Learning High-Dimensional Latent Graphical Models: Girth-Constrained Graph Families

Animashree Anandkumar*and Elchanan Mossel†

November 29, 2011

Abstract

The problem of structure estimation in discrete latent graphical models is considered, where some nodes are latent or hidden. A novel method, termed as LocalCLGrouping, is proposed which attempts to locally reconstruct latent trees and outputs a loopy graph. Correctness of LocalCLGrouping method is established when the underlying graph has a large girth and the model is in the regime of correlation decay, and PAC guarantees for LocalCLGrouping are also derived. For the special case of the Ising model, the number of samples n required for structural consistency scales as $n = \Omega(\theta_{\min}^{-2\delta\eta(\eta+1)-2} \log p)$, where $\theta_{\min}$ is the minimum edge potential, δ is the depth (i.e., distance from a hidden node to the nearest observed nodes), and η is a parameter which depends on the bounds on node and edge potentials of the Ising model. The results are further specialized for the case when the observed nodes are uniformly sampled from the model. Finally, necessary conditions for structural consistency under any algorithm are derived.

Keywords: Graphical models, latent variables, graph estimation, girth-constrained graph families, quartet methods.

1 Introduction

It is widely recognized that the process of fitting the observed samples to a statistical model needs to incorporate latent or hidden factors, which are not directly observed. Such models are popularly known as *latent variable models*, and they relate a set of observed variables (also known as manifest variables) to a set of latent variables. Learning latent variable models from observed samples involves mainly two tasks: discovering relationships between the observed and hidden variables, and estimating the strength of such relationships.

One of the simplest latent variable model is the so-called *latent class model*, incorporating discrete observed and hidden variables, where the observed variables are assumed to be conditionally independent given the state of the latent factor. However this model has several shortcomings: there may be many different latent factors influencing the observed variables in a complex manner, the assumption of conditional independence of *all* observed variables, given a latent factor, may be too restrictive. The class of *latent tree models*, also known as *hidden Markov tree models*, removes

[*] A. Anandkumar is with the Center for Pervasive Communications and Computing, Electrical Engineering and Computer Science Dept., University of California, Irvine, USA 92697. Email: a.anandkumar@uci.edu

[†] E. Mossel is with the Computer Science and the Statistics Dept., University of California, Berkeley. Email: mossel@stat.berkeley.edu

some of these restrictions. It posits that the hidden and observed variables are Markov on a tree structure. Latent tree models have been effective in modeling data from a variety of domains: the evolutionary process which gave rise to the present-day species in bio-informatics (popularly known as *phylogenetic tree models*) [25], for financial and topic modeling [21], and for modeling contextual information for object recognition in computer vision [20].

A major reason behind the popularity of latent tree models is their computational tractability: inference on test data can be carried out efficiently using distributed algorithms such as *belief propagation*, once a tree model is learnt from training data. There has been extensive work on learning latent tree models (e.g. [29, 42]) including some recent and on-going works (e.g. [3, 21]), where it is demonstrated that latent tree models can be learnt efficiently in *high dimensions*. In other words, the number of samples required for consistent learning is much smaller than the number of variables at hand.

However, despite all the above advantages, latent tree models may not be suitable in some scenarios and the assumption of a tree structure may be too restrictive. In this paper, we relax the tree assumption, while retaining many advantages of the latent tree models. Such models which incorporate Markov relationships on a general graph are popularly known as *probabilistic graphical models* or *Markov random fields*. Relaxing the tree assumption leads to non-trivial challenges: in general, learning these models is NP-hard, even when there are no latent variables, and developing methods for learning such fully observed models is itself an area of active research (e.g. [6, 34, 46]). In this paper, we consider structure estimation in latent graphical models Markov on graphs with a large *girth*, which is the length of the smallest cycle in the graph. Such graphs are *locally tree-like*, meaning that local neighborhoods in the graph do not contain cycles. Considering such graphs leads to non-trivial questions: what are the parameter regimes where these models can be learnt efficiently? Can we adapt the existing latent tree learning algorithms to output loopy graphs? How does the sample complexity of the proposed methods scale with the number of dimensions and the graph attributes (e.g. node degrees)? What are some necessary conditions needed for any algorithm to succeed in learning the graph structure? We provide answers to these questions in this paper.

1.1 Summary of Results

Our main contributions in this work are three-fold. We characterize the class of identifiable latent graphical models Markov on graphs of large girth. We propose an algorithm for structure estimation of such models and derive sample complexity guarantees. We derive necessary conditions for any algorithm to succeed in estimating such graphs.

We consider the class of discrete latent graphical models Markov on graphs with a girth constraint and a minimum degree of at least three. We first characterize the conditions on the model class for identifiability. Note that such a characterization is non-trivial due to the presence of latent variables. We establish identifiability in the regime of correlation decay[1], where the pairwise statistics of the model converge locally to a tree limit.

Although girth-constrained graphs are locally tree-like, in general, their global structure makes them hard instances for learning. Specifically, girth-constrained graphs have a large tree-width: it is known that a graph with average degree at least Δ_{avg} and girth at least g has a tree width

[1] The results of this paper are applicable for certain models even beyond the regime of correlation decay since we only require local convergence of pairwise statistics to the tree limit. For instance, in [24], this property is established for attractive Ising models with strictly positive node potentials.

as $\Omega\left(\frac{1}{g+1}(\Delta_{\text{avg}}-1)^{\lfloor(g-1)/2\rfloor}\right)$ [16]. Thus, learning is non-trivial for graphical models Markov on girth-constrained graphs, even when there are no latent variables.

In this paper, we propose a novel method, termed as LocalCLGrouping, for learning graphical models Markov on girth-constrained graphs in the presence of latent models. This method is an extension of CLGrouping, proposed in [21], for learning latent tree models. CLGrouping proceeds by first fitting all the observed variables to a tree model and then iteratively adding hidden variables to local neighborhoods. Consistency and sample complexity results for CLGrouping are derived in [21]. The LocalCLGrouping method for constructing loopy graphs proceeds as follows: it fits local groups of observed variables to trees and merges these trees to obtain a loopy graph. It then iteratively adds latent variables by operating on local neighborhoods of the constructed graph and running a latent tree routine on them. The LocalCLGrouping method is efficient for practical implementation since it inherits all the advantages of CLGrouping method. Due to the "divide and conquer" feature of LocalCLGrouping, the local latent tree building can be parallelized to obtain speedups. For real datasets, a tradeoff between model complexity and fidelity is typically enforced by optimizing scores such as the Bayesian information criterion (BIC) [48]. Such criteria can be easily enforced by implementing a greedy local search in each iteration of the LocalCLGrouping method.

We provide precise conditions for structural consistency of LocalCLGrouping under the probably approximately correct (PAC) model of learning [36]. We simplify the conditions for the Ising model, where each node is a binary random variable, to obtain better intuitions. We establish that for structural consistency, the number of samples is required to scale as $n = \Omega(\theta_{\min}^{-2\delta\eta(\eta+1)-2}\log p)$, where $\theta_{\min}$ is the minimum edge potential, δ is the depth (i.e., distance from a hidden node to the nearest observed nodes), and η is a parameter which depends on the bounds on node and edge potentials of the Ising model. When there are no hidden variables, we have $\delta = 0$, and the sample complexity $n = \Omega(\theta_{\min}^{-2}\log p)$ matches with the sample complexity of other algorithms for learning fully-observed Ising models [6, 34].

A sufficient condition for LocalCLGrouping method to succeed is for the model to be in the regime of correlation decay. This holds for Ising models when $\theta_{\min} \leq \theta_{\max} < 1/\Delta_{\max}$, where $\Delta_{\max}$ is the maximum degree in the graph. This implies that the best-case sample complexity for LocalCLGrouping scales as $n = \Omega(\Delta_{\max}^{2(2\delta+1)}\log p)$. We further specialize the results when nodes are uniformly sampled for observation. We establish that the best-case sample complexity under uniform sampling scales as $n = \Omega(\Delta_{\max}^2\rho^{-4}(\log p)^5)$, where ρ is the fraction of the observed nodes. Thus, we establish that the sample complexity of learning latent Ising models under uniform sampling by LocalCLGrouping method is comparable to other methods for learning fully-observed Ising models (which is $n = \Omega(\Delta_{\max}^2\log p)$) [6,34], under a constant fraction of observed nodes ($\rho = \Theta(1)$). Thus, we establish that structure estimation is tractable even when the number of hidden variables is much larger than the number of observed variable, i.e., the fraction ρ can decay with the number of nodes. Similar observations were previously made for learning locally tree-like graphs under a different measurement model of path-based measurements for network tomography [4].

We also establish necessary conditions on any (deterministic) algorithm for structure estimation of discrete latent graphical models Markov on girth-constrained graphs. We employ the standard counting argument, which maps the observation space to the space of graphs in the specified class. We establish that $n = \Omega(\Delta_{\min}\rho^{-1}\log p)$ samples are necessary for structural consistency, where $\Delta_{\min}$ is the minimum degree and ρ is the fraction of observed nodes. This is comparable to the requirement of LocalCLGrouping, $n = \Omega(\Delta_{\max}^2\rho^{-4}(\log p)^5)$, achieved under uniform node sampling.

Thus, we undertake a detailed study of structure estimation in latent graphical models. Ours is the first work to provide provable guarantees for structure estimation of loopy graphs in discrete latent models (see next Section for a discussion of related work) and it expands the realm of tractable model classes for structure estimation. The class of girth-constrained graphs is attractive due to computational tractability: inference on such graphs can be performed efficiently in the regime of correlation decay, under consideration in this paper. This implies that similar conditions are required for tractability of learning and inference, as has been observed in other works (see [5, 6]).

1.2 Related Work

Our work is at the intersection of latent tree learning, where latent variables are present but a tree structure is imposed, and graphical model selection, where there are no latent variables, but the graph is allowed to have loops. We first discuss previous works on latent models, and then discuss works dealing with graphical model selection.

The classical *latent cluster models* (LCM) consists of multivariate distributions with a single latent variable and the observed variables are conditionally independent under each state of the latent variable [39]. Hierarchical latent class (HLC) models [19, 52, 53] generalize these models by allowing multiple latent variables. However, the proposed learning algorithms are based on greedy local search in a high-dimensional space, which is computationally expensive. Moreover, the algorithms do not have consistency guarantees. Similar shortcomings also hold for expectation-maximization (EM) based approaches [28, 37]. Learning latent trees has been studied extensively before, mainly in the context of phylogenetics. See [26] for a thorough overview. Efficient algorithms with provable performance guarantees are available (e.g. [23, 29]). In [21], an efficient algorithm CLGrouping for latent tree learning was proposed. Our algorithm LocalCLGrouping for learning large girth graphs is based on CLGrouping method.

Works on high-dimensional graphical model selection are more recent. The approaches can be mainly classified into two groups: non-convex local approaches [6, 12, 34, 44] and those based on convex optimization [17, 40, 46, 47]. There is a general agreement that the success of these methods is related to the presence of correlation decay in the model. See [8] for a detailed discussion and analysis. This work makes the connection explicit: it relates the extent of correlation decay (i.e., the convergence rate to the tree limit) with the learning efficiency for latent models on large girth graphs.

The work in [6] considers learning Ising models with no latent variables. The algorithm is based on a series of conditional independence tests, which is simpler than the LocalCLGrouping algorithm proposed in this paper. The class of graphs considered in [6] is based on sparse local separation which is more general than the class of locally tree-like graphs considered in this paper. Similar conditions are imposed for success in [6] and here, and are based on the regime of correlation decay in the model.

This paper is the first work to provide provable guarantees for learning discrete latent models on loopy graphs (which can also be easily be extended to Gaussian models). The work in [18] considers learning latent Gaussian graphical models using a convex relaxation method, by exploiting a sparse-low rank decomposition of the Gaussian precision matrix. However, the method cannot be easily extended to discrete models. Moreover, the "incoherence" conditions required for the success of convex methods are hard to interpret and verify in general. In contrast, our conditions for success are transparent and based on the presence of correlation decay in the model.

2 System Model

2.1 Graphical Models

A *graphical model* is a family of multivariate distributions which are Markov in accordance to a particular undirected graph [38]. Each node in the graph $i \in W$ is associated to a random variable X_i taking value in a set $\mathcal{X}$. We consider discrete graphical models where $\mathcal{X}$ is a finite set. The set of edges E captures the set of conditional independence relations among the random variables. We say that a set of random variables $\mathbf{X}_W := \{X_i, i \in W\}$ with probability mass function (pmf) P is Markov on the graph G if the local Markov property

$$P(x_i|x_{\mathcal{N}(i)}) = P(x_i|x_{W\setminus i}) \tag{1}$$

holds for all nodes $i \in W$, where $\mathcal{N}(i)$ are the neighbors of node i in graph G. More generally, we say that P satisfies the global Markov property, if for all disjoint sets $A, B \subset W$, we have

$$P(\mathbf{x}_A, \mathbf{x}_B|\mathbf{x}_S) = P(\mathbf{x}_A|\mathbf{x}_S)P(\mathbf{x}_B|\mathbf{x}_S). \tag{2}$$

where the set S is a *separator*[2]between A and B. The local and global Markov properties are equivalent under the *positivity* condition, given by $P(\mathbf{x}_W) > 0$, for all $\mathbf{x}_W \in \mathcal{X}^{|W|}$ [38], and we consider such distributions.

The Hammersley-Clifford theorem [11] states that under the positivity condition, a distribution P satisfies the Markov property according to a graph G iff. it factorizes according to the cliques of G, and we can write it in the exponential form as

$$P(\mathbf{x}) = \exp\left(\sum_{c\in\mathcal{C}} \theta_c(\mathbf{x}_c) - A(\boldsymbol{\theta})\right), \tag{3}$$

where $\mathcal{C}$ is the set of cliques of G and $\mathbf{x}_c$ is the set of random variables on clique c. The quantity $A(\boldsymbol{\theta})$ is known as the *log-partition function* and serves to normalize the probability distribution. The functions θ_c are known as *potential* functions and correspond to the *canonical* parameters of the exponential family.

We limit ourselves to the family of pairwise graphical models, which factorize according to the edges of the graph,

$$P(\mathbf{x}_W) = \exp\left(\sum_{e\in E} \theta_e(\mathbf{x}_e) + \sum_{i\in V} \phi_i(x_i) - A(\boldsymbol{\theta})\right), \tag{4}$$

and θ_e and ϕ_i are the edge and the node potentials.

