mathbf{X}_{S_l}=\mathbf{x}_{S_l}] - P[X_i|X_j=-,\mathbf{X}_{S_l}=\mathbf{x}_{S_l}]\|_1 = \widetilde{O}(\alpha^l), \tag{16}$$

where $S_l := \mathcal{S}(i,j;G,l)$ is the minimal l-local separator between i and j, according to Definition 1 in the main paper [2].

We now show that a sufficient condition for the conditional uniqueness condition in (16) to hold for Ising models is for the maximum absolute edge potential to satisfy

$$J_{\max} < J^*, \tag{17}$$

where the threshold $J^* \in \mathbb{R}^+$ is the largest value which satisfies[7], for all $l \in \mathbb{N}$,

$$\max_{i\in V} |\partial B_l(i; T_{\text{saw}}(i; F'_{S_l}))| = \widetilde{O}(\tanh J^*)^{-l}, \tag{18}$$

where $F'_{S_l} := G(V \setminus S_l)$ is the subgraph of G obtained by removing the nodes in S_l, the minimal l-local separator and $T_{\text{saw}}(i; F'_{S_l})$ is the corresponding self-avoiding walk tree rooted at i. Define

$$\alpha := \frac{\tanh J_{\max}}{\tanh J^*} < 1. \tag{19}$$

We now characterize J^* in terms of the self-avoiding walk tree.

LEMMA 2 (Sufficient Conditions for Conditional Uniqueness via $T_{\text{saw}}(i;G)$). *The Ising model satisfying (17) is in the conditional uniqueness regime according to (16) with rate α given by (19), where the threshold J^* is given by (18).*

Proof: Abbreviate the l-local separator, $S := \mathcal{S}(i,j;G,l)$. We have, for $i \in V$,

$$\begin{aligned} &\|P[X_i|X_j=+,\mathbf{X}_S=\mathbf{x}_S] - P[X_i|X_j=-,\mathbf{X}_S=\mathbf{x}_S]\|_1 \\ =&\|P[X_i|\mathbf{X}_{\mathcal{U}(j)}=+,\mathbf{X}_{\mathcal{U}(S)}=\mathbf{x}_{\mathcal{U}(S)},\mathbf{X}_A=\mathbf{x}_A;T_{\text{saw}}(i;G)] \\ &- P[X_i|\mathbf{X}_{\mathcal{U}(j)}=-,\mathbf{X}_{\mathcal{U}(S)}=\mathbf{x}_{\mathcal{U}(S)},\mathbf{X}_A=\mathbf{x}_A;T_{\text{saw}}(i;G)]\|_1 \end{aligned} \tag{20}$$

from Property (1) in Theorem 1 for self-avoiding walk trees, for a certain configuration $\mathbf{x}_A$ over the set of terminal nodes A.

Recall that $\mathcal{U}(j;T_{\text{saw}}(i;G))$ denotes the set of copies of node j in $T_{\text{saw}}(i;G)$. Recall that in $T_{\text{saw}}(i;G)$, each path starting from root node i has exactly one copy of nodes in $S\cup\{j\}$ (if the node is encountered again, a terminal node is added to $T_{\text{saw}}(i;G)$). Denote the set $\mathcal{U}_1(j;T_{\text{saw}}(i;G)) \subset$

[6] In Definition 1, we let l scale as a function of p, albeit under some restrictions depending on the graph ensemble. See Corollary 1 for some examples.

[7] In (18), we let l scale as a function of p, albeit under some restrictions depending on the graph ensemble. This implies that Definition 1 is satisfied for these regimes of l. See Corollary 1 for some examples.

$\mathcal{U}(j; T_{\text{saw}}(i;G))$ as the set, where copies of node j are encountered first before encountering the copies of nodes in S, along the paths from i in $T_{\text{saw}}(i;G)$. Similarly $\mathcal{U}_1(S; T_{\text{saw}}(i;G)) \subset \mathcal{U}(S; T_{\text{saw}}(i;G))$ denotes the set encountered before the copies of j. Let $\mathcal{U}_2(j; T_{\text{saw}}(i;G)) := \mathcal{U}(j; T_{\text{saw}}(i;G)) \setminus \mathcal{U}_1(j; T_{\text{saw}}(i;G))$ and $\mathcal{U}_2(S; T_{\text{saw}}(i;G))$ is defined similarly. See Fig.2. By definition, $\mathbf{X}_{\mathcal{U}_2(j)} - \mathbf{X}_{\mathcal{U}_1(S)} - X_i - \mathbf{X}_{\mathcal{U}_1(j)} - \mathbf{X}_{\mathcal{U}_2(S)}$ forms a Markov chain, and thus,

$$P(X_i|\mathbf{X}_{\mathcal{U}(j)}, \mathbf{X}_{\mathcal{U}(S)}, \mathbf{X}_A; T_{\text{saw}}(i;G)) = P(X_i|\mathbf{X}_{\mathcal{U}_1(j)}, \mathbf{X}_{\mathcal{U}_1(S)}, \mathbf{X}_A; T_{\text{saw}}(i;G)).$$

Substituting this equivalence into (20), we have

$$\begin{aligned}
&\|P[X_i|X_j = +, \mathbf{X}_S = \mathbf{x}_S] - P[X_i|X_j = -, \mathbf{X}_S = \mathbf{x}_S]\|_1 \\
=&\|P[X_i|\mathbf{X}_{\mathcal{U}_1(j)} = +, \mathbf{X}_{\mathcal{U}_1(S)} = \mathbf{x}_{\mathcal{U}_1(S)}, \mathbf{X}_A = \mathbf{x}_A; T_{\text{saw}}(i;G)] \\
&- P[X_i|\mathbf{X}_{\mathcal{U}_1(j)} = -, \mathbf{X}_{\mathcal{U}_1(S)} = \mathbf{x}_{\mathcal{U}_1(S)}, \mathbf{X}_A = \mathbf{x}_A; T_{\text{saw}}(i;G)]\|_1 \\
\overset{(a)}{\leq}&\|\widetilde{P}[X_i|\mathbf{X}_{\mathcal{U}_1(j)} = +; T_{\text{saw}}(i;G)] - \widetilde{P}[X_i|\mathbf{X}_{\mathcal{U}_1(j)} = -; T_{\text{saw}}(i;G)]\|_1, \\
\overset{(b)}{\leq}&\|\widetilde{P}[X_i|\mathbf{X}_{\mathcal{U}_1(j)} = +; T_{\text{saw}}(i;F'_{S_l})] - \widetilde{P}[X_i|\mathbf{X}_{\mathcal{U}_1(j)} = -; T_{\text{saw}}(i;F'_{S_l})]\|_1, \\
\overset{(c)}{\leq}&\|\widetilde{P}[X_i|\mathbf{X}_{\partial B_l(i)} = +; T_{\text{saw}}(i;F'_{S_l})] - \widetilde{P}[X_i|\mathbf{X}_{\partial B_l(i)} = -; T_{\text{saw}}(i;F'_{S_l})]\|_1 \\
\overset{(d)}{\leq}&2|\partial B_l(i; T_{\text{saw}}(i;F'_{S_l}))|(\tanh J_{\max})^l,
\end{aligned}$$

where Inequality (a) is obtained by applying Proposition 1 and involves the symmetric attractive counterpart $\widetilde{P}$ of P, obtained by setting all the node potentials $h_k = 0$ for all $k \in v(T_{\text{saw}}(i;G))$. Note that conditioning on a random variable X_k to be $+$ (resp. $-$) is equivalent to setting its node potential h_j to ∞ (resp. $-\infty$) and erasing the sub-tree beyond node k. Thus dropping conditioning and setting the node potential to zero forms an upper bound in (a).

For Inequality (b), note that in $T_{\text{saw}}(i;G)$, the paths from node i, to $\mathcal{U}_1(j)$ and $\mathcal{U}_1(S)$ are disjoint (except for node i). Thus, the conditional distribution of X_i conditioned on $\mathcal{U}_1(j)$ on $T_{\text{saw}}(i;G)$ is equivalent to a conditional distribution on $T_{\text{saw}}(i;F'_{S_l})$ obtained by marginalizing out the nodes corresponding to paths containing $\mathcal{U}_1(S)$ and suitably changing the node potential of node i. (See [9, Lemma 4.1] for an exact characterization of such a marginalization). Applying Proposition 1, we have an upper bound by setting the node potential in $T_{\text{saw}}(i;F'_{S_l})$ to zero, i.e., given by the model $\widetilde{P}$ on $T_{\text{saw}}(i;F'_{S_l})$.

For Inequality (c), recall that by definition of a l-local separator, the set $\mathcal{U}_1(j)$ has distance at least l from node i. Thus, $X_i - \mathbf{X}_{\partial B_l(i)} - \mathbf{X}_{\mathcal{U}_1(j)}$ forms a Markov chain and in an attractive model $\widetilde{P}$, the inequality (c) holds.

