Lower Bounds on the Size of Markov Equivalence Classes

Erik Jahn[1] **Frederick Eberhardt**[2] **Leonard J. Schulman**[1]

[1]Engineering and Applied Science, California Institute of Technology, Pasadena, California, USA
[2]Humanities and Social Sciences, California Institute of Technology, Pasadena, California, USA

Abstract

Causal discovery algorithms typically recover causal graphs only up to their Markov equivalence classes unless additional parametric assumptions are made. The sizes of these equivalence classes reflect the limits of what can be learned about the underlying causal graph from purely observational data. Under the assumptions of acyclicity, causal sufficiency, and a uniform model prior, Markov equivalence classes are known to be small on average. In this paper, we show that this is no longer the case when any of these assumptions is relaxed. Specifically, we prove exponentially large lower bounds for the expected size of Markov equivalence classes in three settings: sparse random directed acyclic graphs, uniformly random acyclic directed mixed graphs, and uniformly random directed cyclic graphs.

1 INTRODUCTION

One of the most powerful contributions of the theory of causal graphs [Pearl, 2009, Spirtes et al., 2001] is a complete characterization of which causal relationships can be learned from observational data without conducting experiments. This characterization is given by the concept of *Markov equivalence*. Two causal models are Markov-equivalent if they encode the same conditional independence constraints. As a result, even a perfect causal discovery algorithm can only recover the true causal structure up to its Markov equivalence class, unless additional identifiable structure is assumed or happens to be present. Remarkably, Markov equivalence classes for directed acyclic graphs (DAGs) have a simple graphical characterization, due to Verma and Pearl [1991]. This is exploited by some of the most widely used causal discovery algorithms today, all of which return a Markov equivalence class as their output [Spirtes et al., 2001, Chickering, 2003, Raskutti and Uhler, 2018]. Consequently, the size of these equivalence classes is a key measure for how informative the output of such algorithms is, making the challenge of counting the number of Markov equivalent graphs a subject of ongoing research [He et al., 2015, Radhakrishnan et al., 2017, Wienöbst et al., 2023].

The largest Markov equivalence class of DAGs on n variables has size $n!$, consisting of all the fully connected DAGs. However, for small graphs, numerical simulations [Gillispie and Perlman, 2001, 2002] and recursive enumeration [Steinsky, 2003, Gillispie, 2006, Steinsky, 2013] have shown that the average number of DAGs per Markov equivalence class is surprisingly small, even less than four. More recently, Schmid and Sly [2024] proved that for arbitrarily large n, the expected size of the Markov equivalence class of a uniformly random DAG on n variables is bounded by a constant. These results crucially rely on the uniformity assumption, effectively placing almost all the weight on dense DAGs. In practice, however, we expect useful real-world causal structures to be sparse. Indeed, many causal discovery algorithms work best under a sparsity assumption [Kalisch and Bühlmann, 2007, Chickering, 2020] and sparse priors for learning graphical models have been shown to improve practical performance [Huang et al., 2013, Eggeling et al., 2019]. Numerous researchers have posed the question of how large Markov-equivalence classes are for sparse graphs [Chickering, 2002, Talvitie and Koivisto, 2019, Katz et al., 2019].

In this paper, we give a first theoretical answer to this question. We show that for a wide range of sparse random DAG distributions, the expected size of the Markov equivalence classes basically scales exponentially in the inverse edge density. In particular, when the expected degree of each vertex is bounded by a constant, the Markov equivalence classes are exponentially large in the number of vertices of the DAG in expectation. This result reveals a sharp contrast to the uniform setting and also has algorithmic implications: Many greedy search algorithms can be run either in the space of DAGs or in the space of equivalence classes, but

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 244:1819–1836.

with small equivalence class sizes the efficiency gain remained unclear [Chickering, 2002]. Our results suggest that the efficiency gap becomes larger as the input graphs are sparser.

Moreover, we extend our analysis of Markov equivalence classes beyond DAGs. Causal models represented by DAGs inherently make two assumptions: *causal sufficiency*, meaning that there are no unobserved common causes, and *acyclicity*, implying the absence of causal feedback loops. Both assumptions may be violated in applications, motivating the study of more general graphical models: acyclic directed mixed graphs (ADMGs), which allow for unmeasured confounders, and directed cyclic graphs (DCGs), which permit cyclic causal relationships. Just as for DAGs, Markov equivalence has been characterized for both models [Spirtes and Richardson, 1997, Richardson, 1997] and corresponding causal discovery algorithms that output Markov equivalence classes have been developed [Spirtes et al., 1999, Richardson, 1996b]. The characterization of Markov equivalence classes is much more complex in both cases, and little is known about the sizes of these Markov equivalence classes. Here, we establish super-exponential lower bounds for the expected size of Markov equivalence classes of uniformly random ADMGs and DCGs. Our results exploit specific underdetermined substructures, leaving open the possibility that more restrictive model classes, such as maximal ancestral graphs [Richardson and Spirtes, 2002], may exhibit smaller Markov equivalence classes by prohibiting such structures. In general, our results highlight the need to supplement causal discovery with stronger, possibly parametric assumptions, or to perform interventions if one wants to reduce the size of the set of plausible graphs, let alone find a unique causal graph that explains the data.

1.1 OUR RESULTS

In this section, we provide formal statements of our results. For a parameter $p \in [0, 1)$, consider the following natural sampling process for generating a random DAG on n vertices:

- Include each possible directed edge independently with probability p;
- If the graph has any directed cycles, reject and repeat.

We denote the distribution arising from this process by $D(n, p)$ (see also Definition 4 for an alternative description). The parameter p directly controls the edge density of the sampled graph, and one can check that $D(n, 1/2)$ is equal to the uniform distribution on vertex-labeled DAGs. Heckerman et al. [1995] already used this distribution in the context of Bayesian networks and Eggeling et al. [2019] showed that it has desirable properties as a prior for Bayesian inference over causal structures. While the sampling process described above becomes impractical for large n due to a high rejection probability, a more efficient sampling algorithm for $D(n, p)$ has been given by Talvitie et al. [2020].

We study the sizes of Markov equivalence classes of DAGs sampled from $D(n, p)$ in the regime $6/n \leq p \leq o(1/\log n)$. The graphs in this regime have expected vertex degree ranging from constant (whenever $p = C/n$) up to $o(n/\log n)$. It turns out that the number of graphs that are Markov-equivalent to a DAG G sampled from $D(n, p)$ scales at least almost exponentially in the inverse of p.

Theorem 1. *Let $6/n \leq p \leq o\left(\frac{1}{\log n}\right)$ and $G \sim D(n, p)$. Then, we have with probability $1 - o(1)$:*

$$|\operatorname{MEC}(G)| \geq 2^{\Omega\left(\frac{p^{-1}}{\log^2(p^{-1})}\right)}.$$

In particular, this implies:

$$\mathbb{E}[|\operatorname{MEC}(G)|] \geq 2^{\Omega\left(\frac{p^{-1}}{\log^2(p^{-1})}\right)}.$$

For the more general settings of ADMGs and DCGs, we show that the expected size of the Markov equivalence class is super-exponential when the graph is sampled uniformly at random.

Theorem 2. *Let G be a uniformly random ADMG on n vertices, and let H be a uniformly random DCG on n vertices. Then, we have*

(a) $\mathbb{E}[|\operatorname{MEC}(G)|] \geq 2^{\Omega(n^2)}$*;*

(b) $\mathbb{E}[|\operatorname{MEC}(H)|] \geq 2^{\Omega(n^2)}$.

In fact, our proof for Theorem 2 shows that each graph (ADMG or DCG) on average already has exponentially many Markov equivalent graphs that all have different edge adjacencies. This stands in contrast to the DAG setting where all Markov equivalent DAGs have the same adjacencies [Verma and Pearl, 1991]. It raises the question of whether the existence of such underdetermined adjacencies in ADMGs and DCGs is the only reason for the blow-up of the size of their Markov equivalence classes. For DCGs, we answer this question negatively, by building on a result of Richardson [1996a] stating that the direction of cycles is always underdetermined in a Markov equivalence class. However, for ADMGs, this argument does not apply and the question remains open.

Theorem 3. *Let $\operatorname{MEC}^*(H)$ denote the set of graphs that are Markov-equivalent to H and have the same adjacencies as H. Then, for a uniformly random DCG H on n vertices, we have*

$$\mathbb{E}[|\operatorname{MEC}^*(H)|] \geq 2^{\Omega(n)}.$$

The rest of this paper is structured as follows: in Section 2, we provide definitions and notation for graphical models, and state some useful known results. In section 3 we present an outline of our approach towards proving Theorem 1 for DAGs (with full technical proofs in the appendix). The proof of Theorem 2, part (a) for ADMGs can be found in Section 4 and the proof of Theorem 2, part (b) and Theorem 3 for DCGs is given in Section 5. We discuss the implication of our results and further open questions in Section 6.

