On the Unlikelihood of D-Separation

Itai Feigenbaum* ITAI@ITAIFEIGENBAUM.COM
Salesforce AI Research
Lehman College, City University of New York
The Graduate Center, City University of New York

Devansh Arpit DEVANSHARPIT@GMAIL.COM
Shelby Heinecke SHELBY.HEINECKE@SALESFORCE.COM
Juan Carlos Niebles JNIEBLES@SALESFORCE.COM
Weiran Yao WEIRAN.YAO@SALESFORCE.COM
Huan Wang HUAN.WANG@SALESFORCE.COM
Caiming Xiong CXIONG@SALESFORCE.COM
Silvio Savarese SSAVARESE@SALESFORCE.COM
Salesforce AI Research

Editors: J.H.P. Kwisthout & S. Renooij

Abstract

Causal discovery aims to recover a causal graph from data generated by it; constraint-based methods do so by searching for d-separating conditioning sets of nodes. In this paper, we provide analytic evidence that on large graphs, d-separation is a rare phenomenon, even when guaranteed to exist. Our analysis implies poor average-case performance of existing constraint-based methods, except on a vanishingly small class of extremely sparse graphs. We consider a set $V = \{v_1, \ldots, v_n\}$ of nodes, and generate a random DAG $G = (V, E)$ where $v_i \to v_j \in E$ with i.i.d probability p_1 if $i < j$ and probability 0 if $i > j$. For any d-separable pair of nodes v_i and v_j, we provide upper bounds on the probability that a subset of $V \backslash \{v_i, v_j\}$ d-separates the pair, under different subset selection scenarios; our upper bounds decay exponentially fast to 0 as $|V| \to \infty$ for any fixed expected density. We then analyze the average-case performance of constraint-based methods, including the PC Algorithm, a variant of the SGS Algorithm called UniformSGS, and also any constraint-based method limited to small conditioning sets (a limitation which holds in most of existing literature). We show that these algorithms usually suffer from low precision or exponential running time on all but extremely sparse graphs.

Keywords: d-separation, causal discovery

1. Introduction

Causal discovery aims to reverse engineer a *causal graph* from the data it generates. The nonexistence of an edge between two nodes can be shown to be equivalent to the existence of a *d-separating* (Definition 4) *conditioning set* of the other nodes. Constraint-based methods assume access to a d-separation oracle, which deduces d-separation from the data, and discover the graph via a series of oracle calls (Glymour et al., 2019). Perhaps the most well-known constraint-based method is the *PC Algorithm*, which is a specialization of the *SGS Algorithm* (Spirtes and Glymour, 1991). Under some assumptions, for sufficiently sparse

* The author is currently affiliated with the City University of New York.

graphs, the PC Algorithm recovers the undirected skeleton of the causal graph correctly and makes at most a polynomial number of oracle calls. These worst-case guarantees fail on non-sparse graphs, but that still leaves open the possibility of similar average-case guarantees on non-sparse graphs—a possibility we rule out in this paper. **Informally speaking, we show that on almost all large graphs (except for a small class of extremely sparse graphs), for any d-separable node pair, almost all conditioning sets are not d-separating. This implies poor average-case performance of existing constraint-based methods on almost all large graphs, including the PC Algorithm and any algorithm using only small conditioning sets.**

Since searching for d-separation is the core of all constraint-based methods, in this paper we set out to explore the search space. We consider a randomly generated directed acyclic graph (DAG) with $|V|$ nodes and any expected density p_1. For any two d-separable nodes, we provide upper bounds on the probability that a conditioning set d-separates the two nodes, in different scenarios. First, we bound the probability of d-separation when each node is included in the conditioning set with any fixed i.i.d. probability p_2 (Theorem 12); second, we bound the probability that there exists a d-separating set of size at most some linear fraction of $|V|$ (Theorem 14); and third, we bound the probability that a randomly chosen conditioning set of any fixed size is d-separating (Theorem 15). All of our bounds are $O(e^{-|V|})$ for any fixed p_1. In Figure 1, we give an empirical sense of the first scenario. The informal conclusion from our bounds is that for large $|V|$, d-separation is a very rare phenomenon even when guaranteed to occur, unless the graph is extremely sparse; this gives some indication of the difficulty in finding d-separation.

After establishing our bounds, we analyze their implications for the average-case performance of constraint-based methods. We begin by showing that the *extremely sparse* case where $\lim_{|V|\to\infty} p_1 = 0$ includes only a vanishingly small portion of all possible graphs. In addition to this theoretical argument, we note that sparsity is considered an unrealistic assumption in some fields such as epidemiology (Petersen et al., 2023). For all but extremely sparse graphs, we show that with high probability:

(1) Any constraint-based method which is restricted to considering only small conditioning sets (sublinear in $|V|$) suffers from low precision (Corollary 19). This restriction holds in most of literature: in practice the d-separation oracle is replaced with a statistical conditional independence test, and such tests usually perform badly when the conditioning set is not small (Runge et al., 2019; Li and Fan, 2020).

(2) Even without externally imposing a conditioning set size restriction, the PC Algorithm suffers either from low precision or exponential running time (Theorem 20 and Corollary 21). While PC's poor performance on non-sparse graphs has been empirically observed (Petersen et al., 2023), theoretical justification in literature has been limited to worst-case analysis: our average-case results provide a more comprehensive and convincing theoretical justification for this phenomenon.

(3) We consider a variant of the SGS Algorithm we call UniformSGS that samples conditioning sets uniformly at random without replacement. We show that even when there exists a d-separating conditioning set, UniformSGS takes an expected exponential time to find one (Theorem 22).

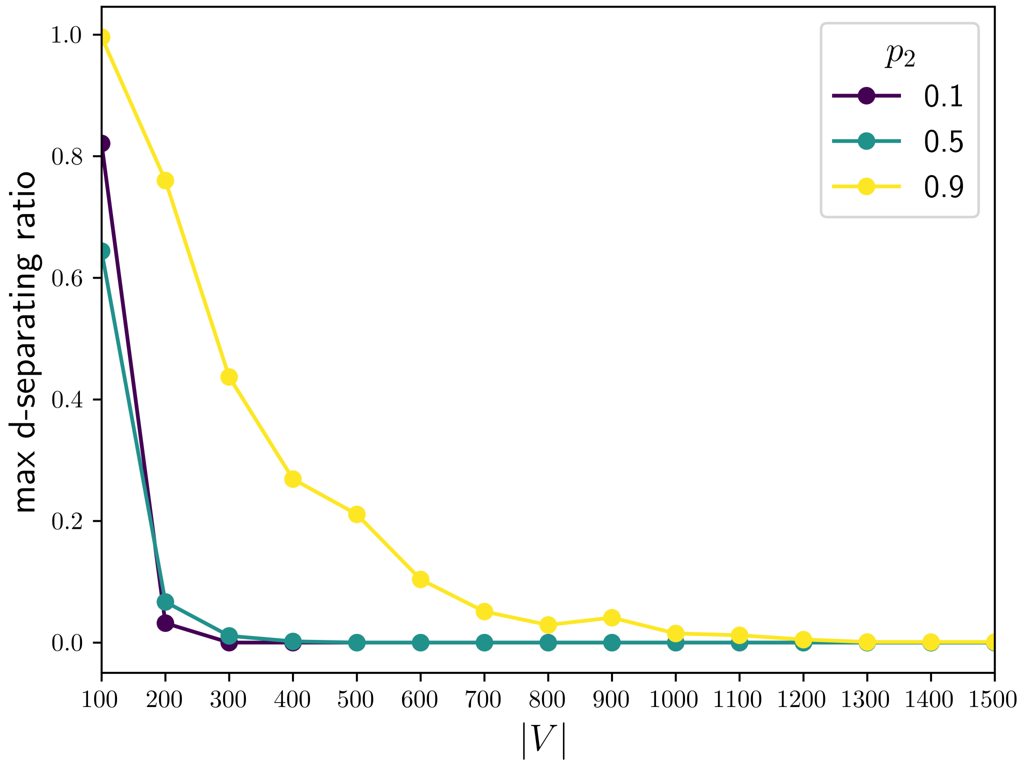


Figure 1: d-separation probability. Given a random DAG with $|V|$ nodes and expected density $p_1 = 0.05$, we sample 100 d-separable variable pairs uniformly w/o replacement from all d-separable pairs. For each pair, we sample 1000 conditioning sets, including each variable (except the pair) in the conditioning set with i.i.d. probability p_2, and calculate the ratio of the number of d-separating sets over all 1000. The y-axis is the maximum ratio over all 100 pairs. Consistently with Theorem 12, d-separation becomes rare fast as $|V| \to \infty$.

