Size of Interventional Markov Equivalence Classes in Random DAG Models

Dmitriy Katz [1] **Karthikeyan Shanmugam**[1] **Chandler Squires**[2] **Caroline Uhler**[2]
[1]IBM Research, NY & MIT-IBM Watson AI Lab [2]MIT, Cambridge, MA

Abstract

Directed acyclic graph (DAG) models are popular for capturing causal relationships. From observational and interventional data, a DAG model can only be determined up to its *interventional Markov equivalence class* (I-MEC). We investigate the size of MECs for random DAG models generated by uniformly sampling and ordering an Erdős-Rényi graph. For constant density, we show that the expected log observational MEC size asymptotically (in the number of vertices) approaches a constant. We characterize I-MEC size in a similar fashion in the above settings with high precision. We show that the asymptotic expected number of interventions required to fully identify a DAG is a constant. These results are obtained by exploiting Meek rules and coupling arguments to provide sharp upper and lower bounds on the asymptotic quantities, which are then calculated numerically up to high precision. Our results have important consequences for experimental design of interventions and the development of algorithms for causal inference.

1 Introduction

Directed acyclic graphs (DAGs) are popular models for capturing causal relationships among a set of variables. This approach has found important applications in various areas including biology, epidemiology and sociology (Gangl, 2010; Lagani et al., 2016). A central problem in these applications is to learn the causal DAG from observations on the nodes. A popular approach is to infer missing edges based on conditional independence information that is learned from the data (Spirtes et al., 2000; Kalisch and Bühlmann, 2007). However, multiple DAGs can encode the same set of conditional independences. Hence in general the causal DAG can only be learned up to a *Markov equivalence class* (MEC) and interventional data is needed in order to identify the causal DAG.

While an MEC may contain a super exponential number of candidate DAGs, Gillispie and Perlman (2001) showed by enumerating all MECs up to 10 nodes that for small graphs (up to 10 nodes) an MEC on average contains about four DAGs and that about a quarter of all MECs consist of a unique DAG. Generalizing these results to larger graphs is critical for estimating the average number of interventional experiments needed for identifying the underlying causal DAG. More generally, given the recent rise in interventional data in genomics enabled by genome editing technologies (Xiao et al., 2015), it is of great interest to understand the average reduction in the size of MECs through the availability of interventional data, i.e., to characterize the average size of an *interventional Markov equivalence class* (I-MEC). Further, such an analysis would also shed light on the number of additional interventions needed to uniquely identify the underlying causal DAG moving away from worst case bounds.

The problem of characterizing the size of an MEC or I-MEC is not only of interest for experimental design of interventions but also from an algorithmic perspective. A popular approach to causal inference is given by score-based methods that assign a score such as the Bayesian Information Criterion (BIC) to each DAG or MEC and greedily optimize over the space of DAGs (Castelo and Kocka, 2003), a combination of permutations and undirected graphs (Teyssier and Koller, 2012; Raskutti and Uhler, 2018; Solus et al., 2017; Mohammadi et al., 2018) or MECs (Meek, 1997; Brenner and Sontag, 2013). Similar score-based approaches have also been developed in the interventional setting (Hauser and Bühlmann, 2012a; Wang et al., 2017; Yang et al., 2018). While a greedy step in the space of graphs can easily be defined (addition, removal or flipping of an edge), a greedy step in the space

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

of Markov equivalence classes is complicated (Meek, 1997). Hence performing a greedy algorithm in the space of MECs only makes sense if the space of MECs is significantly smaller as compared to the space of DAGs. For instance showing that typically occurring MECs or I-MECs are small would imply that graph-based search procedures operate on a similar search space as the ones that use MECs, but can do so using simpler moves.

Motivated by these considerations, in this work, we initiate the study of interventional and observational MECs for random DAG models. We focus on *random order DAGs*, where the skeleton is a random Erdős Rényi graph with constant density ρ and the ordering is a random permutation. We derive tight bounds for the asymptotic versions of various metrics on the I-MECs. More specifically, our contributions are as follows:

1. We derive tight upper and lower bounds on (a) the asymptotic expected number of unoriented edges in an I-MEC given data from $r = 0, 1, 2 \dots$ interventions; (b) the asymptotic probability that the I-MEC is a unique DAG given data from r interventions; (c) the asymptotic number of additional interventions needed to fully discover the DAG given data from r interventions; and (d) the asymptotic expected log-size of the I-MEC given data from r interventions.
2. We also provide tight bounds for the number of unoriented edges in the I-MEC when r interventions have been performed using different algorithms for choosing the interventions given the observational MEC as input.
3. If $M(r)_n$ is the metric of interest of a random order DAG of size n and $r \geq 0$ interventions, then our bounds are of the following form: $\mathbb{E}[M(r)_n] \leq \mathbb{E}[M(r)_\infty] \leq \mathbb{E}[M(r)_n] + \epsilon_n$. Here, $M(r)_\infty$ is the limiting asymptotic metric, which we show is well defined and exists. We also show that ϵ_n decays exponentially fast in n for constant density ρ.
4. We numerically compute $\mathbb{E}[M(r)_n]$ through Monte Carlo simulations for n as large as 110 at which point ϵ_n is a small constant for various parameter regimes.
5. One of the surprising results is that for constant density random order DAGs, all the above metrics tend asymptotically to a constant. Through a combination of analysis of our bounds and numerical computation, we can characterize these constants precisely.
6. As an example of the nature of our results, quite surprisingly, the asymptotic (as $n \to \infty$) expected log-observational MEC size of a random order DAG with density 0.5 is at most 3.497 with probability at least 0.99 (see Theorem 14).

