Finding and Listing Front-door Adjustment Sets

Anonymous Author(s)
Affiliation
Address
email

Abstract

Identifying the effects of new interventions from data is a significant challenge found across a wide range of the empirical sciences. A well-known strategy for identifying such effects is Pearl's *front-door (FD) criterion* [15]. The definition of the FD criterion is declarative, only allowing one to decide whether a specific set satisfies the criterion. In this paper, we present algorithms for finding and enumerating possible sets satisfying the FD criterion in a given causal diagram. These results are useful in facilitating the practical applications of the FD criterion for causal effects estimation and helping scientists to select estimands with desired properties, e.g., based on cost, feasibility of measurement, or statistical power.

1 Introduction

Learning cause and effect relationships is a fundamental challenge across data-driven fields. For example, health scientists developing a treatment for curing lung cancer need to understand how a new drug affects the patient's body and the tumor's progression. The distillation of causal relations is indispensable to understanding the dynamics of the underlying system and how to perform decision-making in a principled and systematic fashion [16, 25, 1, 19].

One of the most common methods for learning causal relations is through *Randomized Controlled Trials* (RCTs, for short) [3]. RCTs are considered as the "gold standard" in many fields of empirical research and are used throughout the health and social sciences as well as machine learning and AI. In practice, however, RCTs are often hard to perform due to ethical, financial, and technical issues. For instance, it may be unethical to submit an individual to a certain condition if such condition may have some potentially negative effects (e.g., smoking). Whenever RCTs cannot be conducted, one needs to resort to analytical methods to infer causal relations from observational data, which appears in the literature as the problem of *causal effect identification* [15, 16].

The causal identification problem asks whether the effect of holding a variable X at a constant value x on a variable Y, written as $P(Y|do(X = x))$, or $P(Y|do(x))$, can be computed from a combination of observational data and causal assumptions. One of the most common ways of eliciting these assumptions is in the form of a causal diagram represented by a directed acyclic graph (DAG), where its nodes and edges describe the underlying data generating process. For instance, in Fig. 1a, three nodes X, Z, Y represent variables, a directed edge $X \rightarrow Z$ indicates that X causes Z, and a dashed-bidirected edge $X \leftrightarrow Y$ represents that X and Y are confounded by unmeasured (latent) factors. Different methods can solve the identification problem, including Pearl's celebrated do-calculus [15] as well as different algorithmic solutions [28, 22, 7].

In practice, researchers often rely on identification strategies that generate well-known identification formulas. One of the arguably most popular strategies is identification by covariate adjustment. Whenever a set Z satisfies the *back-door (BD) criterion* [15] relative to the pair X and Y, where X and Y represent the treatment and outcome variables, respectively, the causal effect $P(Y|do(x))$ can be evaluated through the BD adjustment formula $\sum_z P(y|x, z)P(z)$.

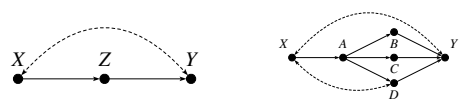
Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Despite the popularity of the covariate adjustment technique for estimating causal effects, there are still settings in which no BD admissible set exists. For example, consider the causal diagram $\mathcal{G}$ in Fig. 1a. There clearly exists no set to block the BD path from X to Y, through the bidirected arrow, $X \leftrightarrow Y$. One may surmise that this effect is not identifiable and the only one of evaluating the interventional distribution is through experimentation. Still, this is not the case. The effect $P(Y|do(x))$ is identifiable from $\mathcal{G}$ and the observed distribution $P(x, y, z)$ over $\{X, Y, Z\}$ by another classic identification strategy known as the *front-door (FD) criterion* [15]. In particular, through the following FD adjustment formula provides the way of evaluating the interventional distribution:

$$P(Y|do(x)) = \sum_{z} P(z|x) \sum_{x'} P(y|x', z)P(x'). \tag{1}$$

We refer to Pearl and Mackenzie [17, Sec. 3.4] for an interesting account of the history of the FD criterion, which was the first graphical generalization of the BD case. The FD criterion is drawing more attention in recent years. For applications of the FD criterion, see, e.g., Hünermund and Bareinboim [8] and Glynn and Kashin [5]. Statistically efficient and doubly robust estimators have recently been developed for estimating the FD estimand in Eq. (1) from finite samples [4], which are still elusive for arbitrary estimands identifiable in a diagram despite recent progress [9, 10, 2, 11].