The remainder of the paper is organized as follows. In Section 1.1, we discuss related literature. In Section 2, we define our model and the relevant terminology. In Section 3, we provide three bounds on the probability that a conditioning set d-separates two nodes, in different scenarios. In Section 4, we show that the extremely sparse case is rare, and analyze the implications of our bounds for constraint-based methods on non extremely sparse graphs. In Section 5 we summarize our results. **The appendix contains experiments supporting our results and omitted proofs.**

1.1. Related Work

The early causal discovery approaches can be broadly categorized into two classes, including constrained-based ones such as PC (Spirtes and Glymour, 1991) and FCI (Spirtes et al., 2013), etc. and scored-based ones such as GES (Chickering, 2002). It was observed that under faithfulness and causal Markov assumptions (Spirtes et al., 1993), a Markov equivalence class of causal graphs could be recovered by exploiting the conditional independence relationships among the observed variables. The causal Markov condition and the faithfulness assumption ensure that there is a correspondence between d-separation properties (Verma and Pearl, 1988) in the underlying causal structure and statistical independence properties in the data. This led to the development of the constraint-based approach to causal discovery, which produces an equivalence class that may contain multiple DAGs or other graphical objects that encode the same conditional independence relationships. Since then, the field of causal discovery has grown significantly. D-Separation and different variations of the PC (and more generally, SGS) Algorithm (Spirtes et al., 1989; Spirtes and Glymour, 1991) are at the heart of constraint-based methods of causal discovery, and are the focus of this paper. Kalisch and Bühlman (2007) simulate PC on large graphs denser than in most of literature, but are still quite sparse: using our terminology, the density of their graphs decays superexponentially with the number of nodes, and their graphs are within the vanishingly small class of extremely sparse graphs defined in this paper.

In contrast, score-based algorithms (Glymour et al., 2019; Spirtes and Zhang, 2016; Heinze-Deml et al., 2018; Vowels et al., 2022) search for the equivalence class yielding the highest score for certain criteria (Chickering, 2002), such as the Bayesian Information Criterion (BIC), the posterior of the graph given the data (Heckerman et al., 2006), or the generalized score functions (Huang et al., 2018). Another set of methods is based on functional causal models (FCMs), which represent the effect as a function of direct causes together with an independent noise term. The causal direction implied by the FCM is recovered by exploiting the model assumptions, such as the independence between the noise and cause, which holds only for the true causal direction and is violated for the wrong direction, or minimal change principles (Ghassami et al., 2018; Schölkopf et al., 2021), which states that with correct causal factorization, only a few factors may change under different contexts. Examples of FCM-based causal discovery algorithms include the linear non-Gaussian acyclic model LiNGAM (Shimizu et al., 2006), the additive noise model ANM (Hoyer et al., 2008), and the post-nonlinear causal model PNL (Zhang and Hyvarinen, 2012).

2. Setup and Definitions

We denote a directed edge from node a to node b as $a \to b$, and the underlying undirected edge with $a - b$: for a set of directed edges X, the statement $a - b \in X$ means "either $a \to b \in X$ or $b \to a \in X$". By *path* we mean an undirected simple path, and by *length* of a path we mean the number of edges in it. Let $V = \{v_1, \dots, v_n\}$. Let $G = (V, E)$ be a random DAG generated as follows: if $i < j$ then $v_i \to v_j \in E$ with probability $0 < p_1 < 1$, and if $i \geq j$ then $v_i \to v_j \notin E$ deterministically. The generation of each edge is independent of the others. Let G^* be the set of all possible DAGs with nodes V which respect the topological order $v_1, \dots, v_n$. Note that G is a discrete random variable with support in G^* (every DAG in G^* is generated with nonzero probability); the special case of $p_1 = 0.5$ corresponds to G being uniform over G^*. For any $v_i, v_j \in V$, define $V_{i,j} = V \setminus \{v_i, v_j\}$.

Next, we define colliders, noncolliders, blocking and d-separation, as well as pseudoblocking and pseudoseparation. Our definitions of the first four are consistent with those of Pearl (2009), while the last two are new simple concepts we are introducing here.

Definition 1 (Collider/Noncollider Path) *Let P be a length 2 path: specifically, P is a digraph with three nodes v_i, v_k, v_j and undirected edges $v_i - v_k$, $v_k - v_j$, which we also denote with $v_i - v_k - v_j$. P is called a collider path (Figure 2(a)) if the directions of the* *edges are $v_i \to v_k$, $v_j \to v_k$ and a noncollider path (Figures 2(b)-2(d)) otherwise. The node v_k is called the middle node of the path.*

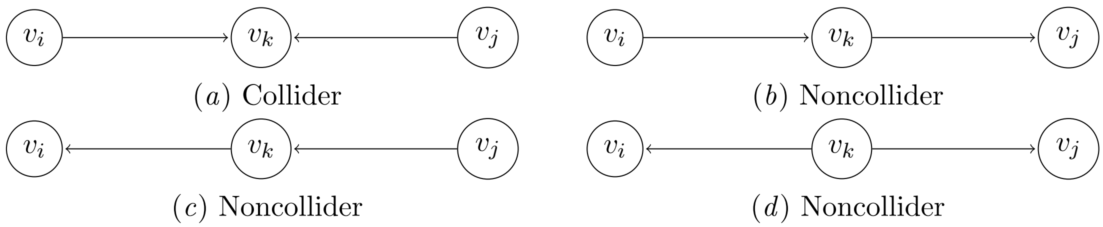


Figure 2: Visualization of colliders and noncolliders.

Definition 2 (Collider/Noncollider Node) *Let P be an arbitrary length path between nodes $a, b \in V$. Every node c in the path s.t. $c \notin \{a, b\}$ is the middle node of a unique length 2 sub-path in P, P', which includes c and the two edges adjacent to c in P. We call c a collider (resp. noncollider) node in P iff P' is a collider (resp. noncollider).*

Definition 3 (Blocking and Pseudoblocking) *Let $Z \subseteq V$. A path is blocked by Z iff at least one of the following holds:*

- *The path contains a collider node s.t. the collider and all of the collider's descendants in G are not in Z.*
- *The path contains a non-collider node in Z.*

A path is pseudoblocked by Z iff at least one of the following holds:

- *The path contains a collider node not in Z.*
- *The path contains a non-collider node in Z.*

Blocking differs from pseudoblocking only in the requirement that the collider's descendants are not in Z. Pseudoblocking is weaker than blocking: every blocked path is pseudoblocked, but not every pseudoblocked path is blocked. Similarly, in the following definition, pseudoseparation is weaker than d-separation. We will only apply the notion of pseudoblocking to length 2 paths, although it is defined for any length.

Definition 4 (D-separation and pseudoseparation) *Let $v_i, v_j \in V$ s.t. $i \neq j$, and assume $Z \subseteq V_{i,j}$. v_i and v_j are d-separated by Z iff every path between v_i and v_j is blocked by Z. v_i and v_j are pseudoseparated by Z iff every length 2 path between v_i and v_j is pseudoblocked. We denote v_i and v_j being d-separated by Z in a graph G' by $v_i \vdash_{Z,G'} v_j$, and pseudoseparated by $v_i \vdash^{ps}_{Z,G'} v_j$.*

Note that d-separation requires blocking every path between v_i and v_j, and pseudoseparation only requires pseudoblocking the length 2 paths. Pseudoseparation is useful because length 2 paths between v_i and v_j are edge disjoint and also middle node disjoint; combined with the fact that pseudoblocking—unlike blocking—ignores colliders' descendants, it allows for a decoupled analysis of the paths. As we will see, despite pseudoseparation being a weaker requirement than d-separation, it allows us to derive fairly strong bounds. Lemma 5 formally establishes the relationship between pseudoseparation and d-separation (the proof is immediate from definitions):

Lemma 5 *For any $v_i, v_j \in V$ and $Z \subseteq V_{i,j}$, $v_i \vdash_{Z,G} v_j \Rightarrow v_i \vdash^{ps}_{Z,G} v_j$.*

The reason d-separation is useful for causal discovery is due to the fact that—under some assumptions—it can be detected via conditional independence testing on the data, as well as the following known result (Pearl, 2009):

Theorem 6 *$x, y \in V$ are d-separable in G iff $x - y \notin E$.*

Throughout this paper, the basic idea is to use Lemma 5 to **establish upper bounds on the probability of d-separation by establishing upper bounds on the probability of pseudoseparation**.

3. Bounds on the Probability of D-Separation

First, some informal intuition: all of our bounds (Theorems 12, 14 and 15) rely on identifying a subset S of length 2 paths s.t. all paths in S must be pseudoblocked for pseudoseparation (and hence, by Lemma 5, for d-separation) to hold. Furthermore, we choose S so that $|S|$ grows linearly with $|V|$. This is useful because the middle node of each path in S either must be included in or must be excluded of the conditioning set for pseudoseparation to hold. Due to the established linear growth of $|S|$, the number of inclusion/exclusion decision combinations for the middle nodes grows exponentially with $|V|$, however only at most one of those combinations can yield pseudoseparation.

3.1. Preliminaries

For this section, let $v_i, v_j \in V$ and assume w.l.o.g. that $i < j$. For every $v_k \in V_{i,j}$, let P_k be the undirected path $v_i - v_k - v_j$, and let $P_{i,j} = \{P_k : v_k \in V_{i,j}\}$ be the set of all potential length 2 paths between v_i and v_j (regardless of whether they actually exist in the graph). We say that a path P_k *exists* in a graph $g = (V, E_g) \in G^*$ if both $v_i - v_k \in E_g$ and $v_j - v_k \in E_g$; when we discuss existence/inexistence of a path without referring to a particular graph, the underlying graph in question is assumed to be G. We prove upper bounds on the probability that a set of nodes Z satisfies $v_i \vdash_{Z,G} v_j$. More precisely, we prove our upper bounds conditional on v_i and v_j being d-separable in G; the implied upper bound on the unconditional probability of d-separation is even tighter (see appendix). We begin with a few lemmas:

Lemma 7 *For each $v_k \in V_{i,j}$, let A_k be the event that P_k exists in G. Then $\mathbb{P}(A_k) = p_1^2$ for all $v_k \in V_{i,j}$. Furthermore, denoting the event that $v_i - v_j \in E$ as $A_{i,j}$, the events $\{A_k : v_k \in V_{i,j}\} \cup \{A_{i,j}\}$ are mutually independent.*

Proof A_k is the event where $v_i - v_k, v_k - v_j \in E$: each of these edges exists with probability p_1 and therefore $\mathbb{P}(A_k) = p_1^2$. All paths in $P_{i,j}$ are edge disjoint, and therefore the existence of edges in each of these paths is independent of the existence of edges in the others; in addition, the edge $v_i - v_j$ is not in any of the paths in $P_{i,j}$, so its existence is independent of the existence of those paths. ■

Lemma 8 *P_k is a collider path iff $k > j$.*

Proof P_k is a collider iff both edges go into v_k, which occurs iff $k > \max\{i,j\} = j$. ■

Next, we introduce a few additional useful definitions.

