STRUCTURE ESTIMATION FOR DISCRETE GRAPHICAL MODELS: GENERALIZED COVARIANCE MATRICES AND THEIR INVERSES

BY PO-LING LOH[1,2] AND MARTIN J. WAINWRIGHT[2]

University of California, Berkeley

We investigate the relationship between the structure of a discrete graphical model and the support of the inverse of a generalized covariance matrix. We show that for certain graph structures, the support of the inverse covariance matrix of indicator variables on the vertices of a graph reflects the conditional independence structure of the graph. Our work extends results that have previously been established only in the context of multivariate Gaussian graphical models, thereby addressing an open question about the significance of the inverse covariance matrix of a non-Gaussian distribution. The proof exploits a combination of ideas from the geometry of exponential families, junction tree theory and convex analysis. These population-level results have various consequences for graph selection methods, both known and novel, including a novel method for structure estimation for missing or corrupted observations. We provide nonasymptotic guarantees for such methods and illustrate the sharpness of these predictions via simulations.

1. Introduction. Graphical models are used in many application domains, running the gamut from computer vision and civil engineering to political science and epidemiology. In many applications, estimating the edge structure of an underlying graphical model is of significant interest. For instance, a graphical model may be used to represent friendships between people in a social network [3] or links between organisms with the propensity to spread an infectious disease [28]. It is a classical corollary of the Hammersley–Clifford theorem [5, 15, 21] that zeros in the inverse covariance matrix of a multivariate Gaussian distribution indicate absent edges in the corresponding graphical model. This fact, combined with various types of statistical estimators suited to high dimensions, has been leveraged by many authors to recover the structure of a Gaussian graphical model when the edge set is sparse (see the papers [8, 27, 31, 38] and the references therein). Recently, Liu et al. [23] and Liu, Lafferty and Wasserman [24] introduced the notion

of a nonparanormal distribution, which generalizes the Gaussian distribution by allowing for monotonic univariate transformations, and argued that the same structural properties of the inverse covariance matrix carry over to the nonparanormal; see also the related work of Xue and Zou [37] on copula transformations.

However, for non-Gaussian graphical models, the question of whether a general relationship exists between conditional independence and the structure of the inverse covariance matrix remains unresolved. In this paper, we establish a number of interesting links between covariance matrices and the edge structure of an underlying graph in the case of discrete-valued random variables. (Although we specialize our treatment to multinomial random variables due to their widespread applicability, several of our results have straightforward generalizations to other types of exponential families.) Instead of only analyzing the standard covariance matrix, we show that it is often fruitful to augment the usual covariance matrix with higher-order interaction terms. Our main result has an interesting corollary for tree-structured graphs: for such models, the inverse of a generalized covariance matrix is always (block) graph-structured. In particular, for binary variables, the inverse of the usual covariance matrix may be used to recover the edge structure of the tree. We also establish more general results that apply to arbitrary (nontree) graphs, specified in terms of graph triangulations. This more general correspondence exploits ideas from the geometry of exponential families [7, 36], as well as the junction tree framework [21, 22].

As we illustrate, these population-level results have a number of corollaries for graph selection methods. Graph selection methods for Gaussian data include neighborhood regression [27, 40] and the graphical Lasso [12, 14, 31, 33], which corresponds to maximizing an ℓ_1-regularized version of the Gaussian likelihood. Alternative methods for selection of discrete graphical models include the classical Chow–Liu algorithm for trees [10]; techniques based on conditional entropy or mutual information [2, 6]; and nodewise logistic regression for discrete graphical models with pairwise interactions [19, 30]. Our population-level results imply that minor variants of the graphical Lasso and neighborhood regression methods, though originally developed for Gaussian data, remain consistent for trees and the broader class of graphical models with singleton separator sets. They also convey a cautionary message, in that these methods will be inconsistent (generically) for other types of graphs. We also describe a new method for neighborhood selection in an arbitrary sparse graph, based on linear regression over subsets of variables. This method is most useful for bounded-degree graphs with correlation decay, but less computationally tractable for larger graphs.

In addition, we show that our methods for graph selection may be adapted to handle noisy or missing data in a seamless manner. Naively applying nodewise logistic regression when observations are systematically corrupted yields estimates that are biased even in the limit of infinite data. There are various corrections available, such as multiple imputation [34] and the expectation-maximization (EM) algorithm [13], but, in general, these methods are not guaranteed to be statistically

consistent due to local optima. To the best of our knowledge, our work provides the first method that is provably consistent under high-dimensional scaling for estimating the structure of discrete graphical models with corrupted observations. Further background on corrupted data methods for low-dimensional logistic regression may be found in Carroll, Ruppert and Stefanski [9] and Ibrahim et al. [17].

The remainder of the paper is organized as follows. In Section 2, we provide brief background and notation on graphical models and describe the classes of augmented covariance matrices we will consider. In Section 3, we state our main population-level result (Theorem 1) on the relationship between the support of generalized inverse covariance matrices and the edge structure of a discrete graphical model, and then develop a number of corollaries. The proof of Theorem 1 is provided in Section 3.4, with proofs of corollaries and more technical results deferred to the supplementary material [26]. In Section 4, we develop consequences of our population-level results in the context of specific methods for graphical model selection. We provide simulation results in Section 4.4 in order to confirm the accuracy of our theoretically-predicted scaling laws, dictating how many samples are required (as a function of graph size and maximum degree) to recover the graph correctly.

2. Background and problem setup. In this section, we provide background on graphical models and exponential families. We then present a simple example illustrating the phenomena and methodology underlying this paper.

2.1. *Undirected graphical models.* An *undirected graphical model* or *Markov random field* (MRF) is a family of probability distributions respecting the structure of a fixed graph. We begin with some basic graph-theoretic terminology. An undirected graph $G = (V, E)$ consists of a collection of vertices $V = \{1, 2, \ldots, p\}$ and a collection of unordered[3] vertex pairs $E \subseteq V \times V$. A *vertex cutset* is a subset U of vertices whose removal breaks the graph into two or more nonempty components [see Figure 1(a)]. A *clique* is a subset $C \subseteq V$ such that $(s, t) \in E$ for all distinct $s, t \in C$. The cliques in Figure 1(b) are all *maximal*, meaning they are not properly contained within any other clique. For $s \in V$, we define the neighborhood $N(s) := \{t \in V \mid (s, t) \in E\}$ to be the set of vertices connected to s by an edge.

For an undirected graph G, we associate to each vertex $s \in V$ a random variable X_s taking values in a space $\mathcal{X}$. For any subset $A \subseteq V$, we define $X_A := \{X_s, s \in A\}$, and for three subsets of vertices, A, B and U, we write $X_A \perp\!\!\!\perp X_B \mid X_U$ to mean that the random vector X_A is conditionally independent of X_B given X_U. The notion of a Markov random field may be defined in terms of certain *Markov properties* indexed by vertex cutsets or in terms of a *factorization property* described by the graph cliques.

---

[3]No distinction is made between the edge (s, t) and the edge (t, s). In this paper, we forbid graphs with self-loops, meaning that $(s, s) \notin E$ for all $s \in V$.
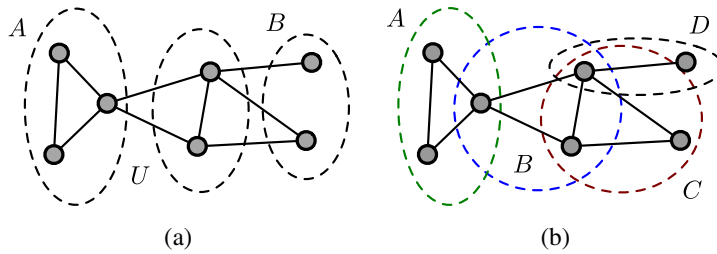
FIG. 1. (a) *Illustration of a vertex cutset*: *when the set U is removed, the graph breaks into two disjoint subsets of vertices A and B.* (b) *Illustration of maximal cliques, corresponding to fully connected subsets of vertices.*

DEFINITION 1 (Markov property). The random vector $X := (X_1, \ldots, X_p)$ is *Markov with respect to the graph G* if $X_A \perp\!\!\!\perp X_B \mid X_U$ whenever U is a vertex cutset that breaks the graph into disjoint subsets A and B.

Note that the neighborhood set $N(s)$ is a vertex cutset for the sets $A = \{s\}$ and $B = V \setminus \{s \cup N(s)\}$. Consequently, $X_s \perp\!\!\!\perp X_{V \setminus \{s \cup N(s)\}} \mid X_{N(s)}$. This property is important for nodewise methods for graphical model selection to be discussed later.

The factorization property is defined directly in terms of the probability distribution q of the random vector X. For each clique C, a *clique compatibility function* ψ_C is a mapping from configurations $x_C = \{x_s, s \in V\}$ of variables to the positive reals. Let $\mathcal{C}$ denote the set of all cliques in G.

DEFINITION 2 (Factorization property). The distribution of X *factorizes according to G* if it may be written as a product of clique functions:

$$(2.1) \qquad q(x_1, \ldots, x_p) \propto \prod_{C \in \mathcal{C}} \psi_C(x_C).$$

The factorization may always be restricted to maximal cliques of the graph, but it is sometimes convenient to include terms for nonmaximal cliques.

2.2. *Graphical models and exponential families.* By the Hammersley–Clifford theorem [5, 15, 21], the Markov and factorization properties are equivalent for any strictly positive distribution. We focus on such strictly positive distributions, in which case the factorization (2.1) may alternatively be represented in terms of an *exponential family* associated with the clique structure of G. We begin by defining this exponential family representation for the special case of binary variables ($\mathcal{X} = \{0, 1\}$), before discussing a natural generalization to m-ary discrete random variables.

Binary variables. For a binary random vector $X \in \{0, 1\}^p$, we associate with each clique C—both maximal and nonmaximal—a sufficient statistic $\mathbb{I}_C(x_C) := \prod_{s \in C} x_s$. Note that $\mathbb{I}_C(x_C) = 1$ if and only if $x_s = 1$ for all $s \in C$, so it is an indicator function for the event $\{x_s = 1, \ \forall s \in C\}$. In the exponential family, this sufficient statistic is weighted by a natural parameter $\theta_C \in \mathbb{R}$, and we rewrite the factorization (2.1) as

$$(2.2) \qquad q_\theta(x_1, \ldots, x_p) = \exp\left\{\sum_{C \in \mathcal{C}} \theta_C \mathbb{I}_C(x_C) - \Phi(\theta)\right\},$$

where $\Phi(\theta) := \log \sum_{x \in \{0,1\}^p} \exp(\sum_{C \in \mathcal{C}} \theta_C \mathbb{I}_C(x_C))$ is the log normalization constant. It may be verified (cf. Proposition 4.3 of Darroch and Speed [11]) that the factorization (2.2) defines a minimal exponential family, that is, the statistics $\{\mathbb{I}_C(x_C), C \in \mathcal{C}\}$ are affinely independent. In the special case of pairwise interactions, equation (2.2) reduces to the classical *Ising model*:

$$(2.3) \qquad q_\theta(x_1, \ldots, x_p) = \exp\left\{\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - \Phi(\theta)\right\}.$$

The model (2.3) is a particular instance of a pairwise Markov random field.

Multinomial variables. In order to generalize the Ising model to nonbinary variables, say, $\mathcal{X} = \{0, 1, \ldots, m-1\}$, we introduce a larger set of sufficient statistics. We first illustrate this extension for a pairwise Markov random field. For each node $s \in V$ and configuration $j \in \mathcal{X}_0 := \mathcal{X} \setminus \{0\} = \{1, 2, \ldots, m-1\}$, we introduce the binary-valued indicator function

$$(2.4) \qquad \mathbb{I}_{s;j}(x_s) = \begin{cases} 1, & \text{if } x_s = j, \\ 0, & \text{otherwise.} \end{cases}$$

We also introduce a vector $\theta_s = \{\theta_{s;j}, j \in \mathcal{X}_0\}$ of natural parameters associated with these sufficient statistics. Similarly, for each edge $(s, t) \in E$ and configuration $(j, k) \in \mathcal{X}_0^2 := \mathcal{X}_0 \times \mathcal{X}_0$, we introduce the binary-valued indicator function $\mathbb{I}_{st;jk}$ for the event $\{x_s = j, x_t = k\}$, as well as the collection $\theta_{st} := \{\theta_{st;jk}, (j, k) \in \mathcal{X}_0^2\}$ of natural parameters. Then any pairwise Markov random field over m-ary random variables may be written in the form

$$(2.5) \quad q_\theta(x_1, \ldots, x_p) = \exp\left\{\sum_{s \in V} \langle \theta_s, \mathbb{I}_s(x_s) \rangle + \sum_{(s,t) \in E} \langle \theta_{st}, \mathbb{I}_{st}(x_s, x_t) \rangle - \Phi(\theta)\right\},$$

where we have used the shorthand $\langle \theta_s, \mathbb{I}_s(x_s) \rangle := \sum_{j=1}^{m-1} \theta_{s;j} \mathbb{I}_{s;j}(x_s)$ and $\langle \theta_{st}, \mathbb{I}_{st}(x_s, x_t) \rangle := \sum_{j,k=1}^{m-1} \theta_{st;jk} \mathbb{I}_{st;jk}(x_s, x_t)$. Equation (2.5) defines a minimal exponential family with dimension $|V|(m-1) + |E|(m-1)^2$ [11]. Note that the family (2.5) is a natural generalization of the Ising model (2.3); in particular, when $m = 2$, we have a single sufficient statistic $\mathbb{I}_{s;1}(x_s) = x_s$ for each vertex and a single sufficient statistic $\mathbb{I}_{st;11}(x_s, x_t) = x_s x_t$ for each edge. (We have omitted the

additional subscript 1 or 11 in our earlier notation for the Ising model, since they are superfluous in that case.)

Finally, for a graphical model involving higher-order interactions, we require additional sufficient statistics. For each clique $C \in \mathcal{C}$, we define the subset of configurations

$$\mathcal{X}_0^{|C|} := \underbrace{\mathcal{X}_0 \times \cdots \times \mathcal{X}_0}_{C \text{ times}} = \{(j_s, s \in C) \in \mathcal{X}^{|C|} : j_s \neq 0 \ \forall s \in C\},$$

a set of cardinality $(m-1)^{|C|}$. As before, $\mathcal{C}$ is the set of all maximal and nonmaximal cliques. For any configuration $J = \{j_s, s \in C\} \in \mathcal{X}_0^{|C|}$, we define the corresponding indicator function

$$(2.6) \qquad \mathbb{I}_{C;J}(x_C) = \begin{cases} 1, & \text{if } x_C = J, \\ 0, & \text{otherwise.} \end{cases}$$

We then consider the general multinomial exponential family

$$q_\theta(x_1, \ldots, x_p) = \exp\left\{\sum_{C \in \mathcal{C}} \langle \theta_C, \mathbb{I}_C \rangle - \Phi(\theta)\right\}$$

$$(2.7) \qquad \qquad \qquad \qquad \text{for } x_s \in \mathcal{X} = \{0, 1, \ldots, m-1\}$$

with $\langle \theta_C, \mathbb{I}_C(x_C) \rangle = \sum_{J \in \mathcal{X}_0^{|C|}} \theta_{C;J} \mathbb{I}_{C;J}(x_C)$. Note that our previous models—namely, the binary models (2.2) and (2.3), as well as the pairwise multinomial model (2.5)—are special cases of this general factorization.