Inequality (d) involves considering a telescoping sum of a sequence of configurations $\lambda^0, \ldots, \lambda^{|\partial B_l(i)|}$ on $\partial B_l(i)$ from all $+$ configuration to all $-$ configuration, where the difference between the vectors λ^i and λ^{i+1} is in a single coordinate, i.e., the configuration at a single node is changed while keeping the others fixed. See [14, Lemma 2.8] for detailed discussion of this step. In particular, by applying Proposition 1, for each term involving λ^i and λ^{i+1}, the conditioning on other nodes can be dropped and we have

$$\|\widetilde{P}[X_i|\mathbf{X}_{\partial B_l(i)} = \lambda^i] - \widetilde{P}[X_i|\mathbf{X}_{\partial B_l(i)} = \lambda^{i+1}]\|_1 \leq 2(\tanh J_{\max})^l.$$

Collecting all the terms we have inequality (c), since there are $|\partial B_l(i)|$ number of terms. By definition of J^* in (17), we have that

$$|\partial B_l(i)| = \widetilde{O}(\tanh J^*)^{-l}.$$

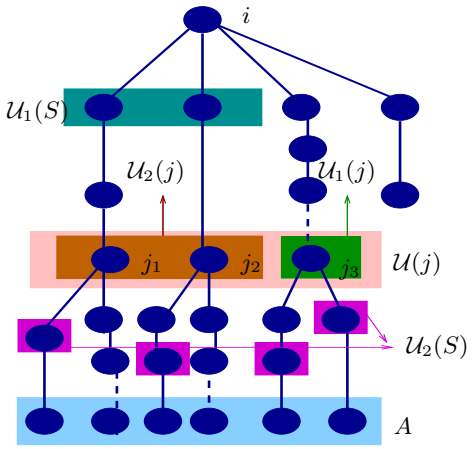


FIG 2. *Illustration of sets on $T_{\text{saw}}(i;G)$, the self-avoiding walk tree at node i corresponding to the graph in Fig.1 in the main paper [2]. $\mathcal{U}(S)$ corresponds to copies of nodes in S in the original graph in Fig.1 in the main paper [2]. The nodes j_1, j_2 and j_3 are the copies of j in $T_{\text{saw}}(i;G)$ and similarly for nodes in S. The set A is the set of terminal nodes in $T_{\text{saw}}(i;G)$. The set $\mathcal{U}_1(j)$ separates $\mathcal{U}_2(S)$ from i and viceversa.*

Now substituting α in the above equation using (19), we have the result. □

We can now obtain the threshold J^* for specific graph ensembles using the above result. Recall that $\mathcal{G}_{\text{Deg}}(p,\Delta)$ denotes a graph ensemble with maximum degree Δ, $\mathcal{G}_{\text{ER}}(p,c/p)$ denotes the Erdős-Rényi ensemble, where an edge between any two nodes occurs with probability c/p and $\mathcal{G}_{\text{Watts}}(p,d,c/p)$ denotes the small-world graph, which is the union of a d-dimensional grid and an Erdős-Rényi graph with parameter c. Recall that $\alpha := \frac{\tanh J_{\max}}{\tanh J^*}$. We have the following result.

COROLLARY 1 (Threshold J^* for Deterministic Graph Families). *We have the following results for various graph families:*

1. *For any graph ensemble $\mathcal{G}_{\text{Deg}}(p,\Delta)$ with maximum degree Δ, (16) holds for all l and (18) simplifies to*

$$J^*_{\text{Deg}} = \infty. \tag{21}$$

In particular, for every Ising model Markov on a Δ-degree bounded graph,

$$\max_{\substack{(i,j)\notin V \\ \mathbf{x}_S \in \mathcal{X}^{|S|}}} \|P[X_i|X_j=+,\mathbf{X}_S=\mathbf{x}_S] - P[X_i|X_j=-,\mathbf{X}_S=\mathbf{x}_S]\|_1 = 0, \tag{22}$$

where S is the exact separator between i and j.

2. *For the girth-bounded ensemble $\mathcal{G}_{\text{Girth}}(p;g,\Delta)$, when $2l<g$, the threshold for (16) is given by*

$$J^*_{\text{Girth}} = \operatorname{atanh}\left(\frac{1}{\Delta}\right). \tag{23}$$

In particular, in this regime, every Ising model Markov on a graph $G \in \mathcal{G}_{\text{Girth}}(p;g,\Delta)$ satisfies

$$\max_{\substack{(i,j)\notin V \\ \mathbf{x}_{S_l} \in \mathcal{X}^{|S_l|}}} \|P[X_i|X_j=+,\mathbf{X}_{S_l}=\mathbf{x}_{S_l}] - P[X_i|X_j=-,\mathbf{X}_{S_l}=\mathbf{x}_{S_l}]\|_1 \le 2\alpha^l, \tag{24}$$

when $2l < g$, where g is the girth of the graph, and $S_l := \mathcal{S}(i,j;G,l)$ is the minimal l-local separator between i and j and satisfies $|S_l| \leq 1$.

We provide probabilistic bounds for random graph families.

COROLLARY 2 (Threshold J^* for Random Graph Families). *We have the following results for various graph families:*

1. *For the random-regular graphs $\mathcal{G}_{\text{Reg}}(p,\Delta)$, (16) is satisfied when $l = O(\log_{\Delta-1} p)$, $\Delta = O(\text{poly}\log p)$, the threshold is given by*

$$J^*_{\text{Reg}} = \text{atanh}\left(\frac{1}{\Delta}\right). \tag{25}$$

In particular, in this regime, for every Ising model Markov on a Δ-random regular graph, when $l < 0.25(0.25p\Delta + 0.5 - \Delta^2)$, with probability at least $1 - \Delta^{16l-2}(p\Delta - 4\Delta^2 - 16l)^{-(8l-1)}$, we have

$$\max_{\substack{(i,j)\notin V \\ \mathbf{x}_{S_l}\in\mathcal{X}^{|S_l|}}} \|P[X_i|X_j = +, \mathbf{X}_{S_l} = \mathbf{x}_{S_l}] - P[X_i|X_j = -, \mathbf{X}_{S_l} = \mathbf{x}_{S_l}]\|_1 \leq 2\alpha^l, \tag{26}$$

where $S_l := \mathcal{S}(i,j;G,l)$ is the minimal l-local separator between i and j and satisfies $|S_l| \leq 2$.

2. *For both the Erdős-Rényi ensemble $\mathcal{G}_{\text{ER}}(p, c/p)$ and the small-world graph ensemble $\mathcal{G}_{\text{Watts}}(p,d,c/p)$, (16) holds when $l \leq \frac{\log p}{4\log c}$ and $c = O(\text{poly}\log p)$ with thresholds given by*

$$J^*_{\text{ER}} = J^*_{\text{Watts}} = \text{atanh}\left(\frac{1}{c}\right). \tag{27}$$

In particular, in this regime, when $l < \frac{\log p}{4\log c}$ and $1 < c = O(\text{poly}\log p)$, with probability at least $1 - le^{\sqrt{125}}p^{-2.5} - l!c^{4l+1}p^{-1}$, we have

$$\max_{\substack{(i,j)\notin V \\ \mathbf{x}_{S_l}\in\mathcal{X}^{|S_l|}}} \|P[X_i|X_j = +, \mathbf{X}_{S_l} = \mathbf{x}_{S_l}] - P[X_i|X_j = -, \mathbf{X}_{S_l} = \mathbf{x}_{S_l}]\|_1 \leq 4l^3\alpha^l \log p, \tag{28}$$

and $S_l := \mathcal{S}(i,j;G,l)$ is the minimal l-local separator between i and j and satisfies $|S_l| \leq 2$ for the Erdős-Rényi ensemble $\mathcal{G}_{\text{ER}}(p,c/p)$ and $|S_l| \leq d+2$ for the small-world graph ensemble $\mathcal{G}_{\text{Watts}}(p,d,c/p)$.

Remarks:

1. Comparing (21), (23) and (25), we note that for the degree-bounded ensemble $J^*_{\text{Deg}} = \infty$ meaning that we do not place any restrictions on the maximum potential $J_{\max}$, while for the girth bounded ensemble and the random regular ensemble $J^*_{\text{Girth}} = J^*_{\text{Reg}} = 1/\Delta$. This is because the minimal l-local separators are different for these two ensembles. For $\mathcal{G}_{\text{Deg}}(p,\Delta)$, it has cardinality Δ and thus, forms an exact separator. On the other hand, for $\mathcal{G}_{\text{Girth}}(p;g,\Delta)$ and $\mathcal{G}_{\text{Reg}}(p,\Delta)$, the minimal l-local separators have cardinalities 1 and 2 when $2l < g$ and $l = O(\log_{\Delta-1} p)$ respectively, and thus, do not form an exact separator. Thus, the threshold J^* depends on whether exact or approximate separators are used for conditioning.