We assume that the edge potential functions $\theta_{i,j}$ and the node potential functions ϕ_i are bounded,

$$\theta_{\min} \leq |\theta_{i,j}(x_i, x_j)| \leq \theta_{\max}, \quad \forall\, (i,j) \in G \tag{5}$$

$$\phi_{\min} \leq |\phi_i(x_i)| \leq \phi_{\max}, \quad \forall i \in V. \tag{6}$$

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

A pairwise model over binary variables with the probability mass function (pmf) given by

$$P(\mathbf{x}_W) = \exp\left(\sum_{e\in E} \theta_{i,j} x_i x_j + \sum_{i\in V} \phi_i(x_i) - A(\boldsymbol{\theta})\right), \tag{7}$$

is known as the *Ising model.* We specialize some of our results to the class of Ising models.

We consider latent graphical models in which a subset of nodes is latent or hidden. Let $H \subset W$ denote the hidden nodes and $V \subset W$ denote the observed nodes. Our goal is to discover the presence of hidden variables $\mathbf{X}_H$ and learn the unknown graph structure $G(W)$ and parameters of the distribution $P(\mathbf{x}_W)$, given n i.i.d. samples from observed variables $\mathbf{X}_V$. Let $p := |V|$ denote the number of observed nodes and $m := |W|$ denote the total number of nodes.

2.2 Identifiability of Latent Graphical Models

Our goal is to learn the unknown latent graphical model $P(\mathbf{x}_W)$ given samples of the observed variables $\mathbf{X}_V$. Before we can design estimators and provide performance guarantees, a fundamental question to be addressed is the identifiability of model parameters, formally defined below.

Definition 1 (Identifiability) *A parametric model $\{P_\theta : \theta \in \Theta\}$ is identifiable with respect to a measure μ if there do not exist two distinct parameters $\theta_1 \neq \theta_2$ such that $P_{\theta_1} = P_{\theta_2}$ almost everywhere with respect to μ.*

Thus, if a model is not identifiable, this is no hope of estimating the model parameters from the samples. In the context of learning graphical models, we are interested in the identifiability of the graph structure and parameters of the model.

It is straightforward to see that the structure and the parameters of a fully observed discrete graphical model are identifiable under the positivity condition. This is because the distribution belongs to the minimal exponential family under these conditions [13]. In the presence of hidden variables, identifiability of a model is more complicated. In this case, there exists an *equivalence class* of models each of which can explain the observed statistics. Thus, identifiability for latent models is defined as identifiability up to the equivalence class.

Given a set of observed variables, a *minimal* model is defined as an element of the equivalence class with the fewest number of hidden variables. The conditions for (structure and parameter) identifiability of the minimal latent tree model (up to re-labeling of the hidden nodes) are well known [43]: (a) Each hidden node $h \in H$ has degree $\mathrm{Deg}(h) \geq 3$, and (b) the Markov model is non-singular. The conditions for identifiability of more general latent graphical models are however not easy to determine. We provide novel conditions for structure and parameter identifiability for a certain class of latent models in Section 5, given a configuration of observed nodes.

3 Background on Latent Tree Models

We first recap the results on latent tree models which will subsequently extended to more general latent graphical models. It is well known that tree-structured graphical models Markov on a tree $T = (W, E)$ have a special form of factorization in (4) given by

$$P(\mathbf{x}_W) = \prod_{i\in W} P_{X_i}(x_i) \prod_{(i,j)\in T} \frac{P_{\mathbf{X}_{i,j}}(x_i, x_j)}{P_{X_i}(x_i) P_{X_j}(x_j)} \tag{8}$$

Comparing with (4), we note that tree distributions are directly parameterized in terms of pairwise marginal distributions on the edges. Similarly, a Markov model can be described on a rooted directed tree $\overrightarrow{T}$ with root $r \in W$, where the edges of $\overrightarrow{T}$ are directed away from the root. Let $\mathrm{Pa}(i)$ denote the (unique) parent of node $i \neq r$ and $P_{X_i|X_{\mathrm{Pa}(i)}}$ denote the corresponding conditional distribution. The Markov model is given by

$$P(\mathbf{x}_W) = P_{X_r}(x_r) \prod_{i \in W, i \neq r} P_{X_i|X_{\mathrm{Pa}(i)}}(x_i|x_{\mathrm{Pa}(i)}). \tag{9}$$

A Markov model is said to be *non-singular* [43,49] if (a) For all $e \in \overrightarrow{T}$, the conditional distributions satisfy $0 < |\det(P_{X_i|X_{\mathrm{Pa}(i)}})| < 1$ and (b) For all $i \in V$, $P_{X_i}(x) > 0$ for all $x \in \mathcal{X}$. A non-singular Markov model on an undirected tree T and its directed counterpart $\overrightarrow{T}$ are equivalent [43,49]. Note that non-singularity is equivalent to positivity (i.e., bounded potential functions) for Markov tree models. In particular, Ising models on trees with bounded node and edge potentials are non-singular.

Latent tree models or phylogenetic tree models are tree-structured graphical models in which a subset of nodes are hidden or latent. Our goal in this paper is to leverage on the techniques developed for learning latent tree models to analyze a more general class of latent graphical models.

3.1 Learning Latent Tree Models

Learning the structure of latent tree models is an extensively studied topic. A majority of structure learning methods (known as distance based methods) rely on the presence of an *additive tree metric* [25, 45]. The additive tree metric can be obtained by considering the pairwise marginal distributions of a tree structured graphical model. For instance, the work in [42] considers the following metric for discrete distributions satisfying the non-singular condition

$$d(i,j) := -\log|\det(P_{\mathbf{X}_{i,j}})|, \quad \forall i,j \in V. \tag{10}$$

By non-singularity assumption, we have that $|\det(P_{\mathbf{X}_{i,j}})| > 0$ for all $i,j \in W^2$. The distance metric further simplifies for some special distributions, e.g. for symmetric Ising models, it is given by the negative logarithm of the correlation between the node pair under consideration [21].

3.1.1 Quartet Based Methods

A popular class of learning methods are based on the construction of *quartets* (e.g., [14, 29, 42]), and various procedures to merge the inferred quartets. A quartet is a structure over four observed nodes, as shown in Fig.1. We now recap the classical quartet test operating on any additive tree metric [25, 45]. The path structure refers to the configuration of paths between the given nodes.

Definition 2 (Quartet or Four-Point Condition on Trees) *Given an additive metric $[d(i,j)]_{i,j \in V}$ on a tree, the tuple of four nodes $a, b, u, v \in V$ has the path structure in Fig.1 iff.*

$$d(a,b) + d(u,v) < \min(d(a,u) + d(b,v), d(b,u) + d(a,v)), \tag{11}$$

and the structure in Fig.1 is denoted by $Q(ab|uv)$.

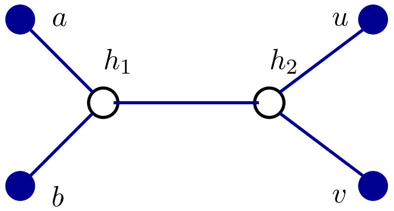


Figure 1: Quartet $Q(ab|uv)$. See (11).

Algorithm 1 $\mathsf{Quartet}(\widehat{\mathbf{d}}^n(V), \Lambda)$ test using distance estimates $\widehat{\mathbf{d}}^n(V) := \{\widehat{d}(i,j)\}_{i,j \in V}$ and confidence interval Λ.

```
Input: Distance estimates between the observed nodes $\widehat{\mathbf{d}}^n(V) := \{\widehat{d}(i,j)\}_{i,j \in V}$ and confidence interval $\Lambda$.
Initialize set of quartets $\mathcal{Q}(V) \leftarrow \emptyset$.
for $\{i, j, i', j'\} \in V$ do
  if $(e^{-\widehat{d}(i,j)} - \Lambda)_+ (e^{-\widehat{d}(i',j')} - \Lambda)_+ > (e^{-\widehat{d}(i,j')} + \Lambda)_+ (e^{-\widehat{d}(i,j)} + \Lambda)_+$ then
    Declare Quartet: $\mathcal{Q}(V) \leftarrow Q(ij|i'j')$.
  end if
  if No quartet declared for $\{i, j, i', j'\}$ then
    $\perp_{i,j,i',j'}$ (Declare null).
  end if
end for
```

It is well known that the set of all quartets uniquely characterize a latent tree. In [29], it was shown that a subset of quartets, termed as *representative quartets*, suffices to uniquely characterize a latent tree. The set of representative quartets consists of one quartet for each edge in the latent tree with shortest distances between the observed nodes. In Section 5, we use this characterization to establish asymptotic identifiability of more general latent graphical models.

3.1.2 Recursive Grouping

We recap the recursive grouping $\mathsf{RG}(\widehat{\mathbf{d}}^n(V), \Lambda, \tau)$ method proposed in [21] (and its refinement in [3]). The method is based on a robust[3] quartet test $\mathsf{Quartet}(\widehat{\mathbf{d}}^n, \Lambda)$ given in Algorithm 1. If the confidence bound is not met, a $\perp$ result is declared. In the first iteration of RG, the algorithm searches for node pairs which occur on the same side of all the quartets, output by the quartet test $\mathsf{Quartet}(\widehat{\mathbf{d}}^n, \Lambda)$ and declares them as siblings and introduces hidden variables. In later iterations of RG, sibling relationships between hidden variables are inferred through quartets consisting of their children. Finally, weak edges are merged and a tree (and more generally a forest) is output. We later use a modified version of recursive grouping method as a routine in our algorithm for estimating locally tree-like graphs. In the end, the neighboring nodes (at least one of which is hidden) are merged based on the threshold τ. See Section 6 for details.

[3] Denote $(\cdot)_+ := \max(\cdot, 0)$.

Algorithm 2 $\mathsf{RG}(\widehat{\mathbf{d}}^n(V), \Lambda, \tau)$ test using distance estimates $\widehat{\mathbf{d}}^n(V) := \{\widehat{d}(i,j)\}_{i,j \in V}$, confidence interval Λ and threshold τ for merging nodes.

```
Input: Distance estimates between the observed nodes $\widehat{\mathbf{d}}^n(V) := \{\widehat{d}(i,j)\}_{i,j \in V}$, confidence interval $\Lambda$ and threshold $\tau$. Let $\mathcal{C}(a)$ denote the children of node $a$.
Initialize $A \leftarrow V$, $\mathcal{C}(i) \leftarrow \{i\}$ for all $i \in V$ and $\mathcal{Q}(V) \leftarrow \mathsf{Quartet}(\widehat{\mathbf{d}}^n(A), \Lambda)$.
while $A \neq \emptyset$ do
  if $\exists\ i, j \in A$ s.t. for each $a \in \mathcal{C}(i)$ and $b \in \mathcal{C}(j)$, $a, b$ are on same side of all quartets in $\mathcal{Q}(V)$ then
    Declare $i, j$ as siblings and introduce hidden node $h$ as parent and $\mathcal{C}(h) \leftarrow \mathcal{C}(i) \cup \mathcal{C}(j)$.
    Remove $i, j$ from $A$ and add $h$ to $A$.
  else
    Sibling relationships cannot be further inferred. Break.
  end if
end while
Form forest $\widehat{T}$ based on sibling and child/parent relationships.
Merge edges in $\widehat{T}$ of length less than $\tau$ and output $\widehat{T}$.
```

3.1.3 Chow-Liu Grouping

An alternative method, known as *Chow-Liu grouping* (CLGrouping), was proposed in [21]. Although the theoretical results for CLGrouping are similar to earlier results (e.g. [29]), experiments on both synthetic and real data sets revealed significant improvement over earlier methods in terms of likelihood fitting and number of hidden variables added.

The CLGrouping method is summarized in Algorithm 3. The CLGrouping method always maintains a candidate tree structure and progressively adds more hidden nodes in local neighborhoods. The initial tree structure is the *minimum spanning tree* (MST) over the observed nodes with respect to the tree metric. The method then considers neighborhood sets on the MST and constructs local subtrees (using quartet based method or any other tree reconstruction algorithm). This local reconstruction property of CLGrouping makes it especially attractive for reconstructing girth-constrained graphs. In this paper, we employ a modified version of Chow-Liu grouping, termed as LocalCLGrouping. See Section 6 for details.

4 Characterization of Tractable Latent Graphical Models

In general, structure estimation of graphical models is NP-hard [10, 35]. We now characterize a tractable class of models for which we can provide guarantees on graph estimation.

4.1 Tractable Graph Families: Girth-Constrained Graphs

We consider the family of graphs with a bound on the *girth*, which is the length of the shortest cycle in the graph. Let $\mathcal{G}_{\mathrm{Girth}}(m; g)$ denote the ensemble of graphs with girth at most g. There are many graph constructions which lead to a bound on girth. For example, the bipartite Ramanujan graph [22, p. 107] and the random Cayley graphs [31] have bounds on the girth. Recently, efficient algorithms have been proposed to generate large girth graphs efficiently [7].

Algorithm 3 CLGrouping($\widehat{\mathbf{d}}^n(V), \Lambda, \tau$) for graph estimation using distance estimates $\widehat{\mathbf{d}}^n(V) := \{\widehat{d}(i,j)\}_{i,j\in V}$, confidence interval Λ and threshold τ.

```
*Input:* Distance estimates between the observed nodes $\widehat{\mathbf{d}}^n(V) := \{\widehat{d}(i,j)\}_{i,j\in V}$, confidence interval $\Lambda$ and threshold $\tau$. Let $\mathrm{MST}(V;\widehat{\mathbf{d}}^n)$ denote the minimum spanning tree over $V$ according to the metric $\widehat{\mathbf{d}}^n(V)$. Given a tree $T$, let $\mathrm{Leaf}(T)$ denote the set of leaves. Let $\mathcal{N}[i;T]$ denote the closed neighborhood of node $i$ in tree $T$.
Initialize $\widehat{T} \leftarrow \mathrm{MST}(V;\widehat{\mathbf{d}}^n)$.
for $v \in V \setminus \mathrm{Leaf}(\widehat{T})$ do
    $A \leftarrow \mathcal{N}[v;\widehat{T}]$.
    $S \leftarrow \mathsf{RG}(\widehat{\mathbf{d}}^n(A), \Lambda, \tau)$.
    $\widehat{T}(A) \leftarrow S$ (Replace subtree over $A$ with $S$ in $\widehat{T}$)
end for
Output $\widehat{T}$.
```

Although girth-constrained graphs are locally tree-like, in general, their global structure makes them hard instances for learning. Specifically, girth-constrained graphs have a large tree-width: it is known that a graph with average degree at least Δ_{avg} and girth at least g has a tree width as $\Omega\left(\frac{1}{g+1}(\Delta_{\mathrm{avg}}-1)^{\lfloor (g-1)/2\rfloor}\right)$ [16]. Thus, learning is non-trivial for graphical models Markov on girth-constrained graphs, even when there are no latent variables.