Recall that an exponential family is *minimal* if no nontrivial linear combination of sufficient statistics is almost surely equal to a constant. The family is *regular* if $\{\theta : \Phi(\theta) < \infty\}$ is an open set. As will be relevant later, the exponential families described in this section are all minimal and regular [11].

2.3. *Covariance matrices and beyond.* We now turn to a discussion of the phenomena that motivate the analysis of this paper. Consider the usual covariance matrix $\Sigma = \text{cov}(X_1, \ldots, X_p)$. When X is jointly Gaussian, it is an immediate consequence of the Hammersley–Clifford theorem that the sparsity pattern of the precision matrix $\Gamma = \Sigma^{-1}$ reflects the graph structure—that is, $\Gamma_{st} = 0$ whenever $(s,t) \notin E$. More precisely, Γ_{st} is a scalar multiple of the correlation of X_s and X_t conditioned on $X_{\backslash\{s,t\}}$ (cf. Lauritzen [21]). For non-Gaussian distributions, however, the conditional correlation will be a function of $X_{\backslash\{s,t\}}$, and it is unknown whether the entries of Γ have any relationship with the strengths of correlations along edges in the graph.

Nonetheless, it is tempting to conjecture that inverse covariance matrices might be related to graph structure in the non-Gaussian case. We explore this possibility by considering a simple case of the binary Ising model (2.3).
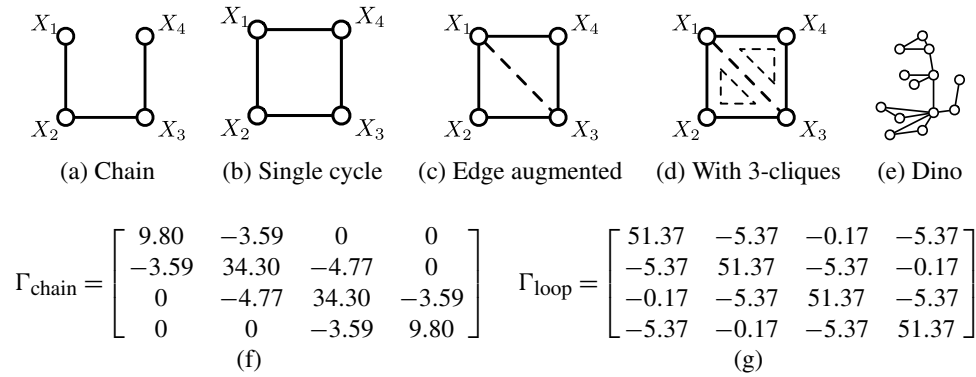
FIG. 2.   (a)–(e) *Different examples of graphical models.* (f) *Inverse covariance for chain graph in* (a). (g) *Inverse covariance for single-cycle graph in* (b).

EXAMPLE 1.   Consider a simple chain graph on four nodes, as illustrated in Figure 2(a). In terms of the factorization (2.3), let the node potentials be $\theta_s = 0.1$ for all $s \in V$ and the edge potentials be $\theta_{st} = 2$ for all $(s, t) \in E$. For a multivariate Gaussian graphical model defined on G, standard theory predicts that the inverse covariance matrix $\Gamma = \Sigma^{-1}$ of the distribution is graph-structured: $\Gamma_{st} = 0$ if and only if $(s, t) \notin E$. Surprisingly, this is also the case for the chain graph with binary variables [see panel (f)]. However, this statement is *not* true for the single-cycle graph shown in panel (b). Indeed, as shown in panel (g), the inverse covariance matrix has no nonzero entries at all. Curiously, for the more complicated graph in (e), we again observe a graph-structured inverse covariance matrix.

Still focusing on the single-cycle graph in panel (b), suppose that instead of considering the ordinary covariance matrix, we compute the covariance matrix of the *augmented* random vector $(X_1, X_2, X_3, X_4, X_1X_3)$, where the extra term X_1X_3 is represented by the dotted edge shown in panel (c). The 5×5 inverse of this generalized covariance matrix takes the form

$$(2.8) \qquad \Gamma_{\mathrm{aug}} = 10^3 \times \begin{bmatrix} 1.15 & -0.02 & 1.09 & -0.02 & -1.14 \\ -0.02 & 0.05 & -0.02 & 0 & 0.01 \\ 1.09 & -0.02 & 1.14 & -0.02 & -1.14 \\ -0.02 & 0 & -0.02 & 0.05 & 0.01 \\ -1.14 & 0.01 & -1.14 & 0.01 & 1.19 \end{bmatrix}.$$

This matrix safely separates nodes 1 and 4, but the entry corresponding to the nonedge $(1, 3)$ is *not* equal to zero. Indeed, we would observe a similar phenomenon if we chose to augment the graph by including the edge $(2, 4)$ rather than $(1, 3)$. This example shows that the usual inverse covariance matrix is not always graph-structured, but inverses of augmented matrices involving higher-order interaction terms may reveal graph structure.

Now let us consider a more general graphical model that adds the 3-clique interaction terms shown in panel (d) to the usual Ising terms. We compute the co-

variance matrix of the augmented vector

$$\Psi(X) = \{X_1, X_2, X_3, X_4, X_1X_2, X_2X_3, X_3X_4,$$
$$X_1X_4, X_1X_3, X_1X_2X_3, X_1X_3X_4\} \in \{0, 1\}^{11}.$$

Empirically, one may show that the 11×11 inverse $(\text{cov}[\Psi(X)])^{-1}$ respects aspects of the graph structure: there are zeros in position (α, β), corresponding to the associated functions $X_\alpha = \prod_{s \in \alpha} X_s$ and $X_\beta = \prod_{s \in \beta} X_\beta$, whenever α and β do not lie within the same maximal clique. [E.g., this applies to the pairs $(\alpha, \beta) = (\{2\}, \{4\})$ and $(\alpha, \beta) = (\{2\}, \{1, 4\})$.]

The goal of this paper is to understand when certain inverse covariances do (and *do not*) capture the structure of a graphical model. At its root is the principle that the augmented inverse covariance matrix $\Gamma = \Sigma^{-1}$, suitably defined, is *always* graph-structured with respect to a graph triangulation. In some cases [e.g., the dino graph in Figure 2(e)], we may leverage the block-matrix inversion formula [16], namely,

$$(2.9) \qquad \Sigma_{A,A}^{-1} = \Gamma_{A,A} - \Gamma_{A,B}\Gamma_{B,B}^{-1}\Gamma_{B,A},$$

to conclude that the inverse of a sub-block of the augmented matrix (e.g., the ordinary covariance matrix) is still graph-structured. This relation holds whenever A and B are chosen in such a way that the second term in equation (2.9) continues to respect the edge structure of the graph. These ideas will be made rigorous in Theorem 1 and its corollaries in the next section.

3. Generalized covariance matrices and graph structure. We now state our main results on the relationship between the zero pattern of generalized (augmented) inverse covariance matrices and graph structure. In Section 4 to follow, we develop some consequences of these results for data-dependent estimators used in structure estimation.

We begin with some notation for defining generalized covariance matrices, stated in terms of the sufficient statistics previously defined (2.6). Recall that a clique $C \in \mathcal{C}$ is associated with the collection $\{\mathbb{I}_{C;J}, J \in \mathcal{X}_0^{|C|}\}$ of binary-valued sufficient statistics. Let $\mathcal{S} \subseteq \mathcal{C}$, and define the random vector

$$(3.1) \qquad \Psi(X; \mathcal{S}) = \{\mathbb{I}_{C;J}, J \in \mathcal{X}_0^{|C|}, C \in \mathcal{S}\},$$

consisting of all the sufficient statistics indexed by elements of $\mathcal{S}$. As in the previous section, the set $\mathcal{C}$ contains both maximal and nonmaximal cliques.

We will often be interested in situations where $\mathcal{S}$ contains all subsets of a given set. For a subset $A \subseteq V$, let $\text{pow}(A)$ denote the collection of all $2^{|A|} - 1$ nonempty subsets of A. We extend this notation to $\mathcal{S}$ by defining

$$\text{pow}(\mathcal{S}) := \bigcup_{C \in \mathcal{S}} \text{pow}(C).$$

3.1. *Triangulation and block structure.* Our first main result concerns a connection between the inverses of generalized inverse covariance matrices associated with the model (2.7) and any triangulation of the underlying graph G. The notion of a triangulation is defined in terms of chordless cycles, which are sequences of distinct vertices $\{s_1, \ldots, s_\ell\}$ such that:

- $(s_i, s_{i+1}) \in E$ for all $1 \le i \le \ell - 1$, and also $(s_\ell, s_1) \in E$;
- no other nodes in the cycle are connected by an edge.

As an illustration, the 4-cycle in Figure 2(b) is a chordless cycle.

DEFINITION 3 (Triangulation). Given an undirected graph $G = (V, E)$, a *triangulation* is an augmented graph $\widetilde{G} = (V, \widetilde{E})$ that contains no chordless cycles of length greater than 3.

Note that a tree is trivially triangulated, since it contains no cycles. On the other hand, the chordless 4-cycle in Figure 2(b) is the simplest example of a nontriangulated graph. By adding the single edge $(1, 3)$ to form the augmented edge set $\widetilde{E} = E \cup \{(1, 3)\}$, we obtain the triangulated graph $\widetilde{G} = (V, \widetilde{E})$ shown in panel (c). One may check that the more complicated graph shown in Figure 2(e) is triangulated as well.

Our first result concerns the inverse Γ of the matrix $\mathrm{cov}(\Psi(X; \widetilde{\mathcal{C}}))$, where $\widetilde{\mathcal{C}}$ is the set of all cliques arising from some triangulation $\widetilde{G}$ of G. For any two subsets $A, B \in \widetilde{\mathcal{C}}$, we write $\Gamma(A, B)$ to denote the sub-block of Γ indexed by all indicator statistics on A and B, respectively. (Note that we are working with respect to the exponential family representation over the triangulated graph $\widetilde{G}$.) Given our previously-defined sufficient statistics (2.6), the sub-block $\Gamma(A, B)$ has dimensions $d_A \times d_B$, where

$$d_A := (m - 1)^{|A|} \quad \text{and} \quad d_B := (m - 1)^{|B|}.$$

For example, when $A = \{s\}$ and $B = \{t\}$, the submatrix $\Gamma(A, B)$ has dimension $(m - 1) \times (m - 1)$. With this notation, we have the following result:

THEOREM 1 (Triangulation and block graph-structure). *Consider an arbitrary discrete graphical model of the form* (2.7), *and let $\widetilde{\mathcal{C}}$ be the set of all cliques in any triangulation of G. Then the generalized covariance matrix $\mathrm{cov}(\Psi(X; \widetilde{\mathcal{C}}))$ is invertible, and its inverse Γ is* block graph-structured:

(a) *For any two subsets $A, B \in \widetilde{\mathcal{C}}$ that are not subsets of the same maximal clique, the block $\Gamma(A, B)$ is identically zero.*

(b) *For almost all parameters θ, the entire block $\Gamma(A, B)$ is nonzero whenever A and B belong to a common maximal clique.*

In part (b), "almost all" refers to all parameters θ apart from a set of Lebesgue measure zero. The proof of Theorem 1, which we provide in Section 3.4, relies on the geometry of exponential families [7, 36] and certain aspects of convex analysis [32], involving the log partition function Φ and its Fenchel–Legendre dual Φ^*. Although we have stated Theorem 1 for discrete variables, it easily generalizes to other classes of random variables. The only difference is the specific choices of sufficient statistics used to define the generalized covariance matrix. This generality becomes apparent in the proof.

To provide intuition for Theorem 1, we consider its consequences for specific graphs. When the original graph is a tree [such as the graph in Figure 2(a)], it is already triangulated, so the set $\widetilde{C}$ is equal to the edge set E, together with singleton nodes. Hence, Theorem 1 implies that the inverse Γ of the matrix of sufficient statistics for vertices and edges is graph-structured, and blocks of nonzeros in Γ correspond to edges in the graph. In particular, we may apply Theorem 1(a) to the subsets $A = \{s\}$ and $B = \{t\}$, where s and t are distinct vertices with $(s, t) \notin E$, and conclude that the $(m - 1) \times (m - 1)$ sub-block $\Gamma(A, B)$ is equal to zero.

When G is not triangulated, however, we may need to invert a larger augmented covariance matrix and include sufficient statistics over pairs $(s, t) \notin E$ as well. For instance, the augmented graph shown in Figure 2(c) is a triangulation of the chordless 4-cycle in panel (b). The associated set of maximal cliques is given by $\overline{C} = \{(1, 2), (2, 3), (3, 4), (1, 4), (1, 3)\}$; among other predictions, our theory guarantees that the generalized inverse covariance Γ will have zeros in the sub-block $\Gamma(\{2\}, \{4\})$.

3.2. *Separator sets and graph structure.*   In fact, it is not necessary to take sufficient statistics over *all* maximal cliques, and we may consider a slightly smaller augmented covariance matrix. (This simpler type of augmented covariance matrix explains the calculations given in Section 2.3.)

By classical graph theory, any triangulation $\widetilde{G}$ gives rise to a *junction tree* representation of G. Nodes in the junction tree are subsets of V corresponding to maximal cliques of $\widetilde{G}$, and the intersection of any two adjacent cliques C_1 and C_2 is referred to as a *separator set* $S = C_1 \cap C_2$. Furthermore, any junction tree must satisfy the *running intersection property*, meaning that for any two nodes of the junction tree—say, corresponding to cliques C and D—the intersection $C \cap D$ must belong to every separator set on the unique path between C and D. The following result shows that it suffices to construct generalized covariance matrices augmented by separator sets:

COROLLARY 1.   *Let $\mathcal{S}$ be the set of separator sets in any triangulation of G, and let Γ be the inverse of $\mathrm{cov}(\Psi(X; V \cup \mathrm{pow}(\mathcal{S})))$. Then $\Gamma(\{s\}, \{t\}) = 0$ whenever $(s, t) \notin \widetilde{E}$.*

Note that $V \cup \mathrm{pow}(\mathcal{S}) \subseteq \widetilde{\mathcal{C}}$, and the set of sufficient statistics considered in Corollary 1 is generally much smaller than the set of sufficient statistics considered in Theorem 1. Hence, the generalized covariance matrix of Corollary 1 has a smaller dimension than the generalized covariance matrix of Theorem 1, which becomes significant when we consider exploiting these population-level results for statistical estimation.

The graph in Figure 2(c) of Example 1 and the associated matrix in equation (2.8) provide a concrete example of Corollary 1 in action. In this case, the single separator set in the triangulation is $\{1, 3\}$, so when $\mathcal{X} = \{0, 1\}$, augmenting the usual covariance matrix with the additional sufficient statistic $\mathbb{I}_{13;11}(x_1, x_3) = x_1 x_3$ and taking the inverse yields a graph-structured matrix. Indeed, since $(2, 4) \notin \widetilde{E}$, we observe that $\Gamma_{\mathrm{aug}}(2, 4) = 0$ in equation (2.8), consistent with the result of Corollary 1.