2. Comparing the thresholds for random regular ensemble in (25) and the Erdős-Rényi ensemble in (27), we see that $J^*_{\text{ER}} \gg J^*_{\text{Reg}}$, if we constrain the maximum degrees in the two ensembles to be the same. Recall that the maximum degree of the Erdős-Rényi ensemble is a.a.s. $\Delta = \Theta(\log p \log c/ \log\log p)$. Thus, by obtaining the threshold J^*_{ER} in terms of the average degree c instead of the maximum degree, we have a larger threshold and thus, can provide guarantees for structure estimation of Erdős-Rényi graphs for a wider regime of edge potentials.
3. Comparing the thresholds for the Erdős-Rényi ensemble $\mathcal{G}_{\text{ER}}(p, c/p)$ and the small-world ensemble $\mathcal{G}_{\text{Watts}}(p, d, c/p)$ in (27), we see that $J^*_{\text{ER}} = J^*_{\text{Watts}}$, but note that the minimal l-local separators are different for these two ensembles. For the Erdős-Rényi ensemble, it has a cardinality of two when $l \leq \frac{\log p}{4\log c}$, as discussed above. For the small-world ensemble, which is the union of a d-dimensional grid and an Erdős-Rényi graph, the minimal l-local separator has a cardinality of $d+2$ when $l \leq \frac{\log p}{4\log c}$ and it forms an exact separator on the grid. Thus, for the small-world graphs, we require a threshold J^*_{Watts} such that the long paths on the Erdős-Rényi subgraph has a decaying effect, leading to the same threshold on the edge potentials ($J^*_{\text{Watts}} = J^*_{\text{ER}}$).

Proof: The result in Eqn. (21) is from the definition of graphical models: the size of the minimal l-local separator for $\mathcal{G}_{\text{Deg}}(p, \Delta)$ ensemble is of size Δ for all $l \in \mathbb{N}$. This implies that $T_{\text{saw}}(i; F'_{S_l})$ has no edges and thus, J^*_{Deg} is infinite.

The result in Eqn. (23) is obtained from the fact that the l-local separator is of size 1 when $2l < g$ since we do not encounter any cycles. In this case, we can bound the neighborhood of $T_{\text{saw}}(i; F'_{S_l})$ via $T_{\text{saw}}(i; G)$ and using Property (5) in Theorem 1, we have the result.

For the result in Eqn. (25), note that the size of minimal l-local separator for $\mathcal{G}_{\text{Reg}}(p, \Delta)$ is 1, when $l = O(\log_{\Delta-1} p)$ [5, p. 107]. In this case, we can bound the neighborhood of $T_{\text{saw}}(i; F'_{S_l})$ via $T_{\text{saw}}(i; G)$ and using Property (5) in Theorem 1, we have the result. For the result in (26), we appeal to [12, Thm. 3] and derive the probability of two cycles each of length at most $2l$ overlapping with one another.

For the result in (27), we appeal to [6, Lemma 1] that with probability at least $1 - le^{\sqrt{125}}p^{-2.5}$, for all $l \in \mathbb{N}$, when $c > 1$,

$$\max_{i \in V} |B_l(i)| \leq 2l^3 c^l \log p. \tag{29}$$

When $l \leq \frac{\log p}{4\log c}$, with probability at least $1 - l!c^{4l+1}p^{-1}$ [1, Lemma 2], there is at most one cycle in $B_l(i)$ for all $i \in V$. From Lemma 1, we have the result. When $c = O(\text{poly}\log p)$, we have $\frac{\log p}{\log c} = \omega(1)$, and thus J^*_{ER} holds.

For the small-world graph ensemble $\mathcal{G}_{watts}(p, d, c/p)$, which is the union of the d-dimensional grid and Erdős-Rényi graph, the size of the minimal l-local separator is $d+2$, when $l \leq \frac{\log p}{4\log c}$. Since F_{S_l} is dominated by the Erdős-Rényi graph, the result holds. □

2.1.1. *Uniqueness Regime.* We now relate the conditional-uniqueness regime to the well-known notion of the *uniqueness regime*[8] of an Ising model.

Intuitively, in the uniqueness regime, as the number of nodes $p \to \infty$, any marginal distribution of variables in a local neighborhood of the graph is asymptotically independent of faraway variables. We formally define it below. Recall that we say $f(p) = \widetilde{O}(g(p))$ if $f(p) \leq Mg(p)\log p$ for some constant M and $p > p_0$ and $F_l(i; G)$ denotes the spanning subgraph of the l-hop neighborhood of node i.

[8]For uniqueness regime, we consider the notion of weak spatial mixing and limit to exponential decay of correlations. Refer to [10, 13] for other notions of correlation decay.

Definition 2 (Uniqueness Regime). *A discrete graphical model P Markov on graph $G \sim \mathcal{G}(p)$ is in the uniqueness regime if there exists $\alpha \in (0,1)$ such that for a.e. G and all $l \in \mathbb{N}$,*

$$\max_{i \in V} \|P[X_i; G] - P[X_i; F_l(i; G)]\|_1 = \widetilde{O}(\alpha^l). \tag{30}$$

Comparing the above definition of the uniqueness regime and the conditional uniqueness regime in Definition 1, we note that the requirement for uniqueness regime is stronger. This is because for uniqueness regime, we require that the "faraway" nodes have a decaying effect on node marginal distributions, while for conditional uniqueness, we only require it upon conditioning on local separators. Note that conditioning itself removes the effect of a subset of "faraway" nodes and thus, conditional uniqueness is a weaker requirement. The notion of uniqueness regime is well-studied (see [10, 13]) and has many implications. For instance, the mixing time of Gibbs sampling is polynomial (in the number of nodes) in the uniqueness regime.

We now note sufficient condition for the uniqueness condition in (30) on lines of analysis in the previous section by requiring the maximum absolute edge potential of the Ising model to satisfy

$$J_{\max} < \widetilde{J}^*, \tag{31}$$

where the threshold $\widetilde{J}^* \in \mathbb{R}^+$ is the largest value which satisfies, for all $l \in \mathbb{N}$,

$$\max_{i \in V} |\partial B_l(i; T_{\text{saw}}(i; G))| = \widetilde{O}(\tanh \widetilde{J}^*)^l. \tag{32}$$

The proof is on similar lines as that of Lemma 2 and is omitted.

On lines of Corollary 1, we can obtain the threshold $\widetilde{J}^*$ in explicit form for many graph families. Recall that $\mathcal{G}_{\text{Deg}}(p, \Delta)$ denotes any graph ensemble with maximum degree Δ and $\mathcal{G}_{\text{ER}}(p, c/p)$ denotes the Erdős-Rényi ensemble, where an edge between any two nodes occurs with probability c/p.

Corollary 3 (Threshold for Uniqueness). *For a degree-bounded graph ensemble $\mathcal{G}_{\text{Deg}}(p, \Delta)$, (32) simplifies to*

$$\widetilde{J}^*_{\text{Deg}} = \operatorname{atanh}\left(\frac{1}{\Delta}\right). \tag{33}$$

The above threshold can be improved for the Erdős-Rényi ensemble $\mathcal{G}_{\text{ER}}(p, c/p)$ as

$$\widetilde{J}^*_{\text{ER}} = \operatorname{atanh}\left(\frac{1}{c}\right), \quad c = O(\text{poly} \log p). \tag{34}$$

Remarks:

Comparing the thresholds J^* and $\widetilde{J}^*$ for conditional uniqueness and uniqueness, we note that $J^* \geq \widetilde{J}^*$. The difference between J^* and $\widetilde{J}^*$ is the largest upon exact separation. For instance, in a $(\Delta - 1)$-regular tree with degree Δ, the uniqueness threshold $\widetilde{J}^* = 1/\Delta$, while the conditional uniqueness is $J^* = \infty$ with $\eta = 1$, since upon (exact) separation, there is no effect of faraway nodes. Thus, our criterion of conditional uniqueness is weaker than the usual notion of uniqueness. This implies that we can guarantee efficient structure estimation in high dimensions for a wide range of models.

2.2. *Conditional Variation Distance Between Non-Neighbors.* Recall that

$$\nu_{i|j;S} := \min_{\mathbf{x}_S \in \{-1,+1\}^{|S|}} \nu(P(X_i|X_j=+,\mathbf{X}_S=\mathbf{x}_S), P(X_i|X_j=-,\mathbf{X}_S=\mathbf{x}_S)), \tag{35}$$

where $\nu(\cdot,\cdot)$ denotes total variation distance. Using the notion of conditional uniqueness regime from Section 2.1, we immediately obtain a bound for the conditional variation distance between non-neighbors of an Ising model, when the conditioning set is a l-local separator.

LEMMA 3 (Conditional Variation Distance Between Non-Neighbors). *Given an Ising model satisfying conditional uniqueness regime according to Definition 1, for graphs satisfying (η,γ)-local separation property with $\eta = O(1)$, we have*

$$\nu_{\max}(p;\eta) := \max_{(i,j)\notin G} \min_{\substack{|S|\le\eta \\ S\subset V\setminus\{i,j\}}} \nu_{i|j;S} = \widetilde{O}(\alpha^{\gamma}). \tag{36}$$

2.3. *Conditional Variation Distance Between Neighbors.* We now provide a lower bound on the conditional variation distance between neighbors. This implies that we can distinguish edges and non-edges through conditional variation distance thresholding. We first provide explicit bounds for special cases such as attractive models. Using analytic theory, this implies that the bound also holds for generic values of edge potentials.

