Minimal Enumeration of All Possible Total Effects in a Markov Equivalence Class

F. Richard Guo and **Emilija Perković**
Department of Statistics, University of Washington, Seattle

Abstract

In observational studies, when a total causal effect of interest is not identified, the set of all possible effects can be reported instead. This typically occurs when the underlying causal DAG is only known up to a Markov equivalence class, or a refinement thereof due to background knowledge. As such, the class of possible causal DAGs is represented by a maximally oriented partially directed acyclic graph (MPDAG), which contains both directed and undirected edges. We characterize the minimal additional edge orientations required to identify a given total effect. A recursive algorithm is then developed to enumerate subclasses of DAGs, such that the total effect in each subclass is identified as a distinct functional of the observed distribution. This resolves an issue with existing methods, which often report possible total effects with duplicates, namely those that are numerically distinct due to sampling variability but are in fact causally identical.

1 INTRODUCTION

We consider identifying total causal effects ("total effects" or simply "effects" throughout) from causal graphs that can be learned from observational data and background knowledge, under the assumption of no latent variables. The full knowledge of the causal system is typically represented by a directed acyclic graph (DAG) (Pearl, 2009). Fig. 1(a) shows an example DAG $\mathcal{D}$. Each node $u \in \{A, Y, V_1, V_2\}$ in $\mathcal{D}$ represents a random variable X_u in a random vector $X = (X_A, X_Y, X_{V_1}, X_{V_2})$. Each edge in $\mathcal{D}$ represents a direct causal relationship between two variables.

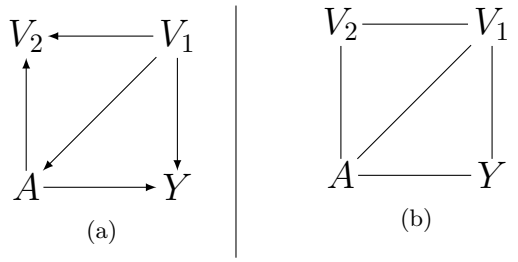

Figure 1: (a) DAG $\mathcal{D}$, (b) CPDAG $\mathcal{C}$.

Given a causal DAG, every total effect can be identified and hence consistently estimated from observational data (Robins, 1986; Pearl, 1995; Pearl and Robins, 1995; Galles and Pearl, 1995). In general, however, one cannot learn a causal DAG from observational data. Instead, under the assumption of no latent variables, one can learn a Markov equivalence class of DAGs that can give rise to the observed distribution. A Markov equivalence class is uniquely represented by a completed partially directed acyclic graph (CPDAG), also known as an essential graph (Meek, 1995; Andersson et al., 1997; Spirtes et al., 2000; Chickering, 2002). Within the equivalence class, one DAG should not be preferred over another based on observational data. Fig. 1(b) shows the CPDAG $\mathcal{C}$ that represents $\mathcal{D}$.

Often, we may have additional background knowledge on the underlying causal system. For example, we may know that A temporally precedes Y and therefore determine (or reveal) the edge orientation $A \to Y$ in CPDAG $\mathcal{C}$. Adding this knowledge to $\mathcal{C}$ results in a maximally oriented partially directed acyclic graph (MPDAG) $\mathcal{G}$, drawn in Fig. 2(a). MPDAGs are a class of graphs that subsumes both CPDAGs and DAGs. They are obtained by (optionally) adding edge orientations to a CPDAG and completing the orientation rules of Meek (1995) (see Fig. 3). As such, the class of DAGs represented by an MPDAG is a refinement of the corresponding Markov equivalence class. For example, the class of DAGs represented by $\mathcal{G}$ is drawn in Fig. 2(b), which consists of all DAGs in the Markov

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

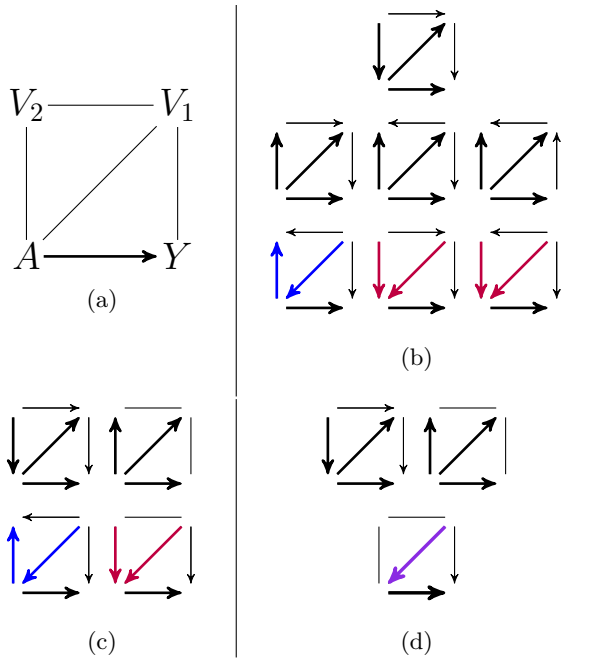

Figure 2: Example 1. (a) MPDAG $\mathcal{G}$, (b) all DAGs represented by $\mathcal{G}$, (c) all MPDAGs with distinct parent sets of A represented by $\mathcal{G}$, (d) all MPDAGs with distinct functionals $f(x_Y|\,\mathrm{do}(x_A))$ represented by $\mathcal{G}$.

equivalence class represented by $\mathcal{C}$ that have $A \to Y$.

The background knowledge of pairwise causal relationships of this type can be derived from field expertise. Moreover, other types of background knowledge, such as tiered orderings (Scheines et al., 1998), non-ancestral background knowledge (Fang and He, 2020), knowledge derived from experimental data (Hauser and Bühlmann, 2012; Wang et al., 2017) as well as certain model restrictions (Hoyer et al., 2008; Rothenhäusler et al., 2018) can also be used to obtain an MPDAG.