Although Theorem 1 and Corollary 1 are clean population-level results, however, forming an appropriate augmented covariance matrix requires prior knowledge of the graph, namely, which edges are involved in a suitable triangulation. This is infeasible in settings where the goal is to recover the edge structure of the graph. Corollary 1 is most useful for edge recovery when G admits a triangulation with only singleton separator sets, since then $V \cup \mathrm{pow}(\mathcal{S}) = V$. In particular, this condition holds when G is a tree. The following corollary summarizes our result:

COROLLARY 2. *For any graph with singleton separator sets, the inverse Γ of the covariance matrix* $\mathrm{cov}(\Psi(X; V))$ *of vertex statistics is graph-structured.* (*This class includes trees as a special case.*)

In the special case of binary variables, we have $\Psi(X; V) = (X_1, \ldots, X_p)$, so Corollary 2 implies that the inverse of the ordinary covariance matrix $\mathrm{cov}(X)$ is graph-structured. For m-ary variables, $\mathrm{cov}(\Psi(X; V))$ is a matrix of dimensions $(m - 1)p \times (m - 1)p$ involving indicator functions for each variable. Again, we may relate this corollary to Example 1—the inverse covariance matrices for the tree graph in panel (a) and the dino graph in panel (e) are exactly graph-structured. Although the dino graph is not a tree, it possesses the nice property that the only separator sets in its junction tree are singletons.

Corollary 1 also guarantees that inverse covariances may be partially graph-structured, in the sense that $\Gamma(\{s\}, \{t\}) = 0$ for any pair of vertices (s, t) separable by a singleton separator set, where $\Gamma = (\mathrm{cov}(\Psi(X; V)))^{-1}$. This is because for any such pair (s, t), we may form a junction tree with two nodes, one containing s and one containing t, and apply Corollary 1. Indeed, the matrix Γ defined over singleton vertices is agnostic to which triangulation we choose for the graph.

In settings where there exists a junction tree representation of the graph with only singleton separator sets, Corollary 2 has a number of useful implications for the consistency of methods that have traditionally only been applied for edge recovery in Gaussian graphical models: for tree-structured discrete graphs, it suffices

to estimate the support of $(\text{cov}(\Psi(X; V)))^{-1}$ from the data. We will review methods for Gaussian graphical model selection and describe their analogs for discrete tree graphs in Sections 4.1 and 4.2.

3.3. *Generalized covariances and neighborhood structure.* Theorem 1 also has a corollary, that is, relevant for nodewise neighborhood selection approaches to graph selection [27, 31], which are applicable to graphs with arbitrary topologies. Nodewise methods use the basic observation that recovering the edge structure of G is equivalent to recovering the neighborhood set $N(s) = \{t \in V : (s, t) \in E\}$ for each vertex $s \in V$. For a given node $s \in V$ and positive integer d, consider the collection of subsets

$$\mathcal{S}(s; d) := \{U \subseteq V \setminus \{s\}, |U| = d\}.$$

The following corollary provides an avenue for recovering $N(s)$ based on the inverse of a certain generalized covariance matrix:

COROLLARY 3 (Neighborhood selection). *For any graph and node $s \in V$ with* $\deg(s) \leq d$, *the inverse Γ of the matrix* $\text{cov}(\Psi(X; \{s\} \cup \text{pow}(\mathcal{S}(s; d))))$ *is s-block graph-structured, that is,* $\Gamma(\{s\}, B) = 0$ *whenever* $\{s\} \neq B \subsetneq N(s)$. *In particular,* $\Gamma(\{s\}, \{t\}) = 0$ *for all vertices* $t \notin N(s)$.

Note that $\text{pow}(\mathcal{S}(s; d))$ is the set of subsets of all candidate neighborhoods of s of size d. This result follows from Theorem 1 (and the related Corollary 1) by constructing a particular junction tree for the graph, in which s is separated from the rest of the graph by $N(s)$. Due to the well-known relationship between the rows of an inverse covariance matrix and linear regression coefficients [27], Corollary 3 motivates the following neighborhood-based approach to graph selection: for a fixed vertex $s \in V$, perform a single *linear regression* of $\Psi(X; \{s\})$ on the vector $\Psi(X; \text{pow}(\mathcal{S}(s; d)))$. Via elementary algebra and an application of Corollary 3, the resulting regression vector will expose the neighborhood $N(s)$ in an arbitrary discrete graphical model; that is, the indicators $\Psi(X; \{t\})$ corresponding to X_t will have a nonzero weight only if $t \in N(s)$. We elaborate on this connection in Section 4.2.

3.4. *Proof of Theorem 1.* We now turn to the proof of Theorem 1, which is based on certain fundamental correspondences arising from the theory of exponential families [4, 7, 36]. Recall that our exponential family (2.7) has binary-valued indicator functions (2.6) as its sufficient statistics. Let D denote the cardinality of this set and let $\mathbb{I} : \mathcal{X}^p \to \{0, 1\}^D$ denote the multivariate function that maps each configuration $x \in \mathcal{X}^p$ to the vector $\mathbb{I}(x)$ obtained by evaluating the D indicator functions on x. Using this notation, our exponential family may be written in the compact form $q_\theta(x) = \exp\{\langle \theta, \mathbb{I}(x) \rangle - \Phi(\theta)\}$, where

$$\langle \theta, \mathbb{I}(x) \rangle = \sum_{C \in \mathcal{C}} \langle \theta_C, \mathbb{I}_C(x) \rangle = \sum_{C \in \mathcal{C}} \sum_{J \in \mathcal{X}_0^{|C|}} \theta_{C;J} \mathbb{I}_{C;J}(x_C).$$

Since this exponential family is known to be minimal, we are guaranteed [11] that

$$\nabla\Phi(\theta) = \mathbb{E}_\theta[\mathbb{I}(X)] \quad \text{and} \quad \nabla^2\Phi(\theta) = \text{cov}_\theta[\mathbb{I}(X)],$$

where $\mathbb{E}_\theta$ and cov_θ denote (resp.) the expectation and covariance taken under the density q_θ [7, 36]. The conjugate dual [32] of the cumulant function is given by

$$\Phi^*(\mu) := \sup_{\theta \in \mathbb{R}^D} \{\langle \mu, \theta \rangle - \Phi(\theta)\}.$$

The function Φ^* is always convex and takes values in $\mathbb{R} \cup \{+\infty\}$. From known results [36], the dual function Φ^* is finite only for $\mu \in \mathbb{R}^D$ belonging to the marginal polytope

$$(3.2) \qquad \mathcal{M} := \left\{\mu \in \mathbb{R}^p \;\middle|\; \exists \text{ some density } q \text{ s.t. } \sum_x q(x)\mathbb{I}(x) = \mu\right\}.$$

The following lemma, proved in the supplementary material [26], provides a connection between the covariance matrix and the Hessian of Φ^*:

LEMMA 1. *Consider a regular, minimal exponential family, and define* $\mu = \mathbb{E}_\theta[\mathbb{I}(X)]$ *for any fixed* $\theta \in \Omega = \{\theta : \Phi(\theta) < \infty\}$. *Then*

$$(3.3) \qquad \qquad (\text{cov}_\theta[\mathbb{I}(X)])^{-1} = \nabla^2\Phi^*(\mu).$$

Note that the minimality and regularity of the family implies that $\text{cov}_\theta[\mathbb{I}(X)]$ is strictly positive definite, so the matrix is invertible.

For any $\mu \in \text{int}(\mathcal{M})$, let $\theta(\mu) \in \mathbb{R}^D$ denote the unique natural parameter θ such that $\nabla\Phi(\theta) = \mu$. It is known [36] that the negative dual function $-\Phi^*$ is linked to the Shannon entropy via the relation

$$(3.4) \qquad -\Phi^*(\mu) = H(q_{\theta(\mu)}(x)) = -\sum_{x \in \mathcal{X}^p} q_{\theta(\mu)}(x) \log q_{\theta(\mu)}(x).$$

In general, expression (3.4) does *not* provide a straightforward way to compute $\nabla^2\Phi^*$, since the mapping $\mu \mapsto \theta(\mu)$ may be extremely complicated. However, when the exponential family is defined with respect to a triangulated graph, Φ^* has an explicit closed-form representation in terms of the mean parameters μ. Consider a junction tree triangulation of the graph, and let $(\overline{\mathcal{C}}, \mathcal{S})$ be the collection of maximal cliques and separator sets, respectively. By the junction tree theorem [20, 22, 36], we have the factorization

$$(3.5) \qquad q(x_1, \ldots, x_p) = \frac{\prod_{C \in \overline{\mathcal{C}}} q_C(x_C)}{\prod_{S \in \mathcal{S}} q_S(x_S)},$$

where q_C and q_S are the marginal distributions over maximal clique C and separator set S. Consequently, the entropy may be decomposed into the sum

$$(3.6) \qquad H(q) = -\sum_{x \in \mathcal{X}^p} q(x) \log q(x) = \sum_{C \in \overline{\mathcal{C}}} H_C(q_C) - \sum_{S \in \mathcal{S}} H_S(q_S),$$

where we have introduced the clique- and separator-based entropies

$$H_S(q_S) := - \sum_{x_S \in \mathcal{X}^{|S|}} q_S(x_S) \log q_S(x_S)$$

and

$$H_C(q_C) := - \sum_{x_C \in \mathcal{X}^{|C|}} q_C(x_C) \log q_C(x_C).$$

Given our choice of sufficient statistics (2.6), we show that q_C and q_S may be written explicitly as "local" functions of mean parameters associated with C and S. For each subset $A \subseteq V$, let $\mu_A \in (m-1)^{|A|}$ be the associated collection of mean parameters, and let

$$\mu_{\text{pow}(A)} := \{\mu_B \mid \varnothing \neq B \subseteq A\}$$

be the set of mean parameters associated with all nonempty subsets of A. Note that $\mu_{\text{pow}(A)}$ contains a total of $\sum_{k=1}^{|A|} \binom{|A|}{k}(m-1)^k = m^{|A|} - 1$ parameters, corresponding to the number of degrees of freedom involved in specifying a marginal distribution over the random vector x_A. Moreover, $\mu_{\text{pow}(A)}$ uniquely determines the marginal distribution q_A:

LEMMA 2. *For any marginal distribution q_A in the $m^{|A|}$-dimensional probability simplex, there is a unique mean parameter vector $\mu_{\text{pow}(A)}$ and matrix M_A such that $q_A = M_A \cdot \mu_{\text{pow}(A)}$.*

For the proof, see the supplementary material [26].

We now combine the dual representation (3.4) with the decomposition (3.6), along with the matrices $\{M_C, M_S\}$ from Lemma 2, to conclude that

$$(3.7) \qquad -\Phi^*(\mu) = \sum_{C \in \overline{\mathcal{C}}} H_C(M_C(\mu_{\text{pow}(C)})) - \sum_{S \in \mathcal{S}} H_S(M_S(\mu_{\text{pow}(S)})).$$

Now consider two subsets $A, B \in \widetilde{\mathcal{C}}$ that are not contained in the same maximal clique. Suppose A is contained within maximal clique C. Differentiating expression (3.7) with respect to μ_A preserves only terms involving q_C and q_S, where S is any separator set such that $A \subseteq S \subseteq C$. Since $B \subsetneq C$, we clearly cannot have $B \subseteq S$. Consequently, all cross-terms arising from the clique C and its associated separator sets vanish when we take a second derivative with respect to μ_B. Repeating this argument for any other maximal clique C' containing A but not B, we have $\frac{\partial^2 \Phi^*}{\partial \mu_A \partial \mu_B}(\mu) = 0$. This proves part (a).

Turning to part (b), note that if A and B are in the same maximal clique, the expression obtained by taking second derivatives of the entropy results in an algebraic expression with only finitely many solutions in the parameters μ (consequently, also θ). Hence, assuming the θ's are drawn from a continuous distribution, the corresponding values of the block $\Gamma(A, B)$ are a.s. nonzero.

4. Consequences for graph structure estimation. Moving beyond the population level, we now state and prove several results concerning the statistical consistency of different methods—both known and some novel—for graph selection in discrete graphical models, based on i.i.d. draws from a discrete graph. For sparse Gaussian models, existing methods that exploit sparsity of the inverse covariance matrix fall into two main categories: global graph selection methods (e.g., [12, 14, 31, 33]) and local (nodewise) neighborhood selection methods [27, 40]. We divide our discussion accordingly.

4.1. *Graphical Lasso for singleton separator graphs.* We begin by describing how a combination of our population-level results and some concentration inequalities may be leveraged to analyze the statistical behavior of log-determinant methods for discrete graphical models with singleton separator sets, and suggest extensions of these methods when observations are systematically corrupted by noise or missing data. Given a p-dimensional random vector $(X_1, \ldots, X_p)$ with covariance Σ^*, consider the estimator

$$(4.1) \qquad \widehat{\Theta} \in \arg\min_{\Theta \geq 0} \left\{ \text{trace}(\widehat{\Sigma}\Theta) - \log\det(\Theta) + \lambda_n \sum_{s \neq t} |\Theta_{st}| \right\},$$

where $\widehat{\Sigma}$ is an estimator for Σ^*. For multivariate Gaussian data, this program is an ℓ_1-regularized maximum likelihood estimate known as the *graphical Lasso* and is a well-studied method for recovering the edge structure in a Gaussian graphical model [3, 14, 33, 39]. Although the program (4.1) has no relation to the MLE in the case of a discrete graphical model, it may still be useful for estimating $\Theta^* := (\Sigma^*)^{-1}$. Indeed, as shown in Ravikumar et al. [31], existing analyses of the estimator (4.1) require only tail conditions such as sub-Gaussianity in order to guarantee that the sample minimizer is close to the population minimizer. The analysis of this paper completes the missing link by guaranteeing that the population-level inverse covariance is in fact graph-structured. Consequently, we obtain the interesting result that the program (4.1)—even though it is ostensibly derived from Gaussian considerations—is a consistent method for recovering the structure of any binary graphical model with singleton separator sets.

In order to state our conclusion precisely, we introduce additional notation. Consider a general estimate $\widehat{\Sigma}$ of the covariance matrix Σ such that

$$(4.2) \qquad \mathbb{P}\left[\|\widehat{\Sigma} - \Sigma^*\|_{\max} \geq \varphi(\Sigma^*)\sqrt{\frac{\log p}{n}} \right] \leq c\exp(-\psi(n, p))$$

for functions φ and ψ, where $\|\cdot\|_{\max}$ denotes the elementwise ℓ_∞-norm. In the case of fully-observed i.i.d. data with sub-Gaussian parameter σ^2, where $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T - \bar{x}\bar{x}^T$ is the usual sample covariance, this bound holds with $\varphi(\Sigma^*) = \sigma^2$ and $\psi(n, p) = c'\log p$.

As in past analysis of the graphical Lasso [31], we require a certain *mutual incoherence* condition on the true covariance matrix Σ^* to control the correlation of nonedge variables with edge variables in the graph. Let $\Gamma^* = \Sigma^* \otimes \Sigma^*$, where $\otimes$ denotes the Kronecker product. Then Γ^* is a $p^2 \times p^2$ matrix indexed by vertex pairs. The incoherence condition is given by

$$(4.3) \qquad \max_{e \in S^c} \|\Gamma^*_{eS}(\Gamma^*_{SS})^{-1}\|_1 \le 1 - \alpha, \qquad \alpha \in (0, 1],$$

where $S := \{(s, t): \Theta^*_{st} \ne 0\}$ is the set of vertex pairs corresponding to nonzero entries of the precision matrix Θ^*, equivalently, the edge set of the graph, by our theory on tree-structured discrete graphs. For more intuition on the mutual incoherence condition, see Ravikumar et al. [31].

With this notation, our global edge recovery algorithm proceeds as follows:

ALGORITHM 1 (Graphical Lasso).