Both the BD and FD criteria are only descriptive, i.e., they specify whether a specific set Z satisfies the criteria or not, but do not provide a way to find an admissible set Z. In addition, in many situations, it is possible that multiple adjustment sets exist. Consider for example the causal diagram in Fig. 1b, and the task of identifying the effect of X on Y. The distribution $P(Y|do(x))$ can indeed be identified by the FD criterion with a set $Z = \{A, B, C\}$ given by the expression in Eq. (1) (with Z replaced with $\{A, B, C\}$). Still, what if the variable B is costly to measure or encodes some personal information about patients which is undesirable to be shared due to ethical concerns? In this case, the set $Z = \{A, C\}$ also satisfies the FD criterion and may be used. Even when both B and C are unmeasured, the set $Z = \{A\}$ is also FD admissible.



(a) $\mathcal{G}$ (b) $\mathcal{G}'$

Figure 1: (a) A canonical example of the FD criterion where $\{Z\}$ satisfies the FD criterion relative to $(\{X\}, \{Y\})$. In (b), four FD adjustment sets relative to $(\{X\}, \{Y\})$ are available: $\{A\}$, $\{A, B\}$, $\{A, C\}$, and $\{A, B, C\}$.

This simple example shows that a target effect can be estimated using different adjustment sets leading to different probability expressions over different set of variables, which has important practical implications. Each variable implies different practical challenges in terms of measurement, such as cost, availability, privacy. Each estimand has different statistical properties in terms of sample complexity, variance, which may play a key role in the study design [20, 6, 21, 24]. Algorithms for finding and listing all possible adjustment sets are hence very useful in practice, which will allow scientists to select an adjustment set that exhibits desirable properties. Indeed, algorithms have been developed in recent years for finding one or listing all BD admissible sets [26, 27, 29, 18, 30]. However, no such algorithm is currently available for finding/listing FD admissible sets.

The goal of this paper is to close this gap to facilitate the practical applications of the FD criterion for causal effects estimation and help scientists to select estimand with certain desired properties. Specifically, the contributions of this paper are as follows:

1. We develop an algorithm that finds an admissible front-door adjustment set $\mathbf{Z}$ in a given causal diagram in polynomial time (if one exists). We solve a variant of the problem that imposes constraints $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$ for given sets $\mathbf{I}$ and $\mathbf{R}$, which allows a scientist to constrain the search to include specific subsets of variables or exclude variables from search perhaps due to cost, availability, or other technical considerations.

2. We develop a sound and complete algorithm that enumerates all front-door adjustment sets with polynomial delay - the algorithm takes polynomial amount of time to return each new admissible set, if one exists, or return failure whenever it exhausted all admissible sets.

2 Preliminaries

Notation. We write a variable in capital letters (X) and its value as small letters (x). Bold letters, $\mathbf{X}$ or $\mathbf{x}$, represent a set of variables or values. We use kinship terminology to denote various relationships in a graph $\mathcal{G}$ and denote the parents, ancestors, and descendants of $\mathbf{X}$ (including $\mathbf{X}$ itself) as $Pa(\mathbf{X})$, $An(\mathbf{X})$, and $De(\mathbf{X})$, respectively. Given a graph $\mathcal{G}$ over a set of variables $\mathbf{V}$, a subgraph $\mathcal{G}_{\mathbf{X}}$ consists of a subset of variables $\mathbf{X} \subseteq \mathbf{V}$ and their incident edges in $\mathcal{G}$. A graph $\mathcal{G}$ can be transformed: $\mathcal{G}_{\overline{\mathbf{X}}}$ is the graph resulting from removing all incoming edges to $\mathbf{X}$, and $\mathcal{G}_{\underline{\mathbf{X}}}$ is the graph with all outgoing edges from $\mathbf{X}$ removed. A DAG $\mathcal{G}$ may be *moralized* into an undirected graph where all directed edges of $\mathcal{G}$ are converted into undirected edges, and for every pair of nonadjacent nodes in $\mathcal{G}$ that share a common child, an undirected edge that connects such pair is added [13].