All omitted proofs can be found in the supplemental material.

Related Work: There is currently only limited work available on counting and characterizing MECs. In (Gillispie and Perlman, 2001), the authors enumerated all MECs on DAGs with $p \leq 10$ nodes and analyzed the total number of MECs, the average size of an MEC, and the proportion of MECs of size one on p nodes. Motivated by this work, Gillispie (2006), Steinsky (2003), and Wagner (2013) provided formulas for counting MECs of a specific size. Supplementing this line of work, He and Yu (2016) developed various methods for counting the size of a given MEC. Finally, Radhakrishnan et al. (2017) addressed these enumerative questions using a pair of generating functions that encode the number and size of MECs for DAGs with a fixed skeleton (i.e. underlying undirected graph) and also applied these results to derive bounds on the MECs for various families of DAGs on trees (Radhakrishnan et al., 2018).

Another line of work (Hu et al., 2014; Hauser and Bühlmann, 2012b; Shanmugam et al., 2015; Eberhardt et al., 2012; Hyttinen et al., 2013; Kocaoglu et al., 2017) aims at characterizing the number of interventions required to learn a causal DAG completely. While some of these works deal with the active learning setting (Shanmugam et al., 2015; Hauser and Bühlmann, 2012b), others choose interventions non-adaptively given the observational MEC (Hu et al., 2014; Eberhardt et al., 2012; Hyttinen et al., 2013; Kocaoglu et al., 2017; Bello and Honorio, 2017) and hence are concerned with the worst-case scenario.

2 Preliminaries and Definitions

In this work, we characterize the asymptotic behavior of different metrics that capture the amount of "causal relationships" which can be inferred from observational and interventional data on random DAG models. In this section, we describe the random orderDAG model, briefly review causal DAG models and Markov equivalence, and introduce the metrics that we will analyze in this work.

2.1 Random Order DAG Model

Let $G = (V, E)$ be a directed acyclic graph (DAG) with vertices $V = [n]$ and directed edges $E \subseteq V \times V$. A random **orderDAG** with density ρ on n vertices is a DAG G_n whose *skeleton* (i.e., underlying undirected graph) is given by an Erdös-Rényi graph on n

vertices with edge probability ρ and whose edges are oriented according to a total ordering which is uniformly sampled among all permutations of n vertices. We denote a graph G_n sampled from this model by $G_n \sim orderDAG(n, \rho)$.

Remark: Our sampling procedure is a standard one used for testing causal inference algorithms. It is for example used in the well known `pcalg` R package[1]. A different sampling scheme would be to sample DAGs uniformly at random from all DAGs in which isomorphic DAGs would not be double counted. However, such a sampling scheme is difficult to perform in practice, while ours has a generative model that is easy and intuitive. Limited prior computational evidence in the observational setting suggests that the two sampling schemes behave similarly (Gillispie and Perlman, 2001).

2.2 Markov Equivalence

A joint distribution P on the variables $(X_v)_{v\in V}$ associated to the vertices of a DAG G is *Markov* with respect to G if for any node $v \in G$, X_v is conditionally independent of its non-descendents given its parents. In this case we say that $P \in \mathcal{M}(G)$. Two directed acyclic graphs G and G' are in the same *Markov equivalence class* (MEC) if and only if $\mathcal{M}(G) = \mathcal{M}(G')$. Two DAGs in the same MEC entail the same set of conditional independence relations. (Meek, 1995).

The MEC of a DAG G can be uniquely represented by a partially directed graph $\mathrm{Ess}(G)$ known as the *essential graph* of G. The skeleton of $\mathrm{Ess}(G)$ is the same as the skeleton of G and the directed edges in $\mathrm{Ess}(G)$ are precisely those edges in G that have the same orientation in all members of the MEC of G. All other edges in $\mathrm{Ess}(G)$ are unoriented (Hauser and Bühlmann, 2012a). The following procedure provides all directed edges in $\mathrm{Ess}(G)$:

1. For every triple of nodes $i, j, k \in V$ if i and j are disconnected in G and the ordered pairs $(i, k), (j, k) \in E$, then both edges (i, k) and (j, k) are also oriented in $\mathrm{Ess}(G)$.
2. Orient edges by successive application of the 'Meek rules' (see (Meek, 1995) or Appendix A) until they cannot be applied anymore to orient any new edge.

2.3 Interventional Markov Equivalence

Let $I \subset V$ and consider the set of single node interventional distributions $(P_i)_{i\in I}$, where node i is set to some constant. Since in P_i, node X_i (a constant) is independent of its parents $X_{\mathrm{Pa}(i)}$, it introduces additional conditional independences in addition to those present in P. Let $G^{(i)}$ denote the intervened DAG obtained by deleting the edges from $\mathrm{Pa}(i)$ to i. If P is Markov with respect to G, then P_i is Markov with respect to $G^{(i)}$. Two DAGs G and G' are in the same *I-Markov equivalence class* (I-MEC) if and only if $G^{(i)}$ and $G'^{(i)}$ are in the same MEC for all $i \in I$ (Hauser and Bühlmann, 2012a).