4.2 Local Convergence to a Tree Limit

This work establishes tractable learning when the graphical model converges locally to a tree limit. A sufficient condition for the existence of such limits is the regime of *correlation decay*[4], which refers to the property that there are no long-range correlations in the model [32, 41, 51]. This regime is also known as the *uniqueness regime* since under such an assumption, the marginal distribution at a node is asymptotically independent of a growing boundary.

We tailor the definition of correlation decay to node neighborhoods and provide the definition below. Given a graph $G=(W,E)$ and a graphical model $P_{\mathbf{X}_W|G}$ Markov on it, and any subset $A \subset W$, let $P_{\mathbf{X}_A|G}$ denote the marginal distribution of variables in A. For some subgraph $F \subset G$, let $P_{\mathbf{X}_A|F}$ denote the marginal distribution on A corresponding to a graphical model Markov on graph F instead of G (i.e., by setting the edge potentials of edges in $G \setminus F$ to zero). Let $\mathcal{N}[i;G] := \mathcal{N}(i;G) \cup i$ denote the closed neighborhood of node i in G. For any two sets $A_1, A_2 \subset W$, let $\mathrm{dist}(A_1,A_2) := \min_{i\in A_1, j\in A_2} \mathrm{dist}(i,j)$ denote the minimum graph distance[5]. Let $B_l(i)$ denote the set of nodes within graph distance l from node i and $\partial B_l(i)$ denote the boundary nodes, i.e., exactly at l from node i. Let $F_l(i;G) := G(B_l(i))$ denote the induced subgraph on $B_l(i)$. For any distributions P, Q, let $\|P-Q\|_1$ denote the total variation distance.

Definition 3 (Correlation Decay) *A graphical model $P_{\mathbf{X}_{W_m}|G}$ Markov on graph $G_m = (W_m, E_m)$*

[4] Technically, correlation decay can be defined in multiple ways [41, p. 520] and the notion we use is the uniqueness or the extremality condition.

[5] We distinguish between the terms *graph distance* and *distances*. The former refers to the least number of hops on the graph, while the latter refers to the quantity in (19).

is said to exhibit correlation decay with a non-increasing rate function $\zeta_m(\cdot) > 0$ if for all $l, m \in \mathbb{N}$,

$$\max_{\substack{i \in W_m \\ A \subset B_l(i)}} \|P_{\mathbf{X}_A|G_m} - P_{\mathbf{X}_A|F_l(i;G_m)}\|_1 = \zeta_m(\text{dist}(A, \partial B_l(i))). \tag{12}$$

In words, the total variation distance between the marginal distribution of a set A of a graphical model Markov on G_m and the corresponding model Markov on subgraph $F_l(i; G_m)$ decays as a function of the graph distance to the boundary. This implies that for a class of functions ζ, the effect of graph configuration beyond l hops from any node i has a decaying effect on the local marginal distributions. A sufficient set of conditions for correlation decay are given by Dobrushin's conditions [32].

For the class of Ising models in (7), the regime of correlation decay can be explicitly characterized, in terms of the maximum edge potential $\theta_{\max}$ of the model. Define

$$\alpha := \frac{\tanh \theta_{\max}}{\tanh \theta^*}, \tag{13}$$

where θ^* is a threshold that can be explicitly characterized for certain graph families.

Lemma 1 (Ising Models) *The class of Ising models is in the regime of correlation decay when $\theta_{\max} < \theta^*$ (or $\alpha < 1$) and the threshold θ^* satisfies*

$$\theta^* > \text{atanh}\left(\frac{1}{\Delta_{\max}}\right), \tag{14}$$

where $\Delta_{\max}$ is the maximum degree in the graph. The rate function $\zeta_m(\cdot)$ for correlation decay in (12) is given by

$$\zeta_m(l) = 2\alpha^l, \quad \forall l \in \mathbb{N}. \tag{15}$$

The threshold θ^* can be improved for random graph families. For instance, $\theta^* = \text{atanh}\,\frac{1}{c}$ for Erdős-Rényi random graphs with average degree c. See [6] for details.

5 Identifiability of Latent Models

We now analyze the task structure estimation of latent graphical models when the underlying graph has a bound the girth. As discussed in Section 2.2, we have to first establish identifiability of latent graphical models under the above assumptions. We establish that the latent graphical model is asymptotically identifiable in the regime of correlation decay subject to the following assumptions.

5.1 Sufficient Conditions for Identifiability

Below we list the set of sufficient conditions on the model parameters, the graph structure and the configuration of observed variables in the model under which we can guarantee identifiability of latent graphical models.

5.1.1 Assumptions on the Model Parameters

We now impose constraints on the parameters of the pairwise latent graphical model.

$(A1)$ **Positivity:** The edge potential functions $\theta_{i,j}$ and the node potential functions ϕ_i of the pairwise graphical model are bounded.

$(A2)$ **Regime of Correlation Decay:** The graphical model satisfies correlation decay according to (12) in Definition 3, i.e.,

$$\max_{\substack{i \in W_m \\ A \subset B_l(i)}} \|P_{\mathbf{X}_A|G_m} - P_{\mathbf{X}_A|F_l(i;G_m)}\|_1 = \zeta_m(\mathrm{dist}(A, \partial B_l(i))). \tag{16}$$

with function ζ_m satisfying[6]

$$\zeta_m(\omega(1)) = o(1). \tag{17}$$

The assumption $(A1)$ is a natural requirement for the identifiability of the model and is also needed for the identifiability of fully observed graphical models. For latent tree models, this corresponds to the notion of non-singularity [43]. The assumption $(A2)$ on correlation decay enables us to bound the effect of long-range cycles and thereby exploit the locally tree-like property in girth-constrained graphs.

5.1.2 Assumptions on the Graph Structure

The following are the structural assumptions:

$(A3)$ **Minimum Degree:** The minimum degree of any hidden node in the graph is $\mathrm{Deg}_{\min}(H) \geq 3$.

$(A4)$ **Girth Constraint:** The family of graphs has a girth $g = \omega(1)$ with respect to the number of nodes.

$(A5)$ **Depth Constraint:** We require that the observed nodes $V \subset W$ be such that the depth δ (defined below in Definition 4) satisfies

$$\frac{g}{2} - \delta = \omega(1), \tag{18}$$

as the number of nodes $m \to \infty$.

The assumption on minimum degree in $(A3)$ is required for identifiability of hidden variables and note that it is identical to the requirement for latent tree models. The assumption on the girth in $(A4)$ helps us exploit the locally tree-like structure for learning.

We now give the definition of the depth. Recall that the notion of quartet[7] was introduced in Section 3.1 and represented in Fig.1.

Definition 4 (Depth of a Graph Configuration) *Given a graph $G = (W, E)$ and a set of observed nodes $V \subset W$, the* representative *quartet of each hidden edge $e \in E$ is given by the quartet whose largest graph distance between the endpoints (which are observed nodes) is minimized and the edge e occurs as the middle edge [29]. The* depth *of a graph configuration is given by the maximum graph distance between the endpoints in the set of representative quartets of the graph.*

[6] The notations $\Omega(\cdot), O(\cdot), \omega(\cdot), o(\cdot)$ are with respect to the number of nodes m.

[7] Although the notion of quartets was introduced in the context of latent trees, we can consider similar configuration on general graphs by considering path disjoint paths to four observed nodes.

5.2 Results on Identifiability

Theorem 1 (Identifiability of Latent Graphical Models) *Under the assumptions* $(A1)$*-*$(A5)$*, the graph structure* $G = (W, E)$ *of a latent graphical model* $P(\mathbf{x}_W)$ *with observed variables* $\mathbf{X}_V$ *has structural and parameter identifiability as the number of nodes* $m \to \infty$*.*

Remarks:

1. As discussed in Section 2.2, the identifiability of the graph is up-to relabeling of hidden nodes.

2. The notion of asymptotic identifiability is more restrictive than the usual notion of identifiability since it only guarantees identifiability as the number of nodes goes to infinity. We relax this condition and provide non-asymptotic conditions for structure identifiability in the next section, in addition to development of methods for structure estimation.

3. The assumption of correlation decay $(A2)$ is crucial to establish the above result, since we can limit the effect of faraway nodes and exploit the locally tree-like behavior of the graph. Identifiability beyond the regime of correlation decay for general models remains an open question.

Proof: The proof is given in Appendix A. We first derive the identifiability condition for local neighborhoods by limiting the structure to a certain distance. Under correlation decay, the effect of nodes beyond this boundary decays and thus, the structure and the parameters of the entire model are asymptotically identifiable. □

6 Method and Guarantees for Structure Estimation

6.1 Overview of LocalCLGrouping Algorithm

We now describe our algorithm, termed as local Chow-Liu grouping (LocalCLGrouping), for structure estimation of latent graphical models Markov on girth-constrained graphs. The algorithm leverages on the Chow-Liu grouping algorithm developed for latent tree models [21], described in Section 3.1. The main intuition for learning a girth-constrained graph is based on reconstructing "local" parts of the graph which are acyclic and piecing them together. However, this approach has many challenges. First, it is not clear if the local acyclic pieces can be learnt efficiently since it requires the presence of an additive tree metric. This is addressed by considering models satisfying correlation decay (see Section 4.2). Second and a harder challenge involves merging the reconstructed local latent trees with provable guarantees due to the introduction of unlabeled latent nodes in different pieces. We circumvent this challenge by leveraging on the Chow-Liu grouping algorithm [21] and merging the different pieces before introducing the latent nodes.

The LocalCLGrouping algorithm is described in Algorithm 4. Let $\widehat{d^n}(i,j)$ denote the estimated distance between i and j according to (10) using the empirical distribution $\widehat{P}^n_{\mathbf{X}_{i,j}}$ computed using n samples, i.e.,

$$\widehat{d^n}(i,j) := -\log|\det(\widehat{P}^n_{\mathbf{X}_{i,j}})|, \quad \forall i,j \in V. \tag{19}$$

The set of distance estimates $\widehat{\mathbf{d}}^n(V) := \{\widehat{d^n}(i,j) : i,j \in V\}$ are input to the algorithm along with a parameter r. For each observed node $i \in V$, the set of nodes $B_r(i; \widehat{\mathbf{d}}^n(V))$ is considered and the minimum spanning tree is constructed. The graph estimate $\widehat{G}^n$ is initialized by taking the union of

Algorithm 4 $\mathsf{LocalCLGrouping}(\widehat{\mathbf{d}}^n(V), \Lambda, \tau, r)$ for graph estimation using distance estimates $\widehat{\mathbf{d}}^n(V) := \{\widehat{d}(i,j)\}_{i,j\in V}$, confidence interval Λ, threshold τ and distance parameter r.

```
Input: Distance estimates between the observed nodes $\widehat{\mathbf{d}}^n(V) := \{\widehat{d}(i,j)\}_{i,j\in V}$, confidence parameter $\Lambda$, threshold $\tau$ and bound $r$ on distances used for local reconstruction. Let $B_r(v;\widehat{\mathbf{d}}^n)$ denote the set of nodes within distance $r$ from $v$, according to the metric $\widehat{\mathbf{d}}^n(V)$. Let $\mathrm{MST}(A;\widehat{\mathbf{d}}^n)$ denote the minimum spanning tree over $A \subset V$ according to the metric $\widehat{\mathbf{d}}^n(A)$. Given a graph $G$, let $\mathrm{Leaf}(G)$ denote the set of nodes with unit degree. Let $\mathcal{N}[i;G]$ denote the closed neighborhood of node $i$ in graph $G$. LatentTreeAlgo$(A, \widehat{\mathbf{d}}^n)$ represents any latent tree learning method over the set of nodes $A$ using distance estimates $\widehat{\mathbf{d}}^n(A)$.
for $v \in V$ do
  $T_v \leftarrow \mathrm{MST}(B_r(v);\widehat{\mathbf{d}}^n)$.
end for
Initialize $\widehat{G}, \widehat{G}_0 \leftarrow \cup_v T_v$.
for $v \in V \setminus \mathrm{Leaf}(\widehat{G}_0)$ do
  $A \leftarrow \mathcal{N}[v;\widehat{G}]$.
  $S \leftarrow \mathsf{RG}(\widehat{\mathbf{d}}^n(A), \Lambda, \tau)$.
  $\widehat{G}(A) \leftarrow S$ (Replace subgraph over $A$ with $S$ in $\widehat{G}$)
end for
Output $\widehat{G}$.
```

all the local minimum spanning trees. The latent nodes are now iteratively added by considering local neighborhoods of $\widehat{G}$ and using any latent tree algorithm for reconstruction (e.g. [21,42]). If a local latent tree cannot be reconstructed in any iteration, the algorithm reports failure and stops. Note that the running time is polynomial (in the number of nodes) as long as polynomial time algorithms are employed for local latent tree reconstruction.

The LocalCLGrouping method is efficient for practical implementation since it inherits all the advantages of CLGrouping method for learning latent trees. Due to the "divide and conquer" feature of LocalCLGrouping, the local latent tree building can be parallelized to obtain speedups. For real datasets, a tradeoff between model complexity and fidelity is typically enforced by optimizing scores such as the Bayesian information criterion (BIC) [48]. Such criteria can be easily enforced through a greedy local search in each iteration of the LocalCLGrouping method.

6.2 Sufficient Conditions for Consistency of LocalCLGrouping

6.2.1 Definition of Local Tree Metric

We first define the notion of a local tree metric $\mathbf{d}_{\mathrm{SP}}(V)$ computed by considering only the shortest paths between the respective node pairs. Given a graph $G = (W, E)$, let $\mathrm{SP}(i,j;G)$ denote the shortest path between i and j (according to graph distance) in G. Recall that $P_{\mathbf{X}_{i,j}|G}$ denotes the pairwise marginal distribution between i and j induced by the graphical model $P(\mathbf{x}_W)$ on graph $G(W)$. Let $P_{\mathbf{X}_{i,j}|\,\mathrm{SP}(i,j)}$ denote the pairwise marginal distribution between i and j induced by considering only the subgraph $\mathrm{SP}(i,j;G) \subset G$. Denote

$$d(i,j;\mathrm{SP}) := -\log|\det P_{\mathbf{X}_{i,j}|\,\mathrm{SP}(i,j)}|. \tag{20}$$

$$d(i,j;G) := -\log|\det P_{\mathbf{X}_{i,j}|G}|. \tag{21}$$

Denote $\mathbf{d}_{\mathrm{SP}}(V) := \{d(i,j;\mathrm{SP}) : i,j \in V\}$ and $\mathbf{d}(V) := \{d(i,j;G) : i,j \in V\}$. Note that for loopy graphs in general, $d(i,j;G)$ is different from $d(i,j;\mathrm{SP})$. The learner has access only to the empirical versions $\widehat{\mathbf{d}}(V)$ of the distances $\mathbf{d}(V)$. However, the performance of the algorithm depends on the local tree metric $\mathbf{d}_{\mathrm{SP}}(V)$ and we list the relevant assumptions before.