2 PRELIMINARIES

2.1 GRAPHICAL MODELS

A *directed graph* G consists of a *vertex set* $V(G)$ and a set of *directed edges* $E(G)$, which are ordered pairs (v, w) of vertices $v \neq w \in V(G)$. A *directed mixed graph* additionally has a set of bidirected edges that are unordered pairs of vertices $\{v, w\}$ for $v \neq w$. We denote directed edges by $v \to w$ and bidirected edges by $v \leftrightarrow w$. For a subset of vertices $W \subseteq V(G)$, we define the *induced* graph $G[W]$ as the graph whose vertex set is W and whose edges are the edges of G that lie entirely within W. A (possibly self-intersecting) *path* in a directed mixed graph G is an ordered list of vertices $(v_1, \ldots, v_k)$ with $v_1 \neq v_k$ such that there is a directed edge (in either direction) or bidirected edge between any two consecutive vertices in the list. A *directed path* is an ordered list of vertices $(v_1, \ldots, v_k)$ with $v_1 \neq v_k$ such that there is a directed edge from v_i to v_{i+1} for $i = 1, \ldots, k-1$. A *cycle* $C = (v_1, \ldots, v_k)$ is a directed path that additionally has a directed edge from v_k to v_1 (all cycles are directed). In the context of a cycle of length k, we usually consider all indices modulo k, in particular, we identify $v_{k+1} = v_1$. The *parents* of a vertex v are the vertices u that have a directed edge $u \to v$, the *children* of v are the vertices w that have a directed edge $v \to w$, and the *descendants* of v are the vertices x such that there is a directed path from v to x. We denote the set of parents of v by $\mathrm{Pa}(v)$. A *source* of the graph G is a vertex without parents. A *matching* of edges is a set of edges that do not have any vertices in common (i.e. their sets of endpoints are pairwise disjoint from each other). A *directed acyclic graph (DAG)* is a directed graph that contains no cycles and a *directed cyclic graph (DCG)* is a directed graph that may or may not contain cycles (this terminology has been used in the literature to clearly distinguish from the more popular DAGs in the context of causal models). An *acyclic directed mixed graph (ADMG)* is a directed mixed graph that contains no (directed) cycles. We study random DAGs under the following probability distribution:

Definition 4. *Fix a positive integer n and a parameter $p \in [0, 1)$. We define the distribution $D(n, p)$ over DAGs on n vertices as the distribution that assigns each DAG G with s edges a probability proportional to $w(G) = \left(\frac{p}{1-p}\right)^s$.*

Equivalently, $D(n, p)$ is the distribution that arises from sampling each possible directed edge independently with probability p (resulting in $\Pr(G) = p^s(1-p)^{\binom{n}{2}-s}$) and then conditioning on acyclicity. The distribution $D(n, 1/2)$ is equal to the uniform distribution over DAGs on n vertices. Note that random DAGs can also be obtained by sampling edges from an upper triangular matrix and uniformly permuting vertex labels. However, this process places a bias on DAGs with many automorphisms and cannot be seen as a natural extension of the uniform distribution. For ADMGs, a uniformly random ADMG is obtained by sampling a uniformly random DAG, and then adding a bidirected edge for each pair of vertices independently with probability $1/2$. A uniformly random DCG is obtained by placing each possible directed edge independently with probability $1/2$.

2.2 MARKOV EQUIVALENCE

The following definitions hold for all three model classes (DAGs, ADMGs, DCGs) alike. Given a path $\pi = (v_1, \ldots, v_k)$, the vertex v_i is a *collider* on the path if there are two incoming arrows from v_{i-1} and v_{i+1} to v_i, that is, one of the following holds true: $v_{i-1} \to v_i \leftarrow v_{i+1}$, $v_{i-1} \to v_i \leftrightarrow v_{i+1}$, $v_{i-1} \leftrightarrow v_i \leftarrow v_{i+1}$, $v_{i-1} \leftrightarrow v_i \leftrightarrow v_{i+1}$. Given two vertices $v, w \in V(G)$ and a *conditioning set* $Z \subseteq V(G) \setminus \{v, w\}$, a path π from v to w is *active given* Z if every non-collider vertex on the path is not in Z and every collider vertex on the path is either in Z or has a descendant in Z. If there is an active path between v, w given Z, then v and w are said to be d-connected given Z; otherwise they are d-separated given Z. Two graphs G_1 and G_2 on the same vertex set V are *Markov-equivalent* if for all $v, w \in V$ and $Z \subseteq V \setminus \{v, w\}$, v, w are d-connected given Z in G_1 if and only if they are d-connected given Z in G_2. The set of all graphs that are Markov-equivalent to a graph G is called the *Markov equivalence class (MEC)* of G. The significance of Markov equivalence in causal inference stems from the fact that two causal models represented by Markov-equivalent graphs encode the same conditional independence structure. The connection between this property and the graphical definition of Markov-equivalence through d-separation has first been formalized for DAGs by Verma and Pearl [1990], Geiger et al. [1990]. Later, it has been shown that the same connection holds for ADMGs [Spirtes et al., 1998, Koster, 1999] and DCGs [Spirtes, 1995] (at least for linear models). Hence, for these models, Markov-equivalent graphs cannot be distinguished based on observing conditional dependence and independence relations.

For DAGs, we will additionally make use of the following result: Let us call an edge $v \to w$ of a DAG G *reversible* if replacing the edge by $w \to v$ results in another DAG G' that is Markov equivalent to G. Then, reversible edges have a simple characterization:

Lemma 5. *[Chickering, 1995] An edge $v \to w$ of a DAG*

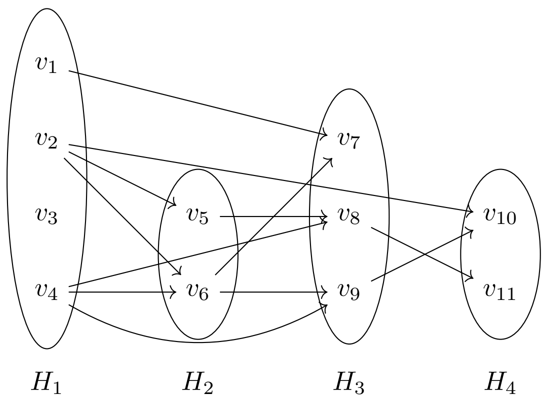


Figure 1: A DAG on 11 vertices with tower decomposition (H_1, H_2, H_3, H_4).

G is reversible if and only if $\mathrm{Pa}(v) = \mathrm{Pa}(w) \setminus \{v\}$.

2.3 TOWER DECOMPOSITION OF DAGS

The definitions and notation in this subsection are from Schmid and Sly [2024]. We define the *tower decomposition* of a directed acyclic graph G as follows: Let $H_1(G)$ be the set of sources in G, $H_2(G)$ the set of sources in $G \setminus H_1(G)$, $H_3(G)$ the set of sources in $G \setminus (H_1(G) \cup H_2(G))$, and so on. This partitions the vertex set $V(G)$ into sets $H_1, H_2 \ldots, H_{s(G)}$, where $s(G)$ is the length of the longest directed path in G (we suppress the dependence on G if it is clear from the context). See Figure 1 for an example. We call the sets H_i the *layers* of G and denote their sizes by $h_i := |H_i|$ for $i = 1, \ldots, s$. The vector $h(G) = (h_1, \ldots, h_s)$ is called the *tower vector* of G. The tower decomposition $(H_1, \ldots, H_s)$ of G is characterized by two properties: first, edges of G can only go from a layer H_i to a layer H_j with $j > i$, in particular, there are no edges within the layers themselves. Secondly, each vertex in H_i must have at least one parent in H_{i-1} (except the vertices in H_1, which are exactly the vertices in G that have no parents).We use tower decompositions to obtain a useful description of the distribution $D(n,p)$:

Lemma 6. *Let $G \sim D(n,p)$, and fix a tower vector $h = (h_1, \ldots, h_s)$ with entries that sum up to n, and a tower decomposition $H = (H_1 \ldots, H_s)$ with $|H_i| = h_i$. Then, the following holds:*

(a) the probability that h is the tower vector of G is proportional to

$$w(h) = \binom{n}{h_1, \ldots, h_s} \prod_{k=2}^{s} \frac{\left(1-(1-p)^{h_{k-1}}\right)^{h_k}}{(1-p)^{h_k \sum_{i=1}^{k-1} h_i}};$$

(b) conditional on the event that H is the tower decomposition of G, the parent sets $\mathrm{Pa}(v)$ are independently distributed for all $v \in V(G)$, and for $v \in H_j$ and $S \subseteq \bigcup_{i=1}^{j-1} H_i$, we have

$$\Pr(\mathrm{Pa}(v) = S) = \begin{cases} C \cdot \left(\frac{p}{1-p}\right)^{|S|} & \text{if } S \cap H_{j-1} \neq \emptyset \\ 0 & \text{otherwise,} \end{cases}$$

where C is the normalization constant.

Here, $\binom{n}{h_1,\ldots,h_s}$ denotes the multinomial coefficient. Lemma 6 generalizes Lemma 3.1 of Schmid and Sly [2024] and essentially follows as a special case of Lemma 3 and equation (2) of Talvitie et al. [2020]. Still, for completeness, we give a full proof of Lemma 6 in Appendix A. Note that tower decompositions have also been studied under different names in the literature, such as root layerings [Talvitie et al., 2020] and DAG partitions [Kuipers and Moffa, 2015, 2017].

For uniformly random DAGs, we will additionally make use of some results of Schmid and Sly [2024]. Let $\sigma(G)$ be the graph obtained from G by keeping only the edges between adjacent layers H_i and H_{i+1} for all i. We call this graph the *tower* of G. Conditioning on the tower of G drastically simplifies the distribution of a uniformly random DAG:

Lemma 7. *[Schmid and Sly, 2024] Let G be a uniformly random DAG on n vertices and condition on the tower $\sigma(G) = \sigma$. Then, the edges of G between two non-adjacent layers occur independently with probability $1/2$ each.*

The following lemma gives a tail bound on the layer sizes in a uniformly random DAG G.

