

RECONSTRUCTION OF MARKOV RANDOM FIELDS FROM SAMPLES: SOME OBSERVATIONS AND ALGORITHMS*

GUY BRESLER[†], ELCHANAN MOSSEL[‡], AND ALLAN SLY[§]

Abstract. Markov random fields are used to model high dimensional distributions in a number of applied areas. Much recent interest has been devoted to the reconstruction of the dependency structure from independent samples from the Markov random fields. We analyze a simple algorithm for reconstructing the underlying graph defining a Markov random field on n nodes and maximum degree d given observations. We show that under mild nondegeneracy conditions it reconstructs the generating graph with high probability using $\Theta(d\epsilon^{-2}\delta^{-4}\log n)$ samples, where ϵ, δ depend on the local interactions. For most local interactions ϵ, δ are of order $\exp(-O(d))$. Our results are optimal as a function of n up to a multiplicative constant depending on d and the strength of the local interactions. Our results seem to be the first results for general models that guarantee that the generating model is reconstructed. Furthermore, we provide explicit $O(n^{d+2}\epsilon^{-2}\delta^{-4}\log n)$ running-time bound. In cases where the measure on the graph has correlation decay, the running time is $O(n^2\log n)$ for all fixed d . We also discuss the effect of observing noisy samples and show that as long as the noise level is low, our algorithm is effective. On the other hand, we construct an example where large noise implies nonidentifiability even for generic noise and interactions. Finally, we briefly show that in some simple cases, models with hidden nodes can also be recovered.

Key words. Markov random fields, algorithms, correlation decay

AMS subject classifications. 68W20, 05C85

DOI. 10.1137/100796029

1. Introduction. In this paper we consider the problem of reconstructing the graph structure of a Markov random field (MRF) from independent and identically distributed samples. MRFs provide a very general framework for defining high dimensional distributions, and the reconstruction of the MRFs from observations has attracted much recent interest, in particular in biology; see, e.g., [8] and a list of related references [9].

1.1. Our results. We give sharp, up to a multiplicative constant, estimates for the number of independent samples needed to infer the underlying graph of an MRF of bounded degree. In Theorem 1 we use a simple information-theoretic argument to show that $\Omega(d\log n)$ samples are required to reconstruct a randomly selected graph on n vertices with maximum degree at most d . Then in Theorems 2 and 3 we propose two algorithms for reconstruction that use only $O(d\epsilon^{-2}\delta^{-4}\log n)$, where ϵ and δ are lower bounds on marginal distributions in the neighborhood of a vertex. Under mild

*Received by the editors May 20, 2010; accepted for publication (in revised form) November 16, 2012; published electronically March 28, 2013. An extended abstract containing some of the results of the current paper appeared in *Proceedings of the 11th International Workshop, APPROX 2008, and 12th International Workshop, RANDOM 2008*, Lecture Notes in Comput. Sci. 5171, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 343–356.

<http://www.siam.org/journals/sicomp/42-2/79602.html>

[†]Department of Electrical Engineering and Computer Sciences, U.C. Berkeley, Berkeley, CA 94720 (gbresler@eecs.berkeley.edu). This author's work was supported by a Vodafone US-Foundation fellowship.

[‡]Department of Statistics and Department of Electrical Engineering and Computer Sciences, U.C. Berkeley, Berkeley, CA 94720 (mossel@stat.berkeley.edu). This author's work was supported by a Sloan fellowship in Mathematics, by NSF Career award DMS-0548249, NSF grant DMS-0528488, and ONR grant N0014-07-1-05-06.

[§]Department of Statistics, U.C. Berkeley, Berkeley, CA 94720 (sly@stat.berkeley.edu). This author's work was supported by NSF grants DMS-0528488 and DMS-0548249.

nondegeneracy conditions, $\epsilon, \delta = \exp(-O(d))$ and for some models $\epsilon, \delta = \text{poly}^{-1}d$. Examples of the latter model include the hardcore model with fugacity $\lambda = \Theta(\frac{1}{d})$. Our main focus is on the reconstruction of sparse MRFs where d is fixed, in which case ϵ and δ are constant. The two theorems differ in their running time and the required nondegeneracy conditions. It is clear that nondegeneracy conditions are needed to ensure that there is a unique graph associated with the observed probability distribution.

In addition to the fully observed setting in which samples of all variables are available, we extend our algorithm in several directions. In section 5 we consider the problem of noisy observations. In subsection 5.1 we show by way of an example that if some of the random variables are perturbed by noise, then it is in general impossible to reconstruct the graph structure with probability approaching 1. Conversely, when the noise is relatively weak as compared to the coupling strengths between random variables, we show that the algorithms used in Theorems 2 and 3 reconstruct the graph with high probability. Furthermore, we study the problem of reconstruction with partial observations, i.e., samples from only a subset of the nodes are available. In Theorem 5 we provide sufficient conditions on the probability distribution for correct reconstruction.

Chickering [3] showed that maximum-likelihood estimation of the underlying graph of an MRF is NP-complete. This does not contradict our results which assume that the data is generated from a model (or a model with a small amount of noise). Although the algorithm we propose runs in time polynomial in the size of the graph, the dependence on degree (the running time is $O(n^{d+2}\epsilon^{-2}\delta^{-4}\log n)$) may impose too high a computational cost for some applications. Indeed, for some MRFs exhibiting a decay of correlation, a vast improvement can be realized: A modified version of the algorithm runs in time $O(dn^2\epsilon^{-2}\delta^{-4}\log n)$. This is proved in Theorem 4.

1.2. Related work. Chow and Liu [4] considered the problem of estimating MRFs whose underlying graphs are trees, and provided an efficient (polynomial-time) algorithm based on the fact that in the tree case maximum-likelihood estimation amounts to the computation of a maximum-weight spanning tree with edge weights equal to pairwise empirical mutual information. Unfortunately, their approach does not generalize to the estimation of MRFs whose graphs have cycles. Much work in mathematical biology is devoted to reconstructing tree Markov fields when there are hidden models. For trees, given data that is generated from the model, the tree can be reconstructed efficiently from samples at a subset of the nodes given mild nondegeneracy conditions. See [7, 10, 5] for some of the most recent and tightest results in this setup.

The most closely related works are [1] and [12]. These can be compared in terms of sampling complexity and running time as well as the generality of the models to which they apply. These are summarized in the table below. The first line refers to the types of models that the method covers: Does the model allow clique interactions of just edge interactions? The next two lines refer to requirements on the strength of interactions: Are they not required to be too weak/are only edges with strong interactions returned? Are they not required to be too strong? The next line refers to the hardness of verifying whether a given model satisfies the conditions of the algorithm (where X denotes that the verification is exponential in the size of the model). The following line refers to the following question: Is there a guarantee that the generating model is returned with high probability? The final two lines refer to computational and sampling complexity, where c_d denotes constants that depend on d .

Method	AKN [1]	WRL [12]	Alg.	High temp. alg.
Cliques	✓	X	✓	✓
No int. low. bd.	✓	X	X	X
No int. upp. bd.	✓	X	✓	X
Verifiable conds.	✓	X	✓	✓
Output gen. model	X	✓	✓	✓
Comp. compl.	$n^{O(d)}$	n^5	$n^{O(d)}$	$c_d n^2 \log n$
Sampl. compl.	$n^{O(d)}$	$\text{poly}(d) \log n$	$c_d \log n$	$c_d \log n$

Abbeel, Koller, and Ng [1] considered the problem of reconstructing graphical models based on factor graphs and proposed a polynomial-time and sample complexity algorithm. However, the goal of their algorithm was not to reconstruct the true structure, but rather to produce a model whose distribution is close in Kullback–Leibler (KL) divergence to the true distribution. In applications it is often of interest to reconstruct the true structure which give some insights into the underlying structure of the inferred model.