A path π from a node X to a node Y in $\mathcal{G}$ is a sequence of edges where X and Y are the endpoints of π. A node W on π is said to be a collider if W has converging arrows into W in π, e.g., $\rightarrow W \leftarrow$ or $\leftrightarrow W \leftarrow$. π is said to be blocked by a set $\mathbf{Z}$ if there exists a node W on π satisfying one of the following two conditions: 1) W is a collider, and neither W nor any of its descendants are in $\mathbf{Z}$, or 2) W is not a collider, and W is in $\mathbf{Z}$ [14]. Given three disjoint sets $\mathbf{X}, \mathbf{Y}$, and $\mathbf{Z}$ in $\mathcal{G}$, $\mathbf{Z}$ is said to d-separate $\mathbf{X}$ from $\mathbf{Y}$ in $\mathcal{G}$ if and only if $\mathbf{Z}$ blocks every path from a node in $\mathbf{X}$ to a node in $\mathbf{Y}$ according to the d-separation criterion [14], and we say that $\mathbf{Z}$ is a *separator* of $\mathbf{X}$ and $\mathbf{Y}$ in $\mathcal{G}$.

Structural Causal Models (SCMs). We use Structural Causal Models (SCMs, for short) [16] as our basic semantical framework. An SCM is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{u}) \rangle$, where 1) $\mathbf{U}$ is a set of exogenous (latent) variables, 2) $\mathbf{V}$ is a set of endogenous (observed) variables, 3) $\mathbf{F}$ is a set of functions $\{f_V\}_{V \in \mathbf{V}}$ that determine the value of endogenous variables, e.g., $v \leftarrow f_V(\mathbf{pa}_V, \mathbf{u}_V)$ is a function with $\mathbf{PA}_V \subseteq \mathbf{V} \setminus \{V\}$ and $\mathbf{U}_V \subseteq \mathbf{U}$, and 4) $P(\mathbf{u})$ is a joint distribution over the exogenous variables $\mathbf{U}$. Each SCM induces a *causal diagram* $\mathcal{G}$ where every variable $v \in \mathbf{V}$ is a vertex and directed edges in $\mathcal{G}$ correspond to functional relationships as specified in $\mathbf{F}$ and dashed bidirected edges represent common exogenous variables between two vertices. Within the structural semantics, performing an intervention and setting $X = x$ is represented through the do-operator, $do(X = x)$, which encodes the operation of replacing the original functions of X (i.e., $f_X(\mathbf{pa}_X, \mathbf{u}_X)$) by the constant x and induces a submodel $\mathcal{M}_x$ and an interventional distribution $P(v|do(x))$.

Classic Causal Effects Identification Criteria. Given a causal diagram $\mathcal{G}$ over $\mathbf{V}$, an effect $P(\mathbf{y}|do(\mathbf{x}))$ is said to be *identifiable* in $\mathcal{G}$ if $P(\mathbf{y}|do(\mathbf{x}))$ is uniquely computable from the observed distribution $P(\mathbf{v})$ in any SCM that induces $\mathcal{G}$ [16, p. 77].

A path between X and Y with an arrow into X is known as a *back-door* path from X to Y. The celebrated back-door (BD) criterion [15] provides a sufficient condition for effect identification from observational data, which states that if a set $\mathbf{Z}$ of non-descendants of $\mathbf{X}$ blocks all BD paths from $\mathbf{X}$ to $\mathbf{Y}$, then the causal effect $P(\mathbf{y}|do(\mathbf{x}))$ is identified by the BD adjustment formula:

$$P(\mathbf{y}|do(\mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{y}|\mathbf{x}, \mathbf{z}) P(\mathbf{z}) \tag{2}$$

Another classic identification condition that is key to the discussion in this paper is known as the front-door criterion, which is defined as follows:

Definition 1. (Front-door (FD) Criterion [15]) A set of variables $\mathbf{Z}$ is said to satisfy the front-door criterion relative to the pair $(\mathbf{X}, \mathbf{Y})$ if

1. $\mathbf{Z}$ intercepts all directed paths from $\mathbf{X}$ to $\mathbf{Y}$,
2. There is no back-door path from $\mathbf{X}$ to $\mathbf{Z}$, and
3. All back-door paths from $\mathbf{Z}$ to $\mathbf{Y}$ are blocked by $\mathbf{X}$, i.e., $\mathbf{X}$ is a separator of $\mathbf{Z}$ and $\mathbf{Y}$ in $\mathcal{G}_{\underline{\mathbf{Z}}}$.