6.2.2 Conditions on the Model Parameters

($B1$) **Non-Singularity:** Given a graphical model $P_{\mathbf{X}_W|G}$ Markov on graph G, the pairwise marginal distribution $P_{\mathbf{X}_{i,j}|\,\mathrm{SP}(i,j)}$ between any two neighbors $(i,j) \in G$ are non-singular and the distances $d(i,j;\mathrm{SP}) := -\log|\det P_{\mathbf{X}_{i,j}|\,\mathrm{SP}(i,j)}|$ satisfy

$$0 < d_{\min} \leq d(i,j;\mathrm{SP}) \leq d_{\max} < \infty, \quad \forall (i,j) \in G(W), \tag{22}$$

for suitable parameters $d_{\min}$ and $d_{\max}$.

($B2$) **Regime of Correlation Decay:** The pairwise statistics of the graphical model converge locally to a tree limit according to the definition in Definition 3 with function $\zeta_m(\cdot)$ in (12) satisfying

$$\zeta\left(\frac{g}{2} - \frac{r}{d_{\min}} - 1\right) < \frac{\upsilon}{|\mathcal{X}|^2}, \tag{23}$$

where g is the girth, r is the distance bound parameter in LocalCLGrouping, $|\mathcal{X}|$ is the dimension of each variable, $d_{\min}$ is the distance bound in (22) and

$$\upsilon := \min\left(0.5 e^{d_{\min} - r}(1 - e^{-d_{\min}}), e^{-0.5 d_{\max}(\frac{r}{d_{\min}}+2)}\right). \tag{24}$$

The above condition ($B1$) is a stronger requirement compared to bounded potentials (($A1$) in Section 5.1) needed for the identifiability of the model. The parameters $d_{\min;\mathrm{SP}}$ and $d_{\max;\mathrm{SP}}$ depend on the bounds on the edge potentials $\theta_{\min}$ and $\theta_{\max}$ and the node potentials $\phi_{\min}$ and $\phi_{\max}$ in (5) and (6). In particular, our formulation allows for models with pairwise marginal independence. The assumption ($B2$) on correlation decay coincides with ($A2$) in Section 5.1, but imposes an additional constraint on the rate function $\zeta(\cdot)$, in terms of the girth of the graph.

6.2.3 Conditions on the Graph Structure

The following are the structural assumptions:

($B3$) **Minimum Degree:** The minimum degree of any hidden node in the graph is $\mathrm{Deg}_{\min}(H) \geq 3$.

($B4$) **Girth Constraint:** The family of graphs has a girth $g = \omega(1)$ with respect to the number of nodes.

($B5$) **Depth Constraint:** We require that the observed nodes $V \subset W$ be such that the depth δ (defined in Definition 4) satisfies

$$\frac{g}{4} d_{\min} - \delta\left(\frac{d_{\max}}{d_{\min}} + 1\right) d_{\max} = \omega(1), \tag{25}$$

where $d_{\min}$ and $d_{\max}$ are bounds according to (22).

Note that the conditions ($B3$) and ($B4$) also coincide with the conditions ($A3$) and ($A4$) employed for identifiability. The condition ($B5$) is slightly stronger compared to ($A5$), meaning that we require the depth to be smaller to establish consistency under LocalCLGrouping.

6.2.4 Conditions on Input Parameters to LocalCLGrouping

Additionally, the input to LocalCLGrouping have to satisfy the following conditions for establishing consistency.

($B6$) **Distance Threshold:** The parameter r in LocalCLGrouping algorithm for building local minimum spanning trees is chosen as

$$r > \delta\left(\frac{d_{\max}}{d_{\min}}+1\right)d_{\max}, \quad \frac{g}{4}d_{\min}-r=\omega(1) \tag{26}$$

where g is the girth of the graph, and $d_{\min}, d_{\max}$ are the bounds on the distances between the neighbors in G.

($B7$) **Confidence Interval for Quartet Test:** The confidence interval in Quartet($\widehat{\mathbf{d}}, \Lambda$) routine in Algorithm 1 is chosen as

$$\Lambda = \exp\left[-\frac{d_{\max}}{2}(\frac{r}{d_{\min}}+2)\right]. \tag{27}$$

($B8$) **Threshold for Merging Nodes:** The threshold τ in RG($\widehat{\mathbf{d}}, \Lambda, \tau$) routine in Algorithm 2 is chosen as

$$\tau = \frac{d_{\min}}{2} - |\mathcal{X}|^2\zeta_m(\frac{g}{2}-1) > 0, \tag{28}$$

where $|\mathcal{X}|$ is the dimension of the variable at each node and $\zeta_m(\cdot)$ is the correlation decay function according to (12).

Intuitively, the constraint in ($B6$) on the distance parameter r used by LocalCLGrouping algorithm implies that r is relatively small compared to the girth of the graph and large enough for every hidden node to be discovered. This enables the LocalCLGrouping algorithm to correct reconstruct latent trees locally. The confidence interval constrain in ($B7$) is based on the concentration bounds for the empirical distances. The threshold for merging nodes in ($B8$) ensures that spurious hidden nodes are not added.

6.3 Guarantees for LocalCLGrouping Algorithm

We now establish that the LocalCLGrouping algorithm is structurally consistent under the above conditions.

Theorem 2 (Structural Consistency of LocalCLGrouping) *Under assumptions ($B1$)-($B8$), the* LocalCLGrouping *algorithm is structurally consistent with probability at least $1-\kappa$, for any $\kappa > 0$, when the number of samples n available for learning satisfies*

$$n > \frac{2|\mathcal{X}|^2}{(\upsilon - |\mathcal{X}|^2\zeta_m(\frac{g}{2}-\frac{r}{d_{\min}}-1))^2}\left(4\log p + |\mathcal{X}|\log 2 - \log\frac{\kappa}{7}\right), \tag{29}$$

where υ is given by (24)*.*

Remarks:

1. Thus, we provide PAC guarantees for reconstructing latent graphical models on girth-constrained graphs. The conditions required for success are mild and transparent, and on lines of the conditions required for learning latent tree models.

2. The conditions imposed on the girth of the graph are relatively mild. We require that the girth be roughly larger than the depth ($B5$) and that the correlation decay function $\zeta_m(\cdot)$ be sufficiently strong ($B2$). Thus, learning girth-constrained graphs is akin to learning latent tree models (in terms of sample and computational complexities) under a wide range of conditions.

3. One notable additional condition required for learning girth-constrained graphs in contrast to latent trees is the requirement of correlation decay ($B2$). However, we note that this is only a sufficient condition, and not necessary for learnability. For instance, the result in [24] establishes that the pairwise statistics converges locally to a tree limit for all attractive Ising models with strictly positive node potentials, but without any additional constraints on the parameters. Our results and analysis hold in such scenarios as well since we only require local convergence to a tree metric.

4. The results above are applicable for discrete models but can be extended to Gaussian models by using the notion of *walk-summability* in place of correlation decay according to (12) (see [5]) and the negative logarithm of correlation coefficient as the distance metric (see [21]). The results can also be extended to more general linear models such as multivariate Gaussian model, Gaussian mixtures and so on, on lines of [3].

Proof: The detailed proof is given in Appendix B. It consists of the following main steps:

1. We first prove correctness of LocalCLGrouping under the tree limit (i.e., distances $\mathbf{d}_{\rm SP}(V) := \{d(i,j;{\rm SP})\}_{i,j\in V}$) and then show sample-based consistency. The latter is based on concentration bounds, along the lines of analysis for latent tree models [29, 42], with an additional distortion introduced due to the presence of a loopy graph.

2. We now briefly describe the proof establishing the correctness of LocalCLGrouping algorithm under $\mathbf{d}_{\rm SP}$ in girth-constrained graphs. Intuitively, the distances $d(i,j;{\rm SP})$ correspond to a tree metric when the graph distance ${\rm dist}(i,j) < g/2$, where g is the girth. Since LocalCLGrouping infers latent trees only locally, it avoids running into cycles and thus correctly reconstructs the local latent trees. The initialization step in LocalCLGrouping corresponds to the correct merge of this local latent trees under the assumptions on parameter r in (26) and the correctness of LocalCLGrouping is established.

□

6.3.1 Results for Ising Models

We now specialize the results for Ising models. The assumption ($B1$) on distance bounds (based on shortest-path distances) $d_{\min}$ and $d_{\max}$ can now be expressed in terms of bounds on edge potentials $\theta_{\min}$ and $\theta_{\max}$, and node potentials $\phi_{\min}$ and $\phi_{\max}$ as

$$e^{-d_{\min}} \leq \frac{2\sinh 2\theta_{\max}}{Z(\theta_{\max},\phi_{\min},\phi_{\max})},\ e^{-d\max} \geq \frac{2\sinh 2\theta_{\min}}{Z(\theta_{\min},\phi_{\min},\phi_{\max})}, \tag{30}$$

where $Z(\theta, \phi_1, \phi_2)$ is the partition function for the Ising model between a pair of nodes with node potentials ϕ_1 and ϕ_2 and edge potential θ. Thus $e^{-d_{\min}} = O(\theta_{\max})$ and $e^{-d_{\max}} = \Omega(\theta_{\min})$.

The assumption $(B2)$ on correlation decay holds when $\theta_{\max} < \Delta_{\max}^{-1}$, where $\Delta_{\max}$ is the maximum degree in the graph, and the rate function $\zeta(\cdot)$ is given by

$$\zeta(l) = 2\alpha^l, \;\; \alpha := \Delta_{\max} \tanh \theta_{\max} < 1. \tag{31}$$

Eqn. (23) holds when the minimum edge potential $\theta_{\min}$ is sufficiently large compared to a function of the girth g, radius r, and the rate α for convergence to the tree limit. A sufficient condition is

$$\theta_{\min} = \Omega\left(\exp\left[0.5r\frac{d_{\max}}{d_{\min}} - (\frac{g}{2} - \frac{r}{d_{\min}})|\log \alpha|\right]\right). \tag{32}$$

Note that similar (but weaker) conditions are imposed in [6] for recovery of Ising models with samples obtained from all the nodes. For the case of Ising models, Theorem 2 simplifies as follows.

Corollary 1 (Guarantees for Ising Models) *Under the assumptions $(B1)$-$(B8)$, the probability that* LocalCLGrouping *method is structurally consistent tends to one, when the number of samples scales as*

$$n = \Omega\left(\theta_{\min}^{-2\left(\delta\frac{d_{\max}}{d_{\min}}\left(1+\frac{d_{\max}}{d_{\min}}\right)+1\right)} \log p\right). \tag{33}$$

Proof: From Theorem 2, we have structural consistency when $n = \Omega(\upsilon^{-2} \log p)$, where υ is given in $(B2)$. We have $\upsilon = \exp[\min(-r, -0.5d_{\max}(r/d_{\min} + 2))]$. Using the constraint on r in $(B6)$ and the fact that $e^{d_{\max}} = \Theta(\theta_{\min})$, we have the result. □

1. Thus, for learning Ising models on girth-constrained graphs, we establish that the sample complexity is dependent both on the minimum edge potential $\theta_{\min}$ and on the depth of the graph δ. For the special case when all the nodes are observed, we have $\delta = 0$, and the sample complexity is $n = \Omega(\theta_{\min}^{-2} \log p)$. This matches the best known sample complexity for learning fully observed Ising models (see [6, 34]). We note that the sample complexity scales exponentially in the depth δ, which matches the sample complexity bounds of algorithms for learning latent tree models [21, 29, 42].

2. For Ising models to satisfy $(B2)$, a sufficient condition is to be in the regime of correlation decay which requires that

   $$\theta_{\min} \le \theta_{\max} < \Delta_{\max}^{-1},$$

   where $\Delta_{\max}$ is the maximum degree in the graph. Thus, the best-case sample complexity for LocalCLGrouping is achieved when the edge potentials are homogeneous ($\theta_{\min} = \theta_{\max}$) and is given by

   $$n = \Omega(\Delta_{\max}^{2(2\delta+1)} \log p). \tag{34}$$

   Thus, the sample complexity is dependent both on the maximum degree[8] $\Delta_{\max}$ and the depth δ of the graph.

[8] Note that for girth constrained graphs, the maximum degree cannot be arbitrarily large. From [27], we have the so-called *Moore bound* (see [2] for a better bound): $\Delta_{\max} \le \lfloor \frac{m}{2} \rfloor^{\frac{2}{g-2}}$, for an m-node graph.

6.3.2 Guarantees under Uniform Sampling

We have so far given guarantees for graph reconstruction, given an arbitrary sampling of nodes in the graph. We now specialize the results to the case when there is a uniform sampling probability of picking a node as an observed node and provide learning guarantees. This analysis provides intuitions on the relationship between fraction of sampled nodes and the resulting learning performance.