Similarly as in the purely observational setting, an I-MEC can be uniquely represented by an *I-essential graph* denoted by $\mathrm{Ess}(G, I)$. The skeleton of $\mathrm{Ess}(G, I)$ is the same as the skeleton of G and the directed edges in $\mathrm{Ess}(G, I)$ are precisely those edges in G that have the same orientation in all members of the I-MEC of G. All other edges in $\mathrm{Ess}(G, I)$ are unoriented. The following procedure provides all directed edges in $\mathrm{Ess}(G, I)$:

1. For every triple of nodes $i, j, k \in V$ with $(i, k), (j, k) \in E$ and if i and j are disconnected in G, then both edges (i, k) and (j, k) are also oriented in $\mathrm{Ess}(G)$.
2. For every edge (i, j) such that either $j \in I$ or $i \in I$, then (i, j) is oriented.
3. Orient further edges by successive application of the four rules in (Hauser and Bühlmann, 2012a) (also given in Appendix A) until it cannot be applied anymore to orient any new edges.

2.4 Metrics of Interest

Suppose that the causal Bayesian network that generates data (both interventional and observational) is an orderDAG G_n. Let $\mathbf{P}_*$ be an associated family of interventional distributions compatible with G_n. In this setting, our work asymptotically characterizes some metrics that reflect identifiable portions of G_n from an observational distribution P nd possibly also interventional distributions.

We denote by uEss an essential graph that is also a DAG, i.e., an essential graph representing an MEC consisting of a unique DAG. Such DAGs are of particular interest since they are identifiable from purely observational data.

In the following, we will measure the degree of identifiability of a random DAG $G_n \sim orderDAG(n, \rho)$ using the following metrics:

1. Let X_n be the number of unoriented edges in $\mathrm{Ess}(G_n)$. We show that $X_\infty := \lim_{n\to\infty} X_n$ exists.
2. Let $isuEss_n$ be an indicator variable that is 1 only if $\mathrm{Ess}(G_n)$ is a DAG. Similarly, the limit is denoted $isuEss_\infty$

[1] https://rdrr.io/rforge/pcalg/man/randomDAG.html

3. Let I_n be the number of single node interventions required to fully orient G_n. Similarly, the limit is denoted I_∞.
4. Let L_n be the size of the (observational) MEC of G_n. The limit is denoted L_∞.
5. Let $X_n(r)$ be the minimum number of unoriented edges in $\text{Ess}(G_n, I)$ optimized over all $I : |I| = r$. The limit is denoted $X_\infty(r)$.
6. Let $isuEss_n(r)$ be an indicator variable that is 1 when $X_n(r) = 0$. The limit is denoted $isuEss_\infty(r)$.
7. Let $L_n(r)$ be the size of the interventional markov equivalence class when the interventions in the set I are performed on G_n, where I minimizes the number of unoriented edges in $\text{Ess}(G_n, I)$ optimized over all $I : |I| = r$. This limit is denoted $L_\infty(r)$.

3 Main Results

We first describe the nature of our results and the approach taken for obtaining these results for X_n. The results for all other metrics follow using a similar approach, although the technical details differ depending on the metric of interest. We show that $\mathbb{E}(X_n) \leq \mathbb{E}(X_\infty) \leq \mathbb{E}(X_n) + \epsilon_n$ and we provide an explicit expression for ϵ_n. As a consequence, tight upper and lower bounds can be constructed on the quantities of interest by numerically computing $\mathbb{E}[X_n]$ using Monte Carlo simulations by generating random order DAGs G_n for large n and averaging.

Formally, we state the main result in our work about the asymptotic quantities of various metrics.

Theorem 1. *We have the following inequalities satisfied by various metrics:*

$$\begin{aligned} E[X_n(r)] \leq E[X_\infty(r)] &\leq E[X_n(r)] + \epsilon_n \\ E[I_n] \leq E[I_\infty] &\leq E[I_n] + \epsilon_n \\ E[\log_2(L_n(r))] \leq E[\log_2(L_\infty(r))] &\leq E[X_\infty(r)] \\ E[isuEss_n(r)] \geq E[isuEss_\infty(r)] &\geq E[isuEss_n(r)] - \epsilon_n \end{aligned}$$

for all $r = 0, 1, 2 \ldots$. Here, ϵ_n is defined as follows:

$$\epsilon_n = \sum_{i \geq n} \text{RHS}(\rho, i) \leq \frac{(1-\rho(1-\rho))^n}{\rho(1-\rho)^2} + n\frac{(1-\rho(1-\rho))^{n-1}}{(1-\rho)}, \qquad (1)$$

where $\text{RHS}(\rho, n) = \rho n * (1-\rho(1-\rho))^{n-1}$ *and ρ is the edge probability when sampling an order DAG.*

We establish the main result on upper and lower bounds through intermediate results as follows (explained taking the example of X_n): a) We first exhibit a coupling between G_n and G_{n+1} such that their respective marginal distributions are preserved. This is done in Section 3.1. b) Using the properties of this specific coupling, we first show that $\mathbb{E}[X_n]$ is a monotonic sequence in n in Section 3.1.1. c) The expression for ϵ_n is obtained by upper bounding the successive differences $\mathbb{E}[X_n] - \mathbb{E}[X_{n+1}]$ again using the properties of order DAG sampling and the coupling. This is explained in Sections 3.1.2 and 3.1.3. Other sections provide additional results on I-MECs obtained through other interventional design algorithms along with numerical and simulation results.