If $\mathbf{Z}$ satisfies the FD criterion relative to the pair $(\mathbf{X}, \mathbf{Y})$, then $P(\mathbf{y}|do(\mathbf{x}))$ is identified by the following FD adjustment formula [15]:

$$P(\mathbf{y}|do(\mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}) \sum_{\mathbf{x}'} P(\mathbf{y}|\mathbf{x}', \mathbf{z}) P(\mathbf{x}'). \tag{3}$$

3 Finding A Front-door Adjustment Set

In this section, we address the following question: given a causal diagram $\mathcal{G}$, is there a set $\mathbf{Z}$ that satisfies the FD criterion relative to the pair $(\mathbf{X}, \mathbf{Y})$ and, therefore, allows us to identify $P(\mathbf{y}|do(\mathbf{x}))$ by the FD adjustment? We solve a more general variant of this question that imposes a constraint $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$ for given sets $\mathbf{I}$ and $\mathbf{R}$. Here, $\mathbf{I}$ are variables that must be included in $\mathbf{Z}$ ($\mathbf{I}$ could be empty) and $\mathbf{R}$ are variables that could be included in $\mathbf{Z}$ ($\mathbf{R}$ could be $\mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$). Note the constraint that variables in $\mathbf{W}$ *cannot* be included can be enforced by excluding $\mathbf{W}$ from $\mathbf{R}$. Solving this version of the problem will allow scientists to put constraints on candidate adjustment sets based on practical considerations. In addition, this version will form a building block for an algorithm that enumerates all FD admissible sets in a given $\mathcal{G}$ - the algorithm LISTFDSETS (shown in Alg. 2 in Section 4) for listing all FD admissible sets will utilize this result during the recursive call.

Algorithm 1 FINDFDSET $(\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R})$

Input: $\mathcal{G}$ a causal diagram; $\mathbf{X}, \mathbf{Y}$ disjoint sets of variables; $\mathbf{I}, \mathbf{R}$ sets of variables.
Output: $\mathbf{Z}$ a set of variables satisfying the front-door criterion relative to $(\mathbf{X}, \mathbf{Y})$ with the constraint $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$.
Step 1:
    $\mathbf{R}' \leftarrow$ GETCAND2NDFDC$(\mathcal{G}, \mathbf{X}, \mathbf{I}, \mathbf{R})$
    if $\mathbf{R}' = \emptyset$ **then: return** $\perp$
Step 2:
    $\mathbf{R}'' \leftarrow$ GETCAND3RDFDC$(\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R}')$
    if $\mathbf{R}'' = \emptyset$ **then: return** $\perp$
Step 3:
    $\mathcal{G}' \leftarrow$ GETCAUSALPATHGRAPH$(\mathcal{G}, \mathbf{X}, \mathbf{Y})$
    if TESTSEP$(\mathcal{G}', \mathbf{X}, \mathbf{Y}, \mathbf{R}'')$ = True **then:**
        return $\mathbf{Z} = \mathbf{R}''$
    else: return $\perp$

We have developed a procedure called FINDFDSET shown in Alg. 1 that outputs a FD adjustment set $\mathbf{Z}$ relative to $(\mathbf{X}, \mathbf{Y})$ satisfying $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$, or outputs $\perp$ if none exists, given a causal diagram $\mathcal{G}$, disjoint sets of variables $\mathbf{X}$ and $\mathbf{Y}$, and two sets of variables $\mathbf{I}$ and $\mathbf{R}$.