A total effect is identified given an equivalence class of DAGs if it can be expressed as a functional of the observed distribution, which is the same for all DAGs in the equivalence class; see Section 2 for the definition. Recently, Perković (2020) gave a necessary and sufficient graphical condition for identifying a total effect given an MPDAG (Theorem 1). When the condition fails, the effect of interest cannot be identified. One example of this case is the effect of A on Y given $\mathcal{G}$ in Fig. 2(a). In such cases, the observational data can still be informative if one identifies a finite set that contains the true effect. To do so, one can enumerate all DAGs represented by $\mathcal{G}$ and estimate the total effect under each, obtaining a set of estimated *possible* total effects. For instance, for the class of DAGs in Fig. 2(b), seven possible total effects can be reported. However, there are two drawbacks to this approach. First, enumerating all DAGs in a Markov equivalence class (or a refinement thereof) is computationally prohibitive unless one has only a few variables. For example, the complete CPDAG of p variables contains $p!$ DAGs; see also Gillispie and Perlman (2002); Steinsky (2013). Second, the number of distinct possible effects can be much smaller than the size of the equivalence class. For example, the effect of A on Y is the same for the three DAGs listed in the second row of Fig. 2(b). That being said, depending on the estimator applied to each DAG, one may obtain three estimates that only *look* different in finite samples, but which only represent different estimators for the same possible effect. These statistical *duplicates* are undesirable as it undermines the interpretability of the estimated set. Hence, to save computation time and deliver causally informative estimates, one should instead enumerate all possible effects that are distinct, or in other words, *minimally*.

Recent works on this topic include the "intervention calculus when the DAG is absent" (IDA) algorithms and joint-IDA algorithms of Maathuis et al. (2009), Nandy et al. (2017), Perković et al. (2017), Witte et al. (2020), Fang and He (2020), and Liu et al. (2020). Given an MPDAG $\mathcal{G}$, these methods enumerate a set of MPDAGs in which the total effect of A on Y is identified, by considering all orientation configurations of edges connected to A. However, this is often not minimal. For instance, to estimate the total effect of A on Y given MPDAG $\mathcal{G}$ in Fig. 2(a), the IDA methods would enumerate four graphs listed in Fig. 2(c) and thus report four estimates. Nevertheless, only three distinct total effects correspond to the MPDAGs listed in Fig. 2(d).

In this paper, we characterize the minimal additional edge orientations needed to identify a given total effect. Based on this characterization, we develop a recursive algorithm that outputs the minimal set of possible total effects and the corresponding MPDAGs. Our results hold nonparametrically, that is, without assuming a particular type of data-generating mechanism such as linearity. Furthermore, our results can be used in conjunction with recent developments on efficient effect estimators (Henckel et al., 2019; Rotnitzky and Smucler, 2020; Guo and Perković, 2020) to produce a set of informative estimates.

2 PRELIMINARIES

Throughout the paper we consider a random vector X, indexed by $V = \{V_1, \ldots, V_p\}$, that is $X = X_V$, such that each variable X_{V_i} is represented by node V_i in a graph $\mathcal{G} = (V, E, U)$.

Graphs, nodes and random variables. A partially directed graph $\mathcal{G} = (V, E, U)$ consists of a set of nodes $\mathbf{V} = \{V_1, \ldots, V_p\}$ for $p \geq 1$, a set of directed ($\to$) edges E and a set of undirected ($-$) edges U.

Induced subgraph. An *induced subgraph* $\mathcal{G}_{V'} = (V', E', U')$ of $\mathcal{G} = (V, E, U)$ consists of $V' \subseteq V$, $E' \subseteq E$, and $U' \subseteq U$ where E' and U' are all edges between nodes in V' that are in E and U respectively.

Paths. A *path* $p = \langle V_1, \ldots, V_k \rangle$, $k > 1$ from $V_1 \in A$ to $V_k \in Y$ in $\mathcal{G}$ is a sequence of distinct nodes, such that V_i and V_{i+1}, $i \in \{1, \ldots, k-1\}$ are adjacent in $\mathcal{G}$. A path of the form $V_1 - \cdots - V_k$ is an undirected path and a path of the form $V_1 \to \cdots \to V_k$ is a causal path. Additionally, p is a *possibly causal path* in $\mathcal{G}$ if no edge $V_i \leftarrow V_j, 0 \leq i < j \leq k$ is in $\mathcal{G}$. Otherwise, p is a *non-causal path* in $\mathcal{G}$ (see Definition 3.1 and Lemma 3.2 of Perković et al., 2017). A path from node set A to Y is *proper* with respect to A when only its first node is in A.

Colliders, shields, and definite status paths. If a path p contains $V_i \to V_j \leftarrow V_k$ as a subpath, then V_j is a *collider* on p. A path $\langle V_i, V_j, V_k \rangle$ is an *unshielded triple* if V_i and V_k are not adjacent. A path is *unshielded* if all successive triples on the path are unshielded. A node V_j is a *definite non-collider* on a path p if the edge $V_i \leftarrow V_j$, or the edge $V_j \to V_k$ is on p, or if $V_i - V_j - V_k$ is a subpath of p and V_i is not adjacent to V_k. A node is of *definite status* on a path if it is a collider, a definite non-collider or an endpoint on the path. A path p is of definite status if every node on p is of definite status.

d-connection, d-separation, and blocking. A definite status path p from node A to node Y is *d-connecting* given a node set Z ($A, Y \notin Z$) if every definite non-collider on p is not in Z, and every collider on p has a descendant in Z. Otherwise, Z *blocks* p. If Z blocks all definite status paths between A and Y in MPDAG $\mathcal{G}$, then A is *d-separated* from Y given Z in $\mathcal{G}$ and we write $A \perp_{\mathcal{G}} Y | Z$ (Lemma C.1 of Henckel et al., 2019).

Ancestral relationships. If $A \to Y$ is in $\mathcal{G}$, then A is a *parent* of Y. If there is a causal path from node A to node Y, then A is an *ancestor* of Y, and Y is a *descendant* of A. If there is a possibly causal path from node A to node Y, then Y is a *possible descendant* of A. We use the convention that every node is a descendant, ancestor, and possible descendant of itself. The sets of parents, ancestors, and possible descendants of a node A in $\mathcal{G}$ are denoted by $\mathrm{Pa}(A, \mathcal{G})$, $\mathrm{An}(A, \mathcal{G})$, $\mathrm{PossDe}(A, \mathcal{G})$ respectively. For a set of nodes $A = \{A_1, \ldots, A_k\}$, we let $\mathrm{Pa}(A, \mathcal{G}) = (\cup_{i=1}^k \mathrm{Pa}(A_i, \mathcal{G})) \setminus A$, $\mathrm{An}(A, \mathcal{G}) = \cup_{i=1}^k \mathrm{An}(A_i, \mathcal{G})$, and $\mathrm{PossDe}(A, \mathcal{G}) = \cup_{i=1}^k \mathrm{PossDe}(A_i, \mathcal{G})$.