Let $\mathcal{G}_{\text{Girth}}(m; g, \Delta_{\min}, \Delta_{\max})$ denote the ensemble of graphs on m nodes with girth at least g and minimum degree $\Delta_{\min} \geq 3$ and maximum degree $\Delta_{\max}$. Let $\rho := \frac{p}{m}$ denote the uniform sampling probability (for observed nodes). We have the following result on the depth δ according to Definition 4. Define a constant $\epsilon_0 > 0$ as

$$\epsilon_0 = -\frac{\log(4m\Delta_{\max}(1-\rho)^{(\Delta_{\min}-1)^{g/2}})}{\log m}. \tag{35}$$

Lemma 2 (Depth Under Uniform Sampling) *Given uniform sampling probability of ρ, for any $\epsilon \leq \max(0, \epsilon_0)$,*

$$\delta < \tfrac{1}{\log(\Delta_{\min}-1)} \left(\log \left[\frac{\log(4m^{1+\epsilon}\Delta_{\max})}{|\log(1-\rho)|} \right] \right) \quad \textit{w.p.} \geq 1 - m^{-\epsilon}. \tag{36}$$

Proof: The proof is by straightforward arguments on binomial random variables and union bound. See Section B.4. □

Remarks:

1. Assuming that the girth satisfies $g > 2\delta(1 + d_{\max}/d_{\min})$ w.h.p., when the sampling probability and the degrees are both constant, then

   $$\rho = \Theta(1),\ \Delta_{\min}, \Delta_{\max} = O(1) \Rightarrow \delta = O(\log\log m) \Rightarrow n = \Omega(\text{poly}(\log m)),\ \text{w.h.p.} \tag{37}$$

   On the other hand, with vanishing sampling probability, for $\beta \in [0, 1)$, we have

   $$\rho = \Theta(m^{\beta-1}),\ \Delta_{\min}, \Delta_{\max} = O(1) \Rightarrow \delta = O(\log m) \Rightarrow n = \Omega(\text{poly}(m)),\ \text{w.h.p.} \tag{38}$$

2. Recall that for Ising models, the best-case sample complexity of $\mathsf{LocalCLGrouping}$ for structural consistency scales as

   $$n = \Omega(\Delta_{\max}^{2(2\delta+1)} \log p).$$

   Thus, under uniform sampling, the sample complexity required for consistency scales as

   $$n = \Omega\left(\Delta_{\max}^2 \left(\frac{\log p}{|\log(1-\rho)|} \right)^{4\frac{\log \Delta_{\max}}{\log(\Delta_{\min}-1)}} \log p \right).$$

   For the special case when the graph is regular ($\Delta_{\min} = \Delta_{\max}$), this reduces to

   $$n = \Omega\left(\Delta_{\max}^2 \rho^{-4} (\log p)^5\right). \tag{39}$$

   Comparing the above bound with the best known sample complexity for structure learning in fully-observed Ising models (with no hidden variables) $n = \Omega(\Delta_{\max}^2 \log p)$, we note that the sample complexity for structure learning latent Ising models is only slightly worse at $n = \Omega(\Delta_{\max}^2 (\log p)^5)$ when a constant fraction of nodes are uniformly sampled ($\rho = \Theta(1)$).

3. The sample complexity of LocalCLGrouping improves as the depth δ decreases. This occurs under uniform sampling when the minimum degree $\Delta_{\min}$ increases and the maximum degree $\Delta_{\max}$ is kept fixed. On the other hand, when $\Delta_{\max}$ increases, the sample complexity is worse, since this limits the range of the edge potentials for correlation decay to hold (assumption ($B2$)). We notice from (39) that this latter effect is dominant under uniform sampling, i.e., the sample complexity degrades as the degree of a (regular) graph is increased.

7 Necessary Conditions for Graph Estimation

We have so far provided sufficient conditions for recovering latent graphical models Markov on girth-constrained graphs. We now provide necessary conditions on the number of samples required by any algorithm to reconstruct the graph. Let $\widehat{G}_n : (\mathcal{X}^{|V|})^n \to \mathcal{G}_m$ denote any deterministic graph estimator using n i.i.d. samples from node set V and $\mathcal{G}_m$ is the set of all possible graphs on m nodes. We first define the notion of the graph edit distance.

Definition 5 (Edit Distance) *Let $G, \widehat{G}$ be two graphs*[9] *with adjacency matrices $\mathbf{A}_G, \mathbf{A}_{\widehat{G}}$, and let V be the set of labeled vertices in both the graphs (with identical labels). Then the edit distance between $G, \widehat{G}$ is defined as*

$$\operatorname{dist}(\widehat{G}, G; V) := \min_{\pi} ||\mathbf{A}_{\widehat{G}} - \pi(\mathbf{A}_G)||_1,$$

where π is any permutation on the unlabeled nodes while keeping the labeled nodes fixed.

In other words, the edit distance is the minimum number of entries that are different in $\mathbf{A}_{\widehat{G}}$ and in any permutation of $\mathbf{A}_G$ over the unlabeled nodes. In our context, the labeled nodes correspond to the observed nodes V while the unlabeled nodes correspond to latent nodes H. We now provide necessary conditions for graph reconstruction up to certain edit distance.

Theorem 3 (Necessary Condition) *For any deterministic estimator $\widehat{G}_m : (\mathcal{X}^{m^\beta})^n \mapsto \mathcal{G}_m$ based on n i.i.d. samples from m^β observed nodes $\beta \in [0, 1]$ of a latent graphical model Markov on graph $G_m \in \mathcal{G}_{\text{Girth}}(m; g, \Delta_{\min}, \Delta_{\max})$ on m nodes with girth g, minimum degree $\Delta_{\min}$ and maximum degree $\Delta_{\max}$, for all $\epsilon > 0$, we have*

$$\mathbb{P}[\operatorname{dist}(\widehat{G}_m, G_m; V) > \epsilon m] \geq 1 - \frac{|\mathcal{X}|^{nm^\beta} m^{(2\epsilon+1)m} 3^{\epsilon m}}{m^{0.5\Delta_{\min} m}(m - g\Delta_{\max}^{g})^{0.5\Delta_{\min} m}}, \tag{40}$$

under any sampling process to choose the observed nodes.

Proof: The proof is based on counting arguments. See Section C for details. □

Remarks:

[9] We consider inexact graph matching where the unlabeled nodes can be unmatched. This is done by adding required number of isolated unlabeled nodes in the other graph, and considering the modified adjacency matrices [15].

1. The above result states that roughly

$$n = \Omega(\Delta_{\min} m^{1-\beta} \log m) = \Omega\left(\frac{\Delta_{\min}}{\rho} \log p\right) \tag{41}$$

samples are required for structural consistency. Thus, when $\beta = 1$ (constant fraction of observed nodes), logarithmic number of samples are necessary while when $\beta < 1$ (vanishing fraction of observed nodes), polynomial number of samples are necessary for reconstruction. From (39), recall that for Ising models, under uniform sampling of observed nodes, the sample complexity of LocalCLGrouping scales as

$$n = \Omega(\Delta_{\max}^2 \rho^{-4} (\log p)^5),$$

and thus, nearly matches the lower bound on sample complexity in (41).

8 Conclusion

In this paper, we considered latent graphical models Markov on girth-constrained graphs. We proposed a novel approach, termed as LocalCLGrouping, for structure estimation in such models. We established the correctness of the method when the model is in the regime of correlation decay and also derived PAC learning guarantees. We compared these guarantees with other methods for graphical model selection, where there are no latent variables. Our findings reveal that latent variables do not add much complexity to the learning process in certain models and regimes, even when the number of hidden variables is large. These findings push the realm of tractable latent models for learning.

Acknowledgement

The first author is supported in part by the setup funds at UCI and the AFOSR Award FA9550-10-1-0310.

A Proofs for Identifiability of Latent Graphical Models

A.1 Recap of Identifiability of Latent Trees

Proposition 1 (Identifiability in a Latent Tree Model) *A latent graphical model Markov on a tree with all variables taking values in $\mathcal{X}$ satisfying positivity condition has identifiable structure and parameters when the hidden variables have degree of at least three.*

Proof: For the class of models on latent trees, for which we can define an additive tree metric, the identifiability of tree topology and the parameters is proven in [14]. For more general latent tree models, identifiability has been considered in [1]. □

We recap the result of [29, Lemma 2] where the concept of representative quartet is introduced (see Definition 4) and reframe it as an identifiability result.

Proposition 2 (Structure Identifiability in Terms of Representative Quartets) *A latent tree structure is uniquely identified by its set of representative quartets.*

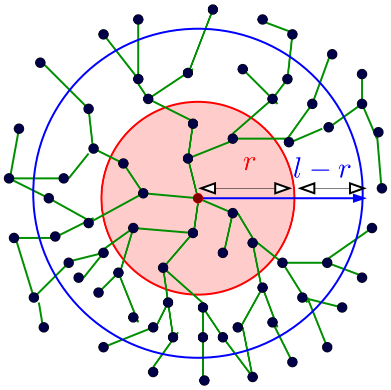


Figure 2: The nodes in $B_r(i;G)$ are at a distance of at least $l - r$ away from the boundary. When $l < g/2$, where g is the girth, no cycles are encountered.

A.2 Proof of Theorem 1

We now prove identifiability of latent graphical models Markov on girth-constrained graphs. Given a graph G, let $B_l(i;G)$ denote the set of nodes within graph distance l from node i and $F_l(i;G)$ be the induced graph on $B_l(i;G)$. Consider a graphical model $P_{\mathbf{X}_W|G}$ Markov on graph G and the set of observed nodes $V \subset W$ satisfying the assumptions in Section 5.1. For any set $A \subset W$, let $P_{\mathbf{X}_A|G}$ denote the corresponding marginal distribution on nodes in A. For some subgraph $F \subset G$, let $P_{\mathbf{X}_A|F}$ denote the distribution corresponding to graphical model Markov on graph F instead of G (i.e., by setting the edge potentials of edges in $G \setminus F$ to zero).

For any node $i \in W$ and $r < l \in \mathbb{N}$, by assumption (A2) in Section 5.1 on correlation decay (see (12)), the marginal distribution on nodes in $B_r(i;G)$ satisfies

$$\|P_{\mathbf{X}_{B_r(i;G)}|F_l(i;G)} - P_{\mathbf{X}_{B_r(i;G)}|G}\|_1 = \zeta_m(l - r). \tag{42}$$

When $l < g/2$, where g is the girth, the subgraph $F_l(i;G)$ is acyclic. Thus, the distribution $P_{\mathbf{X}_{B_r(i;G)}|F_l(i;G)}$ is a latent tree model. We now argue that the structure of $P_{\mathbf{X}_{B_r(i;G)}|F_l(i;G)}$ is identifiable when the depth δ satisfies $\delta < r$ since each hidden edge in $B_r(i;G)$ contains a representative quartet (see Proposition 2). This also implies that the latent tree model $P_{\mathbf{X}_{B_r(i;G)}|F_l(i;G)}$ also has parameter identifiability. When $g/2 - \delta = \omega(1)$ as assumed in (18), we can choose $\delta < r < l < g/2$ such that $l - r = \omega(1)$. Thus, in (42), when $\zeta_m(\omega(1)) = o(1)$ according to (17), we have that the distribution $P_{\mathbf{X}_{B_r(i;G)}|F_l(i;G)}$ converges locally to the latent tree model $P_{\mathbf{X}_{B_r(i;G)}|F_l(i;G)}$ and thus its structure and parameters are also identifiable asymptotically. Since this holds for all $i \in V$, we have the result. See Fig.2.

B Proof of Theorem 2: Structural Consistency of LocalCLGrouping

We first establish that the LocalCLGrouping algorithm proposed in Section 6 recovers the unknown latent graph correctly when statistics corresponding to the tree limit are input. In Section B.2, we then establish that distances based on exact statistics converge locally to their tree limit. Finally, we consider sample-based analysis in Section B.3, and use standard concentration results, along the lines of [30, Section 6].

B.1 Correctness of LocalCLGrouping under Local Tree Metric $\mathbf{d}_{\mathrm{SP}}(V)$

We make the observation that $\mathbf{d}_{\mathrm{SP}}(V)$ form a local tree metric in girth-constrained graphs when the distances are less than the girth of the graph. Recall that $B_l(i;G)$ denotes the set of nodes within graph distance l from node i in graph G. Let $\mathbf{d}_{\mathrm{SP}}(B_l(i;G)) := \{d(u,v;\mathrm{SP}) : u,v \in B_l(i;G)\}$.

Fact 1 (Local Tree Metric) *The distances $\mathbf{d}_{\mathrm{SP}}(B_l(i;G))$ form an additive tree metric for each $i \in V$ when the neighborhood graph distance satisfies $2l < g$, where g is the girth of graph G.*

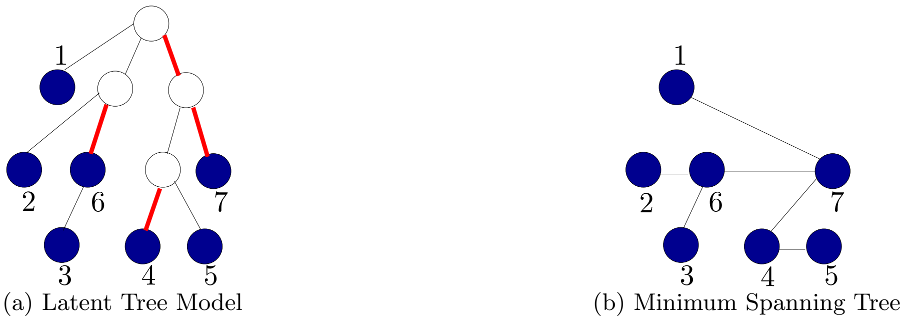
We now establish that the LocalCLGrouping algorithm proposed in Algorithm 4 outputs the correct graph under the assumptions on Theorem 2 when a local tree metric, computed using shortest paths according to (20), $\mathbf{d}_{\mathrm{SP}}(V)$ are input to the algorithm. Note that in practice, we only have access to empirical estimates $\widehat{\mathbf{d}}^n(V)$ of the distances $\mathbf{d}(V)$, and not $\mathbf{d}_{\mathrm{SP}}(V)$. In Section B.2, we establish the local convergence of $\mathbf{d}(V)$ to $\mathbf{d}_{\mathrm{SP}}(V)$ under correlation decay.

B.1.1 Recap of CLGrouping for Learning Latent Trees

We first recap the result from [21, Lemma 8] that relates a latent tree model with the minimum spanning tree over the observed nodes according to a tree metric. Note that in this case, $\mathbf{d}(V)$ coincides with $\mathbf{d}_{\mathrm{SP}}(V)$. For every node $i \in W$ in the latent tree T, define a mapping $\mathrm{Sg} : W \mapsto V$, termed as *surrogate mapping*, as follows:

$$\mathrm{Sg}(i;\mathbf{d}) := \min_{j\in V} d(i,j;T), \quad \forall, i \in W. \tag{43}$$

Thus, observed nodes V are their own surrogates while the hidden nodes H are mapped to the closest observed node according to metric $\mathbf{d}(V)$. See Fig.3 for an example.



(a) Latent Tree Model (b) Minimum Spanning Tree

Figure 3: A latent tree model over $T = (W,E)$ and the corresponding minimum spanning tree $\mathrm{MST}(V;\mathbf{d})$ over the observed nodes $V \subset W$. The observed nodes are shaded while the hidden nodes are unshaded. The thick lines in Fig.3a represent the edge between a hidden node and its surrogate. See Lemma 3.