Note furthermore that two networks that differ only in the neighborhood of one node will have $O(1)$ KL distance. Therefore, even in cases where it is promised that the KL distance between the generating distribution and any other distribution defined by another graph is as large as possible, the lower bound on the KL distance is $\Omega(1)$. Plugging this into the bounds in [1] yields a polynomial sampling complexity in the size of the network in order to find the generating network compared to our logarithmic sampling complexity. For other work based on minimizing the KL divergence, see the references in [1].

The same problem as in the present work (but restricted to the Ising model) was studied by Wainwright, Ravikumar, and Lafferty [12], where an algorithm based on ℓ_1 -regularization was introduced. The algorithm presented is efficient also for dense graphs with running time $O(n^5)$, but is applicable only in very restricted settings. The work applies only to the Ising model and, more importantly, only models with edge interactions (no larger cliques are allowed). The most important restrictions are the two conditions in [12] (A1 and A2). Condition A1 requires (among other things) that the “covariates [spins] do not become overly dependent.” Verifying when the conditions hold seems hard. However, it is easy to see that this condition fails for standard models such as the Ising model on the lattice or on random d -regular graphs when the model is at low temperatures, i.e., for $\beta > \frac{1}{2} \log(1 + \sqrt{2})$ in the case of the two-dimensional Ising model and $\beta > \tanh^{-1}(1/(d-1))$ for random d -regular graphs.

Subsequent to our work being posted on the arXiv e-print service, Santhanam and Wainwright [11] again considered essentially the problem for the Ising model, producing nearly matching lower and upper bounds on the asymptotic sampling complexity. Again their conditions do not apply to the low temperature regime. Another key difference from our work is that they restrict attention to the Ising model, i.e., MRFs with pairwise potentials and where each variable takes two values. Our results are not limited to pairwise interactions and apply to the more general setting of MRFs with potentials on larger cliques.

Since our work was first posted on arXiv, a number of groups continued working on this and related problems. We refer the reader to [2] for a number of theoretical results as well as a survey of many others.

2. Preliminaries. We begin with the definition of Markov random field.

DEFINITION 1. *On a graph $G = (V, E)$, a Markov random field is a distribution X taking values in \mathcal{A}^V for some finite set \mathcal{A} with $|\mathcal{A}| = A$, which satisfies the Markov property*

$$(1) \quad P(X(W), X(U)|X(S)) = P(X(W)|X(S))P(X(U)|X(S))$$

when W , U , and S are disjoint subsets of V such that every path in G from W to U passes through S and where $X(U)$ denotes the restriction of X from \mathcal{A}^V to \mathcal{A}^U for $U \subset V$.

Famously, by the Hammersley–Clifford theorem, such distributions can be written in a factorized form as

$$(2) \quad P(\sigma) = \frac{1}{Z} \exp \left[\sum_a \Psi_a(\sigma_a) \right],$$

where Z is a normalizing constant, a ranges over the cliques in G , and $\Psi_a: \mathcal{A}^{|a|} \rightarrow \mathbb{R} \cup \{-\infty\}$ are functions called *potentials*.

The problem we consider is that of reconstructing the graph G , given k independent samples $\underline{X} = \{X^1, \dots, X^k\}$ from the model. Denote by \mathcal{G}_d the set of labeled graphs with maximum degree at most d . We assume that the graph $G \in \mathcal{G}_d$ is from this class. A structure estimator (or reconstruction algorithm) $\hat{G}: \mathcal{A}^{kn} \rightarrow \mathcal{G}_d$ is a map from the space of possible sample sequences to the set of graphs under consideration. We are interested in the asymptotic relationship between the number of nodes n in the graph, the maximum degree d , and the number of samples k that are required. An algorithm using number of samples $k(n)$ is deemed successful if in the limit of large n the probability of reconstruction error approaches zero.

A special case which is studied extensively is the Ising model on graphs, where $\mathcal{A} = \{\pm 1\}$ and where a in (2) are taken only over edges of a graph so that

$$P(\sigma) = \frac{1}{Z} \exp \left[\sum_{(u,v) \in G} \beta_{u,v} \sigma_u \sigma_v + \sum_u h_u \sigma_u \right]$$

for real numbers $\beta_{u,v}, h_u$.

3. Lower bound on sample complexity. Suppose G is selected uniformly at random from \mathcal{G}_d . The following theorem gives a lower bound of $\Omega(d \log n)$ on the number of samples necessary to reconstruct the graph G . The argument is information-theoretic and follows by comparing the number of possible graphs with the amount of information available from the samples.

THEOREM 1. *Let the graph G be drawn according to the uniform distribution on \mathcal{G}_d . Then there exists a constant $c = c(A) > 0$ such that if $k \leq cd \log n$, then for any estimator $\hat{G}: \mathcal{A}^{kn} \rightarrow \mathcal{G}_d$, the probability of correct reconstruction is $P(\hat{G} = G) = o(1)$.*

Remark 1. Note that the theorem above doesn't need to assume anything about the potentials. The theorem applies for any potentials that are consistent with the generating graph. In particular, it is valid both in cases where the graph is “identifiable” given many samples and in cases where it isn't.

Proof. To begin, we note that the probability of error is minimized by letting \hat{G} be the maximum a posteriori (MAP) decision rule,

$$\hat{G}_{\text{MAP}}(\underline{X}) = \operatorname{argmax}_{g \in \mathcal{G}_d} P[G = g | \underline{X}].$$

By the optimality of the MAP rule, this bounds the probability of error using any estimator. Now, the MAP estimator $\hat{G}_{\text{MAP}}(\underline{X})$ is a deterministic function of \underline{X} . Clearly, if a graph g is not in the range of \hat{G} , then the algorithm always makes an error when $G = g$. Let S be the set of graphs in the range of \hat{G}_{MAP} so that $P(\text{error}|G \in S^c) = 1$. We have

$$\begin{aligned}
 P(\text{error}) &= \sum_{g \in \mathcal{G}} P(\text{error}|G = g)P(G = g) \\
 &= \sum_{g \in S} P(\text{error}|G = g)P(G = g) + \sum_{g \in S^c} P(\text{error}|G = g)P(G = g) \\
 (3) \quad &\geq \sum_{g \in S^c} P(G = g) = 1 - \sum_{g \in S} |\mathcal{G}|^{-1} \\
 &\geq 1 - \frac{A^{nk}}{|\mathcal{G}|},
 \end{aligned}$$

where the last step follows from the fact that $|S| \leq |\underline{X}| \leq A^{nk}$. It remains only to express the number of graphs with maximum degree at most d , $|\mathcal{G}_d|$, in terms of the parameters n, d . The following lemma gives an adequate bound.