Definition 9 *Define $V_{nc} = \{v_k \in V_{i,j} : k < j\}$, $V_c = \{v_k \in V_{i,j} : k > j\}$, $Q_{nc} = \{P_k : v_k \in V_{nc}\}$ and $Q_c = \{P_k : v_k \in V_c\}$. Note that by Lemma 8, $v_k \in V_{nc}$ iff P_k is a noncollider path, and $v_k \in V_c$ iff P_k is a collider path; thus, Q_{nc} and Q_c is the set of all potential noncollider and collider paths respectively. Therefore $|V_{nc}| = |Q_{nc}| = j - 2$ (corresponding to $k < j$ and $k \neq i$) and $|V_c| = |Q_c| = |V| - j$ (corresponding to $k > j$).*

Definition 10 *For $g \in G^*$, denote the number of existing noncollider paths as B^g_{nc} and the number of existing collider paths as B^g_c. When $g = G$, we write B_{nc} and B_c instead of B^G_{nc} and B^G_c.*

Lemma 11 *B_{nc} is a binomial random variable with parameters p_1^2 and $j - 2$, while B_c is a binomial random variable with parameters p_1^2 and $|V| - j$. Furthermore, B_{nc} and B_c are independent of each other and of the event $v_i - v_j \in E$.*

Proof By Lemma 8, there are $j - 2$ potential noncolliders Q_{nc} and $|V| - j$ potential colliders Q_c, and by Lemma 7 each exists with probability p_1^2 independently of the others and of $v_i - v_j \in E$. ■

3.2. Bounds

Next, we establish bounds on d-separation probability. Our first bound is for a random conditioning set which includes each node with a fixed i.i.d. probability. In this proof, S (discussed in the beginning of Section 3) includes all length 2 paths between v_i and v_j in G.

Theorem 12 *Let Z be chosen randomly from $2^{V_{i,j}}$ as follows: for every $v \in V_{i,j}$, we include $v \in Z$ with i.i.d. probability $0 < p_2 < 1$. Then, $\mathbb{P}(v_i \vdash_{Z,G} v_j | v_i - v_j \notin E)$ is upper bounded by $(1 - p_1^2 + (1 - p_2)p_1^2)^{|V|-j}(1 - p_1^2 + p_2 p_1^2)^{j-2}$.*

Proof Let $G^*_{i,j} = \{g = (V, E_g) \in G^* : v_i - v_j \notin E_g\}$ be the set of all graphs in which the edge $v_i - v_j$ is excluded. Let $g = (V, E_g) \in G^*_{i,j}$ be any particular such graph. Consider some k s.t. the path P_k exists in g. Z pseudoblocking P_k is equivalent to $v_k \in Z \Leftrightarrow v_k \in V_{nc}$, that is v_k should be included in Z iff P_k is a noncollider path. We know that $\mathbb{P}(v_k \in Z) = p_2$, and therefore P_k is pseudoblocked with probability p_2 if $P_k \in Q_{nc}$, and with probability $(1 - p_2)$ if $P_k \in Q_c$. As the inclusion/exclusion decision for each node in/from Z is done independently, it follows that:

$$\mathbb{P}(v_i \vdash^{ps}_{Z,G} v_j | G = g) \leq (1 - p_2)^{B^g_c} p_2^{B^g_{nc}}.$$

Our goal, however, is not to bound $\mathbb{P}(v_i \vdash_{Z,G} v_j | G = g)$, but rather $\mathbb{P}(v_i \vdash_{Z,G} v_j | v_i - v_j \notin E)$:

$$\begin{aligned}
&\mathbb{P}(v_i \vdash_{Z,G} v_j | v_i - v_j \notin E) \\
\leq &\mathbb{P}(v_i \vdash^{ps}_{Z,G} v_j | v_i - v_j \notin E) \\
= &\sum_{g \in G^*_{i,j}} \mathbb{P}(v_i \vdash^{ps}_{Z,G} v_j | G = g)\mathbb{P}(G = g | G \in G^*_{i,j}) \\
\leq &\sum_{g \in G^*_{i,j}} (1 - p_2)^{B^g_c} p_2^{B^g_{nc}} \mathbb{P}(G = g | G \in G^*_{i,j}) \\
= &\mathbb{E}[(1 - p_2)^{B_c} p_2^{B_{nc}} | G \in G^*_{i,j}].
\end{aligned}$$

By Lemma 11, B_c and B_{nc} are independent of $v_i - v_j \notin E$, which is equivalent to $G \in G^*_{i,j}$; thus, our bound is equal to $\mathbb{E}[(1-p_2)^{B_c} p_2^{B_{nc}}]$. Also by Lemma 11, B_c is a binomial random variable with parameters $|V| - j$ and p_1^2 while B_{nc} is a binomial random variable with parameters $j-2$ and p_1^2; furthermore, B_c and B_{nc} are independent of each other. Therefore, our bound becomes $\mathbb{E}[(1-p_2)^{B_c} p_2^{B_{nc}}] = \mathbb{E}[(1-p_2)^{B_c}]\mathbb{E}[p_2^{B_{nc}}]$. We can now use the moment generating function of the binomial to get:

$$\begin{aligned}
\mathbb{E}[(1 - p_2)^{B_c}] &= \mathbb{E}[e^{(\ln(1-p_2))B_c}] \\
&= (1 - p_1^2 + p_1^2 e^{\ln(1-p_2)})^{|V|-j} \\
&= (1 - p_1^2 + (1 - p_2)p_1^2)^{|V|-j}, \\
\mathbb{E}[p_2^{B_{nc}}] &= \mathbb{E}[e^{(\ln p_2)B_c}] \\
&= (1 - p_1^2 + p_1^2 e^{\ln p_2})^{j-2} \\
&= (1 - p_1^2 + p_2 p_1^2)^{j-2}.
\end{aligned}$$

Taking the product gives

$$\begin{aligned}
&\mathbb{E}[(1 - p_2)^{B_c}]\mathbb{E}[p_2^{B_{nc}}] \\
= &(1 - p_1^2 + (1 - p_2)p_1^2)^{|V|-j}(1 - p_1^2 + p_2 p_1^2)^{j-2}.
\end{aligned}$$

■

Replacing $(1-p_2)$ and p_2 with $\max\{p_2, (1-p_2)\}$ in the r.h.s., we get the weaker but simpler bound:

Corollary 13 *Let Z be chosen randomly from $2^{V_{i,j}}$ as follows: for every $v \in V_{i,j}$, we include $v \in Z$ with i.i.d. probability $0 < p_2 < 1$. Then, $\mathbb{P}(v_i \vdash_{Z,G} v_j | v_i - v_j \notin E)$ is upper bounded by $(1 - (1 - \max\{p_2, (1-p_2)\})p_1^2)^{|V|-2}$.*

Since $0 < p_2 < 1$, $0 < 1 - \max\{p_2, (1-p_2)\} < 1$, and since $0 < p_1 < 1$, $0 < 1 - (1 - \max\{p_2, (1-p_2)\})p_1^2 < 1$. Thus, for fixed p_1 and p_2, the bound in Corollary 13 decays exponentially fast to 0 as $|V| \to \infty$, and therefore so does the bound in Theorem 12—as it is tighter. Therefore, the probability that a random Z d-separates v_i and v_j quickly becomes very low as $|V| \to \infty$. In fact, the decay to 0 is guaranteed not just for fixed p_1, but rather as long as $p_1 = \Omega(\sqrt{\frac{\log |V|}{|V|}})$ (although, of course, the decay will no longer necessarily be exponential).[1]

Our second bound considers all conditioning sets up to a certain size. S in this proof includes all length 2 noncollider paths between v_i and v_j.