1. Form a suitable estimate $\widehat{\Sigma}$ of the true covariance matrix Σ.
2. Optimize the graphical Lasso program (4.1) with parameter λ_n, and denote the solution by $\widehat{\Theta}$.
3. Threshold the entries of $\widehat{\Theta}$ at level τ_n to obtain an estimate of Θ^*.

It remains to choose the parameters (λ_n, τ_n). In the following corollary, we will establish statistical consistency of $\widehat{\Theta}$ under the following settings:

$$(4.4) \qquad \lambda_n \ge \frac{c_1}{\alpha}\sqrt{\frac{\log p}{n}}, \qquad \tau_n = c_2\left\{\frac{c_1}{\alpha}\sqrt{\frac{\log p}{n}} + \lambda_n\right\},$$

where α is the incoherence parameter in inequality (4.3) and c_1, c_2 are universal positive constants. The following result applies to Algorithm 1 when $\widehat{\Sigma}$ is the sample covariance matrix and (λ_n, τ_n) are chosen as in equations (4.4):

COROLLARY 4. *Consider an Ising model* (2.3) *defined by an undirected graph with singleton separator sets and with degree at most d, and suppose that the mutual incoherence condition* (4.3) *holds. With $n \gtrsim d^2 \log p$ samples, there are universal constants (c, c') such that with probability at least $1 - c \exp(-c' \log p)$, Algorithm 1 recovers all edges (s, t) with $|\Theta^*_{st}| > \tau/2$.*

The proof is contained in the supplementary material [26]; it is a relatively straightforward consequence of Corollary 1 and known concentration properties of $\widehat{\Sigma}$ as an estimate of the population covariance matrix. Hence, if $|\Theta^*_{st}| > \tau/2$ for all edges $(s, t) \in E$, Corollary 4 guarantees that the log-determinant method plus thresholding recovers the full graph exactly.

In the case of the standard sample covariance matrix, a variant of the graphical Lasso has been implemented by Banerjee, El Ghaoui and d'Aspremont [3]. Our

analysis establishes consistency of the graphical Lasso for Ising models on single separator graphs using $n \gtrsim d^2 \log p$ samples. This lower bound on the sample size is unavoidable, as shown by information-theoretic analysis [35], and also appears in other past work on Ising models [2, 19, 30]. Our analysis also has a *cautionary message*: the proof of Corollary 4 relies heavily on the population-level result in Corollary 2, which ensures that Θ^* is graph-structured when G has only singleton separators. For a general graph, we have no guarantees that Θ^* will be graph-structured [e.g., see panel (b) in Figure 2], so the graphical Lasso (4.1) is *inconsistent in general*.

On the positive side, if we restrict ourselves to tree-structured graphs, the estimator (4.1) is attractive, since it relies only on an estimate $\widehat{\Sigma}$ of the population covariance Σ^* that satisfies the deviation condition (4.2). In particular, even when the samples $\{x_i\}_{i=1}^n$ are contaminated by noise or missing data, we may form a good estimate $\widehat{\Sigma}$ of Σ^*. Furthermore, the program (4.1) is always convex regardless of whether $\widehat{\Sigma}$ is positive semidefinite.

As a concrete example of how we may correct the program (4.1) to handle corrupted data, consider the case when each entry of x_i is missing independently with probability ρ, and the corresponding observations z_i are zero-filled for missing entries. A natural estimator is

$$(4.5) \qquad \widehat{\Sigma} = \left(\frac{1}{n} \sum_{i=1}^n z_i z_i^T \right) \div M - \frac{1}{(1-\rho)^2} \bar{z}\bar{z}^T,$$

where $\div$ denotes elementwise division by the matrix M with diagonal entries $(1 - \rho)$ and off-diagonal entries $(1 - \rho)^2$, correcting for the bias in both the mean and second moment terms. The deviation condition (4.2) may be shown to hold w.h.p., where $\varphi(\Sigma^*)$ scales with $(1 - \rho)$ (cf. Loh and Wainwright [25]). Similarly, we may derive an appropriate estimator $\widehat{\Sigma}$ for other forms of additive or multiplicative corruption.

Generalizing to the case of m-ary discrete graphical models with $m > 2$, we may easily modify the program (4.1) by replacing the elementwise ℓ_1-penalty by the corresponding group ℓ_1-penalty, where the groups are the indicator variables for a given vertex. Precise theoretical guarantees follow from results on the group graphical Lasso [18].

4.2. *Consequences for nodewise regression in trees.* Turning to local neighborhood selection methods, recall the neighborhood-based method due to Meinshausen and Bühlmann [27]. In a Gaussian graphical model, the column corresponding to node s in the inverse covariance matrix $\Gamma = \Sigma^{-1}$ is a scalar multiple of $\tilde{\beta} = \Sigma_{\backslash s, \backslash s}^{-1} \Sigma_{\backslash s, s}$, the limit of the linear regression vector for X_s upon $X_{\backslash s}$. Based on n i.i.d. samples from a p-dimensional multivariate Gaussian distribution, the support of the graph may then be estimated consistently under the usual Lasso scaling $n \gtrsim d \log p$, where $d = |N(s)|$.

Motivated by our population-level results on the graph structure of the inverse covariance matrix (Corollary 2), we now propose a method for neighborhood selection in a tree-structured graph. Although the method works for arbitrary m-ary trees, we state explicit results only in the case of the binary Ising model to avoid cluttering our presentation.

The method is based on the following steps. For each node $s \in V$, we first perform ℓ_1-regularized linear regression of X_s against $X_{\backslash s}$ by solving the modified Lasso program

$$(4.6) \qquad \hat{\beta} \in \arg \min_{\|\beta\|_1 \leq b_0 \sqrt{k}} \left\{ \frac{1}{2} \beta^T \widehat{\Gamma} \beta - \hat{\gamma}^T \beta + \lambda_n \|\beta\|_1 \right\},$$

where $b_0 > \|\tilde{\beta}\|_1$ is a constant, $(\widehat{\Gamma}, \hat{\gamma})$ are suitable estimators for $(\Sigma_{\backslash s, \backslash s}, \Sigma_{\backslash s, s})$, and λ_n is an appropriate parameter. We then combine the neighborhood estimates over all nodes via an AND operation [edge (s, t) is present if both s and t are inferred to be neighbors of each other] or an OR operation (at least one of s or t is inferred to be a neighbor of the other).

Note that the program (4.6) differs from the standard Lasso in the form of the ℓ_1-constraint. Indeed, the normal setting of the Lasso assumes a linear model where the predictor and response variables are linked by independent sub-Gaussian noise, but this is not the case for X_s and $X_{\backslash s}$ in a discrete graphical model. Furthermore, the generality of the program (4.6) allows it to be easily modified to handle corrupted variables via an appropriate choice of $(\widehat{\Gamma}, \hat{\gamma})$, as in Loh and Wainwright [25].

The following algorithm summarizes our nodewise regression procedure for recovering the neighborhood set $N(s)$ of a given node s:

ALGORITHM 2 (Nodewise method for trees).

1. Form a suitable pair of estimators $(\widehat{\Gamma}, \hat{\gamma})$ for covariance submatrices $(\Sigma_{\backslash s, \backslash s}, \Sigma_{\backslash s, s})$.
2. Optimize the modified Lasso program (4.6) with parameter λ_n, and denote the solution by $\hat{\beta}$.
3. Threshold the entries of $\hat{\beta}$ at level τ_n, and define the estimated neighborhood set $\widehat{N(s)}$ as the support of the thresholded vector.

In the case of fully observed i.i.d. observations, we choose $(\widehat{\Gamma}, \hat{\gamma})$ to be the recentered estimators

$$(4.7) \qquad (\widehat{\Gamma}, \hat{\gamma}) = \left(\frac{X_{\backslash s}^T X_{\backslash s}}{n} - \bar{x}_{\backslash s} \bar{x}_{\backslash s}^T, \frac{X_{\backslash s}^T X_s}{n} - \bar{x}_s \bar{x}_{\backslash s} \right)$$

and assign the parameters (λ_n, τ_n) according to the scaling

$$(4.8) \qquad \lambda_n \gtrsim \varphi \|\tilde{\beta}\|_2 \sqrt{\frac{\log p}{n}}, \qquad \tau_n \asymp \varphi \|\tilde{\beta}\|_2 \sqrt{\frac{\log p}{n}},$$

where $\tilde{\beta} := \Sigma_{\backslash s, \backslash s}^{-1} \Sigma_{\backslash s, s}$ and φ is some parameter such that $\langle x_i, u \rangle$ is sub-Gaussian with parameter $\varphi^2 \|u\|_2^2$ for any d-sparse vector u, and φ is independent of u. The following result applies to Algorithm 2 using the pairs $(\widehat{\Gamma}, \hat{\gamma})$ and (λ_n, τ_n) defined as in equations (4.7) and (4.8), respectively.

PROPOSITION 1. *Suppose we have i.i.d. observations $\{x_i\}_{i=1}^n$ from an Ising model and that $n \gtrsim \varphi^2 \max\{\frac{1}{\lambda_{\min}(\Sigma_x)}, \|\|\Sigma_x^{-1}\|\|_\infty^2\}d^2 \log p$. Then there are universal constants (c, c', c'') such that with probability greater than $1 - c \exp(-c' \log p)$, for any node $s \in V$, Algorithm 2 recovers all neighbors $t \in N(s)$ for which $|\tilde{\beta}_t| \geq c'' \varphi \|\tilde{\beta}\|_2 \sqrt{\frac{\log p}{n}}$.*

We prove this proposition in the supplementary material [26], as a corollary of a more general theorem on the ℓ_∞-consistency of the program (4.6) for estimating $\tilde{\beta}$, allowing for corrupted observations. The theorem builds upon the analysis of Loh and Wainwright [25], introducing techniques for ℓ_∞-bounds and departing from the framework of a linear model with independent sub-Gaussian noise.

REMARKS. Regarding the sub-Gaussian parameter φ appearing in Proposition 1, note that we may always take $\varphi = \sqrt{d}$, since $|x_i^T u| \leq \|u\|_1 \leq \sqrt{d}\|u\|_2$ when u is d-sparse and x_i is a binary vector. This leads to a sample complexity requirement of $n \gtrsim d^3 \log p$. We suspect that a tighter analysis, possibly combined with assumptions about the correlation decay of the graph, would reduce the sample complexity to the scaling $n \gtrsim d^2 \log p$, as required by other methods with fully observed data [2, 19, 30]. See the simulations in Section 4.4 for further discussion.

For corrupted observations, the strength and type of corruption enters into the factors (φ_1, φ_2) appearing in the deviation bounds (C.2a) and (C.2b) below, and Proposition 1 has natural extensions to the corrupted case. We emphasize that although analogs of Proposition 1 exist for other methods of graph selection based on logistic regression and/or mutual information, the theoretical analysis of those methods does not handle corrupted data, whereas our results extend easily with the appropriate scaling.

In the case of m-ary tree-structured graphical models with $m > 2$, we may perform multivariate regression with the multivariate group Lasso [29] for neighborhood selection, where groups are defined (as in the log-determinant method) as sets of indicators for each node. The general relationship between the best linear predictor and the block structure of the inverse covariance matrix follows from block matrix inversion, and from a population-level perspective, it suffices to perform multivariate linear regression of all indicators corresponding to a given node against all indicators corresponding to other nodes in the graph. The resulting vector of regression coefficients has nonzero blocks corresponding to edges in the graph. We may also combine these ideas with the group Lasso for multivariate regression [29] to reduce the complexity of the algorithm.

4.3. *Consequences for nodewise regression in general graphs.* Moving on from tree-structured graphical models, our method suggests a graph recovery method based on nodewise linear regression for general discrete graphs. Note that by Corollary 3, the inverse of $\mathrm{cov}(\Psi(X; \mathrm{pow}(\mathcal{S}(s; d))))$ is s-block graph-structured, where d is such that $|N(s)| \leq d$. It suffices to perform a single multivariate regression of the indicators $\Psi(X; \{s\})$ corresponding to node s upon the other indicators in $\Psi(X; V \cup \mathrm{pow}(\mathcal{S}(s; d)))$.

We again make precise statements for the binary Ising model ($m = 2$). In this case, the indicators $\Psi(X; \mathrm{pow}(U))$ corresponding to a subset of vertices U of size d' are all $2^{d'} - 1$ distinct products of variables X_u, for $u \in U$. Hence, to recover the d neighbors of node s, we use the following algorithm. Note that knowledge of an upper bound d is necessary for applying the algorithm.

ALGORITHM 3 (Nodewise method for general graphs).

1. Use the modified Lasso program (4.6) with a suitable choice of $(\widehat{\Gamma}, \hat{\gamma})$ and regularization parameter λ_n to perform a linear regression of X_s upon all products of subsets of variables of $X_{\backslash s}$ of size at most d. Denote the solution by $\hat{\beta}$.
2. Threshold the entries of $\hat{\beta}$ at level τ_n, and define the estimated neighborhood set $\widehat{N(s)}$ as the support of the thresholded vector.

Our theory states that at the population level, nonzeros in the regression vector correspond exactly to subsets of $N(s)$. Hence, the statistical consistency result of Proposition 1 carries over with minor modifications. Since Algorithm 3 is essentially a version of Algorithm 4 with the first two steps omitted, we refer the reader to the statement and proof of Corollary 5 below for precise mathematical statements. Note here that since the regression vector has $\mathcal{O}(p^d)$ components, $2^d - 1$ of which are nonzero, the sample complexity of Lasso regression in step (1) of Algorithm 3 is $\mathcal{O}(2^d \log(p^d)) = \mathcal{O}(2^d \log p)$.

For graphs exhibiting correlation decay [6], we may reduce the computational complexity of the nodewise selection algorithm by prescreening the nodes of $V \setminus s$ before performing a Lasso-based linear regression. We define the nodewise correlation according to

$$r_C(s, t) := \sum_{x_s, x_t} \left| \mathbb{P}(X_s = x_s, X_t = x_t) - \mathbb{P}(X_s = x_s)\mathbb{P}(X_t = x_t) \right|$$

and say that the graph exhibits *correlation decay* if there exist constants $\zeta, \kappa > 0$ such that

$$(4.9) \qquad r_C(s, t) > \kappa \qquad \forall (s, t) \in E \quad \text{and} \quad r_C(s, t) \leq \exp\left(-\zeta r(s, t)\right)$$

for all $(s, t) \in V \times V$, where $r(s, t)$ is the length of the shortest path between s and t. With this notation, we then have the following algorithm for neighborhood recovery of a fixed node s in a graph with correlation decay:

ALGORITHM 4 (Nodewise method with correlation decay).

1. Compute the empirical correlations

$$\hat{r}_C(s, t) := \sum_{x_s, x_t} |\widehat{\mathbb{P}}(X_s = x_s, X_t = x_t) - \widehat{\mathbb{P}}(X_s = x_s)\widehat{\mathbb{P}}(X_t = x_t)|$$

between s and all other nodes $t \in V$, where $\widehat{\mathbb{P}}$ denotes the empirical distribution.

2. Let $C := \{t \in V : \hat{r}_C(s, t) > \kappa/2\}$ be the candidate set of nodes with sufficiently high correlation. (Note that C is a function of both s and κ and, by convention, $s \notin C$.)