Example 1. Consider the causal graph $\mathcal{G}'$, shown in Fig. 1b, with $\mathbf{X} = \{X\}$, $\mathbf{Y} = \{Y\}$, $\mathbf{I} = \emptyset$ and $\mathbf{R} = \{A, B, C, D\}$. Then, FINDFDSET outputs $\{A, B, C\}$. With $\mathbf{I} = \{C\}$ and $\mathbf{R} = \{A, C\}$, FINDFDSET outputs $\{A, C\}$. With $\mathbf{I} = \{D\}$ and $\mathbf{R} = \{A, B, C, D\}$, FINDFDSET outputs $\perp$ as no FD adjustment set that contains D is available.

FINDFDSET runs in three major steps. Each step identifies candidate variables that incrementally satisfy each of the conditions of the FD criterion relative to $(\mathbf{X}, \mathbf{Y})$. First, FINDFDSET constructs a set of candidate variables $\mathbf{R}'$, with $\mathbf{I} \subseteq \mathbf{R}' \subseteq \mathbf{R}$, such that every subset $\mathbf{Z}$ with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}'$ satisfies the second condition of the FD criterion (i.e., there is no BD path from $\mathbf{X}$ to $\mathbf{Z}$). Next, FINDFDSET generates a set of candidate variables $\mathbf{R}''$, with $\mathbf{I} \subseteq \mathbf{R}'' \subseteq \mathbf{R}'$, such that for every variable $v \in \mathbf{R}''$, there exists a set $\mathbf{Z}$ with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}'$ and $v \in \mathbf{Z}$ that further satisfies the third condition of the FD criterion, that is, all BD paths from $\mathbf{Z}$ to $\mathbf{Y}$ are blocked by $\mathbf{X}$. Finally, FINDFDSET outputs a set $\mathbf{Z}$ that further satisfies the first condition of the FD criterion - $\mathbf{Z}$ intercepts all causal paths from $\mathbf{X}$ to $\mathbf{Y}$.

function GETCAND2NDFDC$(\mathcal{G}, \mathbf{X}, \mathbf{I}, \mathbf{R})$
    Output: $\mathbf{R}'$ with $\mathbf{I} \subseteq \mathbf{R}' \subseteq \mathbf{R}$, the set of candidate variables consisting of all the variables $v \in \mathbf{R}$ such that there is no BD path from $\mathbf{X}$ to v.
    $\mathbf{R}' \leftarrow \mathbf{R}$
    for all $v \in \mathbf{R}$:
        if TESTSEP$(\mathcal{G}_{\underline{\mathbf{X}}}, \mathbf{X}, v, \emptyset)$ = False **then:**
            if $v \in \mathbf{I}$ **then: return** $\emptyset$
            else: $\mathbf{R}' \leftarrow \mathbf{R}' \setminus \{v\}$
    end for
    return $\mathbf{R}'$
end function

Figure 2: A function that outputs the set of candidate variables satisfying the second condition of the FD criterion.

Step 1 of FINDFDSET

In Step 1, FINDFDSET calls the function GETCAND2NDFDC (presented in Fig. 2) to construct a set $\mathbf{R}'$ that consists of all the variables $v \in \mathbf{R}$ such that there is no BD path from $\mathbf{X}$ to v ($\mathbf{R}'$ is set to empty if there is a BD path from $\mathbf{X}$ to $\mathbf{I}$). Then, there is no BD path from $\mathbf{X}$ to any set $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}'$ since, by definition, there is no BD path from $\mathbf{X}$ to $\mathbf{Z}$ if and only if there is no BD path from $\mathbf{X}$ to any $v \in \mathbf{Z}$.

GETCAND2NDFDC iterates through each variable $v \in \mathbf{R}$ and checks if there exists an open BD path from $\mathbf{X}$ to v by calling the function TESTSEP($\mathcal{G}_{\underline{\mathbf{X}}}, \mathbf{X}, v, \emptyset$) [29]. TESTSEP($\mathcal{G}, \mathbf{A}, \mathbf{B}, \mathbf{C}$) returns True if $\mathbf{C}$ is a separator of $\mathbf{A}$ and $\mathbf{B}$ in $\mathcal{G}$, or False otherwise. Therefore, TESTSEP($\mathcal{G}_{\underline{\mathbf{X}}}, \mathbf{X}, v, \emptyset$) returns True if $\emptyset$ is a separator of $\mathbf{X}$ and v in $\mathcal{G}_{\underline{\mathbf{X}}}$ (i.e., there is no BD path from $\mathbf{X}$ to v), or False otherwise. If TESTSEP returns False, then v is removed from $\mathbf{R}'$ because every set $\mathbf{Z}$ containing v violates the second condition of the FD criterion relative to $(\mathbf{X}, \mathbf{Y})$.