2.3.1. *Attractive Models.* We first carry out the analysis for attractive models ($J_{i,j} \ge 0$ for all $(i,j)\in G$).

PROPOSITION 2 (Variation Distance between Neighbors). *For attractive Ising models Markov on graph G with maximum degree Δ having edge potentials $J_{\max} \ge J_{i,j} \ge J_{\min} > 0$ and node potentials $0 \le h_i \le h_{\max}$, for any set $S \subset V\setminus\{i,j\}$,*

$$\min_{\substack{(i,j)\in G \\ \mathbf{x}_S\in|\mathcal{X}|^{|S|}}} \nu_{i|j;S} \ge \frac{1}{2}\left(\tanh(J_{\min}+h'_{\max}) + \tanh(J_{\min}-h'_{\max})\right), \tag{37}$$

where $h'_{\max}$ is the modified node potential due to conditioning and marginalization.

Proof: Using self-avoiding walk tree construction, we have, for any $\mathbf{x}_S \in \mathcal{X}^{|S|}$,

$$\begin{aligned}
&\nu(P[X_i|X_j=+,\mathbf{x}_S], P[X_i|X_j=-,\mathbf{x}_S]) \\
&\overset{(a)}{=} \nu(P[X_i|\mathbf{X}_{\mathcal{U}(j)}=+,\mathbf{x}_{\mathcal{U}(S)},\mathbf{x}_A;T_{\text{saw}}(i;G)], P[X_i|\mathbf{X}_{\mathcal{U}(j)}=-,\mathbf{x}_{\mathcal{U}(S)},\mathbf{x}_A;T_{\text{saw}}(i;G)]) \\
&\overset{(b)}{\ge} \nu(P[X_i|\mathbf{X}_{j_1}=+,\mathbf{x}_{\mathcal{U}(S)},\mathbf{x}_A;T_{\text{saw}}(i;G)], P[X_i|\mathbf{X}_{j_2}=-,\mathbf{x}_{\mathcal{U}(S)},\mathbf{x}_A;T_{\text{saw}}(i;G)]) \\
&\overset{(c)}{=} \frac{1}{2}\left(\tanh(J_{i,j}+h'_i) + \tanh(J_{i,j}-h'_i)\right),
\end{aligned}$$

where equality (a) is from self-avoiding walk tree construction $T_{\text{saw}}(i;G)$, inequality (b) is true for attractive models and j_1 refers to the copy of node j in $T_{\text{saw}}(i;G)$ occurring as neighbor of i in $T_{\text{saw}}(i;G)$ and equality (c) is from the fact that the effect of terminal nodes A and conditioning set S and marginalization over other nodes is to change the node potential of i to h'_i. □

2.3.2. *Generic Edge Potentials.* When the Ising model is not necessarily attractive, it is harder to obtain lower bounds for conditional variation distance between neighbors, for any conditioning set. Note that the case where the neighbors are marginally independent belongs to the class of non-attractive models, and in this case, our method fails to recover the edge. We now show that such instances, where our method fails, form a set of Lebesgue measure zero, and that the bound established for attractive models also holds for general models under generic edge potentials.

We first note the following result on analytic functions [11, Lemma 2].

LEMMA 4 (Property of Analytic Functions). *For an analytic function $f(\mathbf{y})$ for $\mathbf{y} \in D \subset \mathbb{R}^m$, if f is non-trivial, i.e., there exists $\mathbf{y}_0 \in D$ such that $f(\mathbf{y}_0) \neq 0$, then the set where f vanishes has Lebesgue measure zero.*

Since the conditional variation distance is $\nu_{i|j;S}$ is an analytic function of the edge potentials $\mathbf{J} := [J_{e_1}, \ldots, J_{e_m}]$, we have the following result.

PROPOSITION 3 (Variation Distance under Generic Potentials). *For an Ising models Markov on graph G with edge potentials $|J_{i,j}| \geq J_{\min}$, we have for any $S \subset V \setminus \{i, j\}$,*

$$\min_{\substack{(i,j)\in G \\ \mathbf{x}_S \in |\mathcal{X}|^{|S|}}} \nu_{i|j;S} = \Omega(J_{\min}). \tag{38}$$

Proof: We have that the function $f(\mathbf{J}) := \nu_{i|j;S} - k \min_{i\neq j}(\mathbf{J})$, is an analytic function of the edge potentials $\mathbf{J} := [J_{e_1}, \ldots, J_{e_m}]$, for a suitable constant k. Since $f(\mathbf{J}) > 0$ for an attractive model ($J_{ij} \geq 0$), for a suitable constant $k > 0$, we have that the set of edge potentials $\mathbf{J}$ where $f(\cdot)$ vanishes is of measure zero. Thus, for generic edge potentials, $\nu_{i|j;S} = \Omega(J_{\min})$. □

2.3.3. *Graphs with Local Paths.* In the previous section, we established the bound for generic edge potentials. We now establish a stronger result that the bound holds *for all* edge potentials for a limited set of graphs: the class of graphs $\mathcal{G}_{\text{LP}}(p; \eta, \gamma)$ satisfying the (η, γ)-local paths property. Recall that these graphs have at most η paths of length less than γ.

LEMMA 5 (Variation Distance between Neighbors). *Under assumptions (A2)–(A3) in Section 3.1 of the main paper [2], for an Ising model Markov on a graph $G \sim \mathcal{G}(p; \eta, \gamma)$ satisfying the (η, γ) local-paths property and the model is in the uniqueness regime according to (30), we have*

$$\nu_{i|j;S} = \Omega(J_{\min}), \quad \forall (i, j) \in G, S \subset V \setminus \{i, j\}, |S| = O(1), \tag{39}$$

where $J_{\min} \leq |J_{i,j}| \leq J_{\max}$, for all $(i, j) \in G$, and there exists a constant $\delta > 0$ such that

$$\frac{J_{\min}}{(\eta - 1)J_{\max}^2} > 1 + \delta. \tag{40}$$

Proof: Denote the subset of copies of any node j in the self-avoiding walk tree $T_{\text{saw}}(i; G)$ rooted at a node i with distance smaller than γ as

$$\widetilde{\mathcal{U}}_\gamma(j; T_{\text{saw}}(i; G)) := \{j_k \in \mathcal{U}(j; T_{\text{saw}}(i; G)) : d(i, j_k; T_{\text{saw}}(i; G)) \leq \gamma\}. \tag{41}$$

We now have

$$\nu(P[X_i|X_j = +, \mathbf{x}_S], P[X_i|X_j = -, \mathbf{x}_S])|$$

$$\stackrel{(a)}{=}\nu(P[X_i|\mathbf{X}_{\mathcal{U}(j)}=+,\mathbf{x}_{\mathcal{U}(S)},\mathbf{x}_A;T_{\text{saw}}(i;G)],P[X_i|\mathbf{X}_{\mathcal{U}(j)}=-,\mathbf{x}_{\mathcal{U}(S)},\mathbf{x}_A;T_{\text{saw}}(i;G)])$$
$$\stackrel{(b)}{\geq}\nu(P[X_i|\mathbf{X}_{\widetilde{\mathcal{U}}_\gamma(j)}=+,\mathbf{x}_{\widetilde{\mathcal{U}}_\gamma(S)},\mathbf{x}_{A\cap B_\gamma(i)}],P[X_i|\mathbf{X}_{\widetilde{\mathcal{U}}_\gamma(j)}=-,\mathbf{x}_{\widetilde{\mathcal{U}}_\gamma(S)},\mathbf{x}_{A\cap B_\gamma(i)}])-\widetilde{O}(\alpha^\gamma)$$
$$\stackrel{(c)}{=}\frac{1}{2}\left(\tanh\left[|J_{i,j}+J'_{i,j}|+|h'_i|\right]+\tanh\left[|J_{i,j}+J'_{i,j}|-|h'_i|\right]\right)-\widetilde{O}(\alpha^\gamma)$$
$$\stackrel{(d)}{\geq}\frac{1}{2}\left(\tanh\left[|J_{\min}-(\eta-1)J^2_{\max}|+|h'_i|\right]+\tanh\left[|J_{\min}-(\eta-1)J^2_{\max}|-|h'_i|\right]\right)-\widetilde{O}(\alpha^\gamma)$$
$$\stackrel{(e)}{=}\Omega(\tanh J_{\min})$$

where equality (a) is from the equivalence of conditional distributions on the self-avoiding walk tree (Theorem 1). For equality (b), recall that $\widetilde{\mathcal{U}}(j;\gamma)$ defined in (41), denotes the copies of node j in $T_{\text{saw}}(i;G)$, which are at distance smaller than γ from root i. For equality (b), note that the uniqueness condition, according to (30), states that the effect of nodes beyond $B_\gamma(i)$ decays as $\widetilde{O}(\alpha^l)$. Equality (c) arises from the self-avoiding walk tree configuration. The parameter h'_i is the modified node potential due to conditioning on nodes in $\mathcal{U}(S;\gamma)$ and $A\cap B_\gamma(i)$ and marginalization of the other nodes and is bounded since we condition on finite number of nodes. The parameter $J_{i,j}$ is due to the contribution of the direct path (edge) from i to j while $J'_{i,j}$ is the contribution of all other paths from i to j of length less than γ.