3. Use the modified Lasso program (4.6) with parameter λ_n to perform a linear regression of X_s against $C_d := \Psi(X; V \cup \text{pow}(C(s; d))) \setminus \{X_s\}$, the set of all products of subsets of variables $\{X_c : c \in C\}$ of size at most d, together with singleton variables. Denote the solution by $\hat{\beta}$.

4. Threshold the entries of $\hat{\beta}$ at level τ_n, and define the estimated neighborhood set $\widehat{N(s)}$ as the support of the thresholded vector.

Note that Algorithm 3 is a version of Algorithm 4 with $C = V \setminus s$, indicating the absence of a prescreening step. Hence, the statistical consistency result below applies easily to Algorithm 3 for graphs with no correlation decay.

For fully observed i.i.d. observations, we choose $(\widehat{\Gamma}, \hat{\gamma})$ according to

$$(4.10) \qquad (\widehat{\Gamma}, \hat{\gamma}) = \left(\frac{X_C^T X_C}{n} - \bar{x}_C \bar{x}_C^T, \frac{X_C^T X_s}{n} - \bar{x}_s \bar{x}_C \right)$$

and parameters (λ_n, τ_n) as follows: for a candidate set C, let $x_{C,i} \in \{0, 1\}^{|C_d|}$ denote the augmented vector corresponding to the observation x_i, and define $\Sigma_C := \text{Cov}(x_{C,i}, x_{C,i})$. Let $\tilde{\beta} := \Sigma_C^{-1} \text{Cov}(x_{C,i}, x_{s,i})$. Then set

$$(4.11) \qquad \lambda_n \gtrsim \varphi \|\tilde{\beta}\|_2 \sqrt{\frac{\log |C_d|}{n}}, \qquad \tau_n \asymp \varphi \|\tilde{\beta}\|_2 \sqrt{\frac{\log |C_d|}{n}},$$

where φ is some function such that $\langle x_{C,i}, u \rangle$ is sub-Gaussian with parameter $\varphi^2 \|u\|_2^2$ for any $(2^d - 1)$-sparse vector u, and φ does not depend on u. We have the following consistency result, the analog of Proposition 1 for the augmented set of vectors. It applies to Algorithm 4 with the pairs $(\widehat{\Gamma}, \hat{\gamma})$ and (λ_n, τ_n) chosen as in equations (4.10) and (4.11).

COROLLARY 5. *Consider i.i.d. observations $\{x_i\}_{i=1}^n$ generated from an Ising model satisfying the correlation decay condition* (4.9), *and suppose*

$$(4.12) \qquad n \gtrsim \left(\kappa^2 + \varphi^2 \max\left\{ \frac{1}{\lambda_{\min}(\Sigma_C)}, \|\|\Sigma_C^{-1}\|\|_\infty^2 \right\} 2^{2d} \right) \log |C_d|.$$

Then there are universal constants (c, c', c'') such that with probability at least $1 - c \exp(-c' \log p)$, and for any $s \in V$:

(i) *The set $\mathcal{C}$ from step* (2) *of Algorithm* 4 *satisfies* $|\mathcal{C}| \leq d^{(\log(4/\kappa))/\zeta}$.
(ii) *Algorithm* 4 *recovers all neighbors* $t \in N(s)$ *such that*

$$|\tilde{\beta}_t| \geq c'' \varphi \|\tilde{\beta}\|_2 \sqrt{\frac{\log |\mathcal{C}_d|}{n}}.$$

The proof of Corollary 5 is contained in the supplementary material [26]. Due to the exponential factor 2^d appearing in the lower bound (4.12) on the sample size, this method is suitable only for bounded-degree graphs. However, for reasonable sizes of d, the dimension of the linear regression problem decreases from $\mathcal{O}(p^d)$ to $|\mathcal{C}_d| = \mathcal{O}(|\mathcal{C}|^d) = \mathcal{O}(d^{(d \log(4/\kappa))/\zeta})$, which has a significant impact on the runtime of the algorithm. We explore two classes of bounded-degree graphs with correlation decay in the simulations of Section 4.4, where we generate Erdös–Renyi graphs with edge probability c/p and square grid graphs in order to test the behavior of our recovery algorithm on nontrees. When $m > 2$, corresponding to nonbinary states, we may combine these ideas with the overlapping group Lasso [18] to obtain similar algorithms for nodewise recovery of nontree graphs. However, the details are more complicated, and we do not include them here. Note that our method for nodewise recovery in nontree graphical models is again easily adapted to handle noisy and missing data, which is a clear advantage over other existing methods.

4.4. *Simulations.* In this section we report the results of various simulations we performed to illustrate the sharpness of our theoretical claims. In all cases, we generated data from binary Ising models. We first applied the nodewise linear regression method (Algorithm 2 for trees; Algorithm 3 in the general case) to the method of ℓ_1-regularized logistic regression, analyzed in past work for Ising model selection by Ravikumar, Wainwright and Lafferty [30]. Their main result was to establish that, under certain incoherence conditions of the Fisher information matrix, performing ℓ_1-regularized logistic regression with a sample size $n \gtrsim d^3 \log p$ is guaranteed to select the correct graph w.h.p. Thus, for any bounded-degree graph, the sample size n need grow only logarithmically in the number of nodes p. Under this scaling, our theory also guarantees that nodewise *linear regression* with ℓ_1-regularization will succeed in recovering the true graph w.h.p.

In Figure 3 we present the results of simulations with two goals: (i) to test the $n \approx \log p$ scaling of the required sample size; and (ii) to compare ℓ_1-regularized nodewise linear regression (Algorithms 3 and 4) to ℓ_1-regularized nodewise logistic regression [30]. We ran simulations for the two methods on both tree-structured and nontree graphs with data generated from a binary Ising model, with node weights $\theta_s = 0.1$ and edge weights $\theta_{st} = 0.3$. To save on computation, we employed the neighborhood screening method described in Section 4.3 to prune the candidate neighborhood set before performing linear regression. We selected a candidate neighborhood set of size $\lfloor 2.5d \rfloor$ with highest empirical correlations, then performed a single regression against all singleton nodes and products of sub-
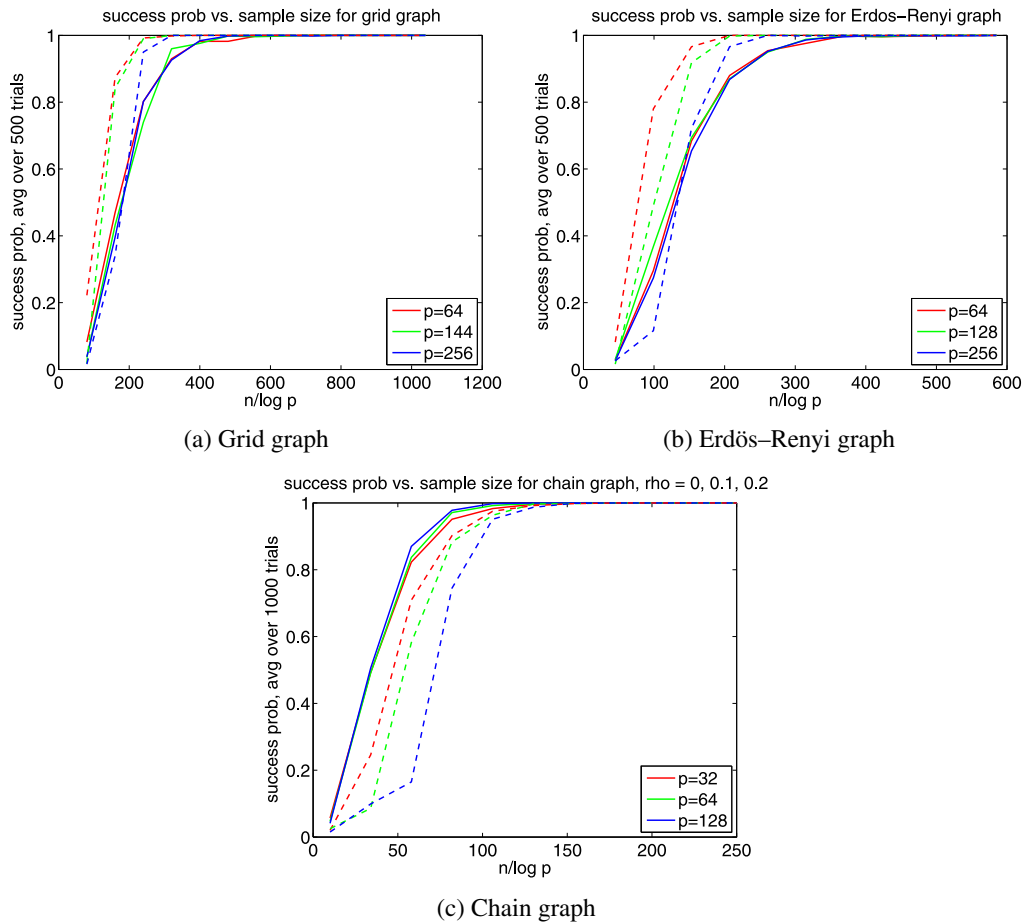
FIG. 3. *Comparison between ℓ_1-regularized logistic vs. linear regression methods for graph recovery. Each panel plots of the probability of correct graph recovery vs. the rescaled sample size $n/\log p$; solid curves correspond to linear regression (method in this paper), whereas dotted curves correspond to logistic regression* [30]. *Curves are based on average performance over* 500 *trials.* (a) *Simulation results for two-dimensional grids with $d = 4$ neighbors, and number of nodes p varying over* {64, 144, 256}. *Consistent with theory, when plotted vs. the rescaled sample size $n/\log p$, all three curves (red, blue, green) are well aligned with one another. Both linear and logistic regression transition from failure to success at a similar point.* (b) *Analogous results for an Erdös–Renyi graph with edge probability $3/p$.* (c) *Analogous results for a chain-structured graph with maximum degree $d = 2$.*

sets of the candidate neighborhood set of size at most d, via the modified Lasso program (4.6). The size of the candidate neighborhood set was tuned through repeated runs of the algorithm. For both methods, the optimal choice of regularization parameter λ_n scales as $\sqrt{\frac{\log p}{n}}$, and we used the same value of λ_n in comparing logistic to linear regression. In each panel we plot the probability of successful graph recovery versus the rescaled sample size $\frac{n}{\log p}$, with curves of different colors corresponding to graphs (from the same family) of different sizes. Solid lines

correspond to linear regression, whereas dotted lines correspond to logistic regression; panels (a), (b) and (c) correspond to grid graphs, Erdös–Renyi random graphs and chain graphs, respectively. For all these graphs, the three solid/dotted curves for different problem sizes are well aligned, showing that the method undergoes a transition from failure to success as a function of the ratio $\frac{n}{\log p}$. In addition, both linear and logistic regression are comparable in terms of statistical efficiency (the number of samples n required for correct graph selection to be achieved).

The main advantage of nodewise linear regression and the graphical Lasso over nodewise logistic regression is that they are straightforward to correct for corrupted or missing data. Figure 4 shows the results of simulations designed to test the be-
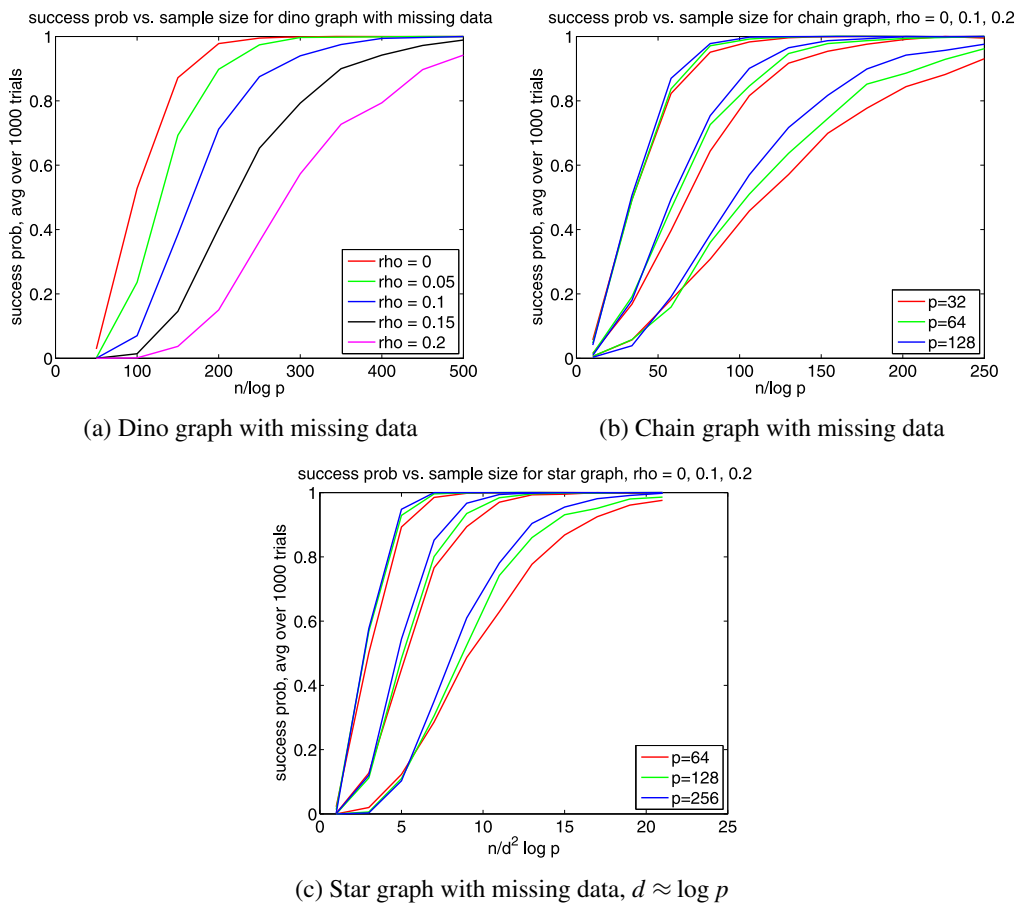

(a) Dino graph with missing data

(b) Chain graph with missing data

(c) Star graph with missing data, $d \approx \log p$

FIG. 4. *Simulation results for global and nodewise recovery methods on binary Ising models, allowing for missing data in the observations. Each point represents an average over* 1000 *trials. Panel* (a) *shows simulation results for the graphical Lasso method applied to the dinosaur graph with the fraction ρ of missing data varying in* $\{0, 0.05, 0.1, 0.15, 0.2\}$. *Panel* (b) *shows simulation results for nodewise regression applied to chain graphs for varying p and ρ. Panel* (c) *shows simulation results for nodewise regression applied to star graphs with maximal node degree $d = \log p$ and varying ρ.*

havior of these corrected estimators in the presence of missing data. Panel (a) shows the results of applying the graphical Lasso method, as described in Section 4.1, to the dino graph of Figure 2(e). We again generated data from an Ising model with node weights 0.1 and edge weights 0.3. The curves show the probability of success in recovering the 15 edges of the graph, as a function of the rescaled sample size $\frac{n}{\log p}$ for $p = 13$. In addition, we performed simulations for different levels of missing data, specified by the parameter $\rho \in \{0, 0.05, 0.1, 0.15, 0.2\}$, using the corrected estimator (4.5). Note that all five runs display a transition from success probability 0 to success probability 1 in roughly the same range, as predicted by our theory. Indeed, since the dinosaur graph has only singleton separators, Corollary 2 ensures that the inverse covariance matrix is exactly graph-structured, so our global recovery method is consistent at the population level. Further note that the curves shift right as the fraction ρ of missing data increases, since the recovery problem becomes incrementally harder.