LEMMA 1. *Suppose that $d \leq n^\alpha$ for $\alpha < 1$. Then the number of graphs with max degree at most d , $|\mathcal{G}_d|$, satisfies*

$$(4) \quad \log |\mathcal{G}_d| = \Omega(nd \log n).$$

Proof. Suppose first that $d \leq n^{1/3}$. To make the dependence on n explicit, let $U_{n,d}$ be the number of graphs with n vertices with maximum degree at most d . We first bound $U_{n+2,d}$ in terms of $U_{n,d}$. Given a graph G with n vertices and degree at most d , add two vertices a and b . Select d distinct neighbors v_1, \dots, v_d for vertex a , with d labeled edges; there are $\binom{n}{d}d!$ ways to do this. If v_i already has degree d in G , then v_i has at least one neighbor u that is not a neighbor of a , since there are only $d - 1$ other neighbors of a . Remove the edge (v_i, u) and place an edge labeled i from vertex b to u . This is done for each vertex v_1, \dots, v_d , so b has degree at most d . The graph G can be reconstructed from the resulting labeled graph on $n + 2$ vertices as follows: Remove vertex a , and return the neighbors of b to their correct original neighbors (this is possible because the edges are labeled).

Removing the labels on the edges from a and b sends at most $d!^2$ edge-labeled graphs of this type on $n + 2$ vertices to the same unlabeled graph. Hence, the number of graphs with maximum degree d on $n + 2$ vertices is lower bounded as

$$U_{n+2,d} \geq U_{n,d} \binom{n}{d} d! \frac{1}{d!^2} = U_{n,d} \binom{n}{d} \frac{1}{d!}.$$

It follows that for n even (and greater than $2d + 4$),

$$(5) \quad U_{n,d} \geq \prod_{i=1}^{n/4} \binom{n-2i}{d} \frac{1}{d!} \geq \left(\binom{n/2}{d} \frac{1}{d!} \right)^{n/4}.$$

If n is odd, it suffices to note that $U_{n+1,d} \geq U_{n,d}$. Taking the logarithm of (5) yields

$$(6) \quad \log U_{n,d} = \Omega(nd(\log n - \log d)) = \Omega(nd \log n),$$

assuming that $d \leq n^{1/3}$. When $d \geq n^{1/3}$, note that with high probability a graph with $dn/4$ edges chosen uniformly at random has maximum degree less than d . The number of such graphs is $\binom{n(n-1)/2}{dn/4}$, and so

$$\log U_{n,d} \geq \Omega \left(\log \binom{n(n-1)}{dn/4} / 2 \right) = \Omega(nd(\log n - \log d)) = \Omega(nd \log n),$$

since $d \leq n^\alpha$ with $\alpha < 1$. \square

Together with (3), Lemma 1 implies that for small enough c , if the number of samples $k \leq cd \log n$, then

$$P(\text{error}) \geq 1 - \frac{A^{nk}}{|\mathcal{G}|} = 1 - o(1).$$

This completes the proof of Theorem 1. \square

4. Reconstruction. We now turn to the problem of reconstructing the graph structure of an MRF from samples. For a vertex v we let $N(v) = \{u \in V - \{v\} : (u, v) \in E\}$ denote the set of neighbors of v . Determining the neighbors of v for every vertex in the graph is sufficient to determine all the edges of the graph and hence reconstruct the graph. We test each candidate neighborhood of size at most d by using the Markov property, which states that for each $w \in V - (N(v) \cup \{v\})$,

$$(7) \quad P(X(v)|X(N(v)), X(w)) = P(X(v)|X(N(v))).$$

We give two theorems for reconstructing networks; they differ in their nondegeneracy conditions and their running time. The first one, Theorem 2, has more stringent nondegeneracy conditions and faster running time.

In the first and more stringent condition (8), we require that an edge (u, v) in the graph will correspond to statistical correlation between the node x_u and x_v . In fact, we require that if u is a neighbor of v , then this correlation holds even when we condition on some additional variables.

In the second and less stringent condition (16), we require that if u is a neighbor of v , then conditioning on all the neighbors of v but u is statistically different from conditioning on all the neighbors of v (including u). Indeed, it is not hard to see that if the conditioning on u has no additional (or small additional) effect, then the MRF is equivalent to (or statistically close to) a model where the edge (u, v) is not present.

The additional conditions (9) and (17) are imposed so that samples exhibiting the required statistical differences will actually be generated with nonnegligible probability.

4.1. Conditional two-point correlation reconstruction.

THEOREM 2. *Suppose the graphical model satisfies the following: There exist $\epsilon, \delta > 0$ such that for all $v \in V$, if $U \subset V - \{v\}$ with $|U| \leq d$ and $N(v) \not\subseteq U$, then there exist values $x_v, x_w, x'_w, x_{u_1}, \dots, x_{u_l}$ such that for some $w \in V - (U \cup \{v\})$,*

$$(8) \quad \begin{aligned} &|P(X(v) = x_v | X(U) = x_U, X(w) = x_w) \\ &- P(X(v) = x_v | X(U) = x_U, X(w) = x'_w)| > \epsilon \end{aligned}$$

and

$$(9) \quad \begin{aligned} &P(X(U) = x_U, X(w) = x_w) > \delta, \\ &P(X(U) = x_U, X(w) = x'_w) > \delta. \end{aligned}$$

Then with the constant $C = \frac{48^2(d+2+C_1)}{\epsilon^2\delta^4 2d}$, when $k > Cd \log n$, there exists an estimator $\hat{G}(\underline{X})$ such that the probability of correct reconstruction is $P(G = \hat{G}(\underline{X})) = 1 - O(n^{-C_1})$. The estimator \hat{G} is efficiently computable in $O(n^{d+2} \log n)$ operations.

Remark 2. Condition (8) captures the notion that each edge should have sufficient strength. Condition (9) is required so that we can accurately calculate the empirical conditional probabilities.

Proof. Let \hat{P} denote the empirical probability measure from the k samples. Azuma's inequality gives that if $Y \sim \text{Bin}(k, p)$, then

$$P(|Y - kp| > \gamma k) \leq 2 \exp(-2\gamma^2 k),$$

and so for any collection $U = \{u_1, \dots, u_l\} \subseteq V$ and $x_1, \dots, x_l \in \mathcal{A}$, we have

$$(10) \quad P\left(\left|\hat{P}(X(U) = x_U) - P(X(U) = x_U)\right| \geq \gamma\right) \leq 2 \exp(-2\gamma^2 k).$$

There are $A^l \binom{n}{l} \leq A^l n^l$ such choices of u_1, \dots, u_l and x_1, \dots, x_l . An application of the union bound implies that with probability at least $1 - A^l n^l 2 \exp(-2\gamma^2 k)$, it holds that

$$(11) \quad \left|\hat{P}(X(U) = x_U) - P(X(U) = x_U)\right| \leq \gamma$$

for all $\{u_i\}_{i=1}^l$ and $\{x_i\}_{i=1}^l$. If we additionally have $l \leq d+2$ and $k \geq C(\gamma)d \log n$, then (11) holds with probability at least $1 - (d+2)A^{d+2}n^{d+2}2/n^{2\gamma^2 C(\gamma)d}$. Choosing $C(\gamma) = \frac{d+2+C_1}{\gamma^2 2d}$, (11) holds with probability at least $1 - 2(d+2)A^{d+2}/n^{C_1}$.