3.1 Probability coupling

In this section, we provide a coupling argument between the distribution of G_n and G_{n+1} such that 'unorientability' properties of certain edges are preserved.

For all $1 \leq i < j \leq n$, let $A_{i,j}$ be a binary random variable that is 1 with probability ρ. Let G_n be the DAG with nodes $v_1 \ldots v_n$ and directed edges between $v_i \to v_j$ if and only if $A_{i,j} = 1$.

Observation 1. *G_n with permutation $v_1, v_2, \ldots, v_n$, has the distribution of a random orderDAG on n vertices with density ρ.*

Remark: Observation 1 says that randomly sampling a symmetric adjacency matrix (undirected graph with edge probability ρ), permuting rows and columns with a random permutation, and then taking the upper triangular part (orienting the graph according to the permutation) is the same as fixing the permutation from 1,2..n and populating the upper triangular part randomly.

Coupling: Motivated by the above observation, we couple G_n and G_{n+1} as follows. We first generate $A_{i,j}$ for $1 \leq i < j \leq n$ as above and use that to orient G_n. Then, we generate additional random variables $A_{i,n+1}$ for all $1 \leq i \leq n$ and orient the edges incident to v_{n+1} accordingly.

The above coupling along with certain structural properties of Meek Rules (given in Appendix A) leads to the following results on orientability of certain edges in G_n and G_{n+1} under the coupling.

Lemma 1. *Under the above coupling, if an edge (i, j) is unorientable in G_n, it is also unorientable in G_{n+1}.*

Lemma 2. *Under the above coupling, if after a set of interventions R on G_n the edge (i, j) is unorientable in G_n, then it is also unorientable in G_{n+1} after the same set of interventions on G_n together with an intervention on v_{n+1}.*

3.1.1 Monotonicity Lemmas

We prove that expected values of all metrics of interest are monotonic in n using the properties of the coupling demonstrated above. First, we show this for observational quantities by appealing to Lemma 1.

Theorem 2. *The following statements hold with probability* 1 *for the coupling between* G_n *and* G_{n+1}*:*
a) $X_{n+1} \geq X_n$.
b) $L_{n+1} \geq L_n$.
c) $I_{n+1} \geq I_n$.
Therefore, $\mathbb{E}(X_{n+1}) \geq \mathbb{E}(X_n)$, $\mathbb{E}(L_{n+1}) \geq \mathbb{E}(L_n)$ *and* $\mathbb{E}(I_{n+1}) \geq \mathbb{E}(I_n)$.

Similar monotonicity properties for interventional quantities are obtained by appealing to Lemma 2. However, note that these proofs are not a straightforward application of Lemma 2. Often, additional arguments need to be made to show the following results.

Theorem 3. $X_{n+1}(r) \geq X_n(r)$ *with probability* 1 *according to the coupling between* G_n *and* G_{n+1}*. Hence,* $\mathbb{E}(X_{n+1}(r)) \geq \mathbb{E}(X_n(r))$.

The previous two theorems directly provide the following result.

Theorem 4. $isuEss_{n+1}(r) \leq isuEss_n(r)$ *for all* $r = 0, 1, 2 \ldots$ *best interventions with probability* 1 *under the coupling between* G_n *and* G_{n+1}*. Hence,* $\mathbb{E}(isuEss_{n+1}(r)) \leq \mathbb{E}(isuEss_n(r))$.

Proof. This follows directly from Theorem 3 and Theorem 2. □

Theorem 5. $L_{n+1}(r) \geq L_n(r)$ *with probability* 1 *under the coupling between* G_n *and* G_{n+1}*. Hence,* $\mathbb{E}(L_{n+1}(r)) \geq \mathbb{E}(L_n(r))$.

The established monotonicity results help prove that the asymptotic versions of these metrics exist.

Theorem 6. $\lim_{n\to\infty} X_n = X_\infty$ *exists and* $\mathbb{E}[X_\infty] = \lim_{n\to\infty} \mathbb{E}[X_n]$.

Remark: Theorem 6 extends to all metrics that have been shown to be monotonic non-decreasing, i.e. metrics in the set $\{X_n(r), I_n, L_n(r)\}$, by analogous arguments. Note that monotonically non-increasing sequences like $isuEss_n(r)$ are bounded below and above and hence the results can be shown again by the same theorem applied to shifted negatives of these variables.

3.1.2 Gap Bounds on Observational Metrics

Using properties of the coupling between G_n and G_{n+1} we can show that the expected difference in the observational metrics for G_n and the asymptotic version is bounded.

Theorem 7. $\mathbb{E}(X_\infty) - \mathbb{E}(X_n) \leq \sum_{i=n}^{\infty} \rho i * (1 - \rho(1-\rho))^{i-1}$.

Theorem 8. $\mathbb{E}(I_\infty) - \mathbb{E}(I_n) \leq \sum_{i=n}^{\infty} \rho i * (1 - \rho(1-\rho))^{i-1}$.

3.1.3 Gap Bounds on Interventional Metrics

In the following, we show that the expected difference in the interventional metrics for G_n and the asymptotic version is bounded again using the properties of the coupling described before.