Inequality (d) arises from the (η,γ)-local paths property, which implies that there are at most η copies of any node in $T_{\text{saw}}(i;G)$ within distance γ from the root (Lemma 1). This implies that the worst-case configuration is when one path from i to a copy of j through the edge (i,j) having a minimum edge potential (i.e., $J_{i,j}=J_{\min}$ and all the other paths to copies of j having the maximum potential but with the opposite sign, i.e., $J'_{i,j}=-(\eta-1)J^2_{\max}$. This is because all the other paths are at least two hops away from i. Equality (e) arises when $\frac{J_{\min}}{(\eta-1)J^2_{\max}}$ is bounded away from one (and larger than one), and from assumption (A3), we have $J_{\min}\alpha^{-\gamma}=\widetilde{\omega}(1)$. □

3. Sample-Based Analysis of CVDT.

3.1. *Concentration of Empirical Variation Distances.* We have so far established bounds on conditional variation distance in graphs with local-separation property. We now provide concentration results for empirical variation distance estimated from samples. We use the following result on empirical distribution [15, Thm. 2.1].

LEMMA 6 (Guarantees for General Empirical Distribution). *The following is true for the empirical distribution $\widehat{P}^n$, obtained using n i.i.d. samples from a discrete distribution P:*

$$\mathbb{P}[\nu(\widehat{P}^n,P)>\epsilon]\leq 2^{|\mathcal{X}|}\exp[-2n\epsilon^2]. \tag{42}$$

LEMMA 7 (Concentration Bounds). *Given n i.i.d. samples from P, we have for all $\delta>0$,*

$$\mathbb{P}\left[\max_{\substack{i,j\in V,|S|\leq\eta\\ S\in V\setminus\{i,j\}}}|\widehat{\nu}^n_{i|j;S}-\nu_{i|j;S}|>\delta\right]\leq 2^{\eta+3}p^{\eta+2}\exp\left[-\frac{nP^2_{\min}\delta^2}{2(\delta+2)^2}\right]. \tag{43}$$

Proof: From Lemma 6,

$$\mathbb{P}\left[\|\widehat{P}^n(X_i,\mathbf{X}_S,X_j)-P(X_i,\mathbf{X}_S,X_j)\|_1>\delta_1\right]\leq 2^{\eta+2}\exp[-n\delta_1^2/2],$$

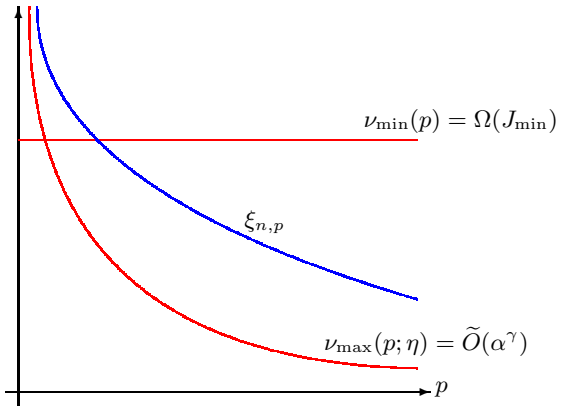


FIG 3. *The threshold $\xi_{n,p}$ in* CVDT *algorithm separates edges and non-edges with high probability. $\nu_{\min}$ and $\nu_{\max}$ are defined in (48) and (50). In the above figure, it is assumed that $\nu_{\min} = O(1)$.*

$$\mathbb{P}\left[\|\widehat{P}^n(\mathbf{X}_S, X_j) - P(\mathbf{X}_S, X_j)\|_1 > \delta_2\right] \leq 2^{\eta+1} \exp[-n\delta_2^2/2].$$

Under the event, that $\|\widehat{P}^n(X_i, \mathbf{X}_S, X_j) - P(X_i, \mathbf{X}_S, X_j)\|_1 \leq \delta_1$ and $\|\widehat{P}^n(\mathbf{X}_S, X_j) - P(\mathbf{X}_S, X_j)\|_1 \leq \delta_2$,

$$\|\widehat{P}^n(X_i|\mathbf{X}_S = \mathbf{x}_S, X_j = x_j) - P(X_i, |\mathbf{X}_S = \mathbf{x}_S, X_j = x_j)\|_1 \leq \frac{\delta_1 + \delta_2}{P_{\min} - \delta_2}.$$

If we require a bound of δ for $\|\widehat{P}^n(X_i|\mathbf{X}_S = \mathbf{x}_S, X_j = x_j) - P(X_i|\mathbf{X}_S = \mathbf{x}_S, X_j = x_j)\|_1$, we can choose $\delta_2 = k\delta P_{\min}$ and $\delta_1 = P_{\min}\delta(1 - k - k\delta)$. Setting $k = 1/(\delta + 2)$ gives the optimal exponent. □

3.2. *Asymptotic Guarantees for* CVDT. We first provide rough asymptotic arguments for recovery under CVDT. We then sharpen them to finite sample complexity results. For any $(i, j) \notin G_p$, define the event

$$\mathcal{F}_1(i, j; \{\mathbf{x}^n\}, G_p) := \left\{\widehat{\nu}_{i|j;S} > \xi_{n,p}\right\}, \tag{44}$$

where $\xi_{n,p}$ is the threshold in (20) in the main file [2]. Similarly for any edge $(i, j) \in G_p$, define the event that

$$\mathcal{F}_2(i, j; \{\mathbf{x}^n\}, G_p) := \left\{\widehat{\nu}_{i|j;S} < \xi_{n,p}\right\}. \tag{45}$$

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

$$\begin{aligned} P[\mathsf{CVDT}(\{\mathbf{x}^n\}; \xi_{n,p}; \eta) \neq G_p] \leq\ & P\left[\bigcup_{(i,j)\in G_p} \mathcal{F}_2(i, j; \{\mathbf{x}^n\}, G_p)\right] \\ & + P\left[\bigcup_{(i,j)\notin G_p} \mathcal{F}_1(i, j; \{\mathbf{x}^n\}, G_p)\right] \end{aligned} \tag{46}$$

For the first term, applying the concentration result in (43) of Lemma 7,

$$P\left[\bigcup_{(i,j)\in G_p} \mathcal{F}_2(i, j; \{\mathbf{x}^n\}, G_p)\right] = O(p^{\eta+2} \exp[-nO(\nu_{\min} - \xi_{n,p})^2]) \tag{47}$$

where

$$\nu_{\min} := \min_{(i,j)\in G_p} \min_{\substack{|S|\leq\eta \\ S\subset V\setminus\{i,j\}}} \nu_{i|j;S} = \Omega(J_{\min}), \tag{48}$$

from Lemma 5. Since $\xi_{n,p} = O(J_{\min})$, (47) is $o(1)$ when $n = \Omega(J_{\min}^{-2}\log p)$. For the second term in (46),

$$P\left[\bigcup_{(i,j)\notin G_p} \mathcal{F}_1(i,j;\{\mathbf{x}^n\},G_p)\right] = O(p^{\eta+2}\exp[-nO(\xi_{n,p}-\nu_{\max})^2]), \tag{49}$$

where

$$\nu_{\max}(p;\eta) := \max_{(i,j)\notin G_p} \min_{\substack{|S|\leq\eta \\ S\subset V\setminus\{i,j\}}} \nu_{i|j;S} = \widetilde{O}(\alpha^\gamma), \tag{50}$$

from (36). For the choice of $\xi_{n,p}$ in (20) in the main paper [2], (49) is $o(1)$. □

3.3. *PAC Guarantees for* CVDT. We now sharpen the results of the previous section to provide finite sample complexity bounds. Recall that

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

Given a fixed $\delta > 0$, recall that we choose threshold $\xi_{n,p}$ as

$$\xi_{n,p}(\delta) = \nu_{\max}(p;\eta) + \delta. \tag{51}$$

On lines of the error events (44) and (45) defined in the previous section and using the concentration bounds in Lemma 7, we have that

$$\mathbb{P}[\mathsf{CVDT}(\{\mathbf{x}^n\};\xi_{n,p}(\delta);\eta) \neq G'_{p;\delta}] \leq 2^{\eta+4}p^{\eta+2}\exp\left[-\frac{2n\delta^2 P_{\min}^2}{(\delta+2)^2}\right].$$