For the remainder of the proof assume (11) holds. For all $v, w \in V$, $U \subset V$, and $x_1, \dots, x_l, x_w, x'_w, x_v \in \mathcal{A}$ such that

$$(12) \quad \begin{aligned} \hat{P}(X(U) = x_U, X(w) = x_w) &> \delta/2, \\ \hat{P}(X(U) = x_U, X(w) = x'_w) &> \delta/2, \end{aligned}$$

and $|U| \leq d$, we have

$$(13) \quad \gamma(\epsilon, \delta) = \epsilon\delta^2/48.$$

Taking this, we can bound the error in conditional probabilities as

$$(14) \quad \begin{aligned} &|\hat{P}(X(v) = x_v | X(U) = x_U) - P(X(v) = x_v | X(U) = x_U)| \\ &= \left| \frac{\hat{P}(X(v) = x_v, X(U) = x_U)}{\hat{P}(X(U) = x_U)} - \frac{P(X(v) = x_v, X(U) = x_U)}{P(X(U) = x_U)} \right| \\ &\leq \left| \frac{\hat{P}(X(v) = x_v, X(U) = x_U)}{P(X(U) = x_U)} - \frac{P(X(v) = x_v, X(U) = x_U)}{P(X(U) = x_U)} \right| \\ &\quad + \left| \frac{1}{\hat{P}(X(U) = x_U)} - \frac{1}{P(X(U) = x_U)} \right| \\ &\leq \frac{\gamma}{\delta/2} + \frac{\gamma}{(\delta/2 - \gamma)\delta/2} \leq \frac{\epsilon\delta^2}{24\delta} + \frac{\epsilon\delta^2}{24(\delta/2 - \epsilon\delta^2/48)\delta} = \frac{\epsilon\delta}{12} + \frac{\epsilon}{(12 - \epsilon\delta/2)} < \frac{\epsilon}{4}. \end{aligned}$$

For each vertex $v \in V$ we consider all candidate neighborhoods for v , subsets $U \subset V - \{v\}$ with $|U| \leq d$. The estimate (14) and the triangle inequality imply that if $N(v) \subseteq U$, then by the Markov property,

$$(15) \quad \begin{aligned} & |\hat{P}(X(v) = x_v | X(U) = x_U, X(w) = x_w) \\ & - \hat{P}(X(v) = x_v | X(U) = x_U, X(w) = x'_w)| < \epsilon/2 \end{aligned}$$

for all $v, w \in V$, $U \subset V$, and $x_1, \dots, x_l, x_w, x'_w, x_v \in \mathcal{A}$ such that (12) holds.

Conversely, by conditions (8) and (9) and the estimate (14), we have that for any U with $N(v) \not\subseteq U$ there exists some $w \in V$ and $x_{u_1}, \dots, x_{u_l}, x_w, x'_w, x_v \in \mathcal{A}$ such that (12) holds but (15) does not hold. Thus, choosing the smallest set U such that (15) holds gives the correct neighborhood.

To summarize, with number of samples

$$k = \left(\frac{48^2(d+2+C_1)}{\epsilon^2 \delta^4 2d} \right) d \log n,$$

the algorithm correctly determines the graph G with probability

$$P(\hat{G}(X) = G) \geq 1 - 2(d+2)A^{d+2}/n^{C_1}.$$

The analysis of the running time is straightforward. There are n nodes, and for each node we consider $O(n^d)$ neighborhoods. For each candidate neighborhood, we check approximately $O(n)$ nodes and perform a correlation test of complexity $O(\log n)$. \square

4.2. General reconstruction. While Theorem 2 applies to a wide range of models, condition (8) may occasionally be too restrictive. One setting in which condition (8) does not apply is if the marginal spin at some vertex v is independent of the marginal spins at all its neighbors (i.e., for all $u \in N(v)$ and all $x, y \in \mathcal{A}$ we have $P(X(v) = x, X(u) = y) = P(X(v) = x)P(X(u) = y)$). In this case the algorithm would incorrectly return the empty set for the neighborhood of v .

As an illustrative example consider the case where the distribution P over $\{0, 1\}^{d \times k}$, where $d \geq 3$ and $k \geq 2$ is such that the parity of the first set of d variables (x_1, \dots, x_d) is even, the parity of the second set of d variables (x_{d+1}, \dots, x_{2d}) is even, etc. It is easy to encode this distribution using an MRF with k cliques of size d . Note that conditioning on the second group of d variables, there is no correlation between x_1 and x_2 . Therefore condition (8) doesn't hold. However, the weaker condition (16) does hold in this case.

In fact, the weaker conditions for Theorem 3 hold on essentially all MRFs. In particular, condition (16) says that the potentials are nondegenerate, which is clearly a necessary condition in order to recover the graph. Condition (17) holds for many models, for example, all models with soft constraints. This additional generality comes at a computational cost, with the algorithm for Theorem 2 having a faster running time, $O(n^{d+2} \log n)$ versus $O(n^{2d+1} \log n)$.

THEOREM 3. For an assignment $x_U = (x_{u_1}, \dots, x_{u_l})$ and $y \in \mathcal{A}$, define

$$x_U^i(y) = (x_{u_1}, \dots, y, \dots, x_{u_l})$$

to be the assignment obtained from x_U by replacing the i th element by y . Suppose there exist $\epsilon, \delta > 0$ such that the following condition holds: for all $v \in V$, if $N(v) =$

$\{u_1, \dots, u_l\}$, then for each $i, 1 \leq i \leq l$, and for any set $W \subset V - (\{v\} \cup N(v))$ with $|W| \leq d$, there exist values $x_v, x_{u_1}, \dots, x_{u_i}, \dots, x_{u_l}, y \in \mathcal{A}$, and $x_W \in \mathcal{A}^{|W|}$ such that

$$(16) \quad \left| P(X(v) = x_v | X(N(v)) = x_{N(v)}) - P(X(v) = x_v | X(N(v)) = x_{N(v)}^i(y)) \right| > \epsilon$$

and

$$(17) \quad \begin{aligned} P(X(N(v)) = x_{N(v)}, X(W) = x_W) &> \delta, \\ P(X(N(v)) = x_{N(v)}^i(y), X(W) = x_W) &> \delta. \end{aligned}$$

Then for some constant $C = C(\epsilon, \delta) > 0$, if $k > Cd \log n$, then there exists an estimator $\hat{G}(\underline{X})$ such that the probability of correct reconstruction is $P(G = \hat{G}(\underline{X})) = 1 - o(1)$. The estimator \hat{G} is computable in time $O(n^{2d+1} \log n)$.

Proof. As in Theorem 2, we can assume that with high probability we have

$$(18) \quad \left| \hat{P}(X(U) = x_U) - P(X(U) = x_U) \right| \leq \gamma$$

for all $U = \{u_i\}_{i=1}^l \subset V$ and $\{x_i\}_{i=1}^l$ when $l \leq 2d + 1$ and $k \geq C(\gamma)d \log n$, so we assume that (18) holds. For each vertex $v \in V$ we consider all candidate neighborhoods for v , subsets $U = \{u_1, \dots, u_l\} \subset V - \{v\}$ with $0 \leq l \leq d$. For each candidate neighborhood U , the algorithm computes a score

$$\begin{aligned} f(v; U) = \min_{W, i} \max_{x_v, x_W, x_U, y} & \left| \hat{P}(X(v) = x_v | X(W) = x_W, X(U) = x_U) \right. \\ & \left. - \hat{P}(X(v) = x_v | X(W) = x_W, X(U) = x_U^i(y)) \right|, \end{aligned}$$

where for each W, i , the maximum is taken over all x_v, x_W, x_U, y , such that

$$(19) \quad \begin{aligned} \hat{P}(X(W) = x_W, X(U) = x_U) &> \delta/2, \\ \hat{P}(X(W) = x_W, X(U) = x_U^i(y)) &> \delta/2, \end{aligned}$$

and $W \subset V - (\{v\} \cup U)$ is an arbitrary set of nodes of size d , $x_W \in \mathcal{A}^d$ is an arbitrary assignment of values to the nodes in W , and $1 \leq i \leq l$.