Theorem 14 *Let $Z^{0.5p_1^2(j-2)} \subseteq 2^{V_{i,j}}$ be the collection of all subsets of size up to at most $0.5p_1^2(j-2)$. Then, we can upper bound $\mathbb{P}(\exists Z \in Z^{0.5p_1^2(j-2)}$ s.t. $v_i \vdash_{Z,G} v_j | v_i - v_j \notin E)$ by $e^{-\frac{0.25p_1^2(j-2)}{2}}$.*

Proof To get pseudoseparation, it is necessary to block all paths from Q_{nc} that exist in the graph, which means include all of their middle nodes in Z. By Lemma 11, the minimum set size needed to do so is at least a binomial random variable B_{nc} with parameters p_1^2 and $j-2$. Hence every subset $Z \subseteq V_{i,j}$ s.t. $|Z| < B_{nc}$ is not d-separating. Using the Chernoff bound for a binomial random variable, we get the bound $\mathbb{P}(B_{nc} \leq 0.5p_1^2(j-2)) \leq e^{-\frac{0.25p_1^2(j-2)}{2}}$. However, in the event that $B_{nc} > 0.5p_1^2(j-2)$, no conditioning set of size at most $0.5p_1^2(j-2)$ is d-separating. ■

For any fixed $0 < p_1 < 1$, $0 < \gamma < 1$, for all $j \geq \gamma|V|$, $0.5p_1^2(j-2) \geq 0.5p_1^2(\gamma|V|-2) \to \infty$ (linearly fast) and $e^{-\frac{0.25p_1^2(j-2)}{2}} \leq e^{-\frac{0.25p_1^2(\gamma|V|-2)}{2}} \to 0$ (exponentially fast) as $|V| \to \infty$. It follows that as $|V|$ grows, for an arbitrarily large fraction of node pairs, the minimum d-separating set size is likely to increase linearly with $|V|$.

Our third bound chooses the conditioning set uniformly at random among all sets of a fixed size. We do not use it in Section 4, but it contributes to our understanding of the search space for d-separation. In the proof (relegated to the appendix), our choice of S is actually done in reverse: we fix the conditioning set, and consider S to be the set of all length 2 paths which would cause the conditioning set to fail pseudoblocking.

Theorem 15 *Let $\alpha \in \{0, 1, 2, \ldots, |V|-2\}$. Let Z be chosen uniformly at random among all subsets of size α of $V_{i,j}$. Then, the probability $\mathbb{P}(v_i \vdash_{Z,G} v_j | v_i - v_j \notin E)$ is upper bounded by $(1 - p_1^2(2-p_1^2)\frac{\alpha}{|V|-2})^{|V|-j}(1-p_1^2)^{j-\alpha-2}$.*

4. Analysis of Constraint-Based Methods

In this section, we discuss the implications of the bounds from Section 3 on the performance of constraint-based methods. We begin by proving that extremely sparse graphs are rare. After that, we define precision and introduce the PC and UniformSGS Algorithms. We then analyze the average-case performance of constraint-based methods on large and non extremely sparse graphs. We first show that constraint-based methods provide low precision when restricted to small conditioning sets (which they usually are in practice). We also show that even without that restriction, the PC algorithm must suffer from either low precision or an exponential number of oracle calls. Finally we analyze UniformSGS, showing it makes an exponential number of oracle calls even when the considered node pair is d-separable.

1. If $p_1 = \Omega(\sqrt{\frac{\log |V|}{|V|}})$, then for large enough $|V|$ we get that $(1 - (1 - \max\{p_2, (1-p_2)\})p_1^2)^{|V|-2} \leq (1 - c\frac{\log |V|}{|V|})^{|V|-2}$ for some constant $c > 0$, and by LHospital's rule this expression goes to 0 as $|V| \to \infty$.

4.1. Extreme Sparsity

In some fields, sparsity is considered to be an unrealistic assumption (Petersen et al., 2023). In this subsection, we show a theoretical sense in which sparsity restricts us to a very small family of graphs. For simplicity of presentation, assume throughout Section 4 that p_1 is a weakly monotonic function of $|V|$. We show that the *extremely sparse* case where $\lim_{|V|\to\infty} p_1 = 0$ is rare in a well-defined sense. Let G^d be the set of graphs in G^* with density at most d. Denoting the density of G as d_1 (d_1 is a random variable with $\mathbb{E}[d_1] = p_1$), we show that the expected ratio $\mathbb{E}[\frac{|G^{d_1}|}{|G^*|}] \to 0$ as $|V| \to \infty$. We provide a deterministic lemma, and convert it into a probabilistic theorem.

Lemma 16 *If $\lim_{|V|\to\infty} d = 0$, then $\lim_{|V|\to\infty} \frac{|G^d|}{|G^*|} = 0$.*

Theorem 17 *Let d_1 be the density of G. If $\lim_{|V|\to\infty} p_1 = 0$, then $\lim_{|V|\to\infty} \mathbb{E}[\frac{|G^{d_1}|}{|G^*|}] = 0$. Note that for any $\alpha > 0$, this implies $\lim_{|V|\to\infty} \mathbb{P}(\frac{|G^{d_1}|}{|G^*|} > \alpha) = 0$.*

4.2. Precision and the Algorithms Considered

Before we discuss the performance of algorithms, we need to define precision. For any causal discovery algorithm, we denote the algorithm's prediction for E as E_{pred}. Since the d-separation oracle is never wrong, $E \subseteq E_{pred}$ for any constraint-based method, as the algorithm would only remove an edge when it finds a d-separating set. Therefore, the only relevant type of mistake is predicting $v_i - v_j \in E_{pred}$ while $v_i - v_j \notin E$, which happens when the algorithm fails to find a d-separating set despite the fact that one exists.

Definition 18 *Let $E_{pred} \in V \times V$ be the output of some causal discovery algorithm A when the underlying causal graph is G. For any $v_i, v_j \in V$, we define the precision of A on v_i, v_j as $\mathbb{P}(v_i - v_j \notin E_{pred} | v_i - v_j \notin E)$.*

Next, we define UniformSGS and PC (Spirtes et al., 1989; Spirtes and Glymour, 1991): see Algorithms 1 and 2. We focus on the skeleton (undirected graph) recovery portion of the algorithms, and ignore their edge orientation phase. Note that in Algorithm 2, we did not specify the order of the selected sets in the for loop.

4.3. Performance

For the remainder of this section, we assume that $\lim_{|V|\to\infty} p_1 \neq 0$: as we showed in Section 4.1, this rules out only a vanishingly small portion of all possible graphs. Due to monotonicity, this implies that p_1 is bounded from below by some positive constant. In practice, constraint based methods must replace the d-separation oracle with a statistical conditional independence test. Most of the tests used in practice are only accurate for small conditioning sets (Runge et al., 2019; Li and Fan, 2020), and thus most constraint-based methods restrict themselves to small conditioning sets in practice. However, a direct application of Theorem 14 shows that for a constraint-based method to achieve high precision, the size of the conditioning sets considered must increase linearly with $|V|$, which for large graphs leads to much larger conditioning sets than what's considered feasible by most statistical tests.

Algorithm 1: The UniformSGS Algorithm

Input: V, the variables
Output: E_{pred}, the predicted edges

$E_{pred} \leftarrow \{x - y : x, y \in V, x \neq y\}$
for $x, y \in V$ **do**
  for $Z \subseteq V \backslash \{x, y\}$ *chosen uniformly at random without replacement* **do**
    if $x \vdash_{Z,G} y$ **then**
      Remove $x - y$ from E_{pred}
      Break
    end
  end
end
return E_{pred}

Algorithm 2: The PC Algorithm

Input: V, the variables, and $C_{\max}$, the maximum conditioning set size considered
Output: E_{pred}, the predicted edges

$E_{pred} \leftarrow \{x - y : x, y \in V, x \neq y\}$
$\forall w \in V, Adj(w) \leftarrow V \backslash \{w\}$
for $x, y \in V$ **do**
  for $Z \in 2^{Adj(x) \backslash \{y\}} \cup 2^{Adj(y) \backslash \{x\}}$ *s.t.* $|Z| \leq C_{\max}$ **do**
    if $x \vdash_{Z,G} y$ **then**
      Remove $x - y$ from E_{pred}
      $Adj(x) \leftarrow Adj(x) \backslash \{y\}$
      $Adj(y) \leftarrow Adj(y) \backslash \{x\}$
      Break
    end
  end
end
return E_{pred}

Note that since δ_1 in Corollary 19 below is arbitrary, the bound applies to any arbitrary fraction of all variable pairs.

Corollary 19 *Let* $0 < \delta_1 < 1$. *Let* A *be a constraint-based method which only calls the oracle with conditioning sets of size less than* $0.5p_1^2(\delta_1|V| - 2)$. *Then, for every node pair* $v_i, v_j \in V$ *where* $i < j$ *and* $j > \delta_1|V|$, *we get that* $\mathbb{P}(v_i - v_j \notin E_{pred} | v_i - v_j \notin E) = O(e^{-|V|})$.

Even when we ignore statistical considerations and allow testing conditioning sets of arbitrary size, our results spell trouble for constraint-based methods. For the PC Algorithm, Corollary 19 already shows that C_{max} must grow linearly with $|V|$ for good precision. However, this doesn't rule out the possibility of a sweet spot: a value of $C_{\max}$ large enough to avoid Corollary 19 yet small enough to enable polynomial running time. Alas, Theorem 20 and its Corollary 21 rule out such a sweet spot. Note that PC restricts the search space to $2^{Adj(x)\setminus\{y\}} \cup 2^{Adj(y)\setminus\{x\}}$ instead of $2^{V\setminus\{x,y\}}$: this does not impact our analysis in part (i) of Theorem 20, but we do need to take it into consideration when we prove part (ii). Theorem 20 continues to hold even if PC searches in $2^{V\setminus\{x,y\}}$ instead of $2^{Adj(x)\setminus\{y\}} \cup 2^{Adj(y)\setminus\{x\}}$.

Theorem 20 *Let $0 < \delta_1 < 1, 0 < \delta_2 < 1$. In the PC Algorithm:*

(i) Assume $C_{\max} < 0.5p_1^2(\delta_1|V| - 2)$. For every $v_i, v_j \in V$ where $i < j$ and $j > \delta_1|V|$, we get $\mathbb{P}(v_i - v_j \notin E_{pred}|v_i - v_j \notin E) = O(e^{-|V|})$.