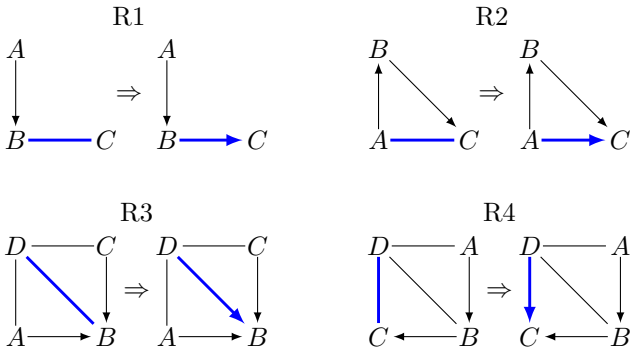

Figure 3: The orientation rules from Meek (1995). If the graph on the left-hand side of a rule is an induced subgraph of a PDAG $\mathcal{G}$, then *orient* the blue undirected edge ($-$) as shown on the right-hand side.

DAGs, PDAGs. A *directed graph* contains only directed edges. A causal path from node A to node Y and $Y \to A$ form a *directed cycle*. A directed graph without directed cycles is a *directed acyclic graph* (DAG). A *partially directed acyclic graph PDAG* is a partially directed graph without directed cycles.

Observational, interventional densities, and causal DAGs. We consider *do-interventions* $\mathrm{do}(X_A = x_a)$ (for $A \subseteq V$), or $\mathrm{do}(x_a)$ for shorthand, which represent outside interventions that set X_a to a fixed value x_a. We call a density f of X under no intervention *an observational density*. An observational density $f(x)$ is *consistent* with a DAG $\mathcal{D} = (V, E, \emptyset)$ if $f(x) = \prod_{i=1}^p f(x_{v_i} | x_{\mathrm{Pa}(v_i, \mathcal{D})})$ (Pearl, 2009).

A density $f(x | \mathrm{do}(x_a))$ under intervention $\mathrm{do}(X_A = x_a)$, $A \subseteq V$ is called an *interventional density*. An interventional density $f(x | \mathrm{do}(x_a))$ is consistent with a DAG $\mathcal{D} = (V, E, \emptyset)$ if there is an observational density f consistent with $\mathcal{D}$ such that

$$f(x | \mathrm{do}(x_a)) = \prod_{\substack{i=1 \\ V_i \notin A}}^{p} f(x_{v_i} | x_{\mathrm{Pa}(v_i, \mathcal{D})}), \qquad (1)$$

for values $x_{\mathrm{Pa}(v_i, \mathcal{D})}$ of $X_{\mathrm{Pa}(V_i, \mathcal{D})}$ that are consistent with x_a. Eq. (1) is known as the truncated factorization formula (Pearl, 2009), manipulated density formula (Spirtes et al., 2000) or the g-formula (Robins, 1986).

A DAG $\mathcal{D} = (V, E, \emptyset)$ is *causal* for a random vector X if the observational and all interventional densities over X are consistent with $\mathcal{D}$.

CPDAGs and MPDAGs. All DAGs that encode the same set of conditional independences are *Markov equivalent* and form a *Markov equivalence class* of DAGs, which can be *represented* by a *completed partially directed acyclic graph* (CPDAG) (Meek, 1995;

Andersson et al., 1997). A PDAG $\mathcal{G}$ is a *maximally oriented* PDAG (MPDAG) if and only if the edge orientations in $\mathcal{G}$ are complete under rules R1-R4 in Figure 3 (Meek, 1995). An MPDAG is also known as CPDAG with background knowledge (Meek, 1995). As such, both a DAG and a CPDAG can be seen as special cases of an MPDAG. Any graph in this paper can hence be labeled an MPDAG.

$\mathcal{G}$ and $[\mathcal{G}]$. A DAG $\mathcal{D} = (V, E, \emptyset)$ is *represented* by MPDAG $\mathcal{G} = (V, E', U')$ if $\mathcal{D}$ and $\mathcal{G}$ have the same adjacencies, same unshielded colliders and if $E' \subseteq E$ (Meek, 1995). If $\mathcal{G}$ is an MPDAG, then $[\mathcal{G}]$ denotes the set of all DAGs represented by $\mathcal{G}$. An MPDAG $\mathcal{G}'$ is said to be *represented* by another MPDAG $\mathcal{G}$ if $[\mathcal{G}'] \subseteq [\mathcal{G}]$.

Causal MPDAGs. An observational or interventional density is consistent with MPDAG $\mathcal{G}$ if it is consistent with a DAG $\mathcal{D}$ in $[\mathcal{G}]$. An MPDAG $\mathcal{G} = (V, E, U)$ is *causal* if it represents the causal DAG.

3 MAIN RESULTS

A total effect of A on Y is generally defined as some functional of the interventional density $f(x_y|\,\mathrm{do}(X_A = x_a))$ (or $f(x_y|\,\mathrm{do}(x_a))$ for short), such as $\mathrm{d}\,\mathbb{E}[X_y|\,\mathrm{do}(x_a)]/\,\mathrm{d}x_a$ for continuous treatments and $\mathbb{E}[X_y|\,\mathrm{do}(x_a = 1)] - \mathbb{E}[X_y|\,\mathrm{do}(x_a = 0)]$ for binary treatments; see, e.g., Hernán and Robins (2020, Ch. 1). For the common definitions, the total effect of A on Y is *identified* in MPDAG $\mathcal{G}$ if and only if $f(x_y|\,\mathrm{do}(x_a))$ can be identified from any observational density $f(x)$ consistent with $\mathcal{G}$ (Galles and Pearl, 1995; Perković, 2020). In this section, we show how to identify all MPDAGs represented by a given MPDAG $\mathcal{G} = (V, E, U)$ that have distinct identification maps for $f(x_Y|\,\mathrm{do}(x_A))$, $A, Y \subseteq V$. Formally, the identification is a map from the space of observational densities that are consistent with $\mathcal{G}$ to the space of conditional kernels:

$$f \in \mathcal{P}(\mathcal{G}) \mapsto f(x_Y|\,\mathrm{do}(x_A)) \in \mathcal{K}(\mathcal{X}_A, \mathcal{X}_Y),$$

where $\mathcal{K}(\mathcal{X}_A, \mathcal{X}_Y)$ is the set of densities on the domain of Y indexed by A. The identification (see Theorem 4) is possible if and only if $\mathcal{G}$ meets the following graphical condition.