The algorithm selects as the neighborhood of v the largest set $U \subset V - \{v\}$ with $f(v; U) > \epsilon/2$. It is necessary to check that if U is the true neighborhood of v , then the algorithm accepts U , and otherwise the algorithm rejects U .

Taking $\gamma(\epsilon, \delta) = \epsilon\delta^2/48$, it follows exactly as in Theorem 2 that the error in each of the relevant empirical conditional probabilities satisfies

$$(20) \quad \begin{aligned} & \left| \hat{P}(X(v) = x_v | X(W) = x_W, X(U) = x_U) \right. \\ & \left. - P(X(v) = x_v | X(W) = x_W, X(U) = x_U) \right| < \frac{\epsilon}{4}. \end{aligned}$$

If $U \not\subseteq N(v)$, choosing $u_i \in U - N(v)$, we have when $N(v) \subset W \cup U$

$$\begin{aligned} & \left| P(X(v) = x_v | X(W) = x_W, X(U) = x_U) - P(X(v) = x_v | X(W) = x_W, X(U) = x_U^i(y)) \right| \\ & = \left| P(X(v) = x_v | X(N(v)) = x_{N(v)}) - P(X(v) = x_v | X(N(v)) = x_{N(v)}) \right| \\ & = 0 \end{aligned}$$

by the Markov property (7). Assuming that (18) holds with γ chosen as in (13), the estimation error in $f(v; U)$ is at most $\epsilon/2$ by (20), and it holds that $f(v; U) < \epsilon/2$ for each $U \not\subseteq N(v)$. Thus all $U \not\subseteq N(v)$ are rejected. If $U = N(v)$, then by the Markov property (7) and the conditions (16) and (17), for any i and $W \subset V$,

$$\begin{aligned} & |P(X(v) = x_v | X(W) = x_W, X(U) = x_U) - P(X(v) = x_v | X(W) = x_W, X(U) = x_U^i(y))| \\ &= |P(X(v) = x_v | X(N(v)) = x_{N(v)}) - P(X(v) = x_v | X(N(v)) = x_{N(v)}^i(y))| \\ &> \epsilon \end{aligned}$$

for some x_v, x_W, x_U, y . The error in $f(v; U)$ is less than $\epsilon/2$ as before; hence $f(v; U) > \epsilon/2$ for $U = N(v)$. Since $U = N(v)$ is the largest set that is not rejected, the algorithm correctly determines the neighborhood of v for every $v \in V$ when (18) holds.

To summarize, with number of samples

$$k = \left(\frac{48^2(2d+1+C_1)}{\epsilon^2 \delta^4 2d} \right) d \log n,$$

the algorithm correctly determines the graph G with probability

$$P(\hat{G}(X) = G) \geq 1 - 2(d+2)A^{2d+1}/n^{C_1}.$$

The analysis of the running time is similar to the previous algorithm. \square

4.3. Nondegeneracy of models. We can expect conditions (16) and (17) to hold in essentially all models of interest. The following proposition shows that they hold for any model with soft constraints.

PROPOSITION 1 (models with soft constraints). *In a graphical model with maximum degree d given by (2), suppose that all the potentials Ψ_{uv} satisfy $\|\Psi_{uv}\|_\infty \leq K$ and*

$$(21) \quad \max_{x_1, x_2, x_3, x_4 \in \mathcal{A}} |\Psi_{uv}(x_1, x_2) - \Psi_{uv}(x_3, x_2) - \Psi_{uv}(x_1, x_4) + \Psi_{uv}(x_3, x_4)| > \gamma$$

for some $\gamma > 0$. Then there exist $\epsilon, \delta > 0$ depending only on d, K , and γ such that the hypothesis of Theorem 3 holds.

In order to understand the meaning of condition (21), consider a model where $\Psi_{uv}(x_1, x_2) \equiv \varphi_1(x_1)\varphi_2(x_2)$. In this case, the edge u, v may be represented by assigning node potentials. In particular, there is no way of reconstructing it. In this case, the expression in (21) is identically 0. The condition (21) guarantees that the edge of the graph cannot be represented in terms of node potentials.

Proof. It is clear that for some sufficiently small $\delta = \delta(d, K) > 0$ we have that for all $u_1, \dots, u_{2d+1} \in V$ and $x_{u_1}, \dots, x_{u_{2d+1}} \in \mathcal{A}$ that

$$(22) \quad P(X(u_1) = x_{u_1}, \dots, X(u_{2d+1}) = x_{u_{2d+1}}) > \delta.$$

Now suppose that u_1, \dots, u_l is the neighborhood of v . Then for any $1 \leq i \leq l$ it follows from (21) that there exist $x_v, x'_v, x_{u_i}, x'_{u_i} \in \mathcal{A}$ such that for any $x_{u_1}, \dots, x_{u_{i-1}}, x_{u_{i+1}}, \dots, x_{u_l} \in \mathcal{A}$,

$$\begin{aligned} & \frac{P(X(v) = x_v | X(u_1) = x_{u_1}, \dots, X(u_i) = x'_{u_i}, \dots, X(u_l) = x_{u_l})}{P(X(v) = x'_v | X(u_1) = x_{u_1}, \dots, X(u_i) = x'_{u_i}, \dots, X(u_l) = x_{u_l})} \\ & \geq e^\gamma \frac{P(X(v) = x_v | X(u_1) = x_{u_1}, \dots, X(u_i) = x_{u_i}, \dots, X(u_l) = x_{u_l})}{P(X(v) = x'_v | X(u_1) = x_{u_1}, \dots, X(u_i) = x_{u_i}, \dots, X(u_l) = x_{u_l})}. \end{aligned}$$

Combining this with (22), condition (16) follows. \square

Although the results to follow hold more generally, for ease of exposition we will keep in mind the example of the Ising model with no external magnetic field,

$$(23) \quad P(\vec{x}) = \frac{1}{Z} \exp \left(\sum_{(u,v) \in E} \beta_{uv} x_u x_v \right),$$

where $\beta_{uv} \in \mathbb{R}$ are coupling constants and Z is a normalizing constant.

The following lemma gives explicit bounds on ϵ, δ in terms of bounds on the coupling constants in the Ising model, showing that the conditions of Theorem 3 can be expected to hold quite generally.