The results of Lemma 1 in the main paper [2] follow from Corollaries 1 and 2. □

4. Necessary Conditions for Structure Estimation.

4.1. *Erdős-Rényi Random Graphs.* This proof is inspired by [4, Thm. 1]. Fix any deterministic estimator $\widehat{G}_p$. Denote $\mathcal{R} := \widehat{G}_p((\mathcal{X}^p)^n)$ as the range of the estimator $\widehat{G}_p$. This is the set of all graphs that can be output by the estimator $\widehat{G}_p$. Then we have the sequence of lower bounds:

$$\begin{aligned}
\mathbb{P}_{\mathbf{X}^n,G_p}(\widehat{G}_p \neq G_p) &\overset{(a)}{=} \sum_{g\in\mathcal{R}^c} \mathbb{P}_{\mathbf{X}|G_p=g}(\widehat{G}_p \neq G_p|G_p=g)\mathbb{P}_{G_p}(G_p=g) \\
&\quad + \sum_{g\in\mathcal{R}} \mathbb{P}_{\mathbf{X}|G_p}(\widehat{G}_p \neq G_p|G_p=g)\mathbb{P}_{G_p}(G_p=g) \\
&\overset{(b)}{\geq} \sum_{g\in\mathcal{R}^c} \mathbb{P}_{\mathbf{X}|G_p}(\widehat{G}_p \neq G_p|G_p=g)\mathbb{P}_{G_p}(G_p=g)
\end{aligned}$$

$$\begin{aligned}
&\overset{(c)}{=} \sum_{g\in\mathcal{R}^c} \mathbb{P}_{G_p}(G_p = g) \\
&\overset{(d)}{=} 1 - \sum_{g\in\mathcal{R}} \mathbb{P}_{G_p}(G_p = g),
\end{aligned} \tag{52}$$

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

Now we provide an asymptotic upper bound for the term

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

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

$$\mathbb{P}_{G_p}(G_p = g) = \left(\frac{c}{p}\right)^{k_g} \left(1 - \frac{c}{p}\right)^{\binom{p}{2} - k_g}. \tag{53}$$

Eqn. (53) says that if the probability of edge appearance $c/p < 1/2$ (which is the case of interest) then $\mathbb{P}(G_p = g)$ is maximized at $k_g = 0$. In fact, we have the general result that for graphs $g_1, g_2 \in \mathcal{G}_p$

$$k_{g_1} \leq k_{g_2} \quad \Rightarrow \quad \mathbb{P}_{G_p}(G_p = g_1) \geq \mathbb{P}_{G_p}(G_p = g_2). \tag{54}$$

It is then straightforward to show that the natural number

$$z := \min\left\{ l \in \mathbb{N} : \sum_{k=1}^{l} \binom{\binom{p}{2}}{k} \geq 2^{nm} \right\} \tag{55}$$

is of the order $nm/\log p$ (by solving for l in (55)). The quantity z defined in (55) 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,

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

where (a) follows from the definition of Υ, (b) follows from rewriting Υ in terms of z, the number of edges and by using (53), (c) follows from (55), 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 $N = \binom{p}{2}$ and $q = c/p$. Finally, we observe from (d) that if $n = ac \log p$ for some $a > 0$, then the term $\Upsilon \to 0$ as $p \to \infty$. Thus, referring back to (52) and noting the arbitrariness of $\widehat{G}_p$, we conclude that if $n \leq \epsilon c \log p$ for sufficiently small $\epsilon > 0$, then $\mathbb{P}_{\mathbf{X}^n, G_p}(\widehat{G}_p \neq G_p) \to 1$. □

4.2. *Other Graph Families. Proof of Lemma 2 of main paper [2]:* The proof is by counting arguments. For girth-bounded graphs, we prove by recursively adding edges. At each stage, one endpoint of the edge can be picked out of p nodes while the other end point cannot be a node within g-hop neighborhood of the first end point. The number of such nodes is at least $\Delta_{\min}^{g}$ and at most $\sum_{i=1}^{g}\Delta_{\max}^{g} \leq g\Delta_{\max}^{g}$. By recursively adding edges we have the result.

We now consider local-paths graphs. Given a graph G, form a partition of nodes such that nodes in the same partition have graph distance at most γ. The number of partitions is at least $m_1 := p/\gamma\Delta_{\max}^{\gamma}$ and at most $m_2 := p/\Delta_{\min}^{\gamma}$. In each partition, the tree excess (additional edges compared to a tree) is $\eta - 1$ from local paths property. Thus, if these edges are removed from all partitions, we obtain a graph with girth γ with number of edges in $[k_1, k_2]$, and use the bound previously derived. We finally note that in each partition, the $\eta - 1$ edges can be chosen arbitrarily given the graph of girth γ.

For augmented graphs, the result is straightforward by noting that there $p\binom{p-1}{d}$ regular graphs of degree d. □

4.3. *Proof of Theorem 4 in Main Paper [2].* Let $\mathfrak{G}_n$ denote the family of undirected labeled graphs with n nodes. Let $\mathcal{G}(n, \frac{c}{n})$ denote the Erdős-Rényi ensemble. A random graph $G \in \mathfrak{G}_n$ is drawn from $\mathcal{G}(n, \frac{c}{n})$. There are also m conditionally i.i.d. samples $\mathbf{X}^m := (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(m)}) \in (\mathcal{X}^n)^m$ drawn from $p_{\mathbf{X}|G}^{m}$. These samples are then used to estimate the underlying random graph G. The alphabet $\mathcal{X}$ is be a finite set $\{1, \ldots, |\mathcal{X}|\}$. Each estimator $\widehat{G}(\,\cdot\,)$ induces a partition of the set of graphs $\mathfrak{G}_n$. That is, we can define *decoding regions*

$$\mathcal{D}(\mathbf{x}^m) := \left\{ G \in \mathfrak{G}_n : \widehat{G}(\mathbf{x}^m) = G \right\} \tag{56}$$

such that $\mathcal{D}(\mathbf{x}^m) \cap \mathcal{D}(\tilde{\mathbf{x}}^m) = \emptyset$ for $\mathbf{x}^m \neq \tilde{\mathbf{x}}^m$ and $\cup_{\mathbf{x}^m \in (\mathcal{X}^n)^m} \mathcal{D}(\mathbf{x}^m) = \mathfrak{G}_n$.

4.3.1. *Typical Graphs.* Given a graph G, we define the *average degree* $\bar{d}(G)$ to be the ratio of the number of edges of G to the total number of nodes n. We define the following set of graphs:

$$\mathcal{T}_{\epsilon}^{(n)} := \left\{ G \in \mathfrak{G}_n : \left| \bar{d}(G) - \frac{c}{2} \right| \leq \frac{c}{2}\epsilon \right\}. \tag{57}$$

The set $\mathcal{T}_{\epsilon}^{(n)}$ is the *ϵ-typical set* of graphs. Every graph $G \in \mathcal{T}_{\epsilon}^{(n)}$ has an average degree that is "close" to the average degree of the graphs in the Erdős-Rényi ensemble. Let $H_{\mathrm{b}}(q) := -q\log_2 q - (1-q)\log_2(1-q)$ be the binary entropy function.

LEMMA 8 (Properties of $\mathcal{T}_{\epsilon}^{(n)}$). *The ϵ-typical set has the following properties:*

1. $\mathbb{P}(\mathcal{T}_{\epsilon}^{(n)}) \to 1$ *as* $n \to \infty$.
2. *For all* $G \in \mathcal{T}_{\epsilon}^{(n)}$, *we have*

$$\exp_2\left[-\binom{n}{2} H_{\mathrm{b}}\left(\frac{c}{n}\right)(1+\epsilon)\right] \leq \mathbb{P}(G) \leq \exp_2\left[-\binom{n}{2} H_{\mathrm{b}}\left(\frac{c}{n}\right)\right]. \tag{58}$$

3. *The cardinality of the ϵ-typical set can be bounded as*

$$(1-\epsilon)\exp_2\left[\binom{n}{2} H_{\mathrm{b}}\left(\frac{c}{n}\right)\right] \leq |\mathcal{T}_{\epsilon}^{(n)}| \leq \exp_2\left[\binom{n}{2} H_{\mathrm{b}}\left(\frac{c}{n}\right)(1+\epsilon)\right] \tag{59}$$

for all n sufficiently large.

Part 1 will be strengthened in Lemma 10. Also, note that part 2 says that all the graphs in the ϵ-typical set have roughly the same probability (in contrast to the family of all graphs $\mathfrak{G}_n$). Furthermore, part 3 says that the number of graphs in the ϵ-typical set is relatively small (compared to $|\mathfrak{G}_n|$). More precisely, the cardinality is of the order $\exp_2[\Theta(n \log n)]$. Note that in contrast to the usual typical sets in information theory, ϵ only appears on one side of the bounds in (58). Lemma 8 is proved along the same lines as the asymptotic equipartition property in [8, Ch. 3].