Theorem 1 (Identifiability condition of Perković 2020)**.** Let $\mathcal{G} = (V, E, U)$ be a causal MPDAG for a random vector X. Further, let A and Y be disjoint node sets in $\mathcal{G}$. The total effect of A on Y is identified in $\mathcal{G}$ if and only if every proper possibly causal path from A to Y starts with a directed edge in $\mathcal{G}$.

It follows that, if a total effect of A on Y is *not* identified given MPDAG $\mathcal{G}$, then there is at least one proper possible causal path from A to Y in $\mathcal{G}$ that starts with an undirected edge $A_1 - V_1$ for $A_1 \in A$ and $V_1 \in V$. Therefore, to identify the total effect, one can enumerate all the valid combinations of orientations just for the undirected edges of this type. In fact, this is the approach taken by the collapsible-IDA algorithm of Liu et al. (2020), for $|A| = 1$.

As an example, consider MPDAG $\mathcal{G}$ in Fig. 2(a). Paths $A - V_1 - Y$ and $A - V_2 - V_1 - Y$ are two proper possibly causal paths from A to Y in $\mathcal{G}$ that start with an undirected edge. There are four ways to orient the two starting edges:

$$R_1 = \{A \leftarrow V_2, A \rightarrow V_1\}, \quad R_2 = \{A \rightarrow V_2, A \rightarrow V_1\}$$
$$R_3 = \{A \rightarrow V_2, A \leftarrow V_1\}, \quad R_4 = \{A \leftarrow V_2, A \leftarrow V_1\}.$$

Using algorithm `MPDAG`$(\mathcal{G}, R_i)$ for $i = 1, \ldots, 4$ (Meek, 1995; Perković et al., 2017; see Algorithm 1 in the Supplement), which adds orientations R_i and then completes the rules of Meek (1995) in Fig. 3, we obtain four MPDAGs as listed in Fig. 2(c). The effect of A on Y can be identified and estimated under each of the four MPDAGs.

The procedure described above already improves over the current standard IDA and joint-IDA algorithms (Maathuis et al., 2009; Nandy et al., 2017; Perković et al., 2017; Witte et al., 2020; Fang and He, 2020) because orientations of fewer edges are considered — IDA and joint-IDA orient *all* undirected edges connected to A. However, we claim that it suffices to consider even fewer edges. The next theorem characterizes the minimal amount of edge orientation needed to identify a total effect. Its proof is left to the Supplement.

Theorem 2. Let $\mathcal{G} = (V, E, U)$ be a causal MPDAG. Let A and Y be disjoint node sets in $\mathcal{G}$ such that the total effect of A on Y is not identified given $\mathcal{G}$. Suppose $p = \langle A_1, V_1, \ldots, Y_1 \rangle$ for $A_1 \in A$, $Y_1 \in Y$ is a *shortest* proper possibly causal path from A to Y such that $A_1 - V_1$. Then the total effect of A on Y is not identified in any MPDAG $\mathcal{G}^*$ that is represented by $\mathcal{G}$ and contains the undirected edge $A_1 - V_1$.

In the above, we say that $\mathcal{G}^*$ is represented by $\mathcal{G}$ if $[\mathcal{G}^*] \subseteq [\mathcal{G}]$. Implicit in Theorem 2 is the fact that when there are multiple paths that violate the identifiability condition (Theorem 1), the order in which we orient edges matters. In particular, the edge on a shortest path should be oriented first.

Consider again $\mathcal{G}$ in Fig. 2(a). Because $A - V_1 - Y$ is shorter than $A - V_2 - V_1 - Y$, edge $A - V_1$ should be oriented first. As soon as this edge is oriented as $A \leftarrow V_1$, the acyclicity of the underlying DAG renders both paths $A \leftarrow V_1 - Y$ and $A - V_2 - V_1 - Y$ as noncausal from A to Y; see the second row of Fig. 2(d).

This characterization naturally leads to a recursive algorithm `IDGraphs` (Algorithm 1). The `IDGraphs` al-

Algorithm 1: `IDGraphs`

input : MPDAG $\mathcal{G}$, disjoint node sets A and Y
output: the minimal set of MPDAGs with identified effects that partition $\mathcal{G}$

if $\mathcal{G}$ *satisfies the condition in Theorem 1* **then**
  return $\mathcal{G}$;
else
  Let $A_1 - V_1$ be an edge in $\mathcal{G}$ that satisfies Theorem 2;
  $\mathcal{G}_1 = \text{MPDAG}(\mathcal{G}, \{A_1 \to V_1\})$;
  $\mathcal{G}_2 = \text{MPDAG}(\mathcal{G}, \{A_1 \leftarrow V_1\})$;
  return $\{\texttt{IDGraphs}(A, Y, \mathcal{G}_1), \texttt{IDGraphs}(A, Y, \mathcal{G}_2)\}$;
end

gorithm takes MPDAG $\mathcal{G}$ and node sets A, Y as input and outputs a finite set of MPDAGs $\{\mathcal{G}_1, \dots, \mathcal{G}_n\}$ that partition $\mathcal{G}$, such that (i) the total effect of A on Y is identified in every $\mathcal{G}_i$ and (ii) the effects identified from $\mathcal{G}_i$ and $\mathcal{G}_j$ are different for $i \neq j$. Property (ii), formally stated in Theorem 3, shows that the enumeration is minimal. Additionally, by construction, it holds that $n \leq 2^{m(\mathcal{G})}$, where $m(\mathcal{G})$ is the number paths that violate the condition in Theorem 1.