PROPOSITION 2. *Consider the Ising model with all parameters satisfying*

$$0 < c < |\beta_{ij}| < C$$

on a graph G with maximum degree at most d . Then conditions (16) and (17) of Theorem 3 are satisfied with

$$\epsilon \geq \frac{\tanh(2c)}{4 \tanh(2C)}$$

and

$$\delta \geq \frac{e^{-4d^2C}}{2^{2d}}.$$

Proof. Fix a vertex $v \in V$, and let $w \in N(v)$ be any vertex in the neighborhood of v . Let $R = N(v) \setminus \{w\}$ be the other neighbors of v . Then

$$(24) \quad \begin{aligned} P(X(v) = 1 | X(R) = x_R, X(w) = x_w) \\ &= \frac{P(X(v) = 1, X(R) = x_R, X(w) = x_w)}{P(X(v) = 1, X(R) = x_R, X(w) = x_w) + P(X(v) = -1, X(R) = x_R, X(w) = x_w)} \\ &= \frac{\exp \left(\sum_{j \in R} x_j \beta_{jv} + x_w \beta_{wv} \right)}{\exp \left(\sum_{j \in R} x_j \beta_{jv} + x_w \beta_{wv} \right) + \exp \left(- \sum_{j \in R} x_j \beta_{jv} - x_w \beta_{wv} \right)}. \end{aligned}$$

Defining

$$A := \exp \left(\sum_{j \in R} x_j \beta_{jv} \right),$$

we have from (24) that

$$\begin{aligned} &|P(X(v) = 1 | X(R) = x_R, X(w) = 1) - P(X(v) = 1 | X(R) = x_R, X(w) = -1)| \\ &= \left| \frac{Ae^{\beta_{wv}}}{Ae^{\beta_{wv}} + A^{-1}e^{-\beta_{wv}}} - \frac{Ae^{-\beta_{wv}}}{Ae^{-\beta_{wv}} + A^{-1}e^{\beta_{wv}}} \right| \\ &= \left| \frac{A^2(e^{2\beta_{wv}} - e^{-2\beta_{wv}})}{A^4 + A^2(e^{2\beta_{wv}} + e^{-2\beta_{wv}}) + 1} \right| \\ &= \frac{A^2(e^{2|\beta_{wv}|} - e^{-2|\beta_{wv}|})}{A^4 + A^2(e^{2|\beta_{wv}|} + e^{-2|\beta_{wv}|}) + 1} \\ &= \frac{(e^{2|\beta_{wv}|} - e^{-2|\beta_{wv}|})}{A^2 + e^{2|\beta_{wv}|} + e^{-2|\beta_{wv}|} + A^{-2}} \geq \frac{\tanh(2|\beta_{wv}|)}{2A^2 + 2A^{-2}}. \end{aligned}$$

It is possible to choose the spins x_R in such a way that $e^{-C} < A < e^C$. Thus the expression above is at least

$$\frac{\tanh(2c)}{4 \tanh(2C)}.$$

Moreover, the probability of any assignment of $2d$ spins can be very crudely bounded as

$$P(X(i_1) = x_{i_1}, \dots, X(i_{2d}) = x_{i_{2d}}) \geq \frac{e^{-4d^2C}}{2^{2d}}. \quad \square$$

4.4. $O(n^2 \log n)$ algorithm for models with correlation decay. The reconstruction algorithm runs in polynomial time $O(dn^{2d+1} \ln n)$. It would be desirable for the degree of the polynomial to be independent of d , and this can be achieved for MRFs with exponential decay of correlations. For two vertices $u, v \in V$ let $d(u, v)$ denote the graph distance and let $d_C(u, v)$ denote the correlation between the spins at u and v defined as

$$d_C(u, v) = \sum_{x_u, x_v \in \mathcal{A}} |P(X(u) = x_u, X(v) = x_v) - P(X(u) = x_u)P(X(v) = x_v)|.$$

If the interactions are sufficiently weak, the graph will satisfy the Dobrushin–Shlosman condition (see, e.g., [6]) and there will be exponential decay of correlations between vertices.

THEOREM 4. *Suppose that G and X satisfy the hypothesis of Theorem 3 and that for all $u, v \in V$, $d_C(u, v) \leq \exp(-\alpha d(u, v))$ and there exists some $\kappa > 0$ such that for all $(u, v) \in E$, $d_C(u, v) > \kappa$. Then for some constant $C = C(\alpha, \kappa, \epsilon, \delta) > 0$, if $k > Cd \log n$, then there exists an estimator $\hat{G}(\underline{X})$ such that the probability of correct reconstruction is $P(G = \hat{G}(\underline{X})) = 1 - o(1)$ and the algorithm running time is $O(nd^{\frac{2d \ln(4/\kappa)}{\alpha}} + dn^2 \ln n)$ with high probability.*

Proof. Denote the correlation neighborhood of a vertex v as $N_C(v) = \{u \in V : \widehat{d}_C(u, v) > \kappa/2\}$, where $\widehat{d}_C(u, v)$ is the empirical correlation of u and v . For large enough C with high probability for all $v \in V$, we have that $N(v) \subseteq N_C(v) \subseteq \{u \in V : d(u, v) \leq \frac{\ln(4/\kappa)}{\alpha}\}$. Now the size of $|\{u \in V : d(u, v) \leq \frac{\ln(4/\kappa)}{\alpha}\}|$ is at most $d^{\frac{\ln(4/\kappa)}{\alpha}}$, which is independent of n .

When reconstructing the neighborhood of a vertex v we modify the algorithm in Theorem 3 to test only candidate neighborhoods U and sets W which are subsets of $N_C(v)$. The algorithm restricted to the smaller range of possible neighborhoods correctly reconstructs the graph with high probability since the true neighborhood of a vertex is in its correlation neighborhood. For each vertex v the total number of choices of candidate neighborhoods U and sets W the algorithm has to check is $O(d^{\frac{2d \ln(4/\kappa)}{\alpha}})$, so running the reconstruction algorithm takes $O(nd^{\frac{2d \ln(4/\kappa)}{\alpha}})$ operations. It takes $O(dn^2 \ln n)$ operations to calculate all the correlations, which for large n dominates the running time. \square

5. Noisy and incomplete observations. More generally there is the problem of reconstructing an MRF from noisy observations. In this setting we observe $\underline{Y} = \{Y^1, \dots, Y^k\}$ instead of $\underline{X} = \{X^1, \dots, X^k\}$, where each Y_i is a noisy version of X_i . The algorithm in Theorem 3 is robust to small amounts of noise, even when the errors in different vertices are not necessarily independent. One sufficient condition is that

there exist $0 < \epsilon' < \epsilon$ and $0 < \delta' < \delta$ such that for any $2d + 1$ vertices v_1, \dots, v_{2d+1} and states x_1, \dots, x_{2d+1} we have that

$$|P(X(v_1) = x_1, \dots, X(v_{2d}) = x_{2d}) - P(Y(v_1) = x_1, \dots, Y(v_{2d}) = x_{2d})| \leq \delta'/2$$

and

$$\begin{aligned} &|P(X(v_{2d+1}) = x_{2d+1} | X(v_1) = x_1, \dots, X(v_{2d}) = x_{2d}) \\ &- P(Y(v_{2d+1}) = x_{2d+1} | Y(v_1) = x_1, \dots, Y(v_{2d}) = x_{2d})| \leq \epsilon'/2. \end{aligned}$$

For some $C' = C'(\epsilon, \epsilon', \delta, \delta') > 0$ with $k = C'd \log n$ samples, the reconstruction algorithm of Theorem 3 correctly reconstructs the graph G with high probability (the same proof holds).