Proposition 3 (Relating Latent Tree and MST) *Given a latent tree $T = (W,E)$, set of observed nodes $V \subset W$ and a tree metric $\mathbf{d}(V)$, the minimum spanning tree $\mathrm{MST}(V;\mathbf{d})$ over the observed nodes satisfies the following properties:*

1. *The* $\mathrm{MST}(V;\mathbf{d})$ *is obtained from the latent tree* T *by merging each hidden node* $h \in H$ *with its surrogate* $\mathrm{Sg}(h;\mathbf{d})$ *and viceversa.*

2. *Let* ξ *denote the maximum graph distance between a hidden node and its surrogate in the latent tree* T *and let* δ *denote the depth of tree* T*. We have*

$$\xi \leq \delta \frac{d_{\max;T}}{d_{\min;T}}, \tag{44}$$

where $d_{\min;T}$ *and* $d_{\max;T}$ *are bounds on the distance in* T*.*

B.1.2 Union of Local MSTs under LocalCLGrouping

Using the results of CLGrouping, we establish properties of the union of local minimum spanning trees for girth-constrained graphs under correlation decay. To this end, consider the choice of parameter r in (26) and bounds $d_{\min;\mathrm{SP}}$ and $d_{\max;\mathrm{SP}}$. Also define

$$r' := \frac{r}{d_{\max;\mathrm{SP}}}, \quad r'' := \frac{r}{d_{\min;\mathrm{SP}}}. \tag{45}$$

Recall that $B_r(i;\mathbf{d}_{\mathrm{SP}})$ denotes the set of observed nodes within distance r according to the metric $\mathbf{d}_{\mathrm{SP}}(V)$. Let $B_{r'}(i;G)$ denote the set of nodes (including hidden nodes) within graph distance r' from node $i \in V$ on graph G. By definition, $B_{r'}(i;G) \subset B_r(i;\mathbf{d}_{\mathrm{SP}}) \subset B_{r''}(i;G)$. In other words, the nodes in $B_r(i;\mathbf{d}_{\mathrm{SP}})$ have graph distance at least r' and at most r''. We have the following result.

Lemma 3 (Properties of Union of Local MSTs under $\mathbf{d}_{\mathrm{SP}}(V)$) *The graph formed by the union of local minimum spanning trees* ($G' := \cup_{i\in V} \mathrm{MST}(B_r(i);\mathbf{d}_{\mathrm{SP}})$) *under* LocalCLGrouping *method using the distance metric* $\mathbf{d}_{\mathrm{SP}}(V)$*, when the parameter* r *is chosen according to* (26)*, satisfies the following properties:*

1. G' *does not contain triangles.*

2. G' *is formed by contracting each hidden node* $h \in H$ *to its surrogate node* $\mathrm{Sg}(h;\mathbf{d}_{\mathrm{SP}})$ *(according to the distance metric* (19)*).*

Proof: The first result is easy to see. We have that for each edge $(i,j) \in G'$, $d(i,j;\mathrm{SP}) \leq r$ since the MSTs are formed on nodes within distance r. By contradiction, assume that a triangle exists between nodes $i,j,k \in V$ in G'. This implies that $d(i,j;\mathrm{SP}), d(j,k;\mathrm{SP}), d(k,i;\mathrm{SP}) \leq r$. For a triangle to exist, we require another node $l \in V$ such that $d(j,l;\mathrm{SP}), d(j,k;\mathrm{SP}), d(k,l;\mathrm{SP}) \leq r$. See Fig.4. Since the maximum graph distance between any two nodes i,j satisfying $d(i,j;\mathrm{SP}) \leq r$ is r'', we have that the maximum length of the cycle containing i,j,k,l is $4r''$. When $4r'' < g$ (which holds for r according to (26)), such a cycle cannot exist and such triangles cannot occur in G'.

For the second result, from Fact 1, the distances $\mathbf{d}_{\mathrm{SP}}(B_{r''}(i;G))$ form a tree metric when $2r'' < g$, where g is the girth of the graph G, which holds for the choice of r in (26). This implies that Lemma 3 is applicable and the minimum spanning tree $\mathrm{MST}(B_r(v);\mathbf{d}))$ is formed as a result of contraction of hidden nodes to their surrogates. When the parameter ξ in (44) satisfies $\xi + \delta < r'$ (which is true under (25)), then every hidden node has a surrogate within some local neighborhood $B_r(v)$ and forms a quartet with its surrogate node. This implies that every hidden node $h \in H$ contracts to its surrogate node in some local MST. □

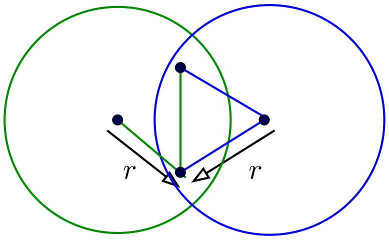


Figure 4: Condition for existence for triangles in ($G' := \cup_{i \in V} \mathrm{MST}(B_r(i); \mathbf{d}_{\mathrm{SP}})$).

Proof of Theorem 2 under $\mathbf{d}_{\mathrm{SP}}(V)$: We now show that the method LocalCLGrouping correctly recovers the graph G when tree-based distances $\mathbf{d}_{\mathrm{SP}}(V)$ are input under the assumptions of Theorem 2. From Lemma 3, we have that in the graph formed from the union of local MSTs ($G' := \cup_{v \in V} \mathrm{MST}(B_r(i); \mathbf{d}_{\mathrm{SP}})$), each hidden node is contracted to its surrogate node. The method LocalCLGrouping proceeds by reversing these contractions by considering neighborhoods on G' and constructing a local latent tree. Since there are no triangles in G', the construction of local latent trees are independent. From the correctness of CLGrouping developed in [21], the local latent trees are correct since the distance metric converges locally to a tree metric. Thus, the correctness of LocalCLGrouping under $\mathbf{d}_{\mathrm{SP}}(V)$ is proven. □

B.2 Local Convergence to a Tree Metric

We have so far analyzed the performance of LocalCLGrouping algorithm under tree-based distances $\mathbf{d}_{\mathrm{SP}}(V)$. We now relate the distances $\mathbf{d}(V)$ computed using exact pairwise statistics with $\mathbf{d}_{\mathrm{SP}}(V)$ under correlation decay according to (12). Let

$$d'_{\max}(l) := l d_{\max;\mathrm{SP}} - \log(1 - e^{l d_{\max;\mathrm{SP}}} |\mathcal{X}| \zeta_m(g/2 - l)),$$

where $d_{\max;\mathrm{SP}}$ is the maximum $d(i, j; \mathrm{SP})$ for any two neighbors i, j on graph G.

Proposition 4 (Local Convergence to a Tree Metric) *When a discrete graphical model satisfies correlation decay with rate α according to* (12)*, we have a.a.s., for nodes $i, j \in W$,*

$$|\exp[-d(i,j;G)] - \exp[-d(i,j;\mathrm{SP})]| \leq |\mathcal{X}| \zeta_m(g/2 - l), \tag{46}$$

where g is the girth of the graph, l is the length of the shortest path $\mathrm{SP}(i,j;G)$, $|\mathcal{X}|$ is the cardinality of the random variable at each node and ζ_m is the correlation decay function in (12)*. Additionally, when $2l < g$,*

$$|d(i,j;G) - d(i,j;\mathrm{SP})| \leq |\mathcal{X}| e^{d'_{\max}(l)} \zeta_m(g/2 - l). \tag{47}$$

Proof: From the definition of correlation decay in (12), we have that

$$\|P_{\mathbf{X}_{i,j}|G} - P_{\mathbf{X}_{i,j}|\,\mathrm{SP}(i,j)}\|_1 = \zeta_m(g/2 - l),$$

since $\mathrm{SP}(i,j;G)$ is the only path between i and j in subgraph $H_{g/2}(i)$ and $g/2 - l$ is the distance from j to the boundary.

From [9, Sec. 20], we have that for any $k \times k$ matrix A,

$$|\det(A + E) - \det(A)| \leq k \max\{\|A\|_q, \|A + E\|_q\}^{k-1} \|E\|_q. \tag{48}$$

Thus, we have that

$$|\det(P_{\mathbf{X}_{i,j}|G}) - \det(P_{\mathbf{X}_{i,j}|\,\mathrm{SP}(i,j)})| \leq |\mathcal{X}| \|P_{\mathbf{X}_{i,j}|G} - P_{\mathbf{X}_{i,j}|\,\mathrm{SP}(i,j)}\|_1 = |\mathcal{X}|\zeta_m(g/2 - l).$$

From Lipschitz continuity, we have that

$$|d(i,j;G) - d(i,j;\mathrm{SP})| \leq e^{d'_{\max}(l)} |\det(P_{\mathbf{X}_{i,j}|G}) - \det(P_{\mathbf{X}_{i,j}|\,\mathrm{SP}(i,j)})|,$$

Let $d_{\max;G}(l)$ be the maximum $d(i,j;G)$ for any two nodes i,j at graph distance l, and similarly for $d_{\max;\mathrm{SP}}(l)$. When $2l < g$, $d(i,j;\mathrm{SP})$ is a tree metric and thus $d_{\max;\mathrm{SP}}(l) = l d_{\max;\mathrm{SP}}(1)$. For $d_{\max;G}(l)$, we note that

$$e^{-d_{\max;G}(l)} \geq e^{-l d_{\max;\mathrm{SP}}(1)} - |\mathcal{X}|\zeta_m(g/2 - l).$$

□

Remark: When

$$e^{l d_{\max;\mathrm{SP}}} |\mathcal{X}|\zeta_m(g/2 - l) = o(1), \tag{49}$$

then

$$|d(i,j;G) - d(i,j;\mathrm{SP})| \leq |\mathcal{X}| e^{l d_{\max;\mathrm{SP}} + o(1)} \zeta_m(g/2 - l) = o(1), \tag{50}$$

B.3 Sample-Based Analysis

B.3.1 Concentration of Distance Estimates

We first derive the concentration bounds for distance estimates on lines of from [30,42]. Let $\widehat{\mathbf{d}}^n(V)$ be the estimated distances using n samples according to (19). We first recap the following result on empirical distribution [50, Thm. 2.1].

Proposition 5 (Guarantees for General Empirical Distribution) *The following is true for the empirical distribution $\widehat{P}^n$, obtained using n i.i.d. samples from a discrete distribution P:*

$$\mathbb{P}[\|\widehat{P}^n - P\|_1 > \epsilon] \leq 2^k \exp[-n\epsilon^2/2], \tag{51}$$

where k is the dimension.

We recap the result from [33, Proposition 19], which yields a better bound in the dimension, and is an application of the McDiarmid's inequality. Let $\|\cdot\|_2$ the ℓ_2 norm.

Proposition 6 (Alternative Bound) *Given empirical estimates $\widehat{P}^n$ of a probability distribution P using n i.i.d. samples, we have*

$$\mathbb{P}[\|\widehat{P}^n - P\|_2 > \epsilon] \leq \exp\left[-n\left(\epsilon - 1/\sqrt{n}\right)^2\right], \quad \epsilon > 1/\sqrt{n}. \tag{52}$$

We limit to the usage of the result in Proposition 5 since it imposes no constraint on the deviation ϵ and is a good bound when the dimension is small (e.g. Ising model), but note that results can also be derived based on the above bound.

Given a graph G, let the graph distance between two nodes i and j under consideration on graph G be l. Recall that $|\mathcal{X}|$ is the dimension of the variable at each node.

Lemma 4 (Concentration of Empirical Distances) *For empirical distance between node i and j at graph distance l, computed according to* (19) *using n samples, we have the following result:*

$$\mathbb{P}\left[|\exp[-\widehat{d}(i,j;G)] - \exp[-d(i,j;G)]| > \epsilon\right] \leq 2^{|\mathcal{X}|} \exp\left[-\frac{n\epsilon^2}{2|\mathcal{X}|^2}\right]. \tag{53}$$

Using Proposition 4, when $\epsilon > |\mathcal{X}|^2\zeta_m(g/2-l)$ and $l < g/2$, we additionally have that

$$\mathbb{P}\left[|\exp[-\widehat{d}(i,j;G)] - \exp[-d(i,j;\mathrm{SP})]| > \epsilon\right] \leq 2^{|\mathcal{X}|} \exp\left[-\frac{n}{2|\mathcal{X}|^2}\left(\epsilon - |\mathcal{X}|^2\zeta_m(\frac{g}{2}-l)\right)^2\right]. \tag{54}$$

Proof: On lines of Proposition 4, using [9, Sec. 20], we have,

$$\mathbb{P}\left[|\det(\widehat{P}^n_{\mathbf{X}_{i,j}|G}) - \det(P_{\mathbf{X}_{i,j}|G})| > \epsilon\right] \leq 2^{|\mathcal{X}|} \exp\left[-\frac{n\epsilon^2}{2|\mathcal{X}|^2}\right]. \tag{55}$$

□

B.3.2 Recap of Sample Analysis of CLGrouping for Learning Latent Trees

We also recap the result [21] that the minimum spanning tree (MST) constructed over observed nodes under CLGrouping method is consistent when the underlying model is a latent tree.

Recall that $p := |V|$ is the number of observed nodes and n is the number of samples. Let η be the maximum graph distance (with respect to the latent tree T) between any two neighbors in $\mathrm{MST}(V, \mathbf{d})$ and $d_{\min}, d_{\max}$ are distance bounds on the edges of the latent tree T.