Example 2. Continuing Example 1. With $\mathbf{I} = \emptyset$ and $\mathbf{R} = \{A, B, C, D\}$, GETCAND2NDFDC outputs a set $\mathbf{R}' = \{A, B, C\}$. D is excluded from $\mathbf{R}'$ since there exists a BD path from $\{X\}$ to $\{D\}$, and any set containing D violates the second condition of the FD criterion relative to $(\{X\}, \{Y\})$.

Lemma 1 (Correctness of GETCAND2NDFDC)**.** GETCAND2NDFDC($\mathcal{G}, \mathbf{X}, \mathbf{I}, \mathbf{R}$) *generates a set of variables* $\mathbf{R}'$ *with* $\mathbf{I} \subseteq \mathbf{R}' \subseteq \mathbf{R}$ *such that* $\mathbf{R}'$ *consists of all and only variables* v *that satisfies the second condition of the FD criterion relative to* $(\mathbf{X}, \mathbf{Y})$*. Further, every subset* $\mathbf{Z} \subseteq \mathbf{R}'$ *satisfies the second condition of the FD criterion relative to* $(\mathbf{X}, \mathbf{Y})$*, and every set* $\mathbf{Z}$ *with* $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$ *that satisfies the second condition of the FD criterion relative to* $(\mathbf{X}, \mathbf{Y})$ *must be a subset of* $\mathbf{R}'$*.*

Step 2 of FINDFDSET

In Step 2, FINDFDSET calls the function GETCAND3RDFDC presented in Fig. 3 to generate a set $\mathbf{R}''$ consisting of all the variables $v \in \mathbf{R}'$ such that there exists a set $\mathbf{Z}$ containing v with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}'$ that further satisfies the third condition of the FD criterion relative to $(\mathbf{X}, \mathbf{Y})$ (i.e., all BD paths from $\mathbf{Z}$ to $\mathbf{Y}$ are blocked by $\mathbf{X}$). In other words, $\mathbf{R}''$ is the union of all $\mathbf{Z}$ with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}'$ that satisfies the third condition of the FD criterion.

GETCAND3RDFDC iterates through each variable $v \in \mathbf{R}'$ and calls the function GETDEP($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \{v\}, \mathbf{R}'$) in line 5. Presented in Fig. 4, GETDEP returns a subset $\mathbf{Z}' \subseteq \mathbf{R}' \setminus \{v\}$ such that all BD paths from $\mathbf{Z} = \{v\} \cup \mathbf{Z}'$ to $\mathbf{Y}$ are blocked by $\mathbf{X}$ (if there exists such $\mathbf{Z}'$). If GETDEP returns $\perp$, then there exists no $\mathbf{Z}$ containing v that satisfies the third condition of the FD criterion relative to $(\mathbf{X}, \mathbf{Y})$, so v is removed from $\mathbf{R}''$.

1: **function** GETCAND3RDFDC($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R}'$)
2: **Output:** $\mathbf{R}''$ consisting of all the variables $v \in \mathbf{R}'$ such that there exists a set $\mathbf{Z}$ containing v with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}'$ that satisfies the third condition of the FD criterion relative to $(\mathbf{X}, \mathbf{Y})$.
3: $\mathbf{R}'' \leftarrow \mathbf{R}'$
4: **for all** $v \in \mathbf{R}'$:
5: **if** GETDEP($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \{v\}, \mathbf{R}'$) $=\perp$ **then:**
6: **if** $v \in \mathbf{I}$ **then: return** $\emptyset$
7: **else:** $\mathbf{R}'' \leftarrow \mathbf{R}'' \setminus \{v\}$
8: **end for**
9: **return** $\mathbf{R}''$
10: **end function**

Figure 3: A function that outputs the set of candidate variables potentially satisfying the second and third conditions of the FD criterion.