5.1. An example of nonidentifiability. Without assumptions on the underlying model or noise, the MRF is not in general identifiable. In other words, a single probability distribution might correspond to two different graph structures. Thus, the problem of reconstruction is not well defined in such a case. The next example shows that even in the Ising model, under unknown noise it is impossible to distinguish between a graph with three vertices and two edges and a graph with three vertices and three edges.

Example 1. Let $V = \{u_1, u_2, u_3\}$ be a set of three vertices, and let G and \tilde{G} be two graphs with vertex set V and edge sets $\{(u_1, u_2), (u_1, u_3)\}$ and $\{(u_1, u_2), (u_1, u_3), (u_2, u_3)\}$, respectively. Let P and \tilde{P} be Ising models on G and \tilde{G} with edge interactions β_{12}, β_{13} and $\tilde{\beta}_{12}, \tilde{\beta}_{13}, \tilde{\beta}_{23}$, respectively; i.e.,

$$\begin{aligned} P[X] &= \frac{1}{Z} \exp(\beta_{12}X(u_1)X(u_2) + \beta_{13}X(u_1)X(u_3)), \\ \tilde{P}[X] &= \frac{1}{Z} \exp(\tilde{\beta}_{12}X(u_1)X(u_2) + \tilde{\beta}_{13}X(u_1)X(u_3) + \tilde{\beta}_{23}X(u_2)X(u_3)). \end{aligned}$$

Suppose that $X'(u_1)$, a noisy version of the spin $X(u_1)$, is observed which is equal to $X(u_1)$ with probability p and $-X(u_1)$ with probability $1 - p$ for some random unknown p while the spins $X(u_2)$ and $X(u_3)$ are observed perfectly. This is equivalent to adding a new vertex u'_1 to G and \tilde{G} with an extra edge (u_1, u'_1) and potential $\Psi_{(u_1, u'_1)} = \beta_{11'}X(u_1)X(u'_1)$. The spin at u'_1 then represents the noisy observation of the spin at u_1 . Suppose that all the β and $\tilde{\beta}$ are chosen independently with $N(0, 1)$ distribution, and let \mathcal{P} and $\tilde{\mathcal{P}}$ be the random noisy distributions on $\mathcal{A}^{\{u'_1, u_2, u_3\}}$. Then the total variation distance between \mathcal{P} and $\tilde{\mathcal{P}}$ is less than 1, and so the graph structure is not identifiable, as we shall show below.

By the symmetry of the Ising model with no external field the random element \mathcal{P} can be parameterized by $(p_{1'2}, p_{1'3}, p_{23}) \in [0, 1]^3$, where $p_{1'2} = P(X_{u'_1} = 1, X_{u_2} = 1)$, $p_{1'3} = P(X_{u'_1} = 1, X_{u_3} = 1)$, $p_{23} = P(X_{u_2} = 1, X_{u_3} = 1)$. These parameters are given by

$$p_{ij} = h(\beta_{1i})h(\beta_{1j}) + h(-\beta_{1i})h(-\beta_{1j}),$$

where $h(\beta) = \frac{e^\beta}{e^\beta + e^{-\beta}}$. Let φ be the function $\varphi : \mathbb{R}^3 \rightarrow [0, 1]^3$ which maps $(\beta_{11'}, \beta_{12}, \beta_{13}) \mapsto (p_{1'2}, p_{1'3}, p_{23})$, and let J_φ be its Jacobian. Then $\det(J_\varphi(1, 1, 1)) > 0$, and by continuity the Jacobian is positive in a neighborhood of $(1, 1, 1)$. It follows that the random vector $(p_{1'2}, p_{1'3}, p_{23})$ has a density with respect to Lebesgue measure in a neighborhood of $(2h(1)^2, 2h(1)^2, 2h(1)^2)$.

Now let $\tilde{\varphi}$ be the function $\tilde{\varphi} : \mathbb{R}^3 \rightarrow [0, 1]^3$ which maps $(\widetilde{\beta_{11'}}, \widetilde{\beta_{12}}, \widetilde{\beta_{13}}, \widetilde{\beta_{23}}) \mapsto (\widetilde{p_{1'2}}, \widetilde{p_{1'3}}, \widetilde{p_{23}})$. If we fix $\widetilde{\beta_{23}} = 0$, then $\tilde{\varphi} = \varphi$ induces a positive density in the random vector $(\widetilde{p_{1'2}}, \widetilde{p_{1'3}}, \widetilde{p_{23}})$ in a neighborhood of $(2h(1)^2, 2h(1)^2, 2h(1)^2)$. By continuity this also holds when $|\widetilde{\beta_{23}}|$ is small enough and so $(\widetilde{p_{1'2}}, \widetilde{p_{1'3}}, \widetilde{p_{23}})$ has a positive density around $(2h(1)^2, 2h(1)^2, 2h(1)^2)$. Hence we have that both \mathcal{P} and $\tilde{\mathcal{P}}$ have positive densities in an overlapping region so their total variation distance is less than 1 and so the graph structure is not identifiable.

5.2. Models with hidden variables. A related question is can we identify if a vertex is missing, and if so, where it fits into the graph. Under the assumption that the vertices all have degree at least 3 and the graph is triangle-free, we can recover missing vertices under mild assumptions.

THEOREM 5. *Suppose that the hypothesis of Theorem 3 holds for some MRF X based on a triangle-free graph with minimum degree at least 3 and maximum degree d' . Let $V^* \subseteq V$ such that for any two points $v, v' \in V - V^*$ we have $d(v, v') \geq 3$ and suppose we are given samples from X^* , the restriction of X to V^* with which to reconstruct G .*

Suppose the following condition also holds: for all $v \in V$ if $v_1, v_2 \in N(v)$ and $U = N(v) \cup N(v_1) - \{v, v_1, v_2\}$ and $W \subset V - (N(v) \cup N(v_1))$ with $|W| \leq 2d'$ then there exists some $x_{v_1}, x_{v_2}, x'_{v_2}, x_U, x_W$ such that

$$(25) \quad \begin{aligned} &|P(X(v_1) = x_{v_1} | X(W) = x_W, X(U) = x_U, X(v_2) = x_{v_2}) \\ &\quad - P(X(v_1) = x_{v_1} | X(W) = x_W, X(U) = x_U, X(v_2) = x'_{v_2})| > \epsilon \end{aligned}$$

and

$$(26) \quad \begin{aligned} &P(X(W) = x_W, X(U) = x_U, X(v_2) = x_{v_2}) > \delta, \\ &P(X(W) = x_W, X(U) = x_U, X(v_2) = x'_{v_2}) > \delta. \end{aligned}$$

Then for some constant $C = C(\epsilon, \delta) > 0$, if $k > Cd \log n$ then there exists an estimator $\hat{G}(\underline{X}^)$ such that the probability of correct reconstruction is $P(G = \hat{G}(\underline{X}^*)) = 1 - o(1)$.*

Proof. We apply the algorithm from Theorem 3 to \underline{X}^* setting the maximum degree as $d = 2d'$. The algorithm will output the graph $G^* = (V^*, E^*)$. If $v, N(v) \subset V^*$ then the algorithm correctly reconstructs the neighborhood $N(v)$. Any vertex in V^* is adjacent to at most one missing vertex so suppose that v_1 is a vertex adjacent to a missing vertex v . Then by condition (25) and (26) we have that the algorithm reconstructs the neighborhood of v_1 as $N(v) \cup N(v_1) - \{v, v_1\}$. So the edge set E^* is exactly all the edges in the induced subgraph of V^* plus a clique connecting all the neighbors of missing vertices. Since G is triangle-free every maximal clique (a clique that cannot be enlarged) of size at least 3 corresponds to a missing vertex.