Lemma 5 (Consistency of MST using CLGrouping for Latent Trees) *Given a latent tree $T = (W, E)$ and observed node set $V \subset W$, the MST constructed by* CLGrouping *method using empirical distances $\widehat{\mathbf{d}}^n(V)$ does not coincide with the true MST based on exact distances $\mathbf{d}(V)$ with probability*

$$\mathbb{P}\left[\mathrm{MST}(V;\widehat{\mathbf{d}}^n) \neq \mathrm{MST}(V;\mathbf{d})\right] \leq 2^{|\mathcal{X}|+1}p^3 \exp\left[-\frac{n}{8|\mathcal{X}|^2}e^{-2\eta d_{\max}}(1-e^{-d_{\min}})^2\right].$$

Proof: From the property of the MST,

$$\begin{aligned}
\mathbb{P}\left[\mathrm{MST}(V;\widehat{\mathbf{d}}^n) \neq \mathrm{MST}(V;\mathbf{d})\right] &\overset{\text{(a)}}{=} \mathbb{P}\left[\bigcup_{\substack{(i,j)\in\mathrm{MST}(V;\mathbf{d})\\(i,j)\in\mathrm{Path}(u,v)}} \left(e^{-\widehat{d}(u,v)} > e^{-\widehat{d}(i,j)}\right)\right], \\
&\overset{\text{(b)}}{\leq} p^3 \max_{\substack{(i,j)\in\mathrm{MST}(V;\mathbf{d})\\(i,j)\in\mathrm{Path}(u,v)}} \mathbb{P}\left[\epsilon_{u,v} - \epsilon_{i,j} > e^{-d(i,j)} - e^{-d(u,v)}\right] \\
&\overset{\text{(c)}}{\leq} p^3 \max_{i,j,u,v\in V} \mathbb{P}\left[\epsilon_{u,v} - \epsilon_{i,j} > e^{-\eta d_{\max}}(1-e^{-d_{\min}})\right] \\
&\overset{\text{(d)}}{\leq} 2p^3 \max_{u,v\in V} \mathbb{P}\left[\epsilon_{u,v} > \frac{e^{-\eta d_{\max}}}{2}(1-e^{-d_{\min}})\right],
\end{aligned}$$

where $\epsilon_{u,v} := \exp[-\widehat{d}(u,v)] - \exp[-d(u,v)]$ and similarly for $\epsilon_{i,j}$. Equality (a) is due to the property of the MST, inequality (b) is the union bound, inequality (c) is obtained by applying bounds on $d(i,j)$ and $d(u,v)$:

$$e^{-d(i,j)} - e^{-d(u,v)} > e^{-d(i,j)}(1 - e^{d(i,j)-d(u,v)}) > e^{-\eta d_{\max}}(1 - e^{-d_{\min}}),$$

since $d(i,j) \leq \eta d_{\max}$ for all $(i,j) \in \text{MST}(V;\mathbf{d})$ and $d(i,j) - d(u,v) \leq -d_{\min}$ for all $(i,j) \in \text{MST}(V;\mathbf{d})$ and $u,v \in V$ containing (i,j) on the path connecting them. Inequality (d) is obtained from the fact that $\epsilon_{u,v} - \epsilon_{i,j} \geq 2\max(\epsilon_{u,v}, \epsilon_{i,j})$ and applying the union bound. The final result is from (55) in Proposition 4. □

B.3.3 Sample Analysis of Union of MSTs under LocalCLGrouping

We now establish consistency under LocalCLGrouping algorithm using the above result and local convergence of the metric $\mathbf{d}(V)$ to tree-based metric $\mathbf{d}_{\text{SP}}(V)$, according to Proposition 4. Recall that $\widehat{\mathbf{d}}^n(V)$ denotes the estimates of the true distances $\mathbf{d}(V)$ according to graph G. Let $\mathbf{d}_{\text{SP}}$ denote the true distances by considering only the shortest paths, defined in Proposition 4. Given empirical distances $\widehat{\mathbf{d}}^n(V)$ and shortest-path distances $\widehat{\mathbf{d}}_{\text{SP}}$ and parameter r according to (26), for each $i \in V$, let $\widehat{A}_i := B_r(i;\widehat{\mathbf{d}}^n)$. Define $\mathcal{L} := \mathbb{N} \cap (r/d_{\min}, g/2)$.

Lemma 6 (Union of Local MSTs under LocalCLGrouping: I) *Given a graphical model Markov on graph $G = (W,E)$ satisfying conditions of Theorem 2 with observed node set $V \subset W$, we have*

$$\begin{aligned}
&\mathbb{P}\left[\bigcup_{i\in V} \text{MST}(\widehat{A}_i;\widehat{\mathbf{d}}^n) \neq \bigcup_{i\in V} \text{MST}(\widehat{A}_i;\mathbf{d}_{\text{SP}})\right] \\
&\leq 2^{|\mathcal{X}|}p^3 \min_{l\in\mathcal{L}} \left(2p\exp\left[-\frac{n}{2|\mathcal{X}|^2}\left(0.5e^{d_{\min}-r}(1-e^{-d_{\min}}) - |\mathcal{X}|^2\zeta_m(\frac{g}{2}-l)\right)^2\right]\right. \\
&\qquad\left. + \exp\left[-\frac{n}{2|\mathcal{X}|^2}\left(ld_{\min} - r - |\mathcal{X}|^2\zeta_m(\frac{g}{2}-l)\right)^2\right]\right). \qquad (56)
\end{aligned}$$

Remark: In the high-dimensional regime, where $p \to \infty$, the first term dominates. Since $\zeta_m(\cdot)$ is monotonically decreasing, we can choose $l = r/d_{\min} + 1$. Roughly, we require $n = \Omega(e^r)$ when the other parameters are bounded, for the error probability to decay.

Proof: On lines of Lemma 5, for each $k \in V$, we have

$$\begin{aligned}
&\mathbb{P}\left[\text{MST}(\widehat{A}_k;\widehat{\mathbf{d}}^n) \neq \text{MST}(\widehat{A}_k;\mathbf{d}_{\text{SP}})\right] \\
=&\mathbb{P}\left[\bigcup_{\substack{(i,j)\in\text{MST}(\widehat{A}_k;\mathbf{d}_{\text{SP}})\\(i,j)\in\text{Path}(u,v)}} \left(e^{-\widehat{d}(u,v)} > e^{-\widehat{d}(i,j)}\right)\right], \\
\leq&p^3 \max_{\substack{(i,j)\in\text{MST}(A_k;\mathbf{d}_{\text{SP}})\\(i,j)\in\text{Path}(u,v)}} \mathbb{P}\left[\epsilon_{u,v} - \epsilon_{i,j} > e^{-d(i,j;\text{SP})} - e^{-d(u,v;\text{SP})}\right]
\end{aligned}$$

$$\overset{(a)}{\leq} p^3 \max_{i,j,u,v\in V} \mathbb{P}\left[\epsilon_{u,v} - \epsilon_{i,j} > (e^{-r} - \epsilon_{i,j})(1 - e^{-d_{\min}})\right]$$
$$\overset{(b)}{\leq} 2p^3 \max_{u,v\in V} \mathbb{P}\left[\epsilon_{u,v} > \frac{e^{d_{\min}-r}}{2}(1 - e^{-d_{\min}})\right],$$

where $\epsilon_{u,v} := \exp[-\widehat{d}(u,v;G)] - \exp[-d(u,v;\mathrm{SP})]$ and similarly for $\epsilon_{i,j}$. Inequality (a) is obtained by applying bounds on $d(i,j)$ and $d(u,v)$:

$$e^{-d(i,j)} - e^{-d(u,v)} > e^{-d(i,j)}(1 - e^{d(i,j)-d(u,v)}) > (e^{-r} - \epsilon_{i,j})(1 - e^{-d_{\min}}),$$

since $\widehat{d}(i,j) \leq r$ for all $(i,j) \in \mathrm{MST}(V;\mathbf{d})$, $e^{-d(i,j)} > e^{-r} - \epsilon_{i,j}$ and $d(i,j) - d(u,v) \leq -d_{\min}$ for all $(i,j) \in \mathrm{MST}(V;\mathbf{d})$ and $u,v \in V$ containing (i,j) on the path connecting them. Inequality (b) is obtained from the fact that $\epsilon_{u,v} - e^{-d_{\min}}\epsilon_{i,j} \geq 2\max(\epsilon_{u,v}, e^{-d_{\min}}\epsilon_{i,j}) \geq 2e^{-d_{\min}}\max(\epsilon_{u,v}, \epsilon_{i,j})$ since $e^{-d_{\min}} < 1$.

Now define $\widehat{l}_{\max}$, the maximum graph distance between any two nodes in any $\widehat{A}_k$, i.e.,

$$\widehat{l}_{\max} := \max_k \left(\mathrm{Diam}(\mathrm{MST}(\widehat{A}_k))\right),$$

where $\mathrm{Diam}(\cdot)$ is the diameter, in terms of graph distance on G. From (54) in Lemma 4, conditioned on $\{\widehat{l}_{\max} = l\}$ and union bound on $k \in V$,

$$\mathbb{P}\left[\bigcup_{i\in V}\mathrm{MST}(\widehat{A}_i;\widehat{\mathbf{d}}^n) \neq \bigcup_{i\in V}\mathrm{MST}(\widehat{A}_i;\mathbf{d}_{\mathrm{SP}})\Big|\{\widehat{l}_{\max} = l\}\right]$$
$$\leq 2^{|\mathcal{X}|+1}p^4 \exp\left[-\frac{n}{2|\mathcal{X}|^2}\left(0.5e^{d_{\min}-r}(1 - e^{-d_{\min}}) - |\mathcal{X}|^2\zeta_m(\frac{g}{2} - l)\right)^2\right] \quad (57)$$

We now derive characterize the event that $\{\widehat{l}_{\max} = l\}$. Note that $\widehat{l}_{\max} \leq \widehat{d}'_{\max}/d_{\min}$, where

$$\widehat{d}'_{\max} := \max_{k\in V}\max_{i,j\in\widehat{A}_k}\widehat{d}(i,j). \quad (58)$$

Thus, we have

$$\mathbb{P}\left[\widehat{l}_{\max} > l\right] \leq \mathbb{P}\left[\bigcup_{\substack{k\in V\\ i,j\in\widehat{A}_k}}\widehat{d}(i,j) > l d_{\min}\right]$$
$$\leq 2^{|\mathcal{X}|}p^3 \exp\left[-\frac{n}{2|\mathcal{X}|^2}\left(l d_{\min} - r - |\mathcal{X}|^2\zeta_m(\frac{g}{2} - l)\right)^2\right].$$

□

We now provide conditions when $\bigcup_{i\in V}\mathrm{MST}(\widehat{A}_i;\mathbf{d}_{\mathrm{SP}})$ coincides with $\bigcup_{i\in V}\mathrm{MST}(A_i;\mathbf{d}_{\mathrm{SP}})$, where $A_i := B_r(i;\mathbf{d}_{\mathrm{SP}})$.

Lemma 7 (Union of Local MSTs under LocalCLGrouping: II) *Given a graphical model Markov on graph $G = (W,E)$ satisfying conditions of Theorem 2 with observed node set $V \subset W$, we have*

$$\mathbb{P}\left[\bigcup_{i\in V}\mathrm{MST}(\widehat{A}_i;\mathbf{d}_{\mathrm{SP}}) \neq \bigcup_{i\in V}\mathrm{MST}(A_i;\mathbf{d}_{\mathrm{SP}})\right]$$

$$
\begin{aligned}
\leq 2^{|\mathcal{X}|}p^3 \Bigg(&\exp\left[-\frac{n}{2|\mathcal{X}|^2}\left(\frac{gd_{\min}}{4} - r - |\mathcal{X}|^2\zeta_m(\frac{g}{2} - r/d_{\min} - 1)\right)^2\right] \\
&+ \exp\left[-\frac{n}{2|\mathcal{X}|^2}\left(r - \delta(\frac{d_{\max}}{d_{\min}} + 1)d_{\max} - |\mathcal{X}|^2\zeta_m(\frac{g}{2} - r/d_{\min} - 1)\right)^2\right]\Bigg).
\end{aligned}
$$

Remark: From the above result, choosing $r = 0.5(g/4 + \delta(\frac{d_{\max}}{d_{\min}} + 1))d_{\max}$ produces the best exponent.

Proof: Define

$$
\widehat{l}_{\max} := \max_k \left(\mathrm{Diam}(\mathrm{MST}(\widehat{A}_k))\right), \tag{59}
$$

$$
\widehat{l}_{\min} := \min_k \left(\mathrm{Diam}(\mathrm{MST}(\widehat{A}_k))\right). \tag{60}
$$

where $\mathrm{Diam}(\cdot)$ is the diameter, in terms of graph distance on G. Conditioned on the event $\{\widehat{l}_{\max} < \frac{g}{4}\} \cap \{\widehat{l}_{\min} > \xi + \delta\}$, the graph satisfies the properties listed in Lemma 3 and thus,

$$
\mathbb{P}\left[\bigcup_{i\in V} \mathrm{MST}(\widehat{A}_i; \mathbf{d}_{\mathrm{SP}}) \neq \bigcup_{i\in V} \mathrm{MST}(A_i; \mathbf{d}_{\mathrm{SP}}) \Big| \{\widehat{l}_{\max} < \frac{g}{4}\} \cap \{\widehat{l}_{\min} > \xi + \delta\}\right] = 0. \tag{61}
$$

Moreover,

$$
\begin{aligned}
\mathbb{P}\left[\{\widehat{l}_{\max} > \frac{g}{4}\} \cup \{\widehat{l}_{\min} < \xi + \delta\}\right] \leq 2^{|\mathcal{X}|}p^3 \Bigg(&\exp\left[-\frac{n}{2|\mathcal{X}|^2}\left(\frac{gd_{\min}}{4} - r - |\mathcal{X}|^2\zeta_m(\frac{g}{2} - r/d_{\min} - 1)\right)^2\right] \\
&+ \exp\left[-\frac{n}{2|\mathcal{X}|^2}\left(r - (\xi + \delta)d_{\max} - |\mathcal{X}|^2\zeta_m(\frac{g}{2} - r/d_{\min} - 1)\right)^2\right]\Bigg),
\end{aligned}
$$

where $\xi := \delta d_{\max}/d_{\min}$ is the worst-case graph distance between a hidden node and its surrogate in G with respect to metric $\mathbf{d}_{\mathrm{SP}}$. This is because the worst-case distance in a quartet containing a hidden node and its surrogate is $(\xi + \delta)d_{\max}$. When the empirical version of this distance exceeds r, then we have a bad event. □

B.3.4 Analysis of the Recursive Grouping

Recall that for each $i \in V$, let $\widehat{A}_i := B_r(i; \widehat{\mathbf{d}}^n)$ and $A_i := B_r(i; \mathbf{d}_{\mathrm{SP}})$. In $\mathsf{LocalCLGrouping}$, the recursive grouping procedure is run on subsets of nodes in each $\widehat{A}_i$. We first analyze the performance of quartet test.