Theorem 9. $\mathbb{E}(X_\infty(r)) - \mathbb{E}(X_n)(r) \leq \sum_{i=n}^{\infty} \rho i * (1 - \rho(1-\rho))^{i-1}$.

Theorem 10. $\mathbb{E}(isuEss_n(r)) - \mathbb{E}(isuEss_\infty(r)) \leq \sum_{i=n}^{\infty} \rho i * (1 - \rho(1-\rho))^{i-1}$.

All these results together allow us to prove the main result (Theorem 1).

Proof of Theorem 1. The theorem follows from results in Sections 3.1.1, 3.1.2, and 3.1.3. We use the fact that $\log_2(L_n(r)) \leq X_n(r)$, since $L_n(r) \leq 2^{X_n(r)}$ by considering all possible orientations of the unoriented edges in the I-essential graph. □

3.1.4 Lower Bound on Successive Differences

The above gap bounds depend on upper bounding successive differences of $\mathbb{E}[X_n]$. In the following, we provide a lower bound on the successive differences which implies that gap bounds that are faster than exponential cannot exist.

Theorem 11.

$$\mathbb{E}(X_n) - \mathbb{E}(X_{n-1}) \geq (n-1)\rho(1-\rho)^{2n-4} \geq \rho(1-\rho)^{2n}.$$

4 Results on I-MECs obtained by Interventional Design Algorithms

In the following, we provide asymptotic convergence rates for the number of undirected edges after r interventions, when the interventions are chosen by an algorithm that has a property that we call *downstream-independence*. Greedy algorithms that choose r interventions sequentially based on the essential graph at the observational stage are downstream-independent. Note that, in this section, we do not consider $X_n(r)$, which is the minimum number of edges left unoriented when r interventions are chosen based on the DAG structure. We are therefore interested in algorithms that optimize the interventions based on the essential graph, which can be inferred from purely observed datasets.

Notation 1. *Let J be a set of interventions. We say that $H = J(G)$ when H is the essential graph that results from performing the interventions J on the underlying causal DAG G. Note that if G' is a subgraph of G, then $J(G')$ is obtained by skipping the interventions on nodes outside of G'.*

Lemma 3. *Let G be a DAG and v_n a vertex of* ess(G) *with no outgoing or undirected edges. Then, $J(G\backslash v_n) = J(G)\backslash v_n$. In other words, interventions do not affect vertices that have no outgoing or undirected edges.*

Lemma 4. *Let G' be an induced subgraph of G consisting of all vertices v_i such that neither v_i nor any descendants of v_i have adjacent undirected edges. Then $J(G\backslash G') = J(G)\backslash G'$.*

Proof. The proof follows by applying Lemma 3 recursively to G. □

Definition 1. *We say that an algorithm A for performing interventions on an essential graph is* ***downstream-independent*** *if the inverventions it performs on G are identical to the ones it performs on $G\backslash G'$.*

Note that $G\backslash G'$ is the result of the following process: starting with G, recursively remove vertices that have no undirected or outgoing edges.

Theorem 12. *Let A be a downstream-independent algorithm. Let $Y(r, A)_i$ be the expected number of undirected edges in the essential graph of the random order DAG G_i after performing r interventions according to algorithm A. Then*

$$|\mathbb{E}(Y(r,A)_{i+1} - \mathbb{E}(Y(r,A)_i)| \leq \rho i * (1-\rho(1-\rho))^{i-1} * i(i+1)/2 \quad (2)$$

Remark: Suppose there is an algorithm A that optimizes some score function based on the essential graphs alone which is a proxy for minimizing the number of expected unoriented edges after r interventions, then such algorithms are likely to be making decisions independent of G' in general due to Lemma 4. An example is the algorithm that greedily picks the intervention that reduces the expected number of unoriented edges where the expectation is over the uniform distribution of DAGs compatible with the essential graph.

Theorem 13. *Let A be an algorithm that is downstream independent and chooses interventions based on* ess(G). *Let $Y(r, A)_n$ be the number of undirected edges after r interventions made by the algorithm A. Then,*

$$\begin{aligned}\mathbb{E}[Y(r,A)_n] &\leq \mathbb{E}[Y(r,A)_\infty] \\ &\leq \mathbb{E}[Y(r,A)_n] + \\ &\quad \sum_{i=n}^{\infty} \rho i^2 (i+1)/2 * (1-\rho(1-\rho))^{i-1}.\end{aligned}$$

Here, $\lim_{n\to\infty} Y(r,A)_n = Y(r,A)_\infty$ and this limit exists.

Proof. This is a direct corollary from the previous results in this section together with analogous arguments regarding monotonicity and existence of limits similar to those for $X_n(r)$. □

5 Discussion of the Results

Theorems 1 and 13 provide upper bounds in terms of quantities computable by Monte-Carlo simulation at finite n from random order DAGs and constants such as ϵ_n that are exponentially small in n. If empirical means of these finite n quantities appearing in these upper bounds can be characterized with very high precision, then we can characterize the constant by which these asymptotic quantities are upper bounded.

In the following section, we plot the empirical means of these finite n quantities or upper bounds to these finite n quantities for very large n and show that when combined with the above bounds, the asymptotic quantities tend to a constant.