Panels (b) and (c) of Figure 4 show the results of the nodewise regression method of Section 4.2 applied to chain and star graphs, with increasing numbers of nodes $p \in \{32, 64, 128\}$ and $p \in \{64, 128, 256\}$, respectively. For the chain graphs in panel (b), we set node weights of the Ising model equal to 0.1 and edge weights equal to 0.3. For the varying-degree star graph in panel (c), we set node weights equal to 0.1 and edge weights equal to $\frac{1.2}{d}$, where the degree d of the central hub grows with the size of the graph as $\lfloor \log p \rfloor$. Again, we show curves for different levels of missing data, $\rho \in \{0, 0.1, 0.2\}$. The modified Lasso program (4.6) was optimized using a form of composite gradient descent due to Agarwal, Negahban and Wainwright [1], guaranteed to converge to a small neighborhood of the optimum even when the problem is nonconvex [25]. In both the chain and star graphs, the three curves corresponding to different problem sizes p at each value of the missing data parameter ρ stack up when plotted against the rescaled sample size. Note that the curves for the star graph stack up nicely with the scaling $\frac{n}{d^2 \log p}$, rather than the worst-case scaling $n \asymp d^3 \log p$, corroborating the remark following Proposition 1. Since $d = 2$ is fixed for the chain graph, we use the rescaled sample size $\frac{n}{\log p}$ in our plots, as in the plots in Figure 3. Once again, these simulations corroborate our theoretical predictions: the corrected linear regression estimator remains consistent even in the presence of missing data, although the sample size required for consistency grows as the fraction of missing data ρ increases.

5. Discussion. The correspondence between the inverse covariance matrix and graph structure of a Gauss–Markov random field is a classical fact with numerous consequences for estimation of Gaussian graphical models. It has been an open question as to whether similar properties extend to a broader class of graphical models. In this paper, we have provided a partial affirmative answer to this question and developed theoretical results extending such relationships to discrete undirected graphical models.

As shown by our results, the inverse of the ordinary covariance matrix is graph-structured for special subclasses of graphs with singleton separator sets. More generally, we have considered inverses of *generalized covariance matrices*, formed by introducing indicator functions for larger subsets of variables. When these subsets are chosen to reflect the structure of an underlying junction tree, the edge structure is reflected in the inverse covariance matrix. Our population-level results have a number of statistical consequences for graphical model selection. We have shown that our results may be used to establish consistency (or inconsistency) of standard methods for discrete graph selection, and have proposed new methods for neighborhood recovery which, unlike existing methods, may be applied even when observations are systematically corrupted by mechanisms such as additive noise and missing data. Furthermore, our methods are attractive in their simplicity, in that they only involve simple optimization problems.

Acknowledgments. Thanks to the Associate Editor and anonymous reviewers for helpful feedback.

SUPPLEMENTARY MATERIAL

Supplementary material for "Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses" (DOI: 10.1214/13-AOS1162SUPP; .pdf). Due to space constraints, we have relegated technical details of the remaining proofs to the supplement [26].

REFERENCES

[1] AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.* **40** 2452–2482. MR3097609

[2] ANANDKUMAR, A., TAN, V. Y. F., HUANG, F. and WILLSKY, A. S. (2012). High-dimensional structure estimation in Ising models: Local separation criterion. *Ann. Statist.* **40** 1346–1375. MR3015028

[3] BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

[4] BARNDORFF-NIELSON, O. E. (1978). *Information and Exponential Families*. Wiley, Chichester.

[5] BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **36** 192–236. MR0373208

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

[7] BROWN, L. D. (1986). *Fundamentals of Statistical Exponential Families*. IMS, Hayward, CA. MR0882001

[8] CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

[9] CARROLL, R. J., RUPPERT, D. and STEFANSKI, L. A. (1995). *Measurement Error in Nonlinear Models*. Chapman & Hall, London. MR1630517

[10] CHOW, C. I. and LIU, C. N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inform. Theory* **14** 462–467.

[11] DARROCH, J. N. and SPEED, T. P. (1983). Additive and multiplicative models and interactions. *Ann. Statist.* **11** 724–738. MR0707924

[12] D'ASPREMONT, A., BANERJEE, O. and EL GHAOUI, L. (2008). First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.* **30** 56–66. MR2399568

[13] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **39** 1–38. MR0501537

[14] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9** 432–441.

[15] GRIMMETT, G. R. (1973). A theorem about random fields. *Bull. Lond. Math. Soc.* **5** 81–84. MR0329039

[16] HORN, R. A. and JOHNSON, C. R. (1990). *Matrix Analysis*. Cambridge Univ. Press, Cambridge. MR1084815

[17] IBRAHIM, J. G., CHEN, M.-H., LIPSITZ, S. R. and HERRING, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *J. Amer. Statist. Assoc.* **100** 332–346. MR2166072

[18] JACOB, L., OBOZINSKI, G. and VERT, J. P. (2009). Group Lasso with overlap and graph Lasso. In *International Conference on Machine Learning* (*ICML*) 433–440. ACM, New York.

[19] JALALI, A., RAVIKUMAR, P. D., VASUKI, V. and SANGHAVI, S. (2011). On learning discrete graphical models using group-sparse regularization. *Journal of Machine Learning Research—Proceedings Track* **15** 378–387.

[20] KOLLER, D. and FRIEDMAN, N. (2009). *Probabilistic Graphical Models*: *Principles and Techniques*. MIT Press, Cambridge. MR2778120

[21] LAURITZEN, S. L. (1996). *Graphical Models*. Oxford Univ. Press, New York. MR1419991

[22] LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **50** 157–224. MR0964177

[23] LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084

[24] LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. MR2563983

[25] LOH, P.-L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.* **40** 1637–1664. MR3015038

[26] LOH, P. and WAINWRIGHT, M. J. (2013). Supplement to "Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses." DOI:10.1214/13-AOS1162SUPP.

[27] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[28] NEWMAN, M. E. J. and WATTS, D. J. (1999). Scaling and percolation in the small-world network model. *Phys. Rev. E* (3) **60** 7332–7342.

[29] OBOZINSKI, G., WAINWRIGHT, M. J. and JORDAN, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Ann. Statist.* **39** 1–47. MR2797839

[30] RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343

[31] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

[32] ROCKAFELLAR, R. T. (1970). *Convex Analysis*. Princeton Univ. Press, Princeton, NJ. MR0274683

[33] ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

[34] RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

[35] SANTHANAM, N. P. and WAINWRIGHT, M. J. (2012). Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Trans. Inform. Theory* **58** 4117–4134. MR2943079

[36] WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305. ISSN 1935-8237.

[37] XUE, L. and ZOU, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.* **40** 2541–2571. MR3097612

[38] YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

[39] YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

[40] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: ploh@berkeley.edu
           wainwrig@stat.berkeley.edu

SUPPLEMENTARY MATERIAL FOR: STRUCTURE ESTIMATION FOR DISCRETE GRAPHICAL MODELS: GENERALIZED COVARIANCE MATRICES AND THEIR INVERSES

BY PO-LING LOH, MARTIN J. WAINWRIGHT

University of California, Berkeley

APPENDIX A: PROOFS OF SUPPORTING LEMMAS FOR THEOREM 1

In this section, we supply the proofs of Lemmas 1 and 2, which are used in the proof of Theorem 1.

A.1. Proof of Lemma 1. By Proposition B.2 of Wainwright and Jordan [8] (cf. Theorems 23.5 and 26.3 of Rockafellar [7]), we know that the dual function Φ^* is differentiable on the interior of the marginal polytope $\mathcal{M}$ defined in equation (3.2), in particular with

$$(A.1) \qquad \nabla\Phi^*(\mu) = (\nabla\Phi)^{-1}(\mu) \quad \text{for all } \mu \in \text{int}(\mathcal{M}).$$

Also, by Theorem 3.4 of Wainwright and Jordan [8], for any $\mu \in \text{int}(\mathcal{M})$, the negative dual function takes the form $\Phi^*(\mu) = -H(q_\theta(\mu))$, where $\theta(\mu) = (\nabla\Phi)^{-1}(\mu)$.

By relation (A.1), we have

$$(\nabla\Phi)(\nabla\Phi^*(\mu)) = \mu \qquad \text{for all } \mu \in \mathcal{M}.$$

Since this equation holds on an open set, we may take derivatives; employing the chain rule yields

$$(\nabla^2\Phi)(\nabla\Phi^*(\mu)) \cdot (\nabla^2\Phi^*(\mu)) = I_{D\times D}.$$

Rearranging yields the relation $\nabla^2\Phi^*(\mu) = (\nabla^2\Phi(\theta))^{-1}\mid_{\theta=\theta(\mu)}$, as claimed.

A.2. Proof of Lemma 2. We induct on the subset size. For sets of size 1, the claim is obvious. Now suppose the claim holds for all subsets up to some size $k > 1$, and consider a subset of size $k + 1$, which we write as $C = \{1, \ldots, k+1\}$, without loss of generality. For any configuration

$J \in \mathcal{X}_0^{|C|}$, the marginal probability $q_C(x_C = J)$ is equal to $\mu_{C;J}$, by construction. Consequently, we need only specify how to determine the probabilities $q_C(x_C = J)$ for configurations $J \in \mathcal{X}^{|C|} \backslash \mathcal{X}_0^{|C|}$. By the definition of $\mathcal{X}_0^{|C|}$, each $j \in J$ has $j_s = 0$ for at least one $s \in \{1, \ldots, k+1\}$.

We show how to express the remaining marginal probabilities sequentially, inducting on the number of positions s for which $j_s = 0$. Starting with the base case in which there is a single zero, suppose without loss of generality that $j_{k+1} = 0$. For each $\ell \in \{1, 2, \ldots, m-1\}$, let J^ℓ be the configuration such that $J_i^\ell = J_i$ for all $i \neq k+1$ and $J_{k+1}^\ell = \ell$. Defining $D := C \backslash \{k+1\}$, we then have

$$(A.2) \qquad q_C(x_C = J) = q_D(x_D = J') - \sum_{\ell=1}^{m-1} q_C(x_C = J^\ell),$$

where $J' \in \mathcal{X}^k$ is the configuration defined by $J_i' = J_i$ for all $i = 1, 2, \ldots, k$. Since $|D| = k$, our induction hypothesis implies that $q_D(x_D = J')$ is a linear function of the specified mean parameters. Moreover, our starting assumption implies that $J^\ell \in \mathcal{X}_0^{|C|}$ for all $\ell = \{1, 2, \ldots, m-1\}$, so we have $q_C(x_C = J^\ell) = \mu_{C;J^\ell}$. This establishes the base case.

Now suppose the sub-claim holds for all configurations with at most t nonzeros, for some $t > 1$. Consider a configuration J with $t+1$ zero entries. Again without loss of generality, we may assume $j_{k+1} = 0$, so equation (A.2) may be derived as before. This time, the configurations J^ℓ are not in $\mathcal{X}_0^{|C|}$ (since they still have $t \geq 1$ zero entries); however, our induction hypothesis implies that the corresponding probabilities may be written as functions of the given mean parameters. This completes the inductive proof of the inner claim, thereby completing the outer induction, as well.

APPENDIX B: PROOFS OF POPULATION-LEVEL COROLLARIES

In this Appendix, we prove Corollaries 1 and 3. (As previously noted, Corollary 2 is an immediate consequence of Corollary 1.)

B.1. Proof of Corollary 1. Recall that $\widetilde{\mathcal{C}}$ denotes the set of all cliques in the triangulation $\widetilde{G}$. The covariance matrix in Theorem 1 is indexed by $\widetilde{\mathcal{C}}$, and our goal is to define appropriate blocks of the matrix and then apply the matrix inversion lemma [3]. Consider the collection $\mathrm{pow}(\mathcal{S})$. We define the collection of singleton subsets $V = \{\{1\}, \{2\}, \ldots, \{p\}\}$, and introduce the disjoint partition

$$\widetilde{\mathcal{C}} = \underbrace{\left(\mathrm{pow}(\mathcal{S}) \cup V \right)}_{\mathcal{U}} \cup \underbrace{\left(\widetilde{\mathcal{C}} \backslash \{ \mathrm{pow}(\mathcal{S}) \cup V \} \right)}_{\mathcal{W}}.$$

The following property of the collection $\mathcal{W}$ is important:

LEMMA 3.   For each maximal clique $C \in \bar{\mathcal{C}}$, define the set collection $\mathcal{F}(C) = \mathrm{pow}(C)\backslash\mathcal{U}$. For any $A \in \mathcal{W}$, we have $A \in \mathcal{F}(C)$ for exactly one C.

PROOF.   We first establish existence. Since $\mathcal{W} \subseteq \widetilde{\mathcal{C}}$, any set $A \in \mathcal{W}$ is contained in some maximal clique C_A. Since $A \notin \mathcal{U}$, we clearly have $A \in \mathcal{F}(C_A)$.

To establish uniqueness, consider a set A belonging to the intersection $C_1 \cap C_2$ of two maximal cliques. If these cliques are adjacent in the junction tree, then A belongs to the separator set $C_1 \cap C_2$, so A cannot belong to $\mathcal{W}$, by definition. Even when C_1 and C_2 are not adjacent, the running intersection property of the junction tree implies that $C_1 \cap C_2$ must belong to every separator set on the unique path between C_1 and C_2 in the junction tree, implying that $A \notin \mathcal{W}$, as before. This is a contradiction, implying that the maximal clique C_A is unique.   □

Define $\Gamma = \big(\mathrm{cov}(\Psi(X;\widetilde{\mathcal{C}}))\big)^{-1}$. By the block-matrix inversion formula [3], we may write

(B.1)
$$\Theta := (\mathrm{cov}\big(\Psi(X;\mathcal{U})\big))^{-1} = \Gamma(\mathcal{U},\mathcal{U}) - \Gamma(\mathcal{U},\mathcal{W})\big(\Gamma(\mathcal{W},\mathcal{W})\big)^{-1}\Gamma(\mathcal{W},\mathcal{U}).$$

We need to show that $\Theta(A, B) = 0$ for any members $A, B \in \mathcal{U}$ that do not belong to the same maximal clique. By Theorem 1(a), we have $\Gamma(A, B) = 0$ whenever A and B do not belong to the same maximal clique, so it remains to show that $\Gamma(A, \mathcal{W})\big(\Gamma(\mathcal{W}, \mathcal{W})\big)^{-1}\Gamma(\mathcal{W}, B) = 0$.