(ii) Assume $C_{\max} > 0.5p_1(\delta_2|V| - 2)$. For every $v_i, v_j \in V$, if $v_i - v_j \in E$ then with probability $1 - O(e^{-|V|})$, PC makes $\Theta(e^{|V|})$ oracle calls for $x = v_i$, $y = v_j$.

Corollary 21 formalizes our claim that for large $|V|$, PC has low precision (Theorem 20(i)) or exponential running time (Theorem 20(ii)). In Corollary 21, note that since we can choose δ_1 as small as we like to begin with, we can make sure that when we end up in case (i), it holds for an arbitrarily large fraction of the pairs.

Corollary 21 *Let $0 < \delta_1 < 1$. In the PC Algorithm, for every $v_i, v_j \in V$ where $i < j$ and $j > \delta_1|V|$, either $\mathbb{P}(v_i - v_j \notin E_{pred}|v_i - v_j \notin E) = O(e^{-|V|})$, or the PC Algorithm makes $\Theta(e^{|V|})$ oracle calls with probability $1 - O(e^{-|V|})$ for $x = v_i$, $y = v_j$ when $v_i - v_j \in E$.*

Finally, we analyze UniformSGS. Since there is no size limit on the conditioning sets considered, then whenever $x = v_i$, $y = v_j$, $v_i - v_j \in E$, UniformSGS trivially requires an exponential number of oracle calls, because it is called on every subset in $2^{V_{i,j}}$. Theorem 22 shows that for large $|V|$, an exponential number of calls is needed even when $v_i - v_j \notin E$.

Theorem 22 *When testing whether $v_i - v_j \in E$, let C be the number of oracle calls made by UniformSGS. Let $\alpha = 1 + \frac{1}{(2-p_1^2)^{|V|-2}}$.*

(i) $\mathbb{E}[C|v_i - v_j \notin E] \geq \frac{1}{\alpha}((\frac{2}{2-p_1^2})^{|V|-2} - 1)$.

(ii) $\mathbb{E}[C] \geq p_1 2^{|V|-2} + (1 - p_1)\frac{1}{\alpha}((\frac{2}{2-p_1^2})^{|V|-2} - 1)$.

5. Conclusion

In this paper, we considered a random DAG model $G = (V, E)$, where each undirected edge is generated i.i.d. with a fixed probability. We have shown that on this model, even when d-separation is guaranteed to exist, only very few conditioning sets are d-separating. We have shown that as $|V| \to \infty$, unless the graph is extremely sparse, a randomly chosen conditioning set is highly unlikely to be d-separating, under different randomization scenarios. Specifically, the probability of d-separation decays exponentially to 0 as the $|V| \to \infty$ when the conditioning set includes each node with any fixed i.i.d. probability, or when it is limited to fixed size, or even when we try all conditioning sets of size up to a certain linear fraction of the nodes. We showed that extremely sparse graphs represent a vanishingly small portion of all possible graphs. We used our bounds to show that in the average case, the PC and UniformSGS Algorithms, as well as any constraint-based method restricted to small conditioning sets, are likely to have poor performance on large and non extremely sparse graphs. Our results indicate that any constraint-based method meant for causal discovery in large graphs must search for d-separating conditioning sets in a non-trivial manner, and must include large conditioning sets in the search.

References

D. M. Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

A. Ghassami, N. Kiyavash, B. Huang, and K. Zhang. Multi-domain causal structure learning in linear systems. *Advances in neural information processing systems*, 31, 2018.

C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

D. Heckerman, C. Meek, and G. Cooper. A bayesian approach to causal discovery. *Innovations in Machine Learning: Theory and Applications*, pages 1–28, 2006.

C. Heinze-Deml, M. H. Maathuis, and N. Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 5:371–391, 2018.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1551–1560, 2018.

M. Kalisch and P. Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

C. Li and X. Fan. On nonparametric conditional independence tests for continuous variables. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(3):e1489, 2020.

J. Pearl. *Causality*. Cambridge university press, 2009.

A. H. Petersen, J. Ramsey, C. T. Ekstrøm, and P. Spirtes. Causal discovery for observational sciences using supervised machine learning. *Journal of Data Science*, 21(2):255–280, 2023.

J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic. Detecting causal associations in large nonlinear time series datasets, sci. adv., 5, eaau4996, 2019.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

P. Spirtes and K. Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pages 1–28. SpringerOpen, 2016.

P. Spirtes, C. Glymour, and R. Scheines. Causality from probability. 1989.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* Spring-Verlag Lectures in Statistics, 1993.

P. L. Spirtes, C. Meek, and T. S. Richardson. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*, 2013.

T. Verma and J. Pearl. *Influence diagrams and d-separation.* UCLA, Computer Science Department, 1988.

M. J. Vowels, N. C. Camgoz, and R. Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

K. Zhang and A. Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

Appendix A. Experiments

Our first experiment (Figure 3) demonstrates Theorem 14 and its implications for any constraint-based algorithm which limits the conditioning set size (Corollary 19), including the PC Algorithm (Theorem 20 and Corollary 21). In our experiment, given a random DAG with $|V|$ nodes and expected density $p_1 \in \{0.05, 0.1\}$, we sample uniformly at random 30 d-separable variable pairs without replacement from all d-separable pairs, and test each pair for d-separation using **all** conditioning sets of size up to $C_{\max} \in \{2, 3\}$. We define the empirical precision to be the percentage of pairs (among the 30) for which a d-separating conditioning set is found. Our results are shown in Figure 3, and indeed already for $|V|$ in the low hundreds, most d-separable pairs do not have d-separating conditioning sets of size up to $C_{\max}$. An equivalent interpretation of our results is as an upper bound on the empirical precision of any constraint-based method limited to conditioning sets of size at most $C_{\max}$. Indeed, consider an idealized such constraint-based method, which always finds a d-separating set of size at most $C_{\max}$ if one exists. The idealized algorithm is free from any issues involving conditional independence test quality, data sample size, or any particular functional form of the structural equation model. Our experiment shows that even this idealized algorithm performs poorly on our test graphs.

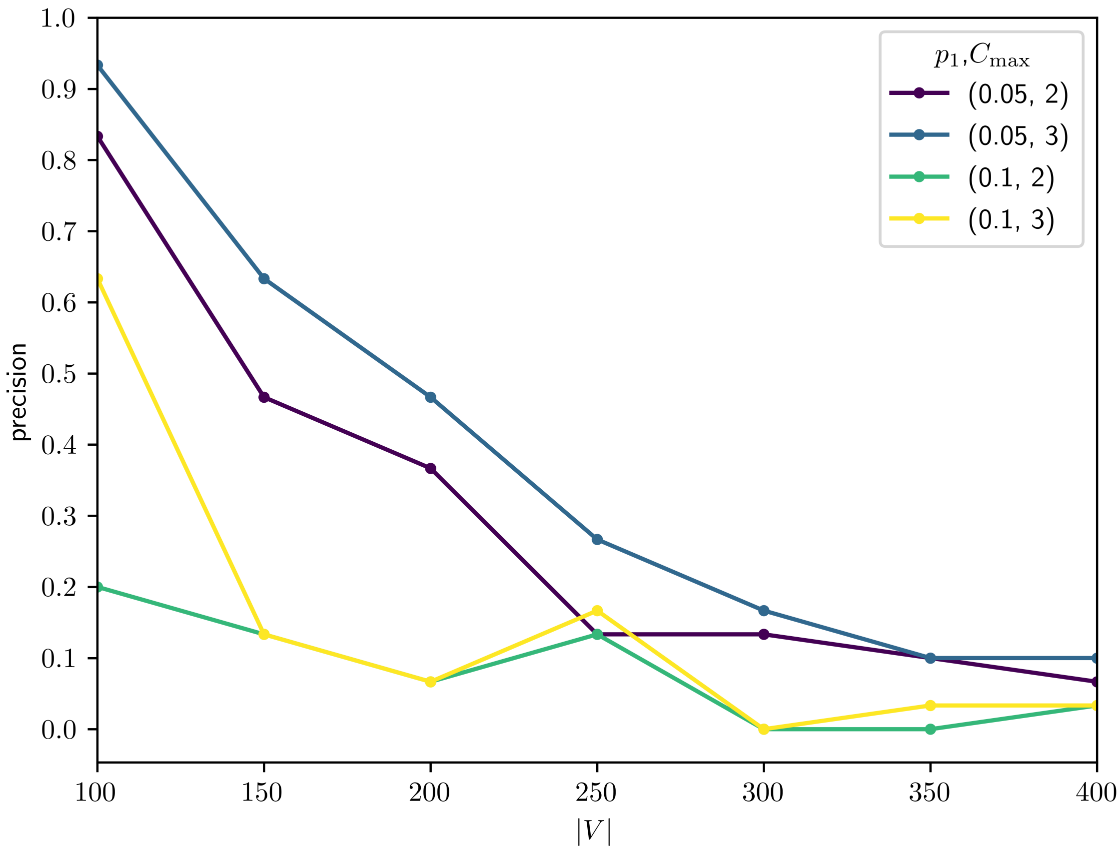
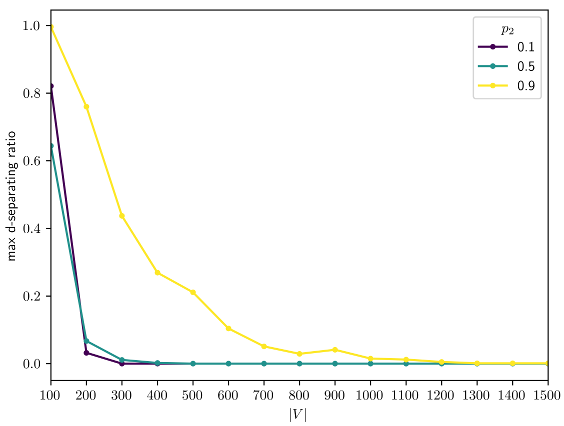
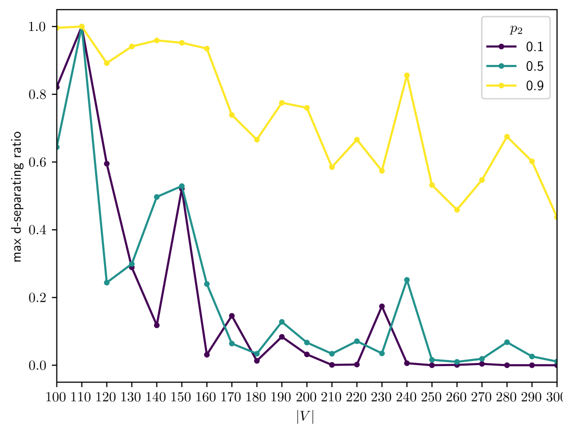