5.1 Precise Calculation of High Confidence Upper Bounds on Asymptotic log-MEC Size for Random Order DAGs of Density $\rho = 0.5$

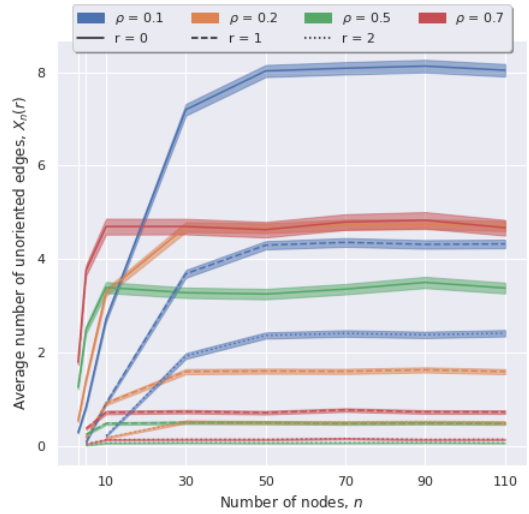
We demonstrate how to obtain confidence intervals on the expected asymptotic mean $E[X_\infty]$ and $E[\log_2(L_\infty)]$ using our bounds and Monte Carlo simulations.

Details of Numerical Experiment: We sampled X_{30} $S = 100000$ times for random order DAGs with $\rho = 0.5$. The sample variance we observed was $V = 7.054$ while the empirical mean was $M = 3.394$.

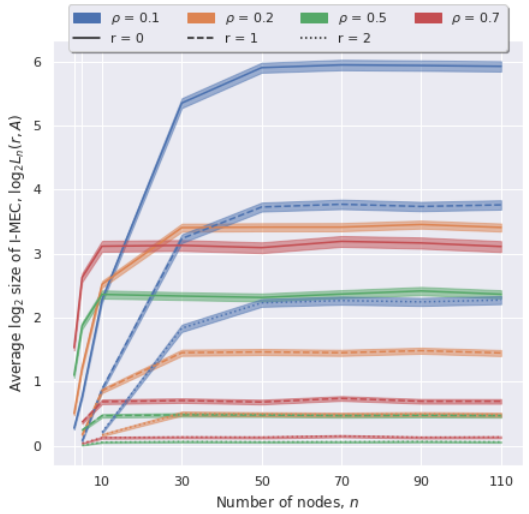
We use an empirical Bernstein bound for $E[X_{30}]$ and show the following bound on expected value of X_∞:

Theorem 14. *With probability at least* 0.99 *over the randomness in our numerical experiments over $S =$* 100000 *samples, we have: $E[\log_2(L_\infty)] \leq E[X_\infty] \leq$* 3.497.

This is an illustration of how our upper bounds, emprirical Bernstein bounds and Monte Carlo simulation can be combined to give highly precise guarantees for all the considered metrics.

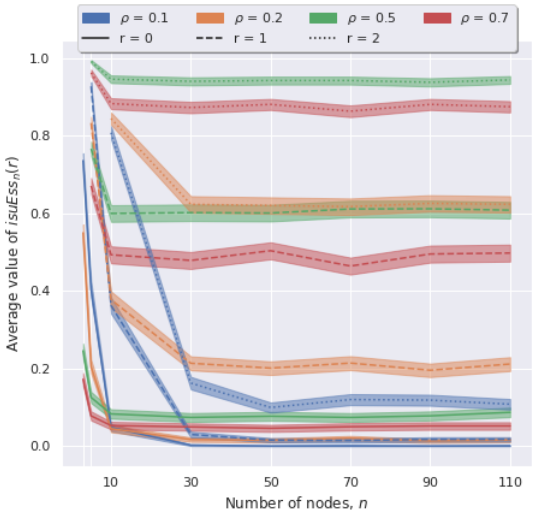


(a) Average number of unoriented edges, $X_n(r)$, in the essential graph associated with orderDAGs of density ρ after r interventions, averaged over 2000 samples; the highlighted region corresponds to points within 2-standard deviations from the mean.

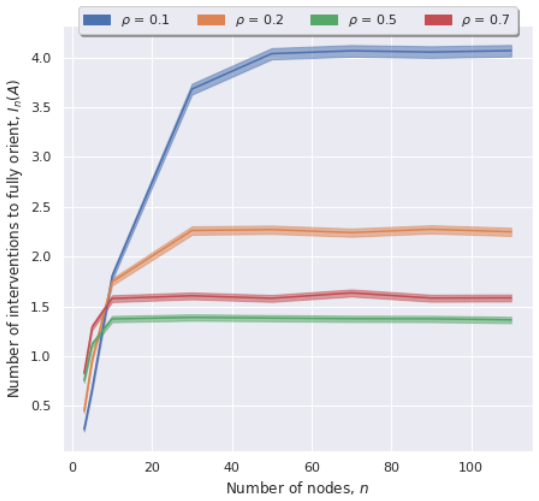


(b) Average logarithm of the size of the I-MEC for order DAGs of density ρ after r interventions, averaged over 2000 samples; the highlighted region corresponds to points within 2-standard deviations from the mean.

Figure 1: We plot Monte-Carlo estimates of $\mathbb{E}[Y_n(r, A)]$, i.e the number of unoriented edges in the essential graph of a random order DAG after r interventions, together with $\mathbb{E}[\log_2 L_n(r, A)]$, i.e. the size of the I-MEC after r interventions.