Lemma 8. *[Schmid and Sly, 2024] Let G be a uniformly random DAG on n vertices. Let $i, \ell \in \mathbb{N}$, $\ell \geq 5$, and n sufficiently large. Then, we have $\Pr[h_i(G) \geq \ell] \leq 2^{-\ell^2/4+2}$.*

3 SPARSE RANDOM DAGS

Our approach towards bounding the size of the Markov equivalence class of a sparse random DAG G is through bounding the number of reversible edges of G. By Lemma 5 an edge $v \to w$ is reversible if and only if the parent sets of v and w align. If we think about G in terms of its tower decomposition, vertices that lie in higher-order layers have many more potential parents, so an edge between such vertices should be less likely to be reversible (in fact, Schmid and Sly [2024] derive upper bounds on the size of MECs in uniformly random DAGs by making this intuition formal). However, if we consider an edge $v \to w$ where w lies in the second layer $H_2(G)$ (i.e. the set of vertices that are children of a source vertex), then v is necessarily a source vertex, so $\mathrm{Pa}(v) = \emptyset$. But this implies that $v \to w$ is reversible whenever w has no parents except v. Since w lies in the second layer of G, it can only have parents in the first layer, so checking that w has no connection to another source vertex suffices to establish reversibility of $v \to w$. In the following, we call an edge $v \to w$ with $w \in H_2(G)$ a *layer-2-edge*. For a uniformly random DAG, the first two layers contain with high probability only a few vertices, however, this is not the case for a sparse random DAG. In fact, we will show that the first two layers of a sparse random DAG G are with high probability quite large, which leads to the existence of many reversible layer-2-edges. This by itself does not yet suffice to give a significant lower bound on the size of the MEC of G. Indeed, two reversible edges need not be independently reversible, that is, after reversing one edge, the other one might become irreversible. However, suppose we have an edge set $S \subseteq E(G)$, which forms a *matching* of reversible edges, i.e. a set of reversible edges with disjoint endpoints. Then, reversing an edge in S only changes the parent sets of its endpoints and leaves all other parent sets the same. But this implies that all other edges in S must remain reversible by Lemma 5. Hence, all possible combinations of edge reversals for edges in S lead to another Markov equivalent DAG. We formulate this as the following observation:

Observation 9. *Let G be a DAG, and let $S \subseteq E(G)$ be a matching of reversible edges in G. Then, the size of the Markov equivalence class of G is at least $2^{|S|}$.*

Now, the key insight of this section is that we can find a large matching of reversible edges in a sparse random graph G by only looking at its layer-2-edges.

Proposition 10. *Let $6/n \leq p \leq o\left(\frac{1}{\log n}\right)$ and $G \sim D(n,p)$. Then, with probability $1 - o(1)$, G has a matching of reversible layer-2-edges of size at least $\frac{p^{-1}}{16e^5 \log^2(p^{-1})}$.*

See Figure 2 for an example of a matching of reversible layer-2-edges. To prove Proposition 10, we crucially rely on the following concentration bounds for the sizes of the first two layers of a sparse random DAG, which are of independent interest:

Lemma 11. *Let $p = o\left(\frac{1}{\log n}\right)$ and $G \sim D(n,p)$. Then, the number of sources in G is less than or equal to $\frac{5}{p}$ with high probability.*

Lemma 12. *Let $\frac{6}{n} \leq p \leq o\left(\frac{1}{\log n}\right)$ and $G \sim D(n,p)$. Then, with high probability, we have $h_1(G) \geq \frac{p^{-1}}{20\log(p^{-1})}$ and $h_2(G) \geq \frac{p^{-1}}{\log^2(p^{-1})}$.*

Lemma 11 follows by showing that a tower vector h with $h_1 > 5p^{-1}$ has exponentially small weight $w(h)$ compared to the vector that is obtained from h by splitting h_1 into equal parts of size p^{-1}. Here, $w(h)$ is given by Lemma 6, part (a). Similarly, Lemma 12 follows from showing that a vector h with $h_1 < p^{-1}/(20\log(p^{-1}))$ has exponentially

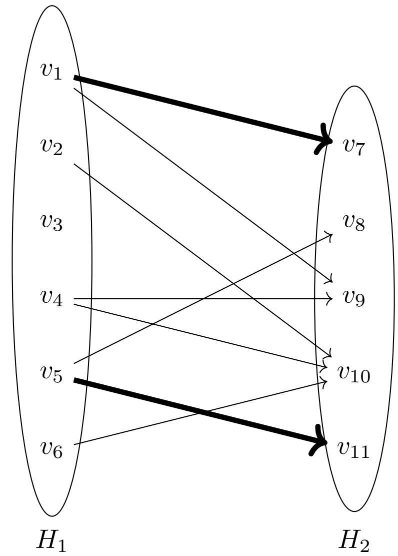


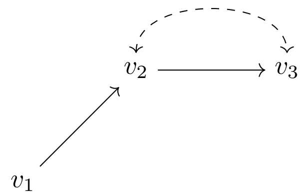
Figure 2: The bold edges form a (maximal, but not unique) matching of reversible layer-2-edges.

small weight compared to the vector that is obtained from h by merging its first layers until their size exceeds $p^{-1}/5$ (and similarly for h_2). We provide the full details of this calculation in Appendix A. In light of these concentration bounds, the statement of Proposition 10 becomes more approachable: Given a DAG G sampled from $D(n,p)$, we simply need to show that there is at least a small constant fraction of vertices in $H_2(G)$ that have only one parent in $H_1(G)$. As these parents should roughly be scattered randomly with just a few overlaps, most of these vertices together with their single parent then form a matching of reversible edges. But $H_2(G)$ contains $p^{-1}/\log^2(p^{-1})$ vertices with high probability, so the matching we found has the desired size. We give a full proof of Proposition 10 in Appendix A, which, together with Observation 9, immediately implies Theorem 1.

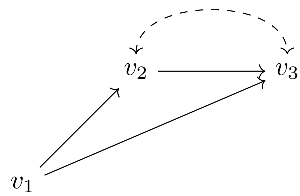
4 ACYCLIC DIRECTED MIXED GRAPHS

For DAGs it is known that any two Markov equivalent graphs must have the same edge adjacencies. This is not true anymore for ADMGs. Let us call an edge $v \to w$ of an ADMG G *underdetermined* if deleting it results in an ADMG that is Markov equivalent to G. Our approach towards proving Theorem 2, part (a) is to show that a uniformly random ADMG G contains many underdetermined edges that can be deleted or included independently of each other while preserving Markov equivalence. The following graph provides an example for an underdetermined edge:

Definition 13. *We define the ADMG S on an ordered set of*



Graph S.



Graph $\overline{S}$.

Figure 3: The edge from v_1 to v_3 is underdetermined: it does not introduce any new d-connections, so S and $\overline{S}$ are Markov equivalent.

three vertices (v_1, v_2, v_3) *as the graph with edges* $v_1 \to v_2$, $v_2 \to v_3$*, and* $v_2 \leftrightarrow v_3$*. Moreover, we denote the graph* $S \cup \{v_1 \to v_3\}$ *as* $\overline{S}$*, see Figure 3.*

It is easy to check that the edge $v_1 \to v_3$ is underdetermined in $\overline{S}$. A key observation is that this remains true even when $\overline{S}$ is part of a larger graph structure.

Lemma 14. *Let* G *be an ADMG that contains* $\overline{S}$ *as a (labeled) subgraph on the vertices* (v_1, v_2, v_3)*. Then, the edge* $v_1 \to v_3$ *is underdetermined in* G*.*

Proof. Let G be an ADMG containing $\overline{S}$ as a labeled subgraph on (v_1, v_2, v_3) and let G' be the graph obtained from deleting the edge $v_1 \to v_3$ from G. Since G' is a subgraph of G, any two vertices a, b that are d-separated given Z in G are also d-separated given Z in G'. Now, assume a and b are d-connected in G given Z. First, note that each vertex has the same descendants in G and G', since G' still contains a directed path from v_1 to v_3. Hence, if π is an active path from a to b given Z in G that does not contain the edge $v_1 \to v_3$, then π is also an active path from a to b given Z in G'. Otherwise, if π is an active path from a to b given Z in G that contains the edge $v_1 \to v_3$, then we distinguish two cases:
First, if $v_2 \notin Z$, then replacing the edge $v_1 \to v_3$ in π by the path $v_1 \to v_2 \to v_3$ gives an active path in G'. Secondly, if $v_2 \in Z$, then replacing the edge $v_1 \to v_3$ in π by the path $v_1 \to v_2 \leftrightarrow v_3$ gives an active path in G'. In all cases, we get that a and b are also d-connected given Z in G', so G and G' are Markov equivalent. $\square$

This leads to the following corollary:

Corollary 15. *Let* G *be an ADMG on* n *vertices and suppose there are* m *different subsets of vertices* $V_1, \ldots, V_m$ *such that* $|V_i| = 3$ *and the induced graph* $G[V_i]$ *contains a copy of* S *or of* $\overline{S}$ *for all* i*. Then,* $|\operatorname{MEC}(G)| \geq 2^{m/(3n)}$*.*

Proof. Given the setup as in the Corollary, note that each subset V_i only intersects with at most $3(n-1)$ other subsets in two vertices. Hence, we can select at least $m/(3n)$ different subsets V_i that pairwise intersect in at most one vertex. However, this implies that the copies of S and $\overline{S}$ on these vertex sets are all edge-disjoint. In particular, deleting or adding the underdetermined edge in these copies does not influence the other copies, so we can add or delete an edge for each of the sub-selected vertex sets independently while preserving Markov-equivalence. This implies $|\operatorname{MEC}(G)| \geq 2^{m/(3n)}$. $\square$

All that is left to do in order to prove Theorem 2, part (a) is to show that in a uniformly random ADMG G, a constant fraction of all possible vertex sets of size 3 is expected to have a copy of S or $\overline{S}$.