Figure 3: Experiment for Theorem 14.

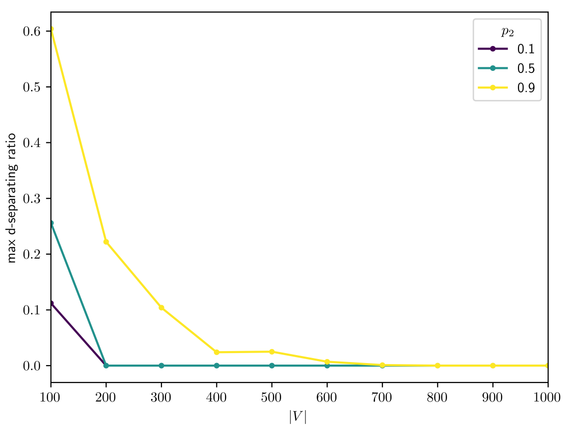
Our second experiment (Figure 4) demonstrates Theorem 12. Given a random DAG with $|V|$ nodes and expected density $p_1 \in \{0.05, 0.1\}$, we sample uniformly at random 100 d-separable variable pairs without replacement from all d-separable pairs. For each pair, we sample 1000 conditioning sets, including each variable (except the pair) in the conditioning set with i.i.d. probability $p_2 \in \{0.1, 0.5, 0.9\}$, and calculate the ratio of the number of d-separating sets over all 1000. We then calculate the maximum ratio over all 100 pairs. Consistently with Theorem 12, d-separation becomes rare fast as $|V| \to \infty$.



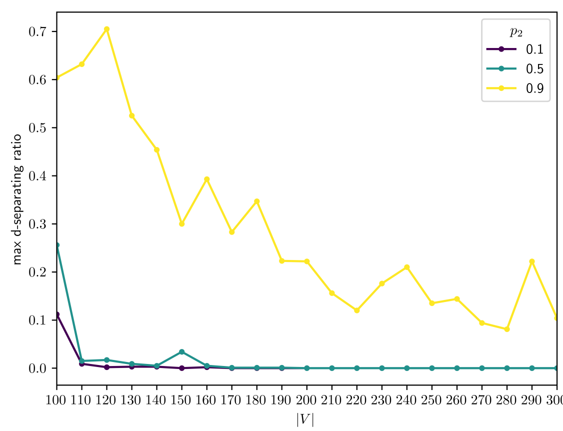
(a) Results for $p_1 = 0.05$.



(b) Fine-grained results for $p_1 = 0.05$.



(c) Results for $p_1 = 0.1$.



(d) Fine-grained results for $p_1 = 0.1$.

Figure 4: Experiment for Theorem 12. Figures 4(b) and 4(d) describe the same scenarios as Figures 4(a) and 4(c) respectively, but more fine-grained on the range $100 \leq |V| \leq 300$.

Appendix B. Proofs

In this appendix, we include the proofs omitted from the main paper.

Corollary 23 *Let Z be chosen randomly from $2^{V_{i,j}}$ as follows: for every $v \in V_{i,j}$, we include $v \in Z$ with i.i.d. probability $0 < p_2 < 1$. Then, $\mathbb{P}(v_i \vdash_{Z,G} v_j | v_i - v_j \notin E)$ is upper bounded by $(1 - (1 - \max\{p_2, (1-p_2)\})p_1^2)^{|V|-2}$.*

Proof Immediate from Theorem 12, since by replacing p_2 and $(1-p_2)$ with their upper bound $\max\{p_2, (1-p_2)\}$, we get:

$$\begin{aligned}
&(1 - p_1^2 + (1-p_2)p_1^2)^{|V|-j}(1 - p_1^2 + p_2 p_1^2)^{j-2} \\
\leq &(1 - p_1^2 + \max\{p_2, (1-p_2)\}p_1^2)^{|V|-j+j-2} \\
= &(1 - p_1^2 + \max\{p_2, (1-p_2)\}p_1^2)^{|V|-2} \\
= &(1 - (1 - \max\{p_2, (1-p_2)\})p_1^2)^{|V|-2}
\end{aligned}$$

∎

Theorem 15 *Let $\alpha \in \{0, 1, 2, \ldots, |V|-2\}$. Let Z be chosen uniformly at random among all subsets of size α of $V_{i,j}$. Then, the probability $\mathbb{P}(v_i \vdash_{Z,G} v_j | v_i - v_j \notin E)$ is upper bounded by $(1 - p_1^2(2-p_1^2)\frac{\alpha}{|V|-2})^{|V|-j}(1-p_1^2)^{j-\alpha-2}$.*

Proof Let $2_\alpha^{V_{i,j}}$ be the set of all subsets of $V_{i,j}$ of size exactly α. Let $z \in 2_\alpha^{V_{i,j}}$. Let $m_{nc}(z) = |V_{nc} \backslash z|$ be the number of nodes from V_{nc} outside of z, and $m_c(z) = |z \cap V_c|$ be the number of nodes from V_c inside z. $m_{nc}(z)$ is the number of paths in Q_{nc} which, if exist, violate pseudoseparation; similarly, $m_c(z)$ is the number of paths in Q_c that, if exist, violate pseudoseparation. Pseudoblocking holds iff all paths counted by $m(z) = m_c(z) + m_{nc}(z)$ do not exist, and by Lemma 7 each of them fails to exist with probability $1 - p_1^2$ independently of the others and of whether $v_i - v_j \in E$. Thus:

$$\mathbb{P}(v_i \vdash_{z,G}^{ps} v_j | v_i - v_j \notin E) = (1 - p_1^2)^{m(z)},$$

Then, using the fact that Z is independent from whether or not $v_i - v_j \in E$, we get:

$$\begin{aligned}
&\mathbb{P}(v_i \vdash_{Z,G}^{ps} v_j | v_i - v_j \notin E) \\
= &\sum_{z \in 2_\alpha^{V_{i,j}}} \mathbb{P}(v_i \vdash_{Z,G}^{ps} v_j | v_i - v_j \notin E, Z = z)\mathbb{P}(Z = z) \\
= &\sum_{z \in 2_\alpha^{V_{i,j}}} (1 - p_1^2)^{m(z)}\mathbb{P}(Z = z) \\
= &\mathbb{E}[(1 - p_1^2)^{m(Z)}]
\end{aligned}$$

Let $M_c = m_c(Z)$, $M_{nc} = m_{nc}(Z)$. M_c determines M_{nc}, as exactly $\alpha - M_c$ slots in Z are given to nodes from V_{nc}, so the remaining $|V_{nc}| - (\alpha - M_c) = j - 2 - (\alpha - M_c) = M_c + (j - \alpha - 2)$ nodes from V_{nc} end up outside of Z: $M_{nc} = M_c + (j - \alpha - 2)$. Therefore,

$m(Z) = M_c + M_{nc} = 2M_c + (j - \alpha - 2)$. Thus, we get:

$$\begin{aligned}
&\mathbb{P}(v_i \vdash^{ps}_{Z,G} v_j | v_i - v_j \notin E)\\
=&\mathbb{E}[(1-p_1^2)^{2M_c+(j-\alpha-2)}]\\
=&\mathbb{E}[((1-p_1^2)^2)^{M_c}]\mathbb{E}[(1-p_1^2)^{j-\alpha-2}]\\
=&\mathbb{E}[((1-p_1^2)^2)^{M_c}](1-p_1^2)^{j-\alpha-2}\\
=&\mathbb{E}[e^{\ln((1-p_1^2)^2)M_c}](1-p_1^2)^{j-\alpha-2}
\end{aligned}$$