To explain how the algorithm works, consider again the example in Fig. 2(a). As we have already seen, `IDGraphs` first orients edge $A - V_1$ in $\mathcal{G}$. We obtain $\mathcal{G}_1 = \texttt{MPDAG}(\mathcal{G}, \{A_1 \to V_1\})$ and $\mathcal{G}_2 = \texttt{MPDAG}(\mathcal{G}, \{A_1 \leftarrow V_1\})$. Note that the effect is already identified in $\mathcal{G}_2$ (second row of Fig. 2(d)) so $\mathcal{G}_2$ appears in the output. The effect in $\mathcal{G}_1$ is still not identified and the algorithm proceeds to orient edge $A - V_2$, which leads to the other two graphs in the output (first row of Fig. 2(d)).

Note that, however, if $A - V_2$ was oriented before $A - V_1$, we would arrive at four MPDAGs (Fig. 2(c)) instead of three! The next theorem summarizes the theoretical guarantees of `IDGraphs`. We prove the minimality constructively; see the Supplement for details.

Theorem 3. Suppose $\mathcal{G}$ is a causal MPDAG and A, Y are two disjoint node sets in $\mathcal{G}$. Let $L = \{\mathcal{G}_1, \dots, \mathcal{G}_n\}$ be the output of $\texttt{IDGraphs}(A, Y, \mathcal{G})$. Then the following statements hold.

(i) The total effect of A on Y is identified in each $\mathcal{G}_i$.

(ii) For any $i \neq j$, there exists an observational density f that is consistent with $\mathcal{G}$ such that the effect identified from f in $\mathcal{G}_i$ is different from the effect identified from f in $\mathcal{G}_j$.

(iii) L is a partition of $\mathcal{G}$ in terms of DAGs represented.

Proof sketch. Here we sketch out the proof of (ii); see the Supplement for details. For two MPDAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ output by `IDGraphs`, we *construct* a density f that factorizes according to $\mathcal{G}_1$ (and $\mathcal{G}_2$, due to Markov equivalence) but such that $\mathbb{E}[X_Y | \text{do}(X_A = \mathbf{1})]$ has different values under $\mathcal{G}_1$ and $\mathcal{G}_2$.

First, we establish some graphical differences between $\mathcal{G}_1$ and $\mathcal{G}_2$ that stem from the application of Thm 3.2 in the `IDGraphs` algorithm. Consider representing the recursion of `IDGraphs` as a binary tree and let $\mathcal{G}^*$ be the lowest common ancestor of $\mathcal{G}_1$ and $\mathcal{G}_2$. Without loss of generality, suppose $A_1 - V_1 \in \mathcal{G}^*$ but $A_1 \to V_1 \in \mathcal{G}_1$, $A_1 \leftarrow V_1 \in \mathcal{G}_2$. Let p be a shortest possibly causal path from A to Y in $\mathcal{G}^*$ that starts with $A_1 - V_1$. The difference between $\mathcal{G}_1$ and $\mathcal{G}_2$ in terms of p can be categorized into two cases given by Lemma B.1 in the Supplement. For each case, we *parametrize* two linear Gaussian DAGs $\mathcal{D}_1 \in [\mathcal{G}_1]$ and $\mathcal{D}_2 \in [\mathcal{G}_2]$ such that their observed distributions are identical (by matching the first two moments) but values of $\mathbb{E}[X_Y | \text{do}(X_A = \mathbf{1})]$ are different. □

3.1 Examples

Suppose that a total effect of interest is not identified in an MPDAG $\mathcal{G}$. To obtain a set of possible total effects, one needs to enumerate the MPDAGs represented by $\mathcal{G}$ in which the total effect is identified. Below, we consider four methods that have appeared in our discussion so far, listed from the most computationally demanding to the least (see also Table 1).

Method 1 List all DAGs represented by $\mathcal{G}$. This is adopted by the global IDA algorithm of Maathuis et al. (2009).

Method 2 List MPDAGs corresponding to all valid orientations for undirected edges attached to A. This is adopted by the local/semi-local IDA algorithms of Maathuis et al. (2009); Perković et al. (2017); Witte et al. (2020); Fang and He (2020) and the joint-IDA algorithm of Nandy et al. (2017).

Method 3 List MPDAGs corresponding to all valid combinations of edge orientations for edges $A_1 - V_1$, $A_1 \in A$, $V_1 \in V$, such that V_1 is on a proper possibly causal path from A to Y in $\mathcal{G}$. For $|A| = 1$, this is adopted by the collapsible-IDA algorithm of Liu et al. (2020).

Method 4 Use $\texttt{IDGraphs}(A, Y, \mathcal{G})$.

These methods are compared through two examples, one for point intervention ($|A| = 1$) and one for joint intervention ($|A| = 2$).

Example 1 ($|A| = 1$)**.** Consider again $\mathcal{G}$ in Fig. 2(a). Recall that **Method 1** lists 7 DAGs shown in Fig. 2(b) and **Method 4** lists 3 MPDAGs shown in Fig. 2(d). And as discussed earlier, both **Method 2** and **Method 3** would orient edges $A - V_2$ and $A - V_1$, and hence list 4 MPDAGs shown in Fig. 2(d).

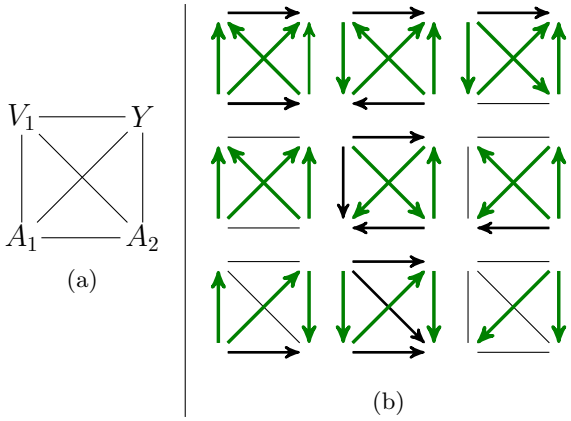

Figure 4: Example 2. (a) MPDAG $\mathcal{G}$, (b) all MPDAGs with distinct causal identification formulas for $f(x_Y|\,\mathrm{do}(x_{A_1}, x_{A_2}))$ represented by $\mathcal{G}$.