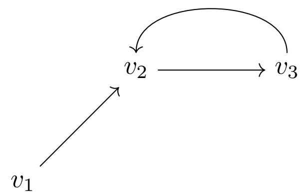
Proof of Theorem 2, part (a). Let G be a uniformly random ADMG on n vertices. Note that the distribution of G can be described by sampling a uniformly random DAG and adding a bidirected edge for each pair of vertices independently with probability $1/2$. We extend the notion of tower decompositions to ADMGs and denote by $\sigma(G)$ the tower of the DAG formed by the directed edges of G. Let $\mathcal{A}$ be the set of towers σ that only have layers of size at most $n/48$. For each $\sigma \in \mathcal{A}$, let $\mathcal{V}_\sigma$ be the set of all vertex subsets of size 3 that do not contain two vertices in adjacent layers. We have $|\mathcal{V}_\sigma| \geq \frac{1}{6} \cdot n \cdot \frac{23}{24} n \cdot \frac{11}{12} n \geq n^3/7$. By Lemma 7, after conditioning on $\sigma(G) = \sigma$, the edges induced by each $W \in \mathcal{V}_\sigma$ occur independently with probability $1/2$ each, so we get $\Pr(W \text{ contains a copy of } S \text{ or } \overline{S} \mid \sigma(G) = \sigma) \geq 1/8$. Then, if X denotes the number of vertex sets $W \in \mathcal{V}_\sigma$ that contain a copy of S or $\overline{S}$, we get $\mathbb{E}[X \mid \sigma(G) = \sigma] \geq n^3/56$. By Corollary 15 and Jensen's inequality, we deduce

$$\begin{aligned}\mathbb{E}\left[|\operatorname{MEC}(G)| \middle| \sigma(G) = \sigma\right] &\geq \mathbb{E}\left[2^{X/(3n)} \middle| \sigma(G) = \sigma\right] \\ &\geq 2^{n^2/168}.\end{aligned}$$

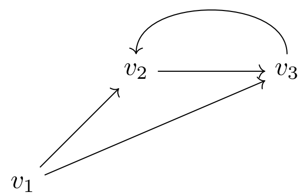
After summing over all $\sigma \in \mathcal{A}$, we conclude using Lemma 8:

$$\begin{aligned}\mathbb{E}\left[|\operatorname{MEC}(G)|\right] &\geq \Pr\left(\sigma(G) \in \mathcal{A}\right) \cdot 2^{n^2/168} \\ &= \Pr\left(h_i \leq n/48 \text{ for all } i\right) \cdot 2^{n^2/168} \\ &\geq \left(1 - n \cdot 2^{-n^2/9216+2}\right) \cdot 2^{n^2/168} \\ &= 2^{\Omega(n^2)}.\end{aligned}$$

$\square$



Graph T.



Graph $\overline{T}$.

Figure 4: Again, the edge from v_1 to v_3 is underdetermined.

We remark that it is straightforward to show tight concentration of the random variable X, counting the occurrence of underdetermined edges in G by Chebychev's inequality. This means that in fact almost all ADMGs on n vertices have at least $2^{\Omega(n^2)}$ Markov-equivalent graphs.

5 DIRECTED CYCLIC GRAPHS

5.1 UNDERDETERMINED EDGES

For DCGs, we can apply essentially the same reasoning as in the previous section. Consider the following DCG:

Definition 16. *We define the DCG T on an ordered set of three vertices (v_1, v_2, v_3) as the graph with edges $v_1 \to v_2$, $v_2 \to v_3$, and $v_3 \to v_2$. Moreover, we denote the graph $T \cup \{v_1 \to v_3\}$ as $\overline{T}$, see Figure 4.*

Again, it is easy to check that the edge $v_1 \to v_3$ is underdetermined in $\overline{T}$. The following statements hold:

Lemma 17. *Let G be a DCG that contains $\overline{T}$ as a (labeled) subgraph on the vertices (v_1, v_2, v_3). Then, the edge $v_1 \to v_3$ is underdetermined in G.*

Corollary 18. *Let G be a DCG on n vertices and suppose there are m different subsets of vertices $V_1, \dots, V_m$ such that $|V_i| = 3$ and $G[V_i]$ contains a copy of T or of $\overline{T}$ for all i. Then, $|\operatorname{MEC}(G)| \geq 2^{m/(3n)}$.*

The proofs of Lemma 17 and Corollary 18 are analogous to the proofs of Lemma 14 and Corollary 15. Now, the proof of Theorem 2, part (b) is simple:

Proof of Theorem 2, part (b). Let G be a uniformly random DCG on n vertices. For each subset W of three vertices, we get $\Pr(W \text{ contains a copy of } T \text{ or } \overline{T}) \geq 1/4$, since each possible directed edge occurs independently with probability $1/2$, and there are at least two edge-disjoint ways to embed T into W. Then, if X denotes the number of vertex sets W of size 3 that contain a copy of T or $\overline{T}$, we get $\mathbb{E}[X] \geq \binom{n}{3} \cdot \frac{1}{4} \geq n^3/25$. By Corollary 15 and Jensen's inequality, we conclude

$$\begin{aligned} \mathbb{E}\left[|\operatorname{MEC}(G)|\right] &\geq \mathbb{E}\left[2^{X/(3n)}\right] \\ &\geq 2^{n^2/75}. \end{aligned}$$

□

5.2 UNDERDETERMINED CYCLES

The existence of underdetermined edges in DCGs is not the only issue that leads to exponential-size Markov equivalence classes. Already Richardson [1996a] noticed that the direction of cycles in a DCG is underdetermined within its Markov equivalence class. Here, we extend this observation by giving a formal construction for how to reverse any cycle in a DCG while preserving Markov equivalence.

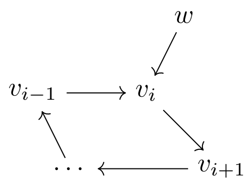
Definition 19. *Let G be a DCG with a cycle $C = (v_1, \dots, v_k)$. We construct the graph $H = \textsc{Reverse}(G, C)$ as follows:*

- *Take a copy of G and reverse the cycle C, i.e. replace the edge $v_i \to v_{i+1}$ by $v_i \leftarrow v_{i+1}$ for all $i \in [k]$;*
- *For each vertex $v_i \in C$, and each vertex $w \in \operatorname{Pa}(v_i) \setminus \{v_{i-1}\}$, delete the edge $w \to v_i$, and replace it by $w \to v_{i-1}$.*

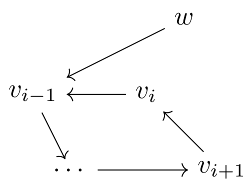
We state the key property of this construction:

Proposition 20. *Let G be a directed cyclic graph containing a cycle C. Then, the graph $H = \textsc{Reverse}(G, C)$ is Markov equivalent to G.*

This result and the entire Reverse construction is based on the following intuition: consider a segment $v_{i-1} \to v_i \to v_{i+1}$ of a cycle C in G together with an incoming edge $w \to v_i$, see Figure 5. If we only reverse the orientation of C without changing the edge $w \to v_i$, then w becomes d-separated from v_{i-1} given $\{v_i, v_{i+1}\}$ in the resulting graph, while it was d-connected to v_{i-1} given $\{v_i, v_{i+1}\}$ in G. This motivates connecting w to v_{i-1} instead of v_i in $H = \textsc{Reverse}(G, C)$. Crucially, this new edge does not introduce any new d-connections, since w and v_{i-1} are in fact d-connected in G given any conditioning set Z: if Z contains a vertex of C, then $w \to v_i \leftarrow v_{i-1}$ is an active path, and if Z and C are disjoint, then $w \to v_i \to v_{i+1} \to \dots v_{i-1}$ (following the cycle) is an active path. By the same argument, w and v_i are d-connected given any set Z in the new graph H, so deleting the edge $w \to v_i$ when going from G to H



Cycle segment of G.



Corresponding segment of the graph $H = \textsc{Reverse}(G, C)$.

Figure 5: Illustration of the REVERSE construction.

does not introduce any new d-separations. By a careful case analysis, one can extend this argumentation to a full proof of Markov equivalence between G and H, see Appendix B.

Now, suppose the DCG G has multiple vertex-disjoint cycles. By construction, the operation REVERSE applied to any of the cycles, leaves all other cycles the same. Hence, one can obtain a new Markov-equivalent DCG for each combination of orientations of these cycles. This results in the following corollary:

Corollary 21. *Let G be a directed cyclic graph that contains k vertex-disjoint cycles of length at least 3. Then, $|\operatorname{MEC}^*(G)| \geq 2^k$.*

This is all we need to prove Theorem 3.

Proof of Theorem 3. Let G be a uniformly random DCG on n vertices. Without loss of generality, we may assume that n is divisible by 3 and partition the vertex set V of G into disjoint sets $V_1, \dots, V_{n/3}$ of size 3 each. Then, $\Pr(V_i \text{ induces a cycle of length } 3) = 1/4$. Let X denote the number of vertex sets V_i that induce a cycle. We have $\mathbb{E}[X] = n/12$. By Corollary 21 and Jensen's inequality, we conclude

$$\begin{aligned} \mathbb{E}\left[|\operatorname{MEC}^*(G)|\right] &\geq \mathbb{E}\left[2^X\right] \\ &\geq 2^{n/12}. \end{aligned}$$

□

Again, we remark that it is straightforward to get tight concentration bounds on the number of underdetermined edges and vertex-disjoint cycles that we are using to bound $|\operatorname{MEC}(G)|$ and $|\operatorname{MEC}^*(G)|$, using standard tools such as Chebyshev's or Chernoff's inequality. This means that almost all DCGs have exponentially large Markov equivalence classes.

6 DISCUSSION

Our results indicate significant underdetermination of causal structure from observational data in various settings. Particularly for DAGs, it might come as a surprise that previously known results on Markov equivalence classes being small for dense graphs do not extend to the sparse setting. However, our results are asymptotic in nature, calling for further numerical studies on how the expected size of Markov equivalence classes scales for small graphs with different edge densities. We hope that our theoretical results shed light on some of the sources of underdetermination in causal graphs: For DAGs, our proof of Theorem 1 could potentially be extended to show that edges are less likely to have determined directions when they appear between earlier layers of the tower decomposition. For ADMGs, we show that, on average, underdetermination arises frequently from small structures that induce underdetermined edges. This issue is addressed in the model class of maximal ancestral graphs (MAGs) through the maximality condition [Richardson and Spirtes, 2002], which forces underdetermined edges to be included in the graph and therefore makes the set of adjacencies unique again within a Markov equivalence class. As a result, MAGs may exhibit smaller Markov equivalence classes on average, which would be an interesting subject of further research (see Wang et al. [2024] for a first approach). However, the analysis becomes challenging as, to our knowledge, the enumeration of MAGs is an open problem. Note that our proof of Theorem 2, part (a) also implies that under the uniform distribution, there is an expected super-exponential number of ADMGs corresponding to just a single MAG. In fact, most edges of a uniformly random ADMG are underdetermined with high probability, raising further questions about the informativeness of MAGs in this setting. For DCGs, addressing underdetermination seems to not only require a maximality condition but also some type of condition for directing cycles in light of our proof of Theorem 3.