Since Z is chosen uniformly at random from $2^{V_{i,j}}_{\alpha}$, M_c can be thought of as drawing α nodes from the population $V_{i,j}$ of nodes, where V_c of the nodes are defined as "success" states (as in, every node from V_c that we draw increases the value of M_c by one). This means that M_c is a hypergeometric random variable with population size $N = |V| - 2$, $K = |V_c| = |V| - j$ success states, and number of draws $d = \alpha$. Thus, $\mathbb{E}[e^{\ln((1-p_1^2)^2)M_c}]$ can be derived from the moment generating function of the hypergeometric distribution. It is known that the moment generating function of the hypergeometric distribution with parameters N, K and d is upper bounded by the moment generating function of the binomial random variable with parameters $\frac{d}{N}$ and K (Hoeffding, 1963). Applying this result, we get:

$$\begin{aligned}
&\mathbb{E}[e^{\ln((1-p_1^2)^2)M_c}](1-p_1^2)^{j-\alpha-2}\\
\leq&((1-\frac{d}{N}) + (1-p_1^2)^2\frac{d}{N})^K(1-p_1^2)^{j-\alpha-2}\\
=&((1-\frac{\alpha}{|V|-2}) + (1-p_1^2)^2\frac{\alpha}{|V|-2})^{|V|-j}(1-p_1^2)^{j-\alpha-2}\\
=&(1-p_1^2(2-p_1^2)\frac{\alpha}{|V|-2})^{|V|-j}(1-p_1^2)^{j-\alpha-2}
\end{aligned}$$

Therefore we have established:

$$\begin{aligned}
&\mathbb{P}(v_i \vdash_{Z,G} v_j | v_i - v_j \notin E)\\
\leq&\mathbb{P}(v_i \vdash^{ps}_{Z,G} v_j | v_i - v_j \notin E)\\
\leq&(1-p_1^2(2-p_1^2)\frac{\alpha}{|V|-2})^{|V|-j}(1-p_1^2)^{j-\alpha-2}.
\end{aligned}$$

■

Lemma 24 *If* $\lim_{|V|\to\infty} d = 0$, *then* $\lim_{|V|\to\infty} \frac{|G^d|}{|G^*|} = 0$.

Proof Assume $\lim_{|V|\to\infty} d = 0$. Consider $|V|$ large enough so that $d < 0.5$, and for convenience assume $\frac{1}{d}$ divides $\binom{|V|}{2}$. The total number of graphs in G^* is

$$A = 2^{\binom{|V|}{2}}.$$

For any n, the number of graphs with exactly n edges is $\binom{\binom{|V|}{2}}{n}$. Graphs in G^* with density at most d have at most $d\binom{|V|}{2}$ edges; the number of such graphs is therefore

$$B = \sum_{n=0}^{d\binom{|V|}{2}} \binom{\binom{|V|}{2}}{n}.$$

We claim that $B \leq \frac{d}{1-2d-\binom{|V|}{2}^{-1}} A$ (it's actually true that $B \leq dA$, but we don't need it); if we can show this, then the ratio of graphs with density at most d to all graphs is

$$\frac{B}{A} \leq \frac{\frac{d}{1-2d-\binom{|V|}{2}^{-1}} A}{A} = \frac{d}{1-2d-\binom{|V|}{2}^{-1}},$$

which goes to 0 as $|V| \to \infty$, so that will prove our claim.

Why does $B \leq \frac{d}{1-2d-\binom{|V|}{2}^{-1}} A$? Note that by definition, $A = \sum_{n=0}^{\binom{|V|}{2}} \binom{\binom{|V|}{2}}{n}$. Therefore, B is the first $d\binom{|V|}{2}$ terms of A. Let $C = \sum_{n=d\binom{|V|}{2}+1}^{(1-d)\binom{|V|}{2}-1} \binom{\binom{|V|}{2}}{n}$. Every term in C is larger than every term in B (binomial coefficients are larger the closer they are to the midpoint), and the number of terms in C is $(1-2d)\binom{|V|}{2} - 1$, while the number of terms in B is $d\binom{|V|}{2}$. It follows that $\frac{B}{A} < \frac{B}{C} < \frac{d}{1-2d-\binom{|V|}{2}^{-1}}$. ■

Theorem 17 *Let d_1 be the density of G. If $\lim_{|V|\to\infty} p_1 = 0$, then $\lim_{|V|\to\infty} \mathbb{E}[\frac{|G^{d_1}|}{|G^*|}] = 0$. Note that for any $\alpha > 0$, this implies $\lim_{|V|\to\infty} \mathbb{P}(\frac{|G^{d_1}|}{|G^*|} > \alpha) = 0$.*

Proof Assume $\lim_{|V|\to\infty} p_1 = 0$. Condition on $d_1 < \sqrt{p_1}$ to get

$$\begin{aligned}
&\mathbb{E}[\frac{|G^{d_1}|}{|G^*|}] \\
=&\mathbb{E}[\frac{|G^{d_1}|}{|G^*|} | d_1 < \sqrt{p_1}]\mathbb{P}(d_1 < \sqrt{p_1}) \\
&+ \mathbb{E}[\frac{|G^{d_1}|}{|G^*|} | d_1 \geq \sqrt{p_1}]\mathbb{P}(d_1 \geq \sqrt{p_1}).
\end{aligned}$$

We bound the first term using $\mathbb{P}(d_1 < \sqrt{p_1}) \leq 1$:

$$\begin{aligned}
&\mathbb{E}[\frac{|G^{d_1}|}{|G^*|} | d_1 < \sqrt{p_1}]\mathbb{P}(d_1 < \sqrt{p_1}) \\
\leq&\mathbb{E}[\frac{|G^{d_1}|}{|G^*|} | d_1 < \sqrt{p_1}] \\
\leq&\frac{|G^{\sqrt{p_1}}|}{|G^*|}.
\end{aligned}$$

Then, we bound the second term via the fact that $\frac{|G^{d_1}|}{|G^*|} \leq 1$ for all possible values of d_1:

$$\begin{aligned}
&\mathbb{E}[\frac{|G^{d_1}|}{|G^*|} | d_1 \geq \sqrt{p_1}]\mathbb{P}(d_1 \geq \sqrt{p_1}) \\
\leq&\mathbb{P}(d_1 \geq \sqrt{p_1}).
\end{aligned}$$

Using Markov's inequality and the fact that $\mathbb{E}[d_1] = p_1$, we get that

$$\mathbb{P}(d_1 \geq \sqrt{p_1}) \leq \frac{\mathbb{E}[d_1]}{\sqrt{p_1}} = \sqrt{p_1}.$$

Combining the bounds, we get $\mathbb{E}[\frac{|G^{d_1}|}{|G^*|}] \leq \frac{|G^{\sqrt{p_1}}|}{|G^*|} + \sqrt{p_1}$. As $p_1 \to 0$, $\frac{|G^{\sqrt{p_1}}|}{|G^*|} \to 0$ by Lemma 16. Since $\lim_{|V|\to\infty} p_1 = 0$, this completes the proof. ■

Corollary 25 *Let $0 < \delta_1 < 1$. Let A be a constraint-based method which only calls the oracle with conditioning sets of size less than $0.5p_1^2(\delta_1|V| - 2)$. Then, for every node pair $v_i, v_j \in V$ where $i < j$ and $j > \delta_1|V|$, we get that $\mathbb{P}(v_i - v_j \notin E_{pred}|v_i - v_j \notin E) = O(e^{-|V|})$.*

Proof Follows from Theorem 14 and the fact that for $v_i - v_j \notin E_{pred}$ it is necessary that there exists a d-separating subset in $\{Z \in 2^{V_{i,j}} : |Z| \leq C_{\max}\}$, where $C_{\max}$ is the maximum size of conditioning set considered by the algorithm. ■

Theorem 20 *Let $0 < \delta_1 < 1, 0 < \delta_2 < 1$. In the PC Algorithm:*

(i) Assume $C_{\max} < 0.5p_1^2(\delta_1|V| - 2)$. For every $v_i, v_j \in V$ where $i < j$ and $j > \delta_1|V|$, we get $\mathbb{P}(v_i - v_j \notin E_{pred}|v_i - v_j \notin E) = O(e^{-|V|})$.

(ii) Assume $C_{\max} > 0.5p_1(\delta_2|V| - 2)$. For every $v_i, v_j \in V$, if $v_i - v_j \in E$ then with probability $1 - O(e^{-|V|})$, PC makes $\Theta(e^{|V|})$ oracle calls for $x = v_i$, $y = v_j$.