4.3.2. *Sparse Random Discrete Graphical Models.*

PROOF. For the sake of convenience, we define the random variable:

$$W = \begin{cases} 1 & G \in \mathcal{T}_\epsilon^{(n)} \\ 0 & G \notin \mathcal{T}_\epsilon^{(n)} \end{cases}. \tag{60}$$

The random variable W indicates whether $G \in \mathcal{T}_\epsilon^{(n)}$. Consider the following sequence of lower bounds:

$$nm \log_2 |\mathcal{X}| \geq H(\mathbf{X}^m) \tag{61}$$

$$\overset{(a)}{\geq} H(\mathbf{X}^m|W) \tag{62}$$

$$= I(\mathbf{X}^m; G|W) + H(\mathbf{X}^m|G, W) \tag{63}$$

$$\overset{(b)}{\geq} I(\mathbf{X}^m; G|W) \tag{64}$$

$$= H(G|W) - H(G|\mathbf{X}^m, W), \tag{65}$$

where (a) is because conditioning does not increase entropy and (b) is because the conditional entropy $H(\mathbf{X}^m|G, W)$ is non-negative. We are going to lower bound the first term in (65) and upper bound the second term in (65). Now consider the first term in the difference in (65):

$$H(G|W) = H(G|W=1)\mathbb{P}(W=1) + H(G|W=0)\mathbb{P}(W=0) \tag{66}$$

$$\overset{(a)}{\geq} H(G|W=1)\mathbb{P}(W=1) \tag{67}$$

$$\overset{(b)}{\geq} H(G|G \in \mathcal{T}_\epsilon^{(n)})(1-\epsilon) \tag{68}$$

$$\overset{(c)}{\geq} (1-\epsilon)\binom{n}{2} H_{\mathrm{b}}\left(\frac{c}{n}\right), \tag{69}$$

where (a) is because the entropy $H(G|W=0)$ and the probability $\mathbb{P}(W=0)$ are both non-negative. Inequality (b) follows for all n sufficiently large from the definition of W as well as Lemma 8 part 1 (law of large numbers). Statement (c) comes from fact that

$$H(G|G \in \mathcal{T}_\epsilon^{(n)}) = -\sum_{g \in \mathcal{T}_\epsilon^{(n)}} \mathbb{P}(g|g \in \mathcal{T}_\epsilon^{(n)}) \log_2 \mathbb{P}(g|g \in \mathcal{T}_\epsilon^{(n)}) \tag{70}$$

$$\overset{(a)}{\geq} -\sum_{g \in \mathcal{T}_\epsilon^{(n)}} \mathbb{P}(g|g \in \mathcal{T}_\epsilon^{(n)}) \left[-\binom{n}{2} H_{\mathrm{b}}\left(\frac{c}{n}\right)\right] = \binom{n}{2} H_{\mathrm{b}}\left(\frac{c}{n}\right) \tag{71}$$

where (a) comes from the upper bound in Lemma 8 part 2. We are now done bounding the first term in the difference in (65).

Now we will attempt to bound the second term in (65). First we will derive a bound on $H(G|\mathbf{X}^m, W=1)$. Consider,

$$P_e^{(n)} := \mathbb{P}(\widehat{G}(\mathbf{X}^m) \neq G) \tag{72}$$

$$\stackrel{(a)}{=} \mathbb{P}(\widehat{G}(\mathbf{X}^m) \neq G|W=1)\mathbb{P}(W=1) + \mathbb{P}(\widehat{G}(\mathbf{X}^m) \neq G|W=0)\mathbb{P}(W=0) \tag{73}$$

$$\stackrel{(b)}{\geq} \mathbb{P}(\widehat{G}(\mathbf{X}^m) \neq G|W=1)\mathbb{P}(W=1) \tag{74}$$

$$\stackrel{(c)}{\geq} \mathbb{P}(\widehat{G}(\mathbf{X}^m) \neq G|G \in \mathcal{T}_\epsilon^{(n)}) \left(\frac{1}{1+\epsilon}\right) \tag{75}$$

$$\stackrel{(d)}{\geq} \frac{H(G|\mathbf{X}^m, G \in \mathcal{T}_\epsilon^{(n)}) - 1}{\log_2 |\mathcal{T}_\epsilon^{(n)}|} \left(\frac{1}{1+\epsilon}\right), \tag{76}$$

where (a) is by the law of total probability, (b) is by the fact that probabilities are non-negative, (c) holds for all n sufficiently large by Lemma 8 part 1 (law of large numbers) and (d) is due to the conditional version of Fano's inequality (see Lemma 9 below). Note that the cardinality of $G|G \in \mathcal{T}_\epsilon^{(n)}$ is exactly $|\mathcal{T}_\epsilon^{(n)}|$ (this is tautological). Then, from (76), we have

$$H(G|\mathbf{X}^m, W=1) \leq P_e^{(n)}(1+\epsilon)\log_2 |\mathcal{T}_\epsilon^{(n)}| + 1 \tag{77}$$

$$\leq P_e^{(n)}(1+\epsilon)\binom{n}{2} H_{\mathrm{b}}\left(\frac{c}{n}\right) + 1. \tag{78}$$

Define the *rate function* $K(c,\epsilon) := \frac{c}{2}[(1+\epsilon)\ln(1+\epsilon) - \epsilon]$. Note that this function is positive whenever $c, \epsilon > 0$. In fact it is monotonically increasing in both parameters. Now we utilize (78) to bound $H(G|\mathbf{X}^m, W)$:

$$H(G|\mathbf{X}^m, W) = H(G|\mathbf{X}^m, W=1)\mathbb{P}(W=1) + H(G|\mathbf{X}^m, W=0)\mathbb{P}(W=0) \tag{79}$$

$$\stackrel{(a)}{\leq} H(G|\mathbf{X}^m, W=1) + H(G|\mathbf{X}^m, W=0)\mathbb{P}(W=0) \tag{80}$$

$$\stackrel{(b)}{\leq} H(G|\mathbf{X}^m, W=1) + H(G|\mathbf{X}^m, W=0)(2e^{-nK(c,\epsilon)}) \tag{81}$$

$$\stackrel{(c)}{\leq} H(G|\mathbf{X}^m, W=1) + n^2(2e^{-nK(c,\epsilon)}) \tag{82}$$

$$\stackrel{(d)}{\leq} P_e^{(n)}(1+\epsilon)\binom{n}{2} H_{\mathrm{b}}\left(\frac{c}{n}\right) + 1 + 2n^2 e^{-nK(c,\epsilon)}, \tag{83}$$

where (a) is because we upper bounded $\mathbb{P}(W=1)$ by unity, (b) follows by Lemma 10, (c) follows by upper bounding the conditional entropy by n^2 (very coarse but is good enough) and (d) follows from (78).

Substituting (69) and (83) back into (65) yields

$$nm\log_2 |\mathcal{X}| \geq (1-\epsilon)\binom{n}{2} H_{\mathrm{b}}\left(\frac{c}{n}\right) - P_e^{(n)}(1+\epsilon)\binom{n}{2} H_{\mathrm{b}}\left(\frac{c}{n}\right) - 1 - 2n^2 e^{-nK(c,\epsilon)} \tag{84}$$

$$= \binom{n}{2} H_{\mathrm{b}}\left(\frac{c}{n}\right)\left[(1-\epsilon) - P_e^{(n)}(1+\epsilon)\right] - \Theta(n^2 e^{-nK(c,\epsilon)}), \tag{85}$$

which implies that

$$m \geq \frac{1}{n\log_2 |\mathcal{X}|}\binom{n}{2} H_{\mathrm{b}}\left(\frac{c}{n}\right)\left[(1-\epsilon) - P_e^{(n)}(1+\epsilon)\right] - \Theta(ne^{-nK(c,\epsilon)}). \tag{86}$$

Note that $\Theta(ne^{-nK(c,\epsilon)}) \to 0$ as $n \to \infty$ since the rate function $K(c,\epsilon) > 0$. If we impose that $P_e^{(n)} \to 0$ as $n \to \infty$, then m has to satisfy (86) by the arbitrariness of ϵ. This completes the proof of the converse. □

LEMMA 9 (Conditional Fano Inequality). *In the above notation, we have*

$$\frac{H(G|\mathbf{X}^m, G \in \mathcal{T}_\epsilon^{(n)}) - 1}{\log_2(|\mathcal{T}_\epsilon^{(n)}| - 1)} \leq \mathbb{P}(\widehat{G}(\mathbf{X}^m) \neq G | G \in \mathcal{T}_\epsilon^{(n)}). \tag{87}$$

PROOF. Define the "error" random variable

$$E = \begin{cases} 1 & \widehat{G}(\mathbf{X}^m) \neq G \\ 0 & \widehat{G}(\mathbf{X}^m) = G \end{cases}. \tag{88}$$

Now consider

$$H(E, G|\mathbf{X}^m, W = 1) = H(E|\mathbf{X}^m, W = 1) + H(G|E, \mathbf{X}^m, W = 1) \tag{89}$$
$$= H(G|\mathbf{X}^m, W = 1) + H(E|G, \mathbf{X}^m, W = 1). \tag{90}$$

The first term in (89) can be bounded above by 1 since the alphabet of the random variable E is of size 2. Since $H(G|E = 0, \mathbf{X}^m, W = 1) = 0$, the second term in (89) can be bounded from above as

$$H(G|E, \mathbf{X}^m, W = 1) = H(G|E = 0, \mathbf{X}^m, W = 1)\mathbb{P}(E = 0|W = 1)$$
$$+ H(G|E = 1, \mathbf{X}^m, W = 1)\mathbb{P}(E = 1|W = 1) \tag{91}$$
$$\leq \mathbb{P}(\widehat{G}(\mathbf{X}^m) \neq G | G \in \mathcal{T}_\epsilon^{(n)}) \log_2(|\mathcal{T}_\epsilon^{(n)}| - 1). \tag{92}$$

The second term in (90) is 0. Hence, we have the desired conclusion. □

LEMMA 10 (Exponential Decay in Probability of Atypical Set). *Define the rate function* $K(c,\epsilon) := \frac{c}{2}[(1+\epsilon)\ln(1+\epsilon) - \epsilon]$. *The probability of the* ϵ*-atypical set decays as*

$$\mathbb{P}((\mathcal{T}_\epsilon^{(n)})^c) = \mathbb{P}(G \notin \mathcal{T}_\epsilon^{(n)}) \leq 2\exp\left(-nK(c,\epsilon)\right) \tag{93}$$

for all $n \geq 1$.