We begin by observing that the matrix $\Gamma(\mathcal{W}, \mathcal{W})$ is block-diagonal with respect to the partition $\{\mathcal{F}(C) : C \in \mathcal{C}\}$ previously defined in Lemma 3. (Indeed, consider two sets $D, E \in \mathcal{W}$ with $D \in \mathcal{F}(C)$ and $E \in \mathcal{F}(C')$ for distinct maximal cliques $C \neq C'$. Two such sets cannot belong to the same maximal clique, so Theorem 1(a) implies that $\Gamma(D, E) = 0$.) Since block-diagonal structure is preserved by matrix inversion, the inverse $\Upsilon = (\Gamma(\mathcal{W}, \mathcal{W}))^{-1}$ shares this property, so for any two members $A, B \in \mathcal{U}$, we may write

$$\Gamma(A, \mathcal{W})\big(\Gamma(\mathcal{W}, \mathcal{W})\big)^{-1}\Gamma(\mathcal{W}, B)$$

(B.2)
$$= \sum_{\mathcal{F}(C), C \in \bar{\mathcal{C}}} \Gamma(A, \mathcal{F}(C))\Upsilon(\mathcal{F}(C), \mathcal{F}(C))\Gamma(\mathcal{F}(C), B).$$

We claim that each of these terms vanishes. For a given maximal clique C', suppose A is not contained within C'; we first claim that $\Gamma(A, \mathcal{F}(C')) = 0$,

or equivalently, for any set $D \in \mathcal{F}(C')$, we have $\Gamma(A, D) = 0$. From Theorem 1(a), it suffices to show that A and D cannot be contained within the same maximal clique. From Lemma 3, we know that A belongs to a unique maximal clique C. Any set $D \in \mathcal{F}(C')$ is contained within C'; if it were also contained within C, then D would be contained in $C \cap C'$. But as argued in the proof of Lemma 3, this implies that D is contained within some separator set, whence it cannot belong to $\mathcal{F}(C')$. We thus conclude that $\Gamma(A, D) = 0$, as claimed.

Taking any two subsets A and B that are not contained in the same maximal clique, we see that for any clique C, we must either have $\Gamma(A, \mathcal{F}(C)) = 0$ or $\Gamma(\mathcal{F}(C), B) = 0$. Hence, each term in the sum (B.2) indeed vanishes, completing the proof.

B.2. Proof of Corollary 3. This corollary follows by a similar argument as in the proof of Corollary 1. As before, let $\mathcal{C}$ denote the set of all cliques in the triangulation $\widetilde{G}$, and let $V = \{\{1\}, \{2\}, \ldots, \{p\}\}$. Define $\mathcal{U} = \mathrm{pow}(\mathcal{S}(s; d)) \cup V$ and $\mathcal{W} = \widetilde{\mathcal{C}} \backslash \mathcal{U}$.

Let $C_s := s \cup N(s)$, and consider a disjoint partition of $\mathcal{W}$ defined by $\mathcal{F}_1 := \mathrm{pow}(C_s) \backslash \mathcal{U}$ and $\mathcal{F}_2 := \mathcal{W} \backslash \mathcal{F}_1$. Note that C_s is the unique maximal clique in $\widetilde{\mathcal{C}}$ containing s. By construction, every clique in $\mathcal{F}_2$ does not contain s and has more than d elements, whereas every clique in $\mathcal{F}_1$ is contained in C_s, with $|C_s| \le d + 1$. It follows that no two cliques $A \in \mathcal{F}_1$ and $B \in \mathcal{F}_2$ can be contained in the same maximal clique. Denoting $\Gamma := (\mathrm{cov}(\Psi(X; \widetilde{\mathcal{C}})))^{-1}$, we conclude via Theorem 1(a) that $\Gamma(\mathcal{W}, \mathcal{W})$ is block-diagonal.

We now use the block matrix-equation formula (B.1). As before, Theorem 1(a) implies that $\Gamma(\mathcal{U}, \mathcal{U})$ is graph-structured according to $\widetilde{G}$. In particular, for any $B \in \mathcal{U}$ with $B \subsetneq C_s$, we have $\Gamma(\{s\}, B) = 0$. (The elements of $\mathcal{U}$ that are subsets of C_s are exactly $\{s\}$ and the nonempty subsets of $N(s)$.) Hence, it remains to show that $\Gamma(\{s\}, \mathcal{W})(\Gamma(\mathcal{W}, \mathcal{W}))^{-1} \Gamma(\mathcal{W}, B) = 0$.

Analogous to equation (B.2), we may write

$$\Gamma(\{s\}, \mathcal{W})(\Gamma(\mathcal{W}, \mathcal{W}))^{-1} \Gamma(\mathcal{W}, B) = \sum_{i=1}^{2} \Gamma(\{s\}, \mathcal{F}_i) \Upsilon(\mathcal{F}_i, \mathcal{F}_i) \Gamma(\mathcal{F}_i, B),$$

where $\Upsilon := (\Gamma(\mathcal{W}, \mathcal{W}))^{-1}$. Applying Theorem 1(a) once more, we see that $\Gamma(\mathcal{F}_1, B) = 0$, since $B \subsetneq C_s$ and $\Gamma(\{s\}, \mathcal{F}_2) = 0$. Therefore, the matrix $\Theta = (\mathrm{cov}(\Psi(X; \mathcal{U})))^{-1}$ appearing in equation (B.1) is indeed s-block graph-structured.

APPENDIX C: PROOF OF PROPOSITION 1

In this section, we provide a proof of our main nodewise recovery result, Proposition 1. For proofs of supporting technical lemmas and all corollaries appearing in the text, see Appendix D.

C.1. Main argument. We derive Proposition 1 as a consequence of a more general theorem. Suppose we have i.i.d. observations $\{(x_i, y_i)\}_{i=1}^{n}$, with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, and we wish to estimate the best linear predictor $\widetilde{\beta} = \Sigma_x^{-1} \operatorname{Cov}(x_i, y_i)$, when $\widetilde{\beta}$ is k-sparse. Loh and Wainwright [5] formulated a modified version of the Lasso based on possibly corrupted observations; however, they assume the linear regression model

$$\text{(C.1)} \qquad y_i = x_i^T \widetilde{\beta} + \epsilon_i,$$

where ϵ_i is sub-Gaussian noise and $\epsilon_i \perp\!\!\!\perp x_i$. Although the model (C.1) holds in the case where y_i is a sample from a single node and x_i is a sample from all other nodes in a Gaussian graphical model, the model (C.1) does *not* hold in a general discrete graphical model. Nonetheless, we show that essentially the same Lasso estimator provides an estimator for $\widetilde{\beta}$ that is consistent for support recovery. Suppose the pair $(\widehat{\Gamma}, \widehat{\gamma})$ in the Lasso program (4.6) satisfies the following deviation bounds:

$$\text{(C.2a)} \qquad \|\widehat{\Gamma}\widetilde{\beta} - \widehat{\gamma}\|_\infty \le \varphi_1 \sqrt{\frac{\log p}{n}},$$

$$\text{(C.2b)} \qquad \|(\widehat{\Gamma} - \Sigma_x)v\|_\infty \le \varphi_2 \|v\|_\infty \sqrt{\frac{k \log p}{n}} \quad \forall v \in \mathbb{B}_1(8k) \cap \mathbb{B}_\infty(1),$$

for some φ_1, φ_2. Also suppose $\widehat{\Gamma}$ satisfies the lower-restricted eigenvalue (RE) condition:

$$\text{(C.3)} \qquad v^T \widehat{\Gamma} v \ge \alpha \|v\|_2^2 \quad \forall v \text{ s.t. } \|v\|_1 \le \sqrt{k} \|v\|_2.$$

Then we have the following technical result:

THEOREM 2. *Suppose the pair $(\widehat{\Gamma}, \widehat{\gamma})$ satisfies the deviation conditions (C.2a) and (C.2b), as well as the lower-RE condition (C.3). Also suppose $n \gtrsim \max\left\{\frac{\varphi_1^2}{\alpha^2(b_0 - \|\widetilde{\beta}\|_2)^2}, \varphi_2^2 \left\|\!\left\|\Sigma_x^{-1}\right\|\!\right\|_\infty^2\right\} k \log p$ and $\lambda_n \gtrsim \varphi_1 \sqrt{\frac{\log p}{n}}$. Then any optimum $\widehat{\beta}$ of the Lasso program (4.6) satisfies*

$$\|\widehat{\beta} - \widetilde{\beta}\|_\infty \le 4\lambda_n \left\|\!\left\|\Sigma_x^{-1}\right\|\!\right\|_\infty.$$

The proof of Theorem 2 is provided in Appendix C.2. In order to prove Proposition 1, we first establish that the deviation conditions (C.2a) and (C.2b) of Theorem 2 hold w.h.p. with $(\varphi_1, \varphi_2) = (\varphi \|\widetilde{\beta}\|_2, \varphi)$, and the lower-RE condition holds with $\alpha = \frac{1}{2}\lambda_{\min}(\Sigma_x)$.

Note that

$$(C.4) \qquad \|\widehat{\Gamma}\widetilde{\beta} - \widehat{\gamma}\|_\infty \leq \|(\widehat{\Gamma} - \Sigma_x)\widetilde{\beta}\|_\infty + \|\operatorname{Cov}(x_i, y_i) - \widehat{\gamma}\|_\infty.$$

Furthermore,

$$\|(\widehat{\Gamma} - \Sigma_x)\widetilde{\beta}\|_\infty \leq \left\|\left(\frac{X^T X}{n} - \mathbb{E}(x_i x_i^T)\right)\widetilde{\beta}\right\|_\infty + \|(\overline{x}\overline{x}^T - \Sigma_x)\widetilde{\beta}\|_\infty$$

and

$$\|\operatorname{Cov}(x_i, y_i) - \widehat{\gamma}\|_\infty \leq \left\|\frac{X^T y}{n} - \mathbb{E}(y_i x_i)\right\|_\infty + \|\overline{y}\overline{x} - \mathbb{E}(y_i)\mathbb{E}(x_i)\|_\infty.$$

As in the analysis of inequality (E.1) below, we may disregard the two second terms involving empirical means, since they concentrate at a fast rate. Since $x_i^T \widetilde{\beta}$ is sub-Gaussian with parameter $\varphi^2 \|\widetilde{\beta}\|_2^2$ by assumption, and $e_j^T x_i$ and y_i are clearly sub-Gaussian with parameter 1, the deviation condition (C.2a) follows with $\varphi_1 = \varphi \|\widetilde{\beta}\|_2$ by standard concentration bounds on an i.i.d. average of products of sub-Gaussians (cf. Lemma 14 of Loh and Wainwright [5]).

For the second deviation bound, we will verify the bound over a more tractable set via the following lemma:

LEMMA 4.   For any constant $c_0 > 0$, we have

$$\mathbb{B}_1(c_0 k) \cap \mathbb{B}_\infty(1) \subseteq (1 + c_0) \operatorname{cl}\{\operatorname{conv}\{\mathbb{B}_0(k) \cap \mathbb{B}_\infty(1)\}\}.$$

Hence, it is sufficient to establish the deviation inequality (C.2b) over the set $\mathbb{B}_0(k) \cap \mathbb{B}_\infty(1)$. We proceed via a discretization argument. Suppose $\{v_1, \ldots, v_M\}$ is a $\frac{1}{2}$-covering of the unit ℓ_∞-ball in $\mathbb{R}^k$ in its own metric. By standard results on metric entropy, we know that such a covering exists with $M \leq c^k$. Writing $\psi(v) = \|(\widehat{\Gamma} - \Sigma_x)v\|_\infty$, we know that there exists v_j such that $\|v - v_j\|_\infty \leq \frac{1}{2}$. Let $\Delta v = v - v_j$. Then

$$\psi(v) = \|(\widehat{\Gamma} - \Sigma_x)(v_j + \Delta v)\|_\infty \leq \psi(v_j) + \psi(\Delta v) \leq \sup_{1 \leq j \leq M} \psi(v_j) + \frac{1}{2} \sup_{\|v\|_\infty \leq 1} \psi(v),$$

simply by rescaling. Taking the sup over $\{\|v\|_\infty \leq 1\}$ on the LHS and rearranging then yields

$$\sup_{\|v\|_\infty \leq 1} \leq 2 \sup_{1 \leq j \leq M} \psi(v_j).$$

Hence, it suffices to establish the bound for a given $v \in \mathbb{B}_1(c_0 k) \cap \mathbb{B}_\infty(1)$, then take a union bound over the $M \leq c^k$ elements in the discretization and the $\binom{p}{k} \leq p^k$ choices of the support set.

For a given k-sparse v, note that $x_i^T v$ has sub-Gaussian parameter $\varphi^2 \|v\|_2^2$ by assumption, and

$$\|v\|_2^2 \leq \|v\|_1 \|v\|_\infty \leq \sqrt{k} \|v\|_2 \|v\|_\infty,$$

so $x_i^T v$ is sub-Gaussian with parameter $\varphi^2 k \|v\|_\infty^2$. Since $e_\ell^T x_i$ is sub-Gaussian with parameter 1, it follows from the same recentering techniques as in inequality (E.1) that

$$\|(\widehat{\Gamma} - \Sigma_x) v\|_\infty = \max_\ell |e_\ell^T (\widehat{\Gamma} - \Sigma_x) v| \leq t,$$

with probability at least $1 - c_1 \exp\left(\frac{-c_2 n t^2}{\varphi^2 k \|v\|_\infty^2}\right)$. Taking a union bound over the discretization and setting $t = c\varphi\sqrt{k}\|v\|_\infty \sqrt{\frac{k \log p}{n}}$ then implies the deviation bound (C.2b) with $\varphi_2 = \varphi$, under the scaling $n \gtrsim \varphi^2 k^2 \log p$.

The lower-RE condition (C.3) may be verified analogously to the results in Loh and Wainwright [5]. The only difference is to use the fact that $x_i^T v$ is sub-Gaussian with parameter $\varphi^2 \|v\|_2^2$ in all the deviation bounds. Then the lower-RE condition holds with probability at least $1 - c_1 \exp(-c_2 k \log p)$, under the scaling $n \gtrsim \varphi^2 k \log p$.

We may take $\lambda_n \asymp \varphi \|\widetilde{\beta}\|_2 \sqrt{\frac{\log p}{n}}$ in Theorem 2 to conclude that w.h.p.,

$$\|\widehat{\beta} - \widetilde{\beta}\|_\infty \precsim \varphi |\widetilde{\beta}\|_2 \sqrt{\frac{\log p}{n}}.$$

Finally, note that the vector $\widetilde{\beta}$ is a scalar multiple of column s of the inverse covariance matrix Γ, as a straightforward consequence of block matrix inversion. Hence, combining Corollary 1 and Theorem 2 implies that thresholding succeeds w.h.p. for neighborhood recovery in a tree graph.