(a) Probability that the essential graph associated with an order DAG of density ρ can be uniquely identified after r interventions, averaged over 2000 samples; the highlighted region corresponds to points within 2-standard deviations from the mean.



(b) Empirical mean of the number of interventions needed to fully identify a random order DAG of density ρ, averaged over 2000 samples; the highlighted region corresponds to points within 2-standard deviations from the mean.

Figure 2: We plot Monte-Carlo estimates of $\mathbb{P}(isuEss_n(r, A))$, i.e. the probability that the essential graph of a random order DAG is equal to the order DAG itself, together with $\mathbb{E}[I_n(r, A)]$, i.e. the number of single-node interventions required to fully orient a random order DAG.

6 Numerical Results

We compute and plot the empirical means of the following observational metrics: a) X_n, b) $isuEss_n$, c) I_n, and d) $\log_2 L_n$. We also plot the empirical mean of the following interventional metrics a) $Y(r, A)_n$, b) $isuEss(r, A)_n$, c) $\log_2 L(r, A)_n$, and d) $I(r, A)_n$. These interventional metrics are obtained on the essential graph $\text{Ess}(G_n, A)$ obtained by the greedy algorithm A that operates as follows: First pick the node I_1 that orients the most edges, then for each consecutive r, pick I_r that orients the most edges in G_n given the $(\{I_1, \ldots, I_{r-1}\}$-)essential graph.

Graph Generation: We generated 2,000 random order DAGs with $n = \{3, 5, 10, 30, \ldots 110\}$ nodes and densities $\rho = \{.1, .2, .5, .7\}$. For each DAG, we used the open-source `causaldag` package in Python to compute the number of DAGs in the ($\mathcal{I}$-)MEC and the number of undirected edges in the ($\mathcal{I}$-)essential graph obtained by applying algorithm A on G_n.

Results Established: The plots serve two purposes - a) The empirical mean plots (Figs. 1a-2b) and the box plots (Figs. 4a-5c) of all the estimated quantities provide an idea of what values the asymptotic quantities are bounded by given the formula for ϵ_n in Theorem 1. For a more refined high confidence upper bound, for large enough n, analysis similar to Theorem 14 can be done. b) They help corroborate the monotonicity results we have derived analytically.

Bounding Interventional Metrics: We observe that the above interventional metrics plotted provide an upper bound to $X_n(r), L_n(r), isuEss_n(r)$ and $I_n(r)$ which are based on the set of optimal interventions for G_n that minimize the number of unoreinted edges given G_n. Therefore, by Theorem 1 they certainly provide valid upper bounds together with ϵ_n. The shaded regions in each plot are the estimates of the 95% confidence intervals as given by the `scipy.stats` function `bayes_mvs`.

Figure 1a plots empirical mean of X_n and $Y(r, A)_n$. We observe that $\bar{X}_n$ increases sharply for $\rho \geq 0.5$ and plateaus near $n = 10$, while $\bar{X}_n$ increases more gradually for $\rho < 0.5$, with a higher limit for sparse graphs. For all densities, the empirical mean of $Y(r, A)_n$ increases more gradually than the observational $\bar{X}_n$.

Figure 1b plots empirical mean of $\log L_n$ and $\log L(r, A)_n$. We again observe sharper increases and lower plateaus for the higher densities, $\rho = 0.5$ and $\rho = 0.7$, compared to more gradual rises and higher plateaus for the lower densities. Whereas in Figure 1a, $\bar{X}_n$ stabilizes at similar values for $\rho = 0.2$ and $\rho = 0.5$, in Figure 1b, the empirical mean of $\log L_n$ is greater for $\rho = 0.2$ than for $\rho = 0.5$. This indicates that each unoriented edge contributes to more MECs when the density is low.

Figure 2a demonstrates the monotonicity of the empirical mean of $isuEss_n$ and $isuEss(r, A)_n$. We observe that the empirical mean of $isuEss_n$ drops sharply for all densities, with $\rho = 0.5$ appearing to have the highest limit. The difference in behavior of the empirical mean of $isuEss(1, A)_n$ and $isuEss(2, A)_n$ for different densities is noteworthy. For sparser graphs, 1 or 2 interventions do not significantly increase the expected ability to identify the DAG; for instance, when $\rho = 0.1$, the expected number of fully identified DAGs barely changes from the observational case after $n = 30$. However, for denser graphs, such as for $\rho = 0.5$ and $\rho = 0.7$, even 1 intervention is sufficient to learn roughly 50% and 60% of the sampled graphs, respectively, and 2 interventions is sufficient to learn nearly all of them, even when $n = 110$. This result can be explained by the fact that sparse graphs often consist of multiple connected components and interventions in one component have no effect on other components. Finally, Figure 2b demonstrates the monotonicity of the empirical mean of I_n. Surprisingly, it takes very few interventions to orient even large, sparse graphs.

7 Conclusion

We provided sharp upper and lower bounds for asymptotic expected log-MEC size and the number of interventions needed to fully orient a random order DAG after $r = 0, 1, 2..$ (constant) number of initial interventions. There are various other metrics associated with I-MECs of random order DAGs that we precisely quantify in this work. Our methods relied on analytical bounds on the asymptotic quantities based on coupling arguments and exploiting the properties of Meek rules. This together with Monte Carlo simulations at finite sizes establishes quantifiable and precise bounds.