Proof Part (i) follows from Theorem 14 and the fact that for $v_i - v_j \notin E_{pred}$ it is necessary that there exists a d-separating subset in $\{Z \in 2^{V_{i,j}} : |Z| \leq C_{\max}\}$. For part (ii), consider $v_i - v_j \in E$. Edges from E never get deleted, so at all times $E \subseteq E_{pred}$. Defining $S_j = Adj(v_j)\backslash\{v_i\}$, $|S_j|$ is therefore always bounded from below by the number of edges in E adjacent to v_j except $v_i - v_j$, which is a binomial random variable with parameters p_1 and $|V| - 2$. Using the Chernoff bound, we get that

$$\mathbb{P}(|S_j| \leq 0.5p_1(|V| - 2)) \leq e^{-\frac{0.25p_1(|V|-2)}{2}} = O(e^{-|V|}).$$

Therefore, with probability $1 - O(e^{-|V|})$,

$$|S_j| > 0.5p_1(|V| - 2) > 0.5p_1(\delta_2|V| - 2),$$

and since also $C_{\max} > 0.5p_1(\delta_2|V| - 2)$, there are at least $2^{0.5p_1(\delta_2|V|-2)} = \Theta(e^{|V|})$ subsets in the search space (we can use Θ instead of Ω because there are at most $2^{|V|-2} = O(e^{|V|})$ subsets). Since $v_i - v_j \in E$, there is no d-separating set, so the oracle will be called on every subset in the search space. ■

Corollary 26 *Let $0 < \delta_1 < 1$. In the PC Algorithm, for every $v_i, v_j \in V$ where $i < j$ and $j > \delta_1|V|$, either $\mathbb{P}(v_i - v_j \notin E_{pred}|v_i - v_j \notin E) = O(e^{-|V|})$, or the PC Algorithm makes $\Theta(e^{|V|})$ oracle calls with probability $1 - O(e^{-|V|})$ for $x = v_i$, $y = v_j$ when $v_i - v_j \in E$.*

Proof Note that no matter how small we choose $\delta_1 > 0$ in Theorem 20, we can always choose $\delta_2 > 0$ s.t. $(p_1\delta_1 - \delta_2)|V| > 2p_1 - 2$, or equivalently $p_1(\delta_1|V| - 2) > \delta_2|V| - 2$. When we do so, $0.5p_1^2(\delta_1|V| - 2) > 0.5p_1(\delta_2|V| - 2)$, and so we are necessarily in either case (i) or (ii) of the theorem. This is the formalization of our claim that for large $|V|$, PC has low precision (case (i)) or exponential running time (case (ii)). ■

Theorem 22 *When testing whether $v_i - v_j \in E$, let C be the number of oracle calls made by UniformSGS. Let $\alpha = 1 + \frac{1}{(2-p_1^2)^{|V|-2}}$.*

(i) $\mathbb{E}[C|v_i - v_j \notin E] \geq \frac{1}{\alpha}((\frac{2}{2-p_1^2})^{|V|-2} - 1)$.

(ii) $\mathbb{E}[C] \geq p_1 2^{|V|-2} + (1 - p_1)\frac{1}{\alpha}((\frac{2}{2-p_1^2})^{|V|-2} - 1)$.

Proof First, assume $v_i - v_j \notin E$. Let $K = |\{z \in 2^{V \setminus \{v_i, v_j\}} : v_i \not\perp_{z,G} v_j\}|$ (the number of subsets that are not d-separating in G). Let $G^*_{i,j} = \{g = (V, E_g) \in G^* : v_i - v_j \notin E_g\}$, and for each $g \in G^*_{i,j}$, let k_g be the value of K when $G = g$. Conditional on $K = k_g$, C is a negative hypergeometric random variable with population size $N = 2^{|V|-2}$ (all subsets), k_g success states, and failure number $r = 1$. The expectation of such a negative hypergeometric random variable is $\frac{k_g}{N-k_g+1}$. Therefore, we get

$$\begin{aligned}
&\mathbb{E}[C|v_i - v_j \notin E] \\
=&\mathbb{E}[C|G \in G^*_{i,j}] \\
=&\sum_{g \in G^*_{i,j}} \mathbb{E}[C|G = g]\mathbb{P}(G = g|G \in G^*_{i,j}) \\
=&\sum_{g \in G^*_{i,j}} \frac{k_g}{N - k_g + 1}\mathbb{P}(G = g|G \in G^*_{i,j}) \\
=&\mathbb{E}[\frac{K}{N - K + 1}|G \in G^*_{i,j}] \\
\geq&\frac{\mathbb{E}[K|G \in G^*_{i,j}]}{N - \mathbb{E}[K|G \in G^*_{i,j}] + 1} \\
=&\frac{\mathbb{E}[K|v_i - v_j \notin E]}{N - \mathbb{E}[K|v_i - v_j \notin E] + 1},
\end{aligned}$$

where the inequality follows from Jensen's inequality.

Suppose Z is chosen as in Theorem 12 with $p_2 = 0.5$, meaning uniformly at random over $2^{V \setminus \{v_i, v_j\}}$. Because Z is chosen uniformly, $\mathbb{P}(v_i \not\perp_{Z,G} v_j | G = g)$ is simply $\frac{k_g}{2^{|V|-2}}$, and so

$k_g = 2^{|V|-2}\mathbb{P}(v_i \not\perp_{Z,G} v_j | G = g)$. Now we can compute:

$$\begin{aligned}
&\mathbb{E}[K|v_i - v_j \notin E] \\
=&\mathbb{E}[K|G \in G^*_{i,j}] \\
=&\sum_{g \in G^*_{i,j}} k_g \mathbb{P}(G = g|G \in G^*_{i,j}) \\
=&\sum_{g \in G^*_{i,j}} 2^{|V|-2}\mathbb{P}(v_i \not\perp_{Z,G} v_j|G = g)\mathbb{P}(G = g|G \in G^*_{i,j}) \\
=&2^{|V|-2} \sum_{g \in G^*_{i,j}} \mathbb{P}(v_i \not\perp_{Z,G} v_j|G = g)\mathbb{P}(G = g|G \in G^*_{i,j}) \\
=&2^{|V|-2}\mathbb{P}(v_i \not\perp_{Z,G} v_j|G \in G^*_{i,j}) \\
=&2^{|V|-2}\mathbb{P}(v_i \not\perp_{Z,G} v_j|v_i - v_j \notin E).
\end{aligned}$$

Applying Theorem 12, we get $\mathbb{P}(v_i \not\perp_{Z,G} v_j|v_i - v_j \notin E) \geq 1 - (1 - 0.5p_1^2)^{|V|-2}$, and therefore we get:

$$\begin{aligned}
&\mathbb{E}[K|v_i - v_j \notin E] \geq 2^{|V|-2}(1 - (1 - 0.5p_1^2)^{|V|-2}) \\
=&2^{|V|-2} - (2 - p_1^2)^{|V|-2},
\end{aligned}$$

and thus, bringing everything together:

$$\begin{aligned}
&\mathbb{E}[C|v_i - v_j \notin E] \\
\geq&\frac{\mathbb{E}[K|v_i - v_j \notin E]}{N - \mathbb{E}[K|v_i - v_j \notin E] + 1} \\
=&\frac{\mathbb{E}[K|v_i - v_j \notin E]}{2^{|V|-2} - \mathbb{E}[K|v_i - v_j \notin E] + 1} \\
\geq&\frac{2^{|V|-2} - (2 - p_1^2)^{|V|-2}}{(2 - p_1^2)^{|V|-2} + 1} \\
=&\frac{2^{|V|-2} - (2 - p_1^2)^{|V|-2}}{(2 - p_1^2)^{|V|-2} + \frac{(2-p_1^2)^{|V|-2}}{(2-p_1^2)^{|V|-2}}} \\
=&\frac{2^{|V|-2} - (2 - p_1^2)^{|V|-2}}{(1 + \frac{1}{(2-p_1^2)^{|V|-2}})(2 - p_1^2)^{|V|-2}} \\
=&\frac{2^{|V|-2} - (2 - p_1^2)^{|V|-2}}{\alpha(2 - p_1^2)^{|V|-2}} \\
=&\frac{1}{\alpha}((\frac{2}{2 - p_1^2})^{|V|-2} - 1).
\end{aligned}$$

We get the unconditional bound in (ii) with the following calculation:

$$\begin{aligned}
&E[C] \\
=&E[C|v_i - v_j \notin E]\mathbb{P}(v_i - v_j \notin E) \\
&+ E[C|v_i - v_j \in E]\mathbb{P}(v_i - v_j \in E) \\
=&E[C|v_i - v_j \notin E](1 - p_1) + 2^{|V|-2}p_1
\end{aligned}$$

■

Appendix C. Unconditional Bounds

Theorems 12 and 14 yield slightly improved bounds when we do not condition on $v_i - v_j \notin E$. These bounds are given in corollaries 27 and 28 below.

Corollary 27 *Let Z be chosen randomly from $2^{V_{i,j}}$ as follows: for every $v \in V_{i,j}$, we include $v \in Z$ with i.i.d. probability $0 < p_2 < 1$. Then, $\mathbb{P}(v_i \vdash_{Z,G} v_j)$ is upper bounded by $(1-p_1)(1-p_1^2+(1-p_2)p_1^2)^{|V|-j}(1-p_1^2+p_2p_1^2)^{j-2}$.*

Proof Recall from Theorem 6 that v_i and v_j are d-separable in G iff $v_i - v_j \notin E$, so $\mathbb{P}(v_i \vdash_{Z,G} v_j | v_i - v_j \in E) = 0$. Therefore, we get:

$$\begin{aligned}
&\mathbb{P}(v_i \vdash_{Z,G} v_j) \\
=&\mathbb{P}(v_i \vdash_{Z,G} v_j | v_i - v_j \notin E)(1-p_1) + 0 \cdot p_1 \\
=&\mathbb{P}(v_i \vdash_{Z,G} v_j | v_i - v_j \notin E)(1-p_1).
\end{aligned}$$

Applying the bound from Theorem 12 completes the proof. ■

Corollary 28 *Let $Z^{0.5p_1^2(j-2)} \subseteq 2^{V_{i,j}}$ be the collection of all subsets of size up to at most $0.5p_1^2(j-2)$. Then, $\mathbb{P}(\exists Z \in Z^{0.5p_1^2(j-2)}$ s.t. $v_i \vdash_{Z,G} v_j)$ is upper bounded by $(1-p_1)e^{-\frac{0.25p_1^2(j-2)}{2}}$.*

Proof Similar to Corollary 27. ■