Lemma 8 (Analysis of Quartet Test) *Given distance estimates $\widehat{\mathbf{d}}^n(\widehat{A}_i)$ over observed nodes in $\widehat{A}_i$, for each $i \in V$, $\mathsf{Quartet}(\widehat{\mathbf{d}}^n(\widehat{A}_i), \Lambda)$ returns the correct set of quartets (and no null results) with probability at least*

$$
\begin{aligned}
&\mathbb{P}[\cup_{i\in V}\{\mathsf{Quartet}(\widehat{\mathbf{d}}^n(\widehat{A}_i), \Lambda) \neq \mathsf{Quartet}(\mathbf{d}_{\mathrm{SP}}(\widehat{A}_i), \Lambda)\}] \\
\leq &2^{|\mathcal{X}|}p^3 \left(p\exp\left[-\frac{n}{2|\mathcal{X}|^2}\left(\exp[-(r/d_{\min} + 2)d_{\max}/2] - |\mathcal{X}|^2\zeta_m(\frac{g}{2} - r/d_{\min} - 1)\right)^2\right]\right.
\end{aligned}
$$

$$+\exp\left[-\frac{n}{2|\mathcal{X}|^2}\left(d_{\min}-|\mathcal{X}|^2\zeta_m(\frac{g}{2}-r/d_{\min}-1)\right)^2\right]\Bigg), \tag{62}$$

when Λ *is chosen as*

$$\Lambda = \exp[-(r/d_{\min}+2)d_{\max}/2]. \tag{63}$$

Proof: For each quartet $Q(v_1v_2|v_3v_4)$ under metric $\mathbf{d}_{\rm SP}(V)$ and $\mathcal{A} := \bigcup_{i=1}^4 v_i$, we have that

$$\mathbb{P}[\mathsf{Quartet}(\widehat{\mathbf{d}}^n(\mathcal{A}),\Lambda) \neq \mathsf{Quartet}(\mathbf{d}_{\rm SP}(\mathcal{A}),\Lambda)\Big| \bigcap_{a,b\in\mathcal{A}} \{|\widehat{d}^n(a,b)-d(a,b;{\rm SP})|<\Lambda\}] = 0, \tag{64}$$

and the test $\mathsf{Quartet}(\widehat{\mathbf{d}}^n(\mathcal{A}),\Lambda)$ does not return null when $\Lambda < \exp[-\max_{a,b\in\mathcal{A}} d(a,b;{\rm SP})/2]$. Considering all sets $\widehat{A}_i$ for $i\in V$, we require $\Lambda < \exp[-\widehat{l}_{\max}d_{\max}/2]$ to not return null, where

$$\widehat{l}_{\max} := \max_k \left(\mathrm{Diam}(\mathrm{MST}(\widehat{A}_k))\right).$$

From Lemma 4, choosing $\Lambda = \exp[-(l+1)d_{\max}/2]$ we that

$$\begin{aligned}&\mathbb{P}\left[\cup_{i\in V}\{\mathsf{Quartet}(\widehat{\mathbf{d}}^n(\widehat{A}_i),\Lambda) \neq \mathsf{Quartet}(\mathbf{d}_{\rm SP}(\widehat{A}_i),\Lambda)\Big|\{\widehat{l}_{\max}<l\}\}\right]\\ &\leq 2^{|\mathcal{X}|}p^4\exp\left[-\frac{n}{2|\mathcal{X}|^2}\left(\exp[-(l+1)d_{\max}/2]-|\mathcal{X}|^2\zeta_m(\frac{g}{2}-l)\right)^2\right].\end{aligned}$$

On lines of analysis in Lemma 7, we have that

$$\mathbb{P}\left[\{\widehat{l}_{\max}>l\}\right] \leq 2^{|\mathcal{X}|}p^3\exp\left[-\frac{n}{2|\mathcal{X}|^2}\left(ld_{\min}-r-|\mathcal{X}|^2\zeta_m(\frac{g}{2}-l)\right)^2\right].$$

Choosing l as $r/d_{\min}+1$, we have the result. □

This yields the following result on recursive grouping.

Lemma 9 (Results for Recursive Grouping) *The recursive grouping method* $\mathsf{RG}(\widehat{\mathbf{d}}(\widehat{A}_i),\Lambda,\tau)$ *returns the same tree as* $\mathsf{RG}(\mathbf{d}_{\rm SP}(\widehat{A}_i),\Lambda,\tau)$ *with probability*

$$\begin{aligned}&\mathbb{P}[\cup_{i\in V}\{\mathsf{RG}(\widehat{\mathbf{d}}^n(\widehat{A}_i),\Lambda,\tau)\neq\mathsf{RG}(\mathbf{d}_{\rm SP}(\widehat{A}_i),\Lambda,\tau)\}]\\ \leq& 2^{|\mathcal{X}|}p^3\left(p\exp\left[-\frac{n}{2|\mathcal{X}|^2}\left(\frac{(\frac{r}{d_{\min}}+2)d_{\max}}{2}-|\mathcal{X}|^2\zeta_m(\frac{g}{2}-\frac{r}{d_{\min}}-1)\right)^2\right]\right.\\ &+\exp\left[-\frac{n}{2|\mathcal{X}|^2}\left(d_{\min}-|\mathcal{X}|^2\zeta_m(\frac{g}{2}-\frac{r}{d_{\min}}-1)\right)^2\right]\\ &\left.+\exp\left[-\frac{n}{2|\mathcal{X}|^2}\left(\frac{d_{\min}}{2}-|\mathcal{X}|^2\zeta_m(\frac{g}{2}-1)\right)^2\right]\right),\end{aligned} \tag{65}$$

when Λ *is chosen as*

$$\Lambda = \exp[-(r/d_{\min}+2)d_{\max}/2]. \tag{66}$$

and τ *is chosen as*

$$\tau = \frac{d_{\min}}{2}-|\mathcal{X}|^2\zeta_m(g/2-1). \tag{67}$$

Proof: On lines of analysis in [3], given the correct set of quartets, the recursive grouping procedure returns the correct tree structure when the nodes are merged correctly with threshold τ. It is easy to see that this happens with probability

$$2^{|\mathcal{X}|}p^2 \exp\left[-\frac{n}{2|\mathcal{X}|^2}(d_{\min}/2 - |\mathcal{X}|^2\zeta_m(g/2-1))^2\right],$$

when the threshold is chosen as (67). □

B.4 Analysis Under Uniform Sampling

Let $\mathcal{A}(e)$ denote the event that an hidden edge (with at least one hidden end point) has a representative quartet in which the end points are at most graph distance $l < g/2$. We have that

$$\mathbb{P}[\mathcal{A}(e)] \leq 4(1-\rho)^{(\Delta_{\min}-1)^{l-1}},$$

since there are at least $(\Delta_{\min}-1)^{l-1}$ nodes in each of the four subtrees from which four observed nodes can be sampled and $\rho := p/m$ is the sampling probability. Taking the union bound, we have the probability that the depth δ is greater than $l < g/2$ as

$$\mathbb{P}[\delta > l] \leq 4m\Delta_{\max}(1-\rho)^{(\Delta_{\min}-1)^{l}}.$$

Thus, the result in (36) holds. □

C Necessary Conditions for Graph Reconstruction

The proof is based on counting arguments on lines of [12, Thm. 1]. For any deterministic estimator $\widehat{G}_m$, let $\mathcal{R} := \widehat{G}_m((\mathcal{X}^{m^\beta})^n)$ as the range of the estimator $\widehat{G}_m$, when the number of observed nodes is $|V| = m^\beta$ for $\beta \in (0,1]$. Thus, we have $|\mathcal{R}| = |\mathcal{X}|^{nm^\beta}$.

For any fixed graph F_m and set of labeled nodes V, denote the set of graphs within graph distance ϵm as

$$\mathcal{D}(F_m;\epsilon m) := \{G_m : \mathrm{dist}(F_m, G_m; V) \leq \epsilon m\}.$$

We note that

$$|\mathcal{D}(F_m;\epsilon m)| \leq m!\binom{m^2}{\delta m} \leq m^{(2\epsilon+1)m}3^{\epsilon m},$$

since we can permute the m vertices and change at most ϵm entries in the adjacency matrix $\mathbf{A}_F$ and we use the bound that $\binom{N}{k} \leq \frac{N^k}{k!} \leq N^k 3^k$.

Let $\mathcal{S}(\widehat{G}_m;\epsilon m)$ denote all the graphs which are within edit distance of ϵm of the graphs in range $\mathcal{R}$. We have that

$$|\mathcal{S}(\widehat{G}_m;\epsilon m)| \leq |\mathcal{X}|^{nm^\beta} m^{(2\epsilon+1)m}3^{\epsilon m}.$$

On lines of [12, Thm. 1], we have that the probability of error should satisfy

$$\mathbb{P}[\mathrm{dist}(\widehat{G}_m, G_m; V) > \epsilon m] \geq 1 - \frac{|\mathcal{S}(\widehat{G}_m;\epsilon m)|}{|\mathcal{G}(m)|},$$

where $|\mathcal{G}(m)|$ is the number of graphs in the family under consideration.

From [6, Lemma 2], we have that for girth-constrained ensembles $\mathcal{G}_{\text{Girth}}(m; g, \Delta_{\min}, \Delta_{\max}, k)$ with girth g, minimum degree $\Delta_{\min}$, maximum degree $\Delta_{\max}$ and number of edges k, we have

$$m^k(m - g\Delta_{\max}^g)^k \leq |\mathcal{G}_{\text{Girth}}(m; g, \Delta_{\min}, \Delta_{\max}, k)| \leq m^k(m - \Delta_{\min}^g)^k, \tag{68}$$

and we have the result. □

References

[1] E.S. Allman and J.A. Rhodes. The identifiability of covarion models in phylogenetics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 76–88, 2008.

[2] N. Alon, S. Hoory, and N. Linial. The moore bound for irregular graphs. *Graphs and Combinatorics*, 18(1):53–57, 2002.

[3] A. Anandkumar, K. Chaudhuri, D. Hsu, S.M. Kakade, L. Song, and T. Zhang. Spectral Methods for Learning Multivariate Latent Tree Structure. *Preprint, ArXiv 1107.1283*, July 2011.

[4] A. Anandkumar, A. Hassidim, and J. Kelner. Topology Discovery of Sparse Random Graphs With Few Participants. *arXiv:1102.5063*, Feb. 2011.

[5] A. Anandkumar, V. Y. F. Tan, and A. S. Willsky. High-Dimensional Gaussian Graphical Model Selection: Tractable Graph Families. *Preprint, ArXiv 1107.1270*, June 2011.

[6] A. Anandkumar, V. Y. F. Tan, and A. S. Willsky. High-Dimensional Structure Learning of Ising Models: Tractable Graph Families. *Preprint, Available on ArXiv 1107.1736*, June 2011.

[7] M. Bayati, A. Montanari, and A. Saberi. Generating random graphs with large girth. In *Proc. of ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2009.

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

[9] R. Bhatia. *Perturbation Bounds for Matrix Eigenvalues (Classics in Applied Mathematics)*. Society for Industrial and Applied Mathematics, 2007.

[10] A. Bogdanov, E. Mossel, and S. Vadhan. The Complexity of Distinguishing Markov Random Fields. *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 331–342, 2008.

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

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

[13] L.D. Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. *Lecture Notes-Monograph Series, Institute of Mathematical Statistics*, 9, 1986.

[14] P. Buneman. The recovery of trees from measures of dissimilarity. *, Mathematics in the Archaeological and Historical Sciences (FR Hodson, DG Kendall, and P. Tautu, eds.)*, 1971.

[15] G. Bunke et al. Inexact graph matching for structural pattern recognition. *Pattern Recognition Letters*, 1(4):245–253, 1983.

[16] L.S. Chandran and CR Subramanian. Girth and treewidth. *J. of combinatorial theory, Series B*, 93(1):23–32, 2005.

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

[18] V. Chandrasekaran, P.A. Parrilo, and A.S. Willsky. Latent Variable Graphical Model Selection via Convex Optimization. *Arxiv preprint*, 2010.

[19] T. Chen, N. L. Zhang, and Y. Wang. Efficient model evaluation in the search based approach to latent structure discovery. In *4th European Workshop on Probabilistic Graphical Models*, 2008.

[20] M.J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *IEEE Conf. on Computer VIsion and Pattern Recognition (CVPR)*, 2010.

[21] M.J. Choi, V.Y.F. Tan, A. Anandkumar, and A. Willsky. Learning Latent Tree Graphical Models. *J. of Machine Learning Research*, 12:1771–1812, May 2011.

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

[23] C. Daskalakis, E. Mossel, and S. Roch. Optimal phylogenetic reconstruction. In *STOC '06: Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 159–168, 2006.

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

[25] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge Univ. Press, 1999.

[26] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge Univ. Press, 1999.

[27] R.D. Dutton and R.C. Brigham. Edges in graphs with large girth. *Graphs and Combinatorics*, 7(4):315–321, 1991.

[28] G. Elidan and N. Friedman. Learning hidden variable networks: The information bottleneck approach. *Journal of Machine Learning Research*, 6:81–127, 2005.

[29] P. L. Erdős, L. A. Székely, M. A. Steel, and T. J. Warnow. A few logs suffice to build (almost) all trees: Part ii. *Theoretical Computer Science*, 221:153–184, 1999.

[30] P.L. Erdös, M.A. Steel, L.A. Székely, and T.J. Warnow. A few logs suffice to build (almost) all trees: part ii. *Theoretical Computer Science*, 221(1-2):77–118, 1999.

[31] A. Gamburd, S. Hoory, M. Shahshahani, A. Shalev, and B. Virag. On the girth of random cayley graphs. *Random Structures & Algorithms*, 35(1):100–117, 2009.

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

[33] D. Hsu, S.M. Kakade, and T. Zhang. A spectral algorithm for learning hidden markov models. *Arxiv preprint arXiv:0811.4413*, 2008.

[34] A. Jalali, C. Johnson, and P. Ravikumar. On learning discrete graphical models using greedy methods. In *Proc. of NIPS*, 2011.

[35] D. Karger and N. Srebro. Learning Markov Networks: Maximum Bounded Tree-width Graphs. In *Proc. of ACM-SIAM symposium on Discrete algorithms*, pages 392–401, 2001.

[36] Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory.* MIT Press., Cambridge, MA, 1994.

[37] C. Kemp and J. B. Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Science*, 105(31):10687–10692, 2008.

[38] S. L. Lauritzen. *Graphical models.* Clarendon Press, 1996.

[39] P. F. Lazarsfeld and N.W. Henry. *Latent structure analysis.* Boston: Houghton Mifflin, 1968.

[40] N. Meinshausen and P. Buehlmann. High Dimensional Graphs and Variable Selection With the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

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

[42] E. Mossel. Distorted metrics on trees and phylogenetic forests. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 108–116, 2007.

[43] E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden markov models. In *Proc. of annual symposium on Theory of computing*, 2005.

[44] P. Netrapalli, S. Banerjee, S. Sanghavi, and S. Shakkottai. Greedy Learning of Markov Network Structure . In *Proc. of Allerton Conf. on Communication, Control and Computing*, Monticello, USA, Sept. 2010.

[45] J. Pearl. *Probabilistic Reasoning in Intelligent Systems—Networks of Plausible Inference.* Morgan Kaufmann, 1988.

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

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

[48] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[49] M. Steel. Recovering a tree from the leaf colourations it generates under a markov model. *Applied Mathematics Letters*, 7(2):19–23, 1994.

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

[51] D. Weitz. Combinatorial Criteria for Uniqueness of Gibbs Measures. *Random Structures & Algorithms*, 27(4):445, 2005.

[52] N. L. Zhang. Hierarchical Latent Class Models for Cluster Analysis. *Journal of Machine Learning Research*, 5:697–723, 2004.

[53] N. L. Zhang and T Kočka. Efficient learning of hierarchical latent class models. In *ICTAI*, 2004.