Example 2 ($|A| = 2$)**.** Consider MPDAG $\mathcal{G}$ in Fig. 4(a) and let $A = \{A_1, A_2\}$. **Method 4** yields 9 MPDAGs listed in Fig. 4(b), which correspond to 9 distinct possible effects of (A_1, A_2) on Y. **Method 2** would consider all valid combinations of orientations for edges $A_1 - A_2$, $A_1 - V_1$, $A_1 - Y$, $A_2 - V_1$, and $A_2 - Y$, resulting in 18 MPDAGs. Lastly, **Method 3** would consider all valid combinations of orientations for edges $A_1 - V_1$, $A_1 - Y$, $A_2 - V_1$, and $A_2 - Y$, resulting in 12 MPDAGs.

3.2 Computational Complexity

The computational complexities of various algorithms are summarized in Table 1. Here $l(\mathcal{G})$ is the number of undirected edges incident to A. For the run time of collapsible IDA (Liu et al., 2020), $O(|V| + |E|)$ reflects finding the neighbors of A that are on a possible causal path to Y and $r(\mathcal{G})$ is the size of that subset. Both local IDA and collapsible IDA are only applicable when $|A| = 1$.

Table 1: Computational complexity

local IDA	$O(2^{l(\mathcal{G})})$
collapsible IDA	$O((\lvert V\rvert + \lvert E\rvert)2^{r(\mathcal{G})})$
semi-local/joint/optimal IDA	$O(2^{l(\mathcal{G})})\mathrm{poly}(\lvert V\rvert)$
`IDGraphs`	$O(2^{m(\mathcal{G})})\mathrm{poly}(\lvert V\rvert)$

For more general settings, our `IDGraphs` algorithm is asymptotically on par with semi-local (Maathuis et al., 2009; Perković et al., 2017), joint (Nandy et al., 2017), and optimal (Witte et al., 2020) variants of the IDA algorithm. The run time of these methods is bounded by $O(2^{l(\mathcal{G})})\mathrm{poly}(|V|)$, where $\mathrm{poly}(|V|)$ time is used to complete the orientation rules of Meek (1995). Similarly, the complexity of `IDGraphs` is $O(2^{m(\mathcal{G})})\mathrm{poly}(|V|)$, where $m(\mathcal{G})$ is the number of undirected edges incident to A on a proper possibly causal path from A to Y (Theorem 1). Clearly, we have $m(\mathcal{G}) \leq l(\mathcal{G})$. The number of recursions is bounded by $2^{m(\mathcal{G})}$ and the time for each recursion by $\mathrm{poly}(|V|)$, which includes completing the orientation rules, checking the condition of Theorem 1, and identifying the shortest path (Theorem 2) if the condition is not met.

In practice, from the simulations in Section 4, we find that `IDGraphs` roughly costs twice the time of IDA type algorithms; see Fig. 5.

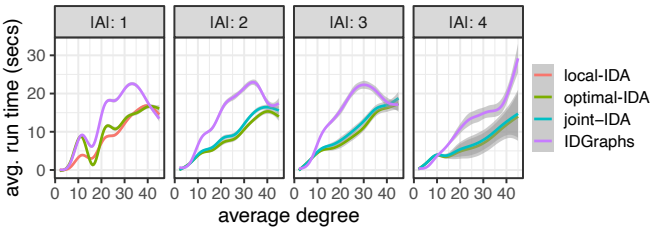

Figure 5: Average run time versus the average node degree of the graph (shade: 95% C.I.).

3.3 Corollaries of the Main Result

Having obtained a set of MPDAGs from the `IDGraphs` algorithm, the total effect can be estimated for each graph. Some recently developed efficient estimators can be employed, including the semiparametric efficient estimator of Rotnitzky and Smucler (2020) and the efficient least-squares estimator of Guo and Perković (2020) under linearity assumptions. This strategy applies to any other causal quantity that is a functional of the interventional density.

Theorem 4 (Causal identification formula, Perković, 2020)**.** Let $\mathcal{G}$ be a causal MPDAG. Suppose $\mathcal{G}$ satisfies the identifiability condition in terms of the total effect of A on Y (Theorem 1). Further, let $B = \mathrm{An}(Y, \mathcal{G}_{V\setminus A}) \setminus Y$ and let $B_1, \ldots, B_k$ be the bucket decomposition of $B \cup Y$ (see Supplement A for its definition). Then for any observational density f that is consistent with $\mathcal{G}$, we have

$$f(x_y|\,\mathrm{do}(x_a)) = \int \prod_{i=1}^{k} f(x_{b_i}|x_{\mathrm{Pa}(b_i,\mathcal{G})})\,\mathrm{d}x_b, \qquad (2)$$

where values of $x_{\mathrm{Pa}(b_i,\mathcal{G})}$ are consistent with x_a.

Let L be the output of `IDGraphs`$(A, Y, \mathcal{G})$.

Corollary 1. There are no two graphs in L that share the same formula Eq. (2).

Another prominent method for identifying the interventional density is through covariate adjustment. The IDA algorithms of Maathuis et al. (2009); Perković et al. (2017); Witte et al. (2020) are based on covariate adjustment for causal linear models. See Perković (2020) for the generalized adjustment criterion, which is necessary and sufficient for covariate adjustment in MPDAGs and generalizes the well-known back-door formula of Pearl (1993).

Corollary 2. There are no two MPDAGs in L that share the same adjustment set relative to (A, Y). Further, if $|A| = |Y| = 1$, then there exists an adjustment set relative to (A, Y) for each MPDAG in L.

A consequence of Corollary 2 and Henckel et al. (2019, Lemma E.7) for $|A| = 1$ is that the optimal IDA algorithm of Witte et al. (2020) returns a minimal set of possible effects.

4 NUMERICAL RESULTS

We present numerical results on estimating possible total effects under a linear causal model (Bollen, 1989). Consider DAG $\mathcal{D}$ in Fig. 7(a) and suppose data is generated using the following causal linear model

$$\begin{aligned} X_{A_1} &= \varepsilon_1, \quad X_{A_2} = \gamma_{12} X_{A_1} + \varepsilon_2, \\ X_V &= \gamma_{13} X_{A_1} + \gamma_{23} X_{A_2} + \varepsilon_3, \\ X_Y &= \gamma_{14} X_{A_1} + \gamma_{24} X_{A_2} + \varepsilon_4. \end{aligned}$$

We further set $\gamma_{14} = \gamma_{23} = 2$ and $\gamma_{12} = \gamma_{13} = \gamma_{24} = 1$. The errors ε_i for $i = 1, \dots, 4$ are drawn independently from $\mathcal{N}(0, 1)$. Suppose the causal DAG is known up to its CPDAG $\mathcal{G}$ (no added background knowledge), which is shown in (b). We consider estimating the total effect of A_1 on Y (point intervention), and the total effect of (A_1, A_2) on Y (joint intervention).