Apart from ruling out certain graphs through additional assumptions, interventions also help to further distinguish between Markov-equivalent graphs. It would be interesting to explore the expected number of interventions required to uniquely recover a random causal graph (see Katz et al. [2019] for a first approach). Based on our results, this question becomes more significant for sparse DAGs, but also for ADMGs and DCGs, where it is not even known how to place interventions to efficiently split the Markov equivalence class. Finally, in practice, statistical uncertainty often prevents the identification of a single Markov equivalence class. Instead, researchers are faced not just with multiple graphs within one equivalence class, but often with sets of equivalence classes. An understanding of which equivalence classes are "close" to each other or how to separate equivalence classes in causal discovery more effectively, would be enormously useful.

Acknowledgements

This research was supported by NSF grant CCF-2321079.

References

David Maxwell Chickering. A transformational characterization of equivalent bayesian network structures. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 87–98, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603859.

David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *J. Mach. Learn. Res.*, 2: 445–498, March 2002. ISSN 1532-4435. URL https://doi.org/10.1162/153244302760200696.

David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(null): 507–554, March 2003. ISSN 1532-4435. URL https://doi.org/10.1162/153244303321897717.

Max Chickering. Statistically efficient greedy equivalence search. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 241–249. PMLR, 03–06 Aug 2020. URL https://proceedings.mlr.press/v124/chickering20a.html.

Ralf Eggeling, Jussi Viinikka, Aleksis Vuoksenmaa, and Mikko Koivisto. On structure priors for learning bayesian networks. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1687–1695. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/eggeling19a.html.

Dan Geiger, Thomas Verma, and Judea Pearl. Identifying independence in bayesian networks. *Networks*, 20(5):507–534, 1990. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/net.3230200504.

Steven B. Gillispie. Formulas for counting acyclic digraph markov equivalence classes. *Journal of Statistical Planning and Inference*, 136(4):1410–1432, 2006. ISSN 0378-3758. URL https://www.sciencedirect.com/science/article/pii/S0378375804003982.

Steven B. Gillispie and Michael D. Perlman. Enumerating markov equivalence classes of acyclic digraph dels. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 171–177, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

Steven B. Gillispie and Michael D. Perlman. The size distribution for markov equivalence classes of acyclic digraph models. *Artificial Intelligence*, 141(1):137–155, 2002. ISSN 0004-3702. URL https://www.sciencedirect.com/science/article/pii/S0004370202002643.

Yangbo He, Jinzhu Jia, and Bin Yu. Counting and exploring sizes of markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 16(79):2589–2609, 2015. URL http://jmlr.org/papers/v16/he15a.html.

David Heckerman, Dan Geiger, and David M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995. URL https://doi.org/10.1023/A:1022623210503.

Shuai Huang, Jing Li, Jieping Ye, Adam Fleisher, Kewei Chen, Teresa Wu, and Eric Reiman. A sparse structure learning algorithm for gaussian bayesian network identification from high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (6):1328–1342, 2013. ISSN 0162-8828. URL https://doi.org/10.1109/TPAMI.2012.129.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(22): 613–636, 2007. URL http://jmlr.org/papers/v8/kalisch07a.html.

Dmitriy Katz, Karthikeyan Shanmugam, Chandler Squires, and Caroline Uhler. Size of interventional markov equivalence classes in random dag models. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3234–3243. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/katz19a.html.

Jan T. A. Koster. On the validity of the markov interpretation of path diagrams of gaussian structural equations systems with correlated errors. *Scandinavian Journal of Statistics*, 26(3):413–431, 1999. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9469.00157.

Jack Kuipers and Giusi Moffa. Uniform random generation of large acyclic digraphs. *Statistics and Computing*, 25(2):227–242, 2015. URL https://doi.org/10.1007/s11222-013-9428-y.

Jack Kuipers and Giusi Moffa. Partition mcmc for inference on acyclic digraphs. *Journal of the American Statistical Association*, 112(517):282–299, 2017. URL https://doi.org/10.1080/01621459.2015.1133426.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, Cambridge, 2005. ISBN 978-0-521-83540-4. URL https://doi.org/10.1017/CBO9780511813603.

J. Pearl. *Causality*. Cambridge, 2nd edition, 2009.

Adityanarayanan Radhakrishnan, Liam Solus, Caroline Uhler, and Martin J. Wainwright. Counting markov equivalence classes by number of immoralities. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, Sydney, Australia, August 11–15 2017. AUAI Press.

Garvesh Raskutti and Caroline Uhler. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1):e183, 2018. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.183.

Thomas Richardson. A polynomial-time algorithm for deciding markov equivalence of directed cyclic graphical models. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, page 462–469, San Francisco, CA, USA, 1996a. Morgan Kaufmann Publishers Inc. ISBN 155860412X.

Thomas Richardson. Discovering cyclic causal structure. 1996b. URL https://doi.org/10.1184/R1/6491426.v1.

Thomas Richardson. A characterization of markov equivalence for directed cyclic graphs. *International Journal of Approximate Reasoning*, 17(2):107–162, 1997. ISSN 0888-613X. URL https://www.sciencedirect.com/science/article/pii/S0888613X97000200. Uncertainty in AI (UAI'96) Conference.

Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962 – 1030, 2002. URL https://doi.org/10.1214/aos/1031689015.

Dominik Schmid and Allan Sly. On the number and size of markov equivalence classes of random directed acyclic graphs, 2024. URL https://arxiv.org/abs/2209.04395.

Peter Spirtes. Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 491–498, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603859.

Peter Spirtes and Thomas S. Richardson. A polynomial time algorithm for determining dag equivalence in the presence of latent variables and selection bias. In David Madigan and Padhraic Smyth, editors, *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, volume R1 of *Proceedings of Machine Learning Research*, pages 489–500. PMLR, 04–07 Jan 1997. URL https://proceedings.mlr.press/r1/spirtes97b.html. Reissued by PMLR on 30 March 2021.

Peter Spirtes, Thomas Richardson, Christopher Meek, Richard Scheines, and Clark Glymour. Using path diagrams as a structural equation modeling tool. *Sociological Methods & Research*, 27(2):182–225, 1998. URL https://doi.org/10.1177/0049124198027002003.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press, 01 2001. ISBN 9780262284158. URL https://doi.org/10.7551/mitpress/1754.001.0001.

Peter L. Spirtes, Christopher Meek, and Thomas S. Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. In *Computation, Causation, and Discovery*. AAAI Press, 05 1999. ISBN 9780262315821. URL https://doi.org/10.7551/mitpress/2006.003.0009.

Bertran Steinsky. Enumeration of labelled chain graphs and labelled essential directed acyclic graphs. *Discrete Mathematics*, 270(1):267–278, 2003. ISSN 0012-365X. URL https://www.sciencedirect.com/science/article/pii/S0012365X02008385.

Bertran Steinsky. Enumeration of labelled essential graphs. *Ars Combinatoria*, 111:485–494, 2013. URL https://combinatorialpress.com/article/ars/Volume%20111/volume-111-paper-40.pdf.

Topi Talvitie and Mikko Koivisto. Counting and sampling markov equivalent directed acyclic graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 7984–7991, Jul. 2019. URL https://ojs.aaai.org/index.php/AAAI/article/view/4799.

Topi Talvitie, Aleksis Vuoksenmaa, and Mikko Koivisto. Exact sampling of directed acyclic graphs from modular distributions. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 965–974. PMLR, 22–25 Jul 2020. URL https://proceedings.mlr.press/v115/talvitie20a.html.

Thomas Verma and Judea Pearl. Causal networks: Semantics and expressiveness. In Ross D. SHACHTER, Tod S.

LEVITT, Laveen N. KANAL, and John F. LEMMER, editors, *Uncertainty in Artificial Intelligence*, volume 9 of *Machine Intelligence and Pattern Recognition*, pages 69–76. North-Holland, 1990. URL https://www.sciencedirect.com/science/article/pii/B9780444886507500111.

TS Verma and Judea Pearl. Equivalence and synthesis of causal models. In M. Henrion, R. Shachter, L. Kanal, and J. Lemmer, editors, *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, page 220–227, 1991.

Tian-Zuo Wang, Wen-Bo Du, and Zhi-Hua Zhou. An efficient maximal ancestral graph listing algorithm. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 50353–50378. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wang24o.html.

Marcel Wienöbst, Max Bannach, and Maciej Liśkiewicz. Polynomial-time algorithms for counting and sampling markov equivalent dags with applications. *Journal of Machine Learning Research*, 24(213):1–45, 2023. URL http://jmlr.org/papers/v24/22-0495.html.

Proofs
(Supplementary Material)

Erik Jahn[1] Frederick Eberhardt[2] Leonard J. Schulman[1]

[1]Engineering and Applied Science, California Institute of Technology, Pasadena, California, USA
[2]Humanities and Social Sciences, California Institute of Technology, Pasadena, California, USA

A PROOFS FOR SPARSE DAGS

Here, we restate and prove our technical results on random DAGs sampled from $D(n,p)$.

Lemma 6. *Let $G \sim D(n,p)$, and fix a tower vector $h = (h_1,\dots,h_s)$ with entries that sum up to n, and a tower decomposition $H = (H_1 \dots, H_s)$ with $|H_i| = h_i$. Then, the following holds:*

(a) the probability that h is the tower vector of G is proportional to

$$w(h) = \binom{n}{h_1,\dots,h_s} \prod_{k=2}^{s} \frac{\left(1-(1-p)^{h_{k-1}}\right)^{h_k}}{(1-p)^{h_k \sum_{i=1}^{k-1} h_i}};$$

(b) conditional on the event that H is the tower decomposition of G, the parent sets $\mathrm{Pa}(v)$ are independently distributed for all $v \in V(G)$, and for $v \in H_j$ and $S \subseteq \bigcup_{i=1}^{j-1} H_i$, we have

$$\Pr\left(\mathrm{Pa}(v) = S\right) = \begin{cases} C \cdot \left(\frac{p}{1-p}\right)^{|S|} & \textit{if } S \cap H_{j-1} \neq \emptyset \\ 0 & \textit{otherwise,} \end{cases}$$

where C is the normalization constant.