Our results mean that a walk over the space of graphs (larger search space but simpler moves) would not be more time consuming than a walk over the space of Markov equivalence classes (more complicated moves) when implementing greedy search for structure learning. This is because the asymptotic log MEC size goes to a constant for dense graphs. In addition, our results imply that in general relatively few interventions are needed to identifying dense causal networks. Investigations like this for random graphs considering various levels of sparsity and relaxing the causal sufficiency assumptions are interesting directions for future work.

Acknowledgements

C. Uhler was partially supported by NSF (DMS-1651995), ONR (N00014-17-1-2147 and N00014-18-1-

2765), IBM, and a Sloan Fellowship.

References

K. Bello and J. Honorio. Learning causal Bayes networks using interventional path queries in polynomial time and sample complexity. *arXiv preprint arXiv:1706.00754*, 2017.

E. Brenner and D. Sontag. Sparsityboost: A new scoring function for learning Bayesian network structure. *arXiv preprint arXiv:1309.6820*, 2013.

R. Castelo and T. Kocka. On inclusion-driven learning of Bayesian networks. *Journal of Machine Learning Research*, 4(Sep):527–574, 2003.

F. Eberhardt, C. Glymour, and R. Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. *arXiv preprint arXiv:1207.1389*, 2012.

M. Gangl. Causal inference in sociological research. *Annual review of sociology*, 36:21–47, 2010.

S. B. Gillispie. Formulas for counting acyclic digraph Markov equivalence classes. *Journal of Statistical Planning and Inference*, 136(4):1410–1432, 2006.

S. B. Gillispie and M. D. Perlman. Enumerating Markov equivalence classes of acyclic digraph models. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 171–177. Morgan Kaufmann Publishers Inc., 2001.

A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(Aug):2409–2464, 2012a.

A. Hauser and P. Bühlmann. Two optimal strategies for active learning of causal networks from interventional data. In *Proceedings of Sixth European Workshop on Probabilistic Graphical Models*, volume 119, page 5, 2012b.

Y. He and B. Yu. Formulas for counting the sizes of Markov equivalence classes of directed acyclic graphs. *arXiv preprint arXiv:1610.07921*, 2016.

H. Hu, Z. Li, and A. R. Vetta. Randomized experimental design for causal graph discovery. In *Advances in Neural Information Processing Systems*, pages 2339–2347, 2014.

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Experiment selection for causal discovery. *The Journal of Machine Learning Research*, 14(1):3041–3071, 2013.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636, 2007.

M. Kocaoglu, A. G. Dimakis, and S. Vishwanath. Cost-optimal learning of causal graphs. *arXiv preprint arXiv:1703.02645*, 2017.

V. Lagani, S. Triantafillou, G. Ball, J. Tegner, and I. Tsamardinos. Probabilistic computational causal discovery for systems biology. In *Uncertainty in Biology*, pages 33–73. Springer, 2016.

A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

C. Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 403–410. Morgan Kaufmann Publishers Inc., 1995.

C. Meek. *Graphical Models: Selecting causal and statistical models*. PhD thesis, PhD thesis, Carnegie Mellon University, 1997.

F. Mohammadi, C. Uhler, C. Wang, and J. Yu. Generalized permutohedra from probabilistic graphical models. *SIAM Journal on Discrete Mathematics*, 32:64–93, 2018.

A. Radhakrishnan, L. Solus, and C. Uhler. Counting Markov equivalence classes by number of immoralities. *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017.

A. Radhakrishnan, L. Solus, and C. Uhler. Counting Markov equivalence classes for dag models on trees. *Discrete Applied Mathematics*, 244:170–185, 2018.

G. Raskutti and C. Uhler. Learning directed acyclic graphs based on sparsest permutations. *Stat*, 7:e183, 2018.

K. Shanmugam, M. Kocaoglu, A. G. Dimakis, and S. Vishwanath. Learning causal graphs with small interventions. In *Advances in Neural Information Processing Systems*, pages 3195–3203, 2015.

L. Solus, Y. Wang, L. Matejovicova, and C. Uhler. Consistency guarantees for permutation-based causal inference algorithms. *arXiv preprint arXiv:1702.03530*, 2017.

P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. *Causation, prediction, and search*. MIT press, 2000.

B. Steinsky. Enumeration of labelled chain graphs and labelled essential directed acyclic graphs. *Discrete Mathematics*, 270(1-3):267–278, 2003.

M. Teyssier and D. Koller. Ordering-based search: A simple and effective algorithm for learning Bayesian networks. *arXiv preprint arXiv:1207.1429*, 2012.

S. Wagner. Asymptotic enumeration of extensional acyclic digraphs. *Algorithmica*, 66(4):829–847, 2013.

Y. Wang, L. Solus, K. Yang, and C. Uhler. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems*, pages 5822–5831, 2017.

Y. Xiao, Y. Gong, Y. Lv, Y. Lan, J. Hu, F. Li, J. Xu, J. Bai, Y. Deng, L. Liu, et al. Gene perturbation atlas (GPA): a single-gene perturbation repository for characterizing functional mechanisms of coding and non-coding genes. *Scientific Reports*, 5:10889, 2015.

K. D. Yang, A. Katcoff, and C. Uhler. Characterizing and learning equivalence classes of causal DAGs under interventions. *Proceedings of Machine Learning Research*, 80:5537–5546, 2018.