Table 2 shows the estimates from 100 samples, where "our method" refers to applying the efficient estimator of Guo and Perković (2020) to each graph returned by `IDGraphs`. We compare against available IDA methods in the R package `pcalg`. The IDA algorithms enumerate possible graphs where the effect is identified and return the estimates as a multiset. The distinct values of the multiset can be taken as the estimates of possible effects. However, as we can see, one possible effect can correspond to more than one distinct values due to sampling variability. Moreover, `NA`'s are produced when applying IDA (optimal) to joint interventions due to the nonexistence of valid adjustment sets (Perković et al., 2018).

To examine these issues in more generality, we simulate random instances and compare the size of estimates to the true number of possible effects. Causal DAG $\mathcal{D}$ is generated by sampling from the Erdős-Rényi model and assigning a random causal ordering. We consider graphs of size $p = 10$ and $p = 50$, where the average degree k is drawn from $\{2, \dots, 8\}$ for the former and $\{2, \dots, 45\}$ for the latter. We take $\mathcal{G}$ to be the CPDAG of $\mathcal{D}$. Treatment variables A and outcome Y are randomly selected such that the total effect of A on Y is unidentified in $\mathcal{G}$. The size of A varies from 1 to 4. For each instance, the possible effects of A on Y are estimated given $\mathcal{G}$ and 500 independent samples generated by a corresponding linear causal model (with random coefficients and errors drawn from $\mathcal{N}(0, 1)$).

The result is summarized in Fig. 6 from roughly 55,000 random instances. For point interventions (left panel), IDA (local) produces duplicates, i.e., more distinct values than the actual number of possible effects, especially when the number of possible effects is small, whereas IDA (optimal) returns the correct number of distinct values. For joint interventions (right panel), the joint-IDA algorithm often suffers from an excessive amount of duplicates, while IDA (optimal), on the other hand, severely under reports the size of possible effects due to too many `NA`'s it produced — note the logarithmic scale of both axes. Therefore, our `IDGraphs` algorithm, in conjunction with a statistically efficient estimator of an identified effect, should be used in place of IDA algorithms in both cases to avoid unnecessary computational overhead and deliver causally informative estimates.

5 DISCUSSION

We have studied the set-identification of a total effect given that the underlying causal DAG is known up to a Markov equivalence class or its refinement. Existing enumerative approaches to this problem are often not minimal, which causes unnecessary computational overheads and undesirable statistical duplicates. We ensure minimal set-identification by focusing on two key steps. First, we use Theorem 1 to locate the set of "problematic" undirected edges. These are undirected edges that need to be oriented in order to identify the causal effect of interest. The second key step is to determine the order in which the "problematic" edges need to be oriented. Perhaps surprisingly, an optimal order can be determined, which is to first orient a "problematic" edge on a shortest proper possibly causal path from A to Y (Theorem 2). This naturally leads to `IDGraphs`, a simple recursive algorithm that applies these two steps and guarantees minimal enumeration (Theorem 3).

Table 2: Estimates of possible effects for the example in Fig. 7. Symbol $(a)^b$ denotes value a with multiplicity b in the multiset. Note that IDA (local) and joint-IDA return more values than the number of possible effects.

	A_1 on Y (Fig. 7(c))	A_1, A_2 on Y (Fig. 7(d))
true effect	3	(2,1)
true possible effects	{3, 2, 1.8, 0}	{(2, 1), (3, 0), (0, 2), (0, 0)}
our method	{2.9, 2.1, 1.9, 0}	{(2.1, 0.9), (2.9, 0), (0, 1.9), (0, 0)}
IDA (optimal) (Witte et al., 2020)	$\{2.9, (2.1)^2, 1.9, 0\}$	$\{(2.1, 0.9)^6, (0, 0)^2, (\texttt{NA}, \texttt{NA})^2\}$
IDA (local) (Maathuis et al., 2009)	{2.9, 2.1, 2.2, 1.9, 0}	—
joint-IDA (Nandy et al., 2017)	—	$\{(2.1, 0.9)^2, (2.2, 0.9), (1.9, 1.1), (2.2, 1.1)^2, (0, 1.9), (2.9, 0), (0, 0)^2\}$

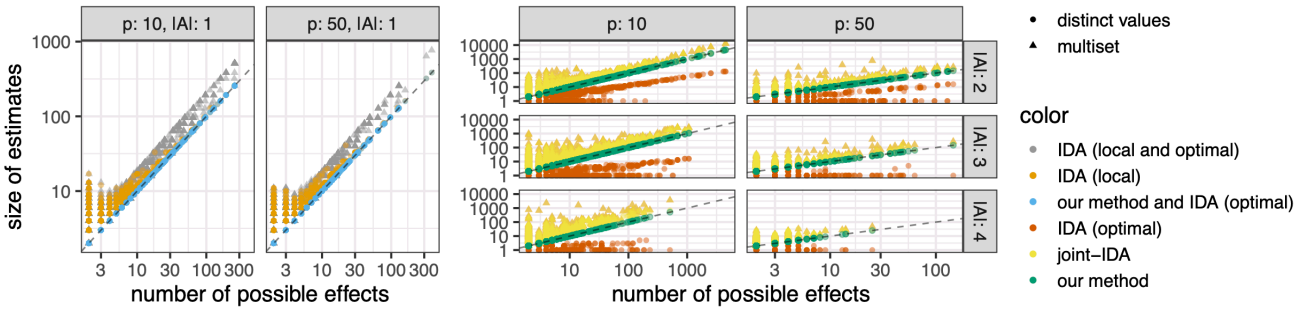

Figure 6: Size of estimates vs. the number of possible effects on random instances (left: $|A| = 1$, right: $|A| > 1$). Both axes are in logarithmic scale. A dot on the graph represents either the size of the multiset (▲) or the number of distinct values (●).