Proof. Fix a tower vector $h = (h_1,\dots,h_s)$ with entries that sum up to n and a tower decomposition $H = (H_1,\dots,H_s)$ with $|H_k| = h_k$. The key observation is that a DAG G has tower decomposition H if and only if for each vertex $v \in H_k$, the parents of v are a subset of $\bigcup_{i=1}^{k-1} H_i$ that intersects H_{k-1} (Lemma 3 of Talvitie et al. [2020]). We define

$$\mathcal{C}_k = \{S \subseteq \bigcup_{i=1}^{k-1} H_i \mid S \cap H_{k-1} \neq \emptyset\};\ \mathcal{C}_{k,s} = \{S \in \mathcal{C}_k \mid |S| = s\},$$

setting $H_0 = \emptyset$ to cover the case $k = 1$. In particular, given that H is the tower decomposition of G, the parent sets of each vertex in H_k can be chosen independently in $\mathcal{C}_k$, which together with Definition 4 of $D(n,p)$ implies part (b) of the Lemma.

Moreover, we get that the probability for $G \sim D(n,p)$ to have tower decomposition H is proportional to

$$\begin{aligned}
w(H) := \sum_{G:H(G)=H} \prod_{k=1}^{s} \prod_{v\in H_k} \left(\frac{p}{1-p}\right)^{|\operatorname{Pa}(v)|} \\
&= \prod_{k=1}^{s} \prod_{v\in H_k} \sum_{P\in\mathcal{C}_k} \left(\frac{p}{1-p}\right)^{|P|} \\
&= \prod_{k=1}^{s} \prod_{v\in H_k} \sum_{s=0}^{n} \left(\frac{p}{1-p}\right)^{s} \cdot |\mathcal{C}_{k,s}| \\
&= \prod_{k=2}^{s} \prod_{v\in H_k} \sum_{s=0}^{n} \left(\frac{p}{1-p}\right)^{s} \cdot \left(\binom{\sum_{i=1}^{k-1} h_i}{s} - \binom{\sum_{i=1}^{k-2} h_i}{s}\right) \\
&= \prod_{k=2}^{s} \prod_{v\in H_k} \left(\left(1+\frac{p}{1-p}\right)^{\sum_{i=1}^{k-1} h_i} - \left(1+\frac{p}{1-p}\right)^{\sum_{i=1}^{k-2} h_i}\right) \\
&= \prod_{k=2}^{s} (1-p)^{-h_k \sum_{i=1}^{k-1} h_i} \cdot \left(1-(1-p)^{h_{k-1}}\right)^{h_k}.
\end{aligned}$$

Now part (a) of Lemma 6 follows from summing over all $\binom{n}{h_1,\ldots,h_s}$ choices of tower decompositions with layer sizes $h_1,\ldots,h_s$. ☐

Lemma 11. *Let $p = o\left(\frac{1}{\log n}\right)$ and $G \sim D(n,p)$. Then, the number of sources in G is less than or equal to $\frac{5}{p}$ with high probability.*

Proof. Let $p = o\left(\frac{1}{\log n}\right)$. Without loss of generality, we assume that p^{-1} is an integer. For each index j, we define $\mathcal{H}_j$ to be the set of tower vectors $h = (h_1,\ldots,h_s)$ with entries that sum up to n and $h_j \geq 5p^{-1} \geq h_{j+1}$. Consider the functions φ_j that map $h = (h_1,\ldots,h_s) \in \mathcal{H}_j$ to the vector $\varphi_j(h) = (h_1,\ldots,h_{j-1},\ell_1,\ldots,\ell_r,h_{j+1},\ldots,h_s)$, where r is equal to $h_j p$ rounded to the nearest integer, $\ell_2 = \ell_3 = \cdots = \ell_r = p^{-1}$ and $\ell_1 = h_j - \sum_{i=2}^{r} \ell_i$. Note that the entries of $\varphi_j(h)$ are still positive integers that sum up to n, so it is a valid tower vector. Moreover, $\ell_1 \in [p^{-1}/2, 3p^{-1}/2]$. We are now going to bound the ratio of the weights of h and $\varphi_j(h)$. By Lemma 6,

$$\frac{w(h)}{w(\varphi_j(h))} = \frac{\prod_{i=1}^{r} \ell_i!}{h_j!}(1-p)^{\sum_{1\leq i<k\leq r} \ell_i\ell_k} \cdot \frac{(1-(1-p)^{h_{j-1}})^{h_j-\ell_1}}{\prod_{i=2}^{r}(1-(1-p)^{\ell_{i-1}})^{\ell_i}} \cdot \left(\frac{1-(1-p)^{h_j}}{1-(1-p)^{\ell_r}}\right)^{h_{j+1}}.$$

For the edge cases of this formula to be correct, we set $h_0 = \infty$ and $h_{s+1} = 0$. We bound each term separately: First, since $\sum_{i=1}^{r} \ell_i = h_j$, we have $\prod_{i=1}^{r} \ell_i! \leq h_j!$.
Secondly, since $r \geq 5$, we get the estimate

$$\begin{aligned}
\sum_{1\leq i<k\leq r} \ell_i\ell_k &= \ell_r(h_j-\ell_r) + \ell_{r-1}(h_j-\ell_r-\ell_{r-1}) + \sum_{1\leq i<k\leq r-2} \ell_i\ell_k \\
&= 2h_jp^{-1} - 3p^{-1} + \sum_{1\leq i<k\leq r-2} \ell_i\ell_k \geq 2h_jp^{-1}.
\end{aligned}$$

Thirdly, since $\ell_i \geq (2p)^{-1}$, we may use

$$1-(1-p)^{\ell_i} \geq 1-(1-p)^{1/(2p)} \geq 1-e^{-1/2} \geq 7/20,$$

and

$$1-(1-p)^{\ell_r} = 1-(1-p)^{1/p} \geq 1-e^{-1} \geq 1/2.$$

Finally, using the condition $h_j \geq 5p^{-1} \leq h_{j+1}$, we obtain

$$\frac{w(h)}{w(\varphi_j(h))} \leq (1-p)^{2h_jp^{-1}} \cdot (20/7)^{\sum_{i=2}^{r} \ell_i} \cdot 2^{h_{j+1}} \leq e^{-2h_j+\log(20/7)\cdot h_j+\log(2)\cdot h_j} \leq e^{-h_j/4} \leq e^{-p^{-1}}.$$

Now, let $G \sim D(n,p)$. Note that each function φ_j maps at most np vectors to the same image (the extreme case is when $\varphi_1(h) = (p^{-1}, \ldots, p^{-1})$). We calculate

$$\Pr\left(h_j(G) \geq 5p^{-1} \geq h_{j+1}(G)\right) = C \cdot \sum_{h \in \mathcal{H}_j} w(h) \leq C \cdot \sum_{h \in \mathcal{H}_j} w(\varphi(h)) \cdot e^{-p^{-1}} \leq np \cdot e^{-p^{-1}},$$

where $C = \left(\sum_h w(h)\right)^{-1}$ is the normalization constant. We conclude

$$\begin{aligned}\Pr\left(h_1(G) \geq 5p^{-1}\right) &\leq \Pr\left(h_1(G) \geq 5p^{-1} \geq h_2(G)\right) + \Pr\left(h_2(G) \geq 5p^{-1}\right) \leq \ldots \\ &\leq \sum_{i=1}^{T(G)-1} \Pr\left(h_j(G) \geq 5p^{-1} \geq h_{j+1}(G)\right) \leq n^2 p \cdot e^{-p^{-1}} = o(1).\end{aligned}$$

□

We will need the following bound in the proof of the next Lemma:

Lemma 22. *Let $h_1, \ldots, h_r$ be positive integers that sum up to n. Then,*

$$\binom{n}{h_1, \ldots, h_r} \leq \frac{n^n}{h_r^{h_1} \cdot \prod_{i=2}^{r} h_{i-1}^{h_i}}$$

Proof. The lemma simply follows from:

$$n^n = \left(\sum_{i=1}^{r} h_i\right)^n = \sum_{j_1+\cdots+j_r=n} \binom{n}{j_1, \ldots, j_r} h_1^{j_1} \cdot \ldots \cdot h_r^{j_r} \geq \binom{n}{h_2, \ldots, h_r, h_1} h_1^{h_2} \cdot \ldots \cdot h_{r-1}^{h_r} \cdot h_r^{h_1}.$$

□

Lemma 12. *Let $\frac{6}{n} \leq p \leq o\left(\frac{1}{\log n}\right)$ and $G \sim D(n,p)$. Then, with high probability, we have $h_1(G) \geq \frac{p^{-1}}{20\log(p^{-1})}$ and $h_2(G) \geq \frac{p^{-1}}{\log^2(p^{-1})}$.*

Proof. Let $6/n \leq p \leq o\left(\frac{1}{\log n}\right)$. First, we define $\mathcal{H}_1$ to be the set of tower vectors $h = (h_1, \ldots, h_s)$ with entries that sum up to n and $h_1 \leq p^{-1}/(20\log(p^{-1}))$. Consider the function φ_1 that maps a tower vector $h = (h_1, \ldots, h_s) \in \mathcal{H}_1$ to the vector $\varphi_1(h) = (L, h_{r+1}, \ldots, h_s)$, where $r = \min\{k : \sum_{i=1}^{k} h_i > p^{-1}/5\}$ and $L = \sum_{i=1}^{r} h_i$. Note that $\varphi_1(h)$ is again a valid tower vector, and by Lemma 6, we get

$$\begin{aligned}\frac{w(h)}{w(\varphi_1(h))} &= \frac{L!}{\prod_{i=1}^{r} h_i!}(1-p)^{-\sum_{1\leq i<k\leq r} h_i h_k} \cdot \prod_{i=2}^{r}(1-(1-p)^{h_{i-1}})^{h_i} \cdot \left(\frac{1-(1-p)^{h_r}}{1-(1-p)^L}\right)^{h_{r+1}} \\ &\leq \frac{L!}{\prod_{i=1}^{r} h_i!}(1-p)^{-\sum_{1\leq i<k\leq r} h_i h_k} \cdot \prod_{i=2}^{r}(1-(1-p)^{h_{i-1}})^{h_i}. \qquad (1)\end{aligned}$$