C.2. Proof of Theorem 2. We begin by establishing ℓ_1- and ℓ_2- error bounds, which will be used in the sequel:

LEMMA 5.   Suppose the deviation condition (C.2a) holds and $\widehat{\Gamma}$ satisfies the lower-RE condition (C.3). Also suppose $\lambda_n \gtrsim \varphi_1 \sqrt{\frac{\log p}{n}}$. Then any global

optimum $\widehat{\beta}$ of the Lasso program (4.6) satisfies the bounds

(C.5) $$\|\widehat{\beta} - \widetilde{\beta}\|_2 \leq \frac{c_0\sqrt{k}}{\alpha_\ell} \max\{\varphi_1 \sqrt{\frac{\log p}{n}}, \lambda_n\},$$

(C.6) $$\|\widehat{\beta} - \widetilde{\beta}\|_1 \leq \frac{8c_0 k}{\alpha_\ell} \max\{\varphi_1 \sqrt{\frac{\log p}{n}}, \lambda_n\}.$$

We now argue that for suitable scaling $n \gtrsim k \log p$, any optimum $\widehat{\beta}$ lies in the interior of $\mathbb{B}_1(b_0\sqrt{k})$:

LEMMA 6. Suppose $\widehat{\beta}$ is an optimum of the Lasso program (4.6). Then under the scaling $n \gtrsim \left(\frac{\varphi_1}{\alpha(b_0 - \|\widetilde{\beta}\|_2)}\right)^2 k \log p$, we have

$$\widehat{\beta} \notin \partial\mathbb{B}_1(b_0\sqrt{k}).$$

By Lemma 6, we are guaranteed that $\widehat{\beta}$ is an interior point of the feasible set. Consequently, by Proposition 2.3.2 of Clarke [2], we are guaranteed that 0 is a generalized gradient of the objective function at $\widehat{\beta}$. By Proposition 2.3.3 of Clarke [2], there must exist a vector $\hat{z} \in \partial\|\beta\|_1 \mid_{\beta=\widehat{\beta}}$ such that

$$\widehat{\Gamma}\widehat{\beta} - \widehat{\gamma} + \lambda_n \hat{z} = 0.$$

Denoting the loss function $\mathcal{L}(\beta) = \frac{1}{2}\beta^T \widehat{\Gamma}\beta - \widehat{\gamma}^T \beta$, we have $\nabla\mathcal{L}(\beta) = \widehat{\Gamma}\beta - \widehat{\gamma}$, so

$$\nabla\mathcal{L}(\widetilde{\beta}) - \nabla\mathcal{L}(\widehat{\beta}) = \nabla\mathcal{L}(\widetilde{\beta}) + \lambda_n \hat{z} = \widehat{\Gamma}\widetilde{\beta} - \widehat{\gamma} + \lambda_n \hat{z}.$$

Then

(C.7) $\quad \|\nabla\mathcal{L}(\widetilde{\beta}) - \nabla\mathcal{L}(\widehat{\beta})\|_\infty \leq \|\widehat{\Gamma}\widetilde{\beta} - \widehat{\gamma}\|_\infty + \lambda_n \|\hat{z}\|_\infty \leq \|\widehat{\Gamma}\widetilde{\beta} - \widehat{\gamma}\|_\infty + \lambda_n.$

Using the deviation bound (C.2a) again, we have

$$\|\widehat{\Gamma}\widetilde{\beta} - \widehat{\gamma}\|_\infty \leq \varphi_1 \sqrt{\frac{\log p}{n}}.$$

It follows from equation (C.7) that if $\lambda_n \geq \varphi_1 \sqrt{\frac{\log p}{n}}$, then

(C.8) $$\|\nabla\mathcal{L}(\widetilde{\beta}) - \nabla\mathcal{L}(\widehat{\beta})\|_\infty \leq 2\lambda_n.$$

Finally, we lower-bound

$$\|\nabla\mathcal{L}(\widetilde{\beta}) - \nabla\mathcal{L}(\widehat{\beta})\|_\infty = \|\widehat{\Gamma}\widehat{\nu}\|_\infty$$
$$\geq \|\Sigma_x \widehat{\nu}\|_\infty - \|(\widehat{\Gamma} - \Sigma_x)\widehat{\nu}\|_\infty$$
(C.9) $$\geq \left\|\left|\Sigma_x^{-1}\right|\right\|_\infty^{-1} \|\widehat{\nu}\|_\infty - \|(\widehat{\Gamma} - \Sigma_x)\widehat{\nu}\|_\infty.$$

Now note that $\|\widehat{\nu}\|_1 \leq 8\sqrt{k}\|\widehat{\nu}\|_2$, as shown in Loh and Wainwright [5], so we have

$$\|\widehat{\nu}\|_2^2 \leq \|\widehat{\nu}\|_\infty \|\widehat{\nu}\|_1 \leq 8\sqrt{k}\|\widehat{\nu}\|_\infty \|\widehat{\nu}\|_2,$$

so $\|\widehat{\nu}\|_2 \leq 8\sqrt{k}\|\widehat{\nu}\|_\infty$. In particular, $\|\widehat{\nu}\|_1 \leq 8k\|\widehat{\nu}\|_\infty$. Applying inequality (C.2b) to $v = \frac{\widehat{\nu}}{\|\widehat{\nu}\|_\infty}$ then gives

$$\|(\widehat{\Gamma} - \Sigma_x)\widehat{\nu}\|_\infty \leq c\varphi_2\|\widehat{\nu}\|_\infty \sqrt{\frac{k \log p}{n}}.$$

Combining this with inequality (C.9), we have

(C.10) $$\|\widehat{\Gamma}\widehat{\nu}\|_\infty \geq \|\widehat{\nu}\|_\infty \left(\frac{1}{\left\|\left|\Sigma_x^{-1}\right|\right\|_\infty} - c\varphi_2\sqrt{\frac{k \log p}{n}} \right),$$

so when $n \gtrsim \varphi_2^2 \left\|\left|\Sigma_x^{-1}\right|\right\|_\infty^2 k \log p$, we have

$$\|\widehat{\Gamma}\widehat{\nu}\|_\infty \geq \frac{1}{2} \frac{\|\widehat{\nu}\|_\infty}{\left\|\left|\Sigma_x^{-1}\right|\right\|_\infty}.$$

Finally, combining with inequality (C.8) yields the result of the theorem.

APPENDIX D: PROOF OF SUPPORTING LEMMAS TO PROPOSITION 1

D.1. Proof of Lemma 4. We denote the left-hand set by A and the right-hand set by B. It suffices to show that $\varphi_A(z) \leq \varphi_B(z)$ for all z, where φ is the support function.

For a given z, let S be the set of indices of coordinates of z with highest absolute value. We may write

$$\varphi_A(z) = \sup_{\theta \in A}\langle \theta, \, z \rangle$$
$$= \sup_{\theta \in A}\langle \theta_S, \, z_S \rangle + \langle \theta_{S^c}, \, z_{S^c} \rangle$$

(D.1) $$\leq \|z_S\|_1 + c_0 k\|z_{S^c}\|_\infty,$$

since

$$\langle \theta_S, \, z_S \rangle \leq \|\theta_S\|_\infty \|z_S\|_1 \leq \|\theta\|_\infty \|z_S\|_1 \leq \|z_S\|_1$$

and

$$\langle \theta_{S^c}, \, z_{S^c} \rangle \leq \|\theta_{S^c}\|_1 \|z_{S^c}\|_\infty \leq c_0 k\|z_{S^c}\|_\infty$$

for $\theta \in A$. Furthermore, $k\|z_{S^c}\|_\infty \leq \|z_S\|_1$. Hence, inequality (D.1) becomes

$$\varphi_A(z) \leq (1 + c_0)\|z_S\|_1.$$

Finally, note that

$$\varphi_B(z) = (1 + c_0) \max_{|U| \leq k} \sup_{\|\theta_U\|_\infty \leq 1} \langle \theta_U, \, z_U \rangle = (1 + c_0)\|z_S\|_1,$$

establishing the desired result.

D.2. Proof of Lemma 5.   The proof is essentially the same as in the case of a standard linear model [5]. From the fact that $\widetilde{\beta}$ is feasible and $\widehat{\beta}$ is optimal, we obtain a basic inequality. Furthermore, defining $\widehat{\nu} = \widehat{\beta} - \widetilde{\beta}$, we may verify the cone condition $\|\widehat{\nu}\|_1 \leq c\sqrt{k}\|\widehat{\nu}\|_2$. We will not repeat the arguments here.

D.3. Proof of Lemma 6.   Note that

$$\|\widehat{\beta} - \widetilde{\beta}\|_1 \geq \|\widehat{\beta}\|_1 - \|\widetilde{\beta}\|_1 \geq \|\widehat{\beta}\|_1 - \sqrt{k}\|\widetilde{\beta}\|_2.$$

Hence, if $\widehat{\beta} \in \partial\mathbb{B}_1(b_0\sqrt{k})$, we have

$$(D.2) \qquad \|\widehat{\beta} - \widetilde{\beta}\|_1 \geq b_0\sqrt{k} - \|\widetilde{\beta}\|_2\sqrt{k} = (b_0 - \|\widetilde{\beta}\|_2)\sqrt{k}.$$

On the other hand, Theorem 1 in Loh and Wainwright [5] guarantees that under deviation condition (C.2a) and the lower-RE condition (C.3), we have the ℓ_1-bound

$$(D.3) \qquad \|\widehat{\beta} - \widetilde{\beta}\|_1 \leq \frac{c\varphi_1 k}{\alpha}\sqrt{\frac{\log p}{n}}.$$

Combining inequalities (D.2) and (D.3) gives

$$(b_0 - \|\widetilde{\beta}\|_2)\sqrt{k} \leq \frac{c\varphi_1 k}{\alpha}\sqrt{\frac{\log p}{n}},$$

contradicting the assumption that $n > \left(\frac{c\varphi_1}{\alpha(b_0 - \|\widetilde{\beta}\|_2)}\right)^2 k \log p$.

APPENDIX E: PROOFS OF SAMPLE-BASED COROLLARIES

Here, we provide proofs for the remaining corollaries involved in sample-based approaches to graph selection.

E.1. Proof of Corollary 4. As noted by Liu et al. [4], the proof of this corollary hinges only on the deviation condition condition (4.2) being satisfied w.h.p.; the rest of the proof follows from the analysis of Ravikumar et al. [6]. We verify inequality (4.2) with $\varphi(\Sigma^*) = c_1$ and $\psi(n,p) = c' \log p$.

Note that

$$
\|\widehat{\Sigma} - \Sigma\|_{\max} = \left\| \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i^T - \overline{x}\overline{x}^T \right) - \Sigma \right\|_{\max}
$$

$$
\text{(E.1)} \qquad \leq \left\| \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T - \mathbb{E}(x_i x_i^T) \right\|_{\max} + \left\| \overline{x}\overline{x}^T - \mathbb{E}(x_i)\mathbb{E}(x_i)^T \right\|_{\max},
$$

where we have used the triangle inequality and the fact that $\Sigma = \mathbb{E}(x_i x_i^T) - \mathbb{E}(x_i)\mathbb{E}(x_i)^T$ in the second line. Noting that $\|Y\|_{\max} = \max_{j,k} |e_j^T Y e_k|$ for a matrix Y, and the random variables $e_j^T x_i$ are i.i.d. Bernoulli (sub-Gaussian parameter 1) for each fixed j, we conclude by standard sub-Gaussian tail bounds (cf. Lemma 14 in Loh and Wainwright [5]) that the first term is bounded by $\sqrt{\frac{\log p}{n}}$, with probability at least $1 - c\exp(-c' \log p)$. For the second term, we may further bound

$$
\|\overline{x}\overline{x}^T - \mathbb{E}(x_i)\mathbb{E}(x_i)^T\|_{\max} \quad \leq \quad \|(\overline{x} - \mathbb{E}(x_i))(\overline{x} - \mathbb{E}(x_i))^T\|_{\max}
$$
$$
+ 2\|\mathbb{E}(x_i)\|_{\infty}\|\overline{x} - \mathbb{E}(x_i)\|_{\infty},
$$

by way of the triangle inequality. Note that $e_j^T(\overline{x} - \mathbb{E}(x_i))$ is an average of i.i.d. sub-Gaussian variables with parameter 1, hence has sub-Gaussian parameter $\frac{1}{n}$. Therefore, we have the even tighter bound $\frac{1}{n}\sqrt{\frac{\log p}{n}}$ for this term. Combining the bounds for the two terms in inequality (E.1) establishes the deviation condition (4.2).

By the machinery of Ravikumar et al. [6], we then have the elementwise bound

$$
\mathbb{P}\big[\|\widehat{\Theta} - \Theta^*\|_{\max} \geq \tau_n\big] \leq c\exp(-c' \log p).
$$

The statement about thresholding $\widehat{\Theta}$ to obtain a consistent estimate of Θ^* follows immediately.

E.2. Proof of Corollary 5. The analysis borrows techniques from the paper [1]. We first prove that under the scaling $n \gtrsim \kappa^2 \log p$, we have $|r_C(s,t) - \widehat{r}_C(s,t)| \leq \frac{\kappa}{4}$ for all $(s,t) \in V \times V$, with probability at least $1 - c_1 \exp(-c_2 \log p)$. First fix a pair (s,t) and a corresponding pair of values (x_s, x_t). By a simple application of Hoeffding's inequality, we have

$$
\mathbb{P}\left(|\mathbb{P}(X_s = x_s, X_t = x_t) - \widehat{\mathbb{P}}(X_s = x_s, X_t = x_t)| \geq \epsilon \right) \leq c\exp(-c'n\epsilon^2),
$$

and similarly for the marginal deviations $|\mathbb{P}(X_s = x_s) - \widehat{\mathbb{P}}(X_s = x_s)|$ and $|\mathbb{P}(X_t = x_t) - \widehat{\mathbb{P}}(X_t = x_t)|$. Note that

$$|r_C(s,t) - \widehat{r}_C(s,t)| \leq \sum_{x_s,x_t} \left(|\mathbb{P}(X_s = x_s, X_t = x_t) - \widehat{\mathbb{P}}(X_s = x_s, X_t = x_t)| \right.$$
$$\left. + |\mathbb{P}(X_s = x_s)\mathbb{P}(X_t = x_t) - \widehat{\mathbb{P}}(X_s = x_s)\widehat{\mathbb{P}}(X_t = x_t)| \right).$$

Furthermore,

$$|\mathbb{P}(X_s = x_s)\mathbb{P}(X_t = x_t) - \widehat{\mathbb{P}}(X_s = x_s)\widehat{\mathbb{P}}(X_t = x_t)|$$
$$\leq |\mathbb{P}(X_s = x_s) - \widehat{\mathbb{P}}(X_s = x_s)| \cdot \mathbb{P}(X_t = x_t)$$
$$+ |\mathbb{P}(X_t = x_t) - \widehat{\mathbb{P}}(X_t - x_t)| \cdot \widehat{\mathbb{P}}(X_s = x_s)$$
$$\leq 2\epsilon,$$

so taking a union bound over all pairs (s,t) and all values (x_s, x_t), we have $|r_C(s,t) - \widehat{r}_C(s,t)| \leq 3m^2\epsilon$ for all $(s,t) \in V \times V$, with probability at least $1 - cm^2p^2 \exp(-c'n\epsilon^2)$. Finally, taking $\epsilon = \frac{\kappa}{12m^2}$ and using the fact that $n \succsim \kappa^2 \log p$ gives the desired bound, with probability at least $1 - c_1 \exp(-c_2 \log p)$.

In particular, it follows that

$$N(s) \subseteq \mathcal{C} \subseteq \left\{ t \in V : r_C(s,t) \geq \frac{\kappa}{4} \right\},$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$. Since the last subset has cardinality at most $d^{\frac{\log(4/\kappa)}{\zeta}}$ by the correlation decay condition, we also have $|\mathcal{C}| \leq d^{\frac{\log(4/\kappa)}{\zeta}}$, as claimed.

The remainder of the proof is identical to the proof of Proposition 1, and is a consequence of Theorem 2.

REFERENCES

[1] G. Bresler, E. Mossel, and A. Sly. Reconstruction of Markov random fields from samples: Some observations and algorithms. In *APPROX-RANDOM*, pages 343–356, 2008.
[2] F. H. Clarke. *Optimization and Nonsmooth Analysis*. Wiley-Interscience, New York, 1983.
[3] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.
[4] H. Liu, F. Han, M. Yuan, J.D. Lafferty, and L.A. Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *Annals of Statistics*, 40(4):2293–2326, 2012.

[5] P. Loh and M.J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics*, 40(3):1637–1664, 2012.

[6] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 4:935–980, 2011.

[7] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[8] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, January 2008. ISSN 1935-8237.

Department of Statistics
University of California, Berkeley
Berkeley, California 94720
USA
E-mail: ploh@berkeley.edu
            wainwrig@stat.berkeley.edu