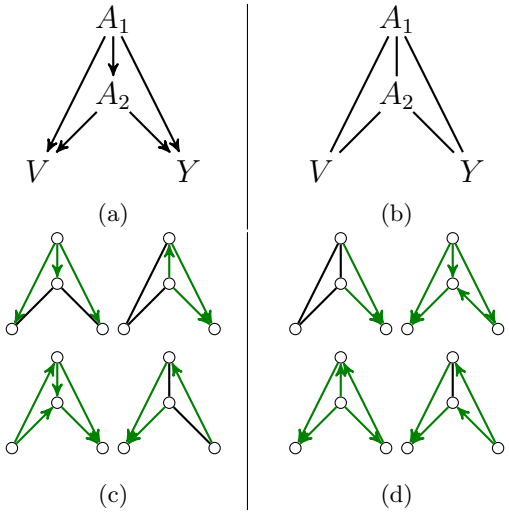

Figure 7: Estimating possible total effects: (a) the underlying causal DAG $\mathcal{D}$, (b) CPDAG $\mathcal{G}$ of $\mathcal{D}$, (c) $\texttt{IDGraphs}(A_1, Y, \mathcal{G})$, (d) $\texttt{IDGraphs}(\{A_1, A_2\}, Y, \mathcal{G})$.

`IDGraphs` can be readily used in conjunction with recent developments in efficient estimation (Henckel et al., 2019; Rotnitzky and Smucler, 2020; Guo and Perković, 2020) to deliver informative estimates of the true causal effect. From this perspective, our result can be viewed as *separating* two sources of uncertainty — identification and estimation — the two crucial steps in causal inference.

An advantage of the local IDA algorithms over our approach is that they only require knowledge of the parent set (or parent sets) of A instead of the whole MPDAG. Yet, it is unclear whether there are many cases where one knows the neighborhood of A without knowing more structure.

Finally, one may wonder whether this approach can be extended to allow latent variables. The latent variable IDA (LV-IDA) algorithm of Malinsky and Spirtes (2017) employs a strategy similar to the IDA algorithm given a partial ancestral graph (PAG). A PAG represents a Markov equivalence class of maximal ancestral graphs (MAGs), and each MAG can be obtained from a DAG by marginalizing out latent variables. One obstacle to applying an approach like ours in this setting is that it is unknown how to incorporate background knowledge of edge orientations into a PAG. In addition, Jaber et al. (2019) recently showed that if an effect is not identified given a PAG, then there is at least one MAG in the Markov equivalence class in which the effect is still not identified (see their Theorem 4) — the same enumeration strategy will no longer work.

Acknowledgment FRG acknowledges the support from ONR Grant N000141912446.

References

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25:505–541.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley, New York.

Chickering, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, 2:445–498.

Fang, Z. and He, Y. (2020). IDA with background knowledge. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 270–279.

Galles, D. and Pearl, J. (1995). Testing identifiability of causal effects. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 185–195.

Gillispie, S. B. and Perlman, M. D. (2002). The size distribution for Markov equivalence classes of acyclic digraph models. *Artificial Intelligence*, 141(1-2):137–155.

Guo, F. R. and Perković, E. (2020). Efficient least squares for estimating total effects under linearity and causal sufficiency. *arXiv preprint arXiv:2008.03481*.

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

Henckel, L., Perković, E., and Maathuis, M. H. (2019). Graphical criteria for efficient total effect estimation via adjustment in causal linear models. arXiv preprint arXiv:1907.02435.

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What if*. Boca Raton: Chapman & Hall/CRC.

Hoyer, P. O., Hyvarinen, A., Scheines, R., Spirtes, P. L., Ramsey, J., Lacerda, G., and Shimizu, S. (2008). Causal discovery of linear acyclic models with arbitrary distributions. In *Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, pages 282–289.

Jaber, A., Zhang, J., and Bareinboim, E. (2019). Causal identification under Markov equivalence: completeness results. In *Proceedings of the International Conference on Machine Learning (ICML-19)*, volume 97, pages 2981–2989.

Liu, Y., Fang, Z., He, Y., and Geng, Z. (2020). Collapsible IDA: Collapsing parental sets for locally estimating possible causal effects. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 290–299.

Maathuis, M. H., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164.

Malinsky, D. and Spirtes, P. (2017). Estimating bounds on causal effects in high-dimensional and possibly confounded systems. *International Journal of Approximate Reasoning*.

Meek, C. (1995). Causal inference and causal explanation with background knowledge. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 403–410.

Nandy, P., Maathuis, M. H., and Richardson, T. S. (2017). Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *The Annals of Statistics*, 45(2):647–674.

Pearl, J. (1993). Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2009). *Causality*. Cambridge University Press, Cambridge, 2nd edition.

Pearl, J. and Robins, J. M. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 444–453.

Perković, E. (2020). Identifying causal effects in maximally oriented partially directed acyclic graphs. In *Proceedings of the 36th Annual Conference on Uncertainty in Artificial Intelligence (UAI-20)*.

Perković, E., Kalisch, M., and Maathuis, M. H. (2017). Interpreting and using CPDAGs with background knowledge. In *Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence (UAI-17)*.

Perković, E., Textor, J., Kalisch, M., and Maathuis, M. H. (2018). Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18(220):1–62.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512.

Rothenhäusler, D., Ernest, J., and Bühlmann, P. (2018). Causal inference in partially linear structural equation models: identifiability and estimation. *Annals of Statistics*, 46:2904–2938.

Rotnitzky, A. and Smucler, E. (2020). Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21(188):1–86.

Scheines, R., Spirtes, P., Glymour, C., Meek, C., and Richardson, T. (1998). The TETRAD project: constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition.

Steinsky, B. (2013). Enumeration of labelled essential graphs. *Ars Combinatoria*, 111:485–494.

Wang, Y., Solus, L., Yang, K. D., and Uhler, C. (2017). Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems 30*, pages 5822–5831.

Witte, J., Henckel, L., Maathuis, M. H., and Didelez, V. (2020). On efficient adjustment in causal graphs. *Journal of Machine Learning Research*, 21(246):1–45.