We are interested in further bounding expression (1) under the constraint $L = \sum_{i=1}^{r} h_i$. First of all, note that increasing h_r by 1, while leaving $h_1, \ldots, h_{r-1}$ the same, changes expression (1) by a factor of

$$\frac{L+1}{h_r+1} \cdot (1-p)^{-\sum_{i=1}^{r-1} h_i} \cdot (1-(1-p)^{h_{r-1}}) \leq \frac{L}{h_r} \cdot (1-p)^{-p^{-1}/5} \cdot p h_{r-1} \leq \frac{L}{h_r} \cdot e^{1/5} \cdot \frac{1}{5}.$$

Here, we used the fact that $\sum_{i=1}^{r-1} h_i \leq p^{-1+\varepsilon}$. This factor is strictly less than 1 if, $h_r > L/4$, hence, we may assume $h_r \leq L/4$, and therefore $L \leq 4p^{-1}/15$. Now, we can bound (1) using Lemma 22:

$$\begin{aligned}\frac{w(h)}{w(\varphi_1(h))} &\leq \frac{L^L}{h_r^{h_1} \cdot \prod_{i=2}^{r} h_{i-1}^{h_i}}(1-p)^{-L^2} \prod_{i=2}^{r}(h_{i-1}p)^{h_i} \\ &\leq L^L e^{pL^2} p^{L-h_1} \leq \left(\frac{4}{15}\right)^L p^{-L} e^{4L/15} p^L p^{p^{-1}/(20\log p)} \\ &\leq e^{L(\log(4/15)+4/15))+p^{-1}/20} \leq e^{p^{-1}/5\cdot(\log(4/15)+4/15+1/4)} \leq \left(\frac{5}{11}\right)^{p^{-1}/5}.\end{aligned}$$

Note that for a fixed tower vector of the form $\varphi_1(h) = (L, h_{r+1}, \ldots, h_s)$, the entries $h_1, \ldots, h_{r-1}$ of the preimage must sum up to a number less than $p^{-1}/5$, for which there are less than $2^{p^{-1}/5}$ choices. After choosing $h_1, \ldots, h_{r-1}$, the entry $h_r = L - \sum_{i=1}^{r-1}$ and the rest of the preimage is uniquely defined. Hence, the function φ_1 sends at most $2^{p^{-1}/5}$ tower vectors to the same image. Now, let $G \sim D(n,p)$ and $C = (\sum_h w(h))^{-1}$. We get

$$\begin{aligned}\Pr\left(h_1(G) \leq \frac{p^{-1}}{20\log(p^{-1})}\right) &= C \cdot \sum_{h \in \mathcal{H}_1} w(h) \leq C \cdot \sum_{h \in \mathcal{H}_1} w(\varphi_1(h)) \cdot \left(\frac{5}{11}\right)^{p^{-1}/5} \\ &\leq 2^{p^{-1}/5} \cdot \left(\frac{5}{11}\right)^{p^{-1}/5} = \left(\frac{10}{11}\right)^{p^{-1}/5} = o(1).\end{aligned}$$

Now, let $\mathcal{H}_2$ be the set of tower vectors $h = (h_1, \ldots, h_s)$ with the following properties:

(a) the entries h_i sum up to n;

(b) $\frac{p^{-1}}{20\log(p^{-1})} \leq h_1 \leq 5p^{-1}$;

(c) $h_2 \leq \frac{p^{-1}}{\log^2(p^{-1})}$.

Consider the function φ_2 that maps a tower vector $h = (h_1, \ldots, h_s) \in \mathcal{H}_2$ to the vector $\varphi_2(h) = (h_1, L, h_{r+1}, \ldots, h_s)$, where $r = \min\{k : \sum_{i=2}^{k} h_i > p^{-1}/(800\log(p^{-1}))\}$ and $L = \sum_{i=2}^{r} h_i$. This is well-defined, as $p \geq 6/n$, so $\sum_{i=2}^{s} h_i \geq n - 5p^{-1} \geq n/6 > p^{-1}/(800\log(p^{-1}))$. Moreover, $\varphi_2(h)$ is again a valid tower vector, and by Lemma 6, we get

$$\frac{w(h)}{w(\varphi_2(h))} \leq \frac{L!}{\prod_{i=2}^{r} h_i!}(1-p)^{-\sum_{2\leq i<k\leq r} h_i h_k} \cdot \frac{\prod_{i=2}^{r}(1-(1-p)^{h_{i-1}})^{h_i}}{(1-(1-p)^{h_1})^{L}}.$$

Here, increasing h_r by 1, while leaving $h_1, \ldots, h_{r-1}$ the same, changes the expression above by a factor of

$$\frac{L+1}{h_r+1} \cdot (1-p)^{-\sum_{i=2}^{r-1} h_i} \cdot \frac{1-(1-p)^{h_{r-1}}}{1-(1-p)^{h_1}} \leq \frac{L}{h_r} \cdot (1-p)^{-p^{-1}/800} \cdot \frac{6h_{r-1}}{h_1} \leq \frac{L}{h_r} \cdot \frac{1}{4}.$$

Here, we used the Bernoulli-type inequalities $1 - pm \leq (1-p)^m \leq 1 - pm/(1+pm)$ together with the fact $h_1 \leq 5p^{-1}$. The factor above is strictly less than 1 if, $h_r > L/4$, hence, we may assume $h_r \leq L/4$, and therefore $L \leq p^{-1}/(600\log(p^{-1}))$. Reusing the bound on (1) from before, we get:

$$\begin{aligned}\frac{w(h)}{w(\varphi_1(h))} &\leq L^L e^{pL^2}\left(\frac{6p}{h_1 p}\right)^{L-h_2} \leq L^L e^{L/600}\left(\frac{6}{30L}\right)^{L-h_2} \\ &\leq \left(\frac{1}{4}\right)^{L} \cdot e^{\log(5L)\cdot h_2} \leq \left(\frac{1}{4}\right)^{p^{-1}/(800\log(p^{-1}))} e^{\log(5L)\cdot p^{-1}/\log^2(p^{-1})} \leq \left(\frac{1}{3}\right)^{p^{-1}/(800\log(p^{-1}))},\end{aligned}$$

where the last step follows when p is small enough. Now, the function φ_2 sends at most $2^{p^{-1}/(800\log(p^{-1}))}$ tower vectors to the same image. Let $G \sim D(n,p)$ and $C = (\sum_h w(h))^{-1}$. We conclude using Lemma 11 and the result above:

$$\begin{aligned}\Pr\left(h_2(G) \leq \frac{p^{-1}}{\log^2(p^{-1})}\right) &\leq \Pr\left(h_2(G) \leq \frac{p^{-1}}{\log^2(p^{-1})} \wedge \frac{p^{-1}}{20\log(p^{-1})} \leq h_1 \leq 5p^{-1}\right) \\ &\quad + \Pr\left(h_1(G) < \frac{p^{-1}}{20\log(p^{-1})}\right) + \Pr\left(h_1(G) > 5p^{-1}\right) \\ &\leq C \cdot \sum_{h \in \mathcal{H}_2} w(h) + o(1) \\ &\leq C \cdot \sum_{h \in \mathcal{H}_2} w(\varphi_2(h)) \cdot \left(\frac{1}{3}\right)^{p^{-1}/(800\log(p^{-1}))} + o(1) \\ &\leq \left(\frac{2}{3}\right)^{p^{-1}/(800\log(p^{-1}))} + o(1) = o(1).\end{aligned}$$

□

In the final proof of this section, we will make use of the Chernoff bound:

Lemma 23. *(Chernoff bound, see Mitzenmacher and Upfal [2005], Theorem 4.5) Let $X_1, \dots, X_n$ be independent random variables with $\Pr(X_i = 1) = p$ and $\Pr(X_i = 0) = 1 - p$ and define $X = \sum_{i=1}^{n} X_i$. Then, for any $\delta \in (0, 1)$, we have*

$$\Pr\left(X \le (1-\delta)np\right) \le e^{-np\delta^2/2}.$$

Proposition 10. *Let $6/n \le p \le o\left(\frac{1}{\log n}\right)$ and $G \sim D(n,p)$. Then, with probability $1 - o(1)$, G has a matching of reversible layer-2-edges of size at least $\frac{p^{-1}}{16e^5 \log^2(p^{-1})}$.*

Proof. Let $6/n \le p \le o\left(\frac{1}{\log n}\right)$ and $G \sim D(n,p)$. For technical reasons that will become apparent later, let us define $\widetilde{H_2}(G)$ to be a subset of $H_2(G)$ of at most $h_1(G)/2$ vertices. This subset can be picked by an arbitrary but fixed rule (for instance, take the $h_1(G)/2$ vertices in $H_2(G)$ with the smallest indices, according to some indexing of the vertex set $V(G)$), and $\widetilde{H_2}(G) = H_2(G)$ in the case $h_2(G) \le h_1(G)/2$. Now, let $S(G) \subseteq \widetilde{H_2}(G)$ be the set of vertices in $\widetilde{H_2}(G)$ that have exactly one parent. Furthermore, let $X(G)$ be the set of vertices in S whose parent is different from all the other parents of vertices in S. By definition, the edges that connect vertices in $X(G)$ to their parents form a matching of reversible layer-2-edges. Hence, it suffices to bound $|X(G)|$. First, note that conditional on the tower decomposition $H(G) = H = (H_1, \dots, H_s)$ and on the event $S(G) = S$, the parent vertex of each vertex $v \in S$ is distributed uniformly and independently in H_1. Uniformity follows from the symmetry of the $D(n,p)$ distribution with respect to permuting vertices, and independence follows from Lemma 6, part (b). By a union bound, we get for $v \in S$:

$$\begin{aligned}
&\Pr\left(v \in X(G) \mid H(G) = H, S(G) = S\right) \\
&\ge 1 - \sum_{w \in S, w \ne v} \Pr\left(\mathrm{Pa}(v) = \mathrm{Pa}(w)\right) \\
&= 1 - \frac{|S|}{h_1} \ge 1 - \frac{|\widetilde{H_2}|}{h_1} \ge \frac{1}{2}.
\end{aligned}$$