Note the non-asymptotic nature of the bound in (93). The rate function satisfies $K(c,\epsilon) = O(\epsilon^2)$. The proof uses standard Chernoff bounding techniques but the scaling in n is somewhat different from the vanilla Chernoff (Cramér) upper bound.

PROOF. For simplicity, we will use $M := \binom{n}{2}$. Let $Y_i, i = 1, \ldots, M$ be independent Bernoulli random variables such that $\mathbb{P}(Y_i = 1) = c/n$. Then the probability in question can be bounded as

$$\mathbb{P}(G \notin \mathcal{T}_\epsilon^{(n)}) = \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{M} Y_i - \frac{c}{2}\right| > \frac{c\epsilon}{2}\right) \tag{94}$$
$$\overset{(a)}{\leq} 2\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{M} Y_i > (1+\epsilon)\frac{c\epsilon}{2}\right) \tag{95}$$

$$= 2\mathbb{E}\left[\mathbb{I}\left\{\frac{1}{n}\sum_{i=1}^{M} Y_i - (1+\epsilon)\frac{c\epsilon}{2} > 0\right\}\right] \tag{96}$$

$$\overset{(b)}{\leq} 2\mathbb{E}\left[\exp\left(t\sum_{i=1}^{M} Y_i - nt\frac{c}{2}(1+\epsilon)\right)\right] \tag{97}$$

$$= 2\exp\left(-nt\frac{c}{2}(1+\epsilon)\right)\prod_{i=1}^{M}\mathbb{E}[\exp(tY_i)], \tag{98}$$

where (a) comes from the union bound, (b) comes from an application of Markov's inequality. Note that $t \geq 0$ in (97). Now the moment generating function of a Bernoulli random variable with probability of success q is $qe^t + (1-q)$. Using this fact, we can further upper bound (98) as follows:

$$\mathbb{P}(G \notin \mathcal{T}_\epsilon^{(n)}) = 2\exp\left(-nt\frac{c}{2}(1+\epsilon) + M\ln(\frac{c}{n}e^t + (1-\frac{c}{n})\right) \tag{99}$$

$$\overset{(a)}{\leq} 2\exp\left(-nt\frac{c}{2}(1+\epsilon) + \frac{n(n-1)}{2}\frac{c}{n}(e^t-1)\right) \tag{100}$$

$$\overset{(b)}{\leq} 2\exp\left(-n\left[t\frac{c}{2}(1+\epsilon) - \frac{c}{2}(e^t-1)\right]\right), \tag{101}$$

where in (a), we used the fact that $\ln(1+z) \leq z$ and in (b) we upper bounded $n-1$ by n. Now, we differentiate the exponent in square brackets with respect to $t \geq 0$ to find the tightest bound. We observe that the optimal parameter is $t^* = \ln(1+\epsilon)$. Substituting this back into (101) completes the proof. □

5. Properties of Power-law Graphs. We briefly note the local-paths property of power-law random graphs. Recall that the ensemble $\mathcal{G}_{\mathrm{LP}}(p;\eta,\gamma)$ has at most η paths of length at most γ in G between any two nodes or equivalently, there are at most $\eta-1$ number of overlapping cycles of length smaller than 2γ. We now describe the power-law random graph model. For details, refer to [7, Ch. 5].

For a given sequence $\mathbf{w} = (w_1, w_2, \ldots, w_p)$, the random-graph $G = (V, E)$ with $V = \{1, \ldots, p\}$ is generated as follows: for any two nodes $i, j \in V$, the probability of edge (i,j) occurs with probability $w_i w_j \rho$, independent of other edges, where $\rho := (\sum_j w_j)^{-1}$ is the normalization factor. The sequence $\mathbf{w}$ is the sequence of expected degrees in the random-graph model. A power-law random graph ensemble $\mathcal{G}_{\mathrm{PL}}(p, \overline{w}, \beta, \Delta)$ has an expected degree sequence given by

$$w_i = \alpha i^{-\frac{1}{\beta-1}}, \quad \forall i \geq i_0,$$
$$\alpha := \frac{(\beta-2)}{(\beta-1)}\overline{w}p^{\frac{1}{\beta-1}}, \quad i_0 = p\left(\frac{\overline{w}(\beta-2)}{\Delta(\beta-1)}\right)^{\beta-1},$$

where $\overline{w}$ is the average degree, Δ is the maximum degree and $\beta > 0$ is the exponent of the power law. We immediately see that a special case of the above parameterization is the Erdős-Rényi ensemble $G \sim \mathcal{G}_{\mathrm{ER}}(p, c/p)$ where $w_i = c$ for all $i \in V$, implying that $\overline{w} = c$ and $\beta = \infty$.

Proposition 4 (Local-Paths Property of Power-Law Graphs). *The power-law random graph ensemble $\mathcal{G}_{\mathrm{PL}}(p, \overline{w}, \beta, \Delta)$ satisfies the (η,γ)-local paths property a.a.s. when*

$$\overline{w} = o(p^{\frac{\eta-1}{2\eta\gamma} - \frac{2}{\beta-1}}), \tag{102}$$

Proof: Let $F = (V_F, E_F)$ be a graph which is the union of at least η cycles of length less than 2γ. We see that $|E_F| = |V_F| + \eta - 1$ and $|E_F| < 2\gamma\eta$. By a counting argument, the expected number of subgraphs F in $G \sim \mathcal{G}_{\mathrm{PL}}(p, \overline{w}, \beta, \Delta)$ is bounded by

$$\binom{p}{|V_F|} \alpha^{2|E_F|} \rho^{|E_F|} \leq p^{|V_F|} \alpha^{2|E_F|} \rho^{|E_F|} \leq \overline{w}^{|E_F|} p^{\frac{2|E_F|}{\beta-1} - \eta + 1},$$

by substituting for α and ρ and using the fact that $|E_F| = |V_F| + \eta - 1$. Thus, the expected number of subgraphs F in $G \sim \mathcal{G}_{\mathrm{PL}}(p, \overline{w}, \beta, \Delta)$ is $o(1)$ when (102) holds by noting that $|E_F| < 2\gamma\eta$. By Markov's inequality, the subgraph F does not occur in G a.a.s. □

Thus, we have a relationship between the average degree $\overline{w}$, the power-law exponent β, the number of local paths η and the threshold γ on the length of the paths. We note that in the special case of Erdős-Rényi ensemble $\mathcal{G}_{\mathrm{ER}}(p, c/p)$, the (η, γ)-local path property is satisfied when

$$(103)\quad \eta = 2,\ \gamma < \frac{\log p}{4 \log c},$$

by substituting $\overline{w} = c$ and $\beta = \infty$.

References.

[1] Anandkumar, A., Hassidim, A. and Kelner, J. (2012). Topology Discovery of Sparse Random Graphs With Few Participants. *Accepted to J. of Random Structures and Algorithms*.

[2] Anandkumar, A., Tan, V. Y. F., Huang, F. and Willsky, A. S. (2012). High-Dimensional Structure Learning of Ising Models: Local Separation Criterion. *Accepted to Annals of Statistics*.

[3] Berger, N., Kenyon, C., Mossel, E. and Peres, Y. (2005). Glauber dynamics on trees and hyperbolic graphs. *Probability Theory and Related Fields* **131** 311–340.

[4] 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.

[5] Chung, F. R. K. (1997). *Spectral graph theory*. Amer Mathematical Society.

[6] Chung, F. and Lu, L. (2001). The diameter of sparse random graphs. *Advances in Applied Mathematics* **26** 257–279.

[7] Chung, F. R. K. and Lu, L. (2006). *Complex graphs and network*. Amer. Mathematical Society.

[8] Cover, T. and Thomas, J. (2006). *Elements of Information Theory*. John Wiley & Sons, Inc.

[9] Dembo, A. and Montanari, A. (2010). Ising Models on Locally Tree-like Graphs. *Annals of Applied Probability*.

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

[11] Jiang, T., Sidiropoulos, N. D. and ten Berge, J. M. F. (2001). Almost-sure identifiability of multidimensional harmonic retrieval. *Signal Processing, IEEE Transactions on* **49** 1849–1859.

[12] McKay, B. D., Wormald, N. C. and Wysocka, B. (2004). Short cycles in random regular graphs. *The Electronic Journal of Combinatorics* **11** 1.

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

[14] 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.

[15] Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S. and Weinberger, M. L. (2003). Inequalities for the l_1 deviation of the empirical distribution Technical Report, Hewlett-Packard Labs.

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