So to reconstruct G from G^* we simply replace every maximal clique in G^* with a vertex connected to all the vertices in the clique. This exactly reconstructs the graph with high probability. \square

Remark 3. The condition that missing vertices are at distance at least 3 is not necessary, but this assumption simplifies the algorithm because the cliques corresponding to missing vertices are disjoint. A slightly more involved algorithm is able to reconstruct graphs where the missing vertices have $d(v, v') = 2$.

The following lemma shows that the conditions for recovery of missing vertices in Theorem 5 are satisfied for a ferromagnetic Ising model satisfying the assumptions of Lemma 2.

LEMMA 2. Consider the ferromagnetic Ising model where all coupling parameters satisfy

$$0 < c < \beta_{ij} < C$$

on a triangle-free graph G with minimum degree 3. Then the conditions of Theorem 5 are satisfied with

$$\epsilon \geq \frac{\tanh(2c)}{64e^{2(d+1)C} \tanh(2C)}$$

and

$$\delta \geq \frac{e^{-4d^2C}}{2^{2d}}.$$

Proof. To check the first condition we write

$$\begin{aligned} & |P(X(v_1) = 1 | X(N \cup W) = x_{N \cup W}, X(v_2) = 1) \\ & \quad - P(X(v_1) = 1 | X(N \cup W) = x_{N \cup W}, X(v_2) = -1)| \\ &= |P(X(v_1) = 1 | X(N) = x_N, X(v_2) = 1, X(v) = 1)P(v = 1 | X(N) = x_N, X(v_2) = 1) \\ & \quad + P(X(v_1) = 1 | X(N) = x_N, X(v_2) = 1, X(v) = -1)P(v = -1 | X(N) = x_N, X(v_2) = 1) \\ & \quad - P(X(v_1) = 1 | X(N) = x_N, X(v_2) = -1, X(v) = 1)P(v = 1 | X(N) = x_N, X(v_2) = -1) \\ & \quad - P(X(v_1) = 1 | X(N) = x_N, X(v_2) = -1, X(v) = -1)P(v = -1 | X(N) = x_N, X(v_2) = -1)| \\ &= |P(X(v_1) = 1 | X(N) = x_N, X(v) = 1)P(v = 1 | X(N) = x_N, X(v_2) = 1) \\ & \quad + P(X(v_1) = 1 | X(N) = x_N, X(v) = -1)P(v = -1 | X(N) = x_N, X(v_2) = 1) \\ & \quad - P(X(v_1) = 1 | X(N) = x_N, X(v) = 1)P(v = 1 | X(N) = x_N, X(v_2) = -1) \\ & \quad - P(X(v_1) = 1 | X(N) = x_N, X(v) = -1)P(v = -1 | X(N) = x_N, X(v_2) = -1)|, \end{aligned}$$

where $N = N(v) \cup N(v_1) - \{v, v_1, v_2\}$ and where the last step follows by the Markov property (since all paths from v_1 to v_2 pass through vertices in N or through v). Continuing, we have that the above is equal to

$$\begin{aligned} & |(P(X(v_1) = 1 | X(N), X(v) = 1) - P(X(v_1) = 1 | X(N), X(v) = -1)) \\ & \quad \cdot (P(X(v) = 1 | X(N), X(v_2) = 1) - P(X(v) = 1 | X(N), X(v_2) = -1))|. \end{aligned}$$

But by Lemma 2,

$$|(P(X(v_1) = 1 | X(N), X(v) = 1) - P(X(v_1) = 1 | X(N), X(v) = -1))| > \frac{\tanh(2c)}{4 \tanh(2C)}.$$

By the ferromagnetic assumption, the second factor can be lower bounded as

$$|(P(X(v) = 1 | X(N), X(v_2) = 1) - P(X(v) = 1 | X(N), X(v_2) = -1))| > \frac{1}{16e^{2(d+1)C}}.$$

Hence the first condition is satisfied with

$$\epsilon > \frac{\tanh(2c)}{64e^{2(d+1)C} \tanh(2C)}.$$

The second condition, by the same argument as in Lemma 2, is satisfied with

$$\delta \geq \frac{e^{-4d^2C}}{2^{2d}}. \quad \square$$

Acknowledgments. E. M. thanks Marek Biskup for helpful discussions on models with hidden variables. The authors greatly appreciate Miklos Racz's many comments on the paper.

REFERENCES

- [1] P. ABBEEL, D. KOLLER, AND A. NG, *Learning factor graphs in polynomial time and sample complexity*, J. Mach. Learn. Res., 7 (2006), pp. 1743–1788.
- [2] J. BENTO AND A. MONTANARI, *On the Trade-off Between Complexity and Correlation Decay in Structural Learning Algorithms*, preprint, arXiv:1110.1769v1 [statML], 2011.
- [3] D. M. CHICKERING, *Learning Bayesian networks is NP-complete*, in Learning from Data: Of AI and Statistics, D. Fisher and H. J. Lenz, eds., Springer, New York, 1996, pp. 121–130.
- [4] C. K. CHOW AND C. N. LIU, *Approximating discrete probability distributions with dependence trees*, IEEE Trans. Inform. Theory, 14 (1968), pp. 462–467.
- [5] C. DASKALAKIS, E. MOSSEL, AND S. ROCH, *Optimal phylogenetic reconstruction*, in Proceedings of the 38th Annual ACM Symposium on Theory of Computing (STOC'06), ACM, New York, 2006, pp. 159–168.
- [6] R. L. DOBRUSHIN AND S. B. SHLOSMAN, *Completely analytical Gibbs fields*, in Statistical Physics and Dynamical Systems, J. Fritz, A. Jaffe, and D. Szasz, eds., Birkhäuser Boston, Boston, 1985, pp. 371–403.
- [7] P. L. ERDŐS, M. A. STEEL, L. A. SZÉKELY, AND T. WARNOW, *A few logs suffice to build (almost) all trees. I*, Random Structures Algorithms, 14 (1999) pp. 153–184.
- [8] N. FRIEDMAN, *Inferring cellular networks using probabilistic graphical models*, Science, 303 (2004), pp. 799–805.
- [9] S. KASIF, *Bayes Networks and Graphical Models in Computational Molecular Biology and Bioinformatics, Survey of Recent Research*, <http://genomics10.bu.edu/bioinformatics/kasif/bayes-net.html> (2007).
- [10] E. MOSSEL, *Distorted metrics on trees and phylogenetic forests*, IEEE/ACM Trans. Comput. Biol. Bioinform., 4 (2007), pp. 108–116.
- [11] N. SANTHANAM AND M. J. WAINWRIGHT, *Information-theoretic limits of graphical model selection in high dimensions*, in Proceedings of IEEE International Symposium on Information Theory (ISIT), 2008, pp. 2136–2140.
- [12] M. J. WAINWRIGHT, P. RAVIKUMAR, AND J. D. LAFFERTY, *High-dimensional graphical model selection using ℓ_1 -regularized logistic regression*, in Advances in Neural Processing Systems (NIPS), MIT Press, Cambridge, MA, 2006, pp. 1465–1472.