Hence, $\mathbb{E}[|X(G)| \mid H(G) = H, S(G) = S] \ge |S|/2$, and by a standard Chernoff concentration bound (see Lemma 23), we get

$$\begin{aligned}
&\Pr\left(|X(G)| \le |S|/4 \mid H(G) = H, S(G) = S\right) \\
&\le e^{-|S|/16}.
\end{aligned} \tag{2}$$

Next, we turn our attention to bounding $|S(G)|$. Conditional on the tower distribution $H(G) = H = (H_1, \dots, H_s)$, we have for each vertex $v \in \widetilde{H_2}(G)$ by Lemma 6, part (b):

$$\begin{aligned}
&\Pr\left(v \in S \mid H(G) = H\right) = \Pr\left(|\mathrm{Pa}(v)| = 1 \mid H(G) = H\right) \\
&= \frac{h_1 \cdot p(1-p)^{h_1 - 1}}{\sum_{i=1}^{s} \binom{h_1}{s} p^s (1-p)^{h_1 - s}} = \frac{h_1 \cdot p(1-p)^{h_1-1}}{1 - (1-p)^{h_1}} \\
&\ge \frac{h_1 \cdot p(1-p)^{h_1-1}}{h_1 \cdot p} \ge (1-p)^{h_1}.
\end{aligned}$$

Here, we used Bernoulli's inequality in the third line of the derivation. This implies $\mathbb{E}[|S(G)| \mid H(G) = H] \ge (1-p)^{h_1} \cdot |\widetilde{H_2}|$. Since the events $v \in S(G)$ are again independent for different $v \in \widetilde{H_2}$ conditional on the tower distribution, we get by the Chernoff bound (Lemma 23):

$$\begin{aligned}
&\Pr\left(|S(G)| \le \frac{(1-p)^{h_1}}{2} \cdot |\widetilde{H_2}| \,\middle|\, H(G) = H\right) \\
&\le e^{-(1-p)^{h_1} |\widetilde{H_2}|/8}.
\end{aligned} \tag{3}$$

Now, let $\mathcal{H}$ be the set of tower decompositions $H = (H_1, \dots, H_s)$ satisfying $|H_1| \le 5p^{-1}$ and $|H_2| \ge p^{-1}/\log^2(p^{-1})$. For $H \in \mathcal{H}$, we get

$$(1-p)^{h_1} \cdot |\widetilde{H_2}| \ge (1-p)^{5/p} \cdot \frac{p^{-1}}{\log^2(p^{-1})} \ge \frac{e^{-5}}{2} \cdot \frac{p^{-1}}{\log^2(p^{-1})},$$

where the last steps holds when p is small enough. Set $\alpha(p) := p^{-1}/(4e^5 \log^2(p^{-1}))$. Then, for $H \in \mathcal{H}$, equation (3) implies

$$\Pr\left(|S(G)| \le \alpha(p) \mid H(G) = H\right) \le e^{-\alpha(p)/2}.$$

Summing over all $H \in \mathcal{H}$ gives

$$\Pr\left(|S(G)| \le \alpha(p) \wedge H(G) \in \mathcal{H}\right) \le e^{-\alpha(p)/2}.$$

And by Lemma 12 and 11, we get

$$\begin{aligned}\Pr\left(|S(G)| \le \alpha(p)\right) &\le e^{-\alpha(p)/2} + \Pr\left(H(G) \notin \mathcal{H}\right)\\ &= o(1).\end{aligned}$$

Now, summing equation (2) over all graphs with $S(G) \ge \alpha(p)$ gives

$$\Pr\left(|X(G)| \le \alpha(p)/4 \wedge S(G) \ge \alpha(p)\right) \le e^{-\alpha(p)/16}.$$

So, we finally obtain

$$\begin{aligned}\Pr\left(|X(G)| \le \alpha(p)/4\right) &\le e^{-\alpha(p)/16} + \Pr\left(S(G) \le \alpha(p)\right)\\ &= o(1).\end{aligned}$$

Hence, with high probability, G has a matching of reversible layer-2-edges of size at least $\alpha(p)/4$. □

B PROOFS FOR DIRECTED CYCLIC GRAPHS

The following statement is a simple observation of the properties of the construction given in Definition 19.

Lemma 24. *Let G be a directed cyclic graph containing a cycle C. Let $H = \textsc{Reverse}(G, C)$ and let $\overline{C}$ be the reversed version of C that occurs in H. We have the following properties:*

1. *each vertex has the same descendants in G as in H;*
2. *$G = \textsc{Reverse}(H, \overline{C})$.*

Now, we prove the key property of the construction in Definition 19.

Proposition 20. *Let G be a directed cyclic graph containing a cycle C. Then, the graph $H = \textsc{Reverse}(G, C)$ is Markov equivalent to G.*

Proof of Proposition 20. Suppose $C = (v_1, \dots, v_k)$ in G and $H = \textsc{Reverse}(G, C)$. Fix an arbitrary conditioning set $Z \subseteq V(G)$. Let $\pi = (w_1, \dots, w_t)$ be a simple active path between w_1 and w_t given Z in G. The first part of the proof is to show that there exists an active path from w_1 to w_t given Z in H.

Case 1: Suppose some vertex of C is in Z, i.e. $C \cap Z \neq \emptyset$. Consider the path π' in H obtained as follows:

- whenever $w \to v_i$ occurs in π for $v_i \in C$, replace it with $w \to v_{i-1} \leftarrow v_i$ (similarly, if $v_i \leftarrow w$ occurs in π, replace it with $v_i \to v_{i-1} \leftarrow w$).
- make the resulting path simple by deleting any self-loops if they exist.

By definition of $H = \textsc{Reverse}(G, C)$, π' is a valid simple path from w_1 to w_t in H. We show that π' is also an active path given Z.

First, consider any segment $a - b - c$ in π' with $b \notin C$. If b is a collider in the segment, i.e., we have $a \to b \leftarrow c$, then $a \to b \leftarrow c$ must occur in π as well, since incoming edges to $b \notin C$ are never changed between π and π'. This means b has a G-descendant in Z, and by part 1 of Lemma 24, b must also have an H-descendant in Z, so $a \to b \leftarrow c$ is active in π'. If b

is a non-collider in $a - b - c$, then it must also occur as a non-collider in π (but perhaps in a different segment). This is because outgoing edges from b in π' that do not exist in π must have been a replacement for another outgoing edge from b in π. Since π is active, this means $b \notin Z$, and hence the segment $a - b - c$ is active in π'.

Since C contains a vertex of Z, any collider in C is automatically active. To show that the entire path π' is active, it now suffices to show that any vertex $v_i \in C$ that is a non-collider in π' must have been a non-collider in π as well. Indeed, if v_i is a non-collider in π' with an outgoing edge $v_i \to w$ with $w \neq v_{i-1}$, then this outgoing edge must exist in π or must be a replacement for another outgoing edge in π from v_i, so v_i is also a non-collider in π. If v_i is a non-collider in a circle segment $v_{i+1} \to v_i \to v_{i-1}$ in π', then v_i must occur as a non-collider in the segment $v_{i+1} \leftarrow v_i \leftarrow v_{i-1}$ in π. Finally, we could have $w \to v_i \to v_{i-1}$ in π' for $w \neq v_{i+1}$, but this is only possible if $w \to v_{i+1} \leftarrow v_i \leftarrow v_{i-1}$ occurred in π. In each case, v_i is also a non-collider in π, which concludes the argument.

Case 2: Suppose none of the vertices in C is part of the conditioning set Z, i.e. $C \cap Z = \emptyset$. In this case, we have to construct the path π' in H in a different way:

- whenever $w \to v_i$ occurs in π for $v_i \in C$, replace it with $w \to v_{i-1} \to v_{i-2} \to \cdots \to v_i$ (similarly, if $v_i \leftarrow w$ occurs in π, replace it with $v_i \leftarrow v_{i+1} \leftarrow \cdots \leftarrow v_{i-1} \leftarrow w$).
- make the resulting path simple by deleting any self-loops if they exist.

By definition of $H = \textsc{Reverse}(G, C)$, π' is a valid simple path from w_1 to w_t in H. We show that π' is also an active path given Z.

By the same argument as in the first case, any segment $a - b - c$ of π' with $b \notin C$ must be active. Since C does not contain any vertex in Z, any non-collider of π' in C is active too. Hence, it suffices to show that every collider of π' in C is active. If π' contains a segment $w \to v_i \leftarrow u$ with $w, u \neq v_{i+1}$, then π must contain the segment $w \to v_{i+1} \leftarrow u$. Since π is active, v_{i+1} must have a G-descendant in Z, and by part 1 of Lemma 24, this is also an H-descendant. Hence, also v_i has an H-descendant in Z, so $w \to v_i \leftarrow u$ is active in π'. Finally, note that π' cannot contain a segment of the form $w \to v_i \leftarrow v_{i+1}$, by construction. This completes the first part of the proof.

We have shown that, whenever a and b are d-separated in $H = \textsc{Reverse}(G, C)$ given Z, they also must be d-separated in G given Z. However, since $G = \textsc{Reverse}(H, \overline{C})$ by part 2 of Lemma 24, the converse must also hold. Hence, G and H are Markov equivalent. $\square$