Example 3. Continuing Example 2. Given $\mathbf{I} = \emptyset$ and $\mathbf{R}' = \{A, B, C\}$, GETCAND3RDFDC outputs $\mathbf{R}'' = \{A, B, C\}$ because for each variable $v \in \mathbf{R}''$, GETDEP finds a set $\mathbf{Z}'$ such that $\{v\} \cup \mathbf{Z}'$ satisfies the third condition of the FD criterion relative to $(\{X\}, \{Y\})$. For $v = A$, $\mathbf{Z}' = \emptyset$, for $v = B$, $\mathbf{Z}' = \{A\}$, and for $v = C$, $\mathbf{Z}' = \{A\}$.

Next, we explain how the function GETDEP($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{T}, \mathbf{R}'$) works. First, GETDEP constructs an undirected graph $\mathcal{M}$ in a way that the paths from $\mathbf{T}$ to $\mathbf{Y}$ in $\mathcal{M}$ represent all BD paths from $\mathbf{T}$ to $\mathbf{Y}$ that cannot be blocked by $\mathbf{X}$ in $\mathcal{G}$. The auxiliary function MORALIZE($\mathcal{G}$) moralizes a given graph $\mathcal{G}$ into an undirected graph. The moralization is performed on the subgraph over $An(\mathbf{T} \cup \mathbf{X} \cup \mathbf{Y})$ instead of $\mathcal{G}$ based on the following property: $\mathbf{T}$ and $\mathbf{Y}$ are d-separated by $\mathbf{X}$ in $\mathcal{G}$ if and only if $\mathbf{X}$ is a $\mathbf{T}$-$\mathbf{Y}$ node cut (i.e., removing $\mathbf{X}$ disconnects $\mathbf{T}$ and $\mathbf{Y}$) in $\mathcal{G}' = \text{MORALIZE}(\mathcal{G}_{An(\mathbf{T} \cup \mathbf{X} \cup \mathbf{Y})})$ [12].

GETDEP performs Breadth-First Search (BFS) from $\mathbf{T}$ to $\mathbf{Y}$ on $\mathcal{M}$ and incrementally constructs a subset $\mathbf{Z}' \subseteq \mathbf{R}' \setminus \mathbf{T}$ such that, after BFS terminates, there will be no BD path from $\mathbf{Z} = \mathbf{T} \cup \mathbf{Z}'$ to $\mathbf{Y}$ that cannot be blocked by $\mathbf{X}$ in $\mathcal{G}$. While constructing $\mathbf{Z}'$, GETDEP calls the function GETNEIGHBORS($u, \mathcal{M}$) (presented in Fig. 8, Appendix) to obtain all observed neighbors of u in $\mathcal{M}$.

The BFS starts from each variable $v \in \mathbf{T}$. Whenever a non-visited node u is encountered, the set $\mathbf{NR}$, observed neighbors of u that belong to $\mathbf{R}'$, is computed. $\mathbf{NR}$ can be added to $\mathbf{Z}'$ because removing all outgoing edges of $\mathbf{NR}$ may contribute to disconnecting some BD paths Π from $\mathbf{T}$ to $\mathbf{Y}$ that cannot be blocked by $\mathbf{X}$ in $\mathcal{G}$. In other words, in $\mathcal{G}_{\underline{\mathbf{T} \cup \mathbf{Z}' \cup \mathbf{NR}}}$, Π could be disconnected from $\mathbf{T}$ to

function GETDEP($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{T}, \mathbf{R}'$)
  Output: $\mathbf{Z}' \subseteq \mathbf{R}' \setminus \mathbf{T}$, a set of variables such that $\mathbf{T} \cup \mathbf{Z}'$ satisfies the third condition of the FD criterion relative to $(\mathbf{X}, \mathbf{Y})$.
  $\mathcal{G}' \leftarrow \mathcal{G}_{An(\mathbf{T} \cup \mathbf{X} \cup \mathbf{Y})}$
  $\mathcal{G}' \leftarrow \mathcal{G}'$ with all bidirected edges $A \leftrightarrow B$ replaced by a latent node L_{AB} and two edges $L_{AB} \rightarrow A$ and $L_{AB} \rightarrow B$
  $\mathcal{G}'' \leftarrow \mathcal{G}'_{\underline{\mathbf{T}}}$
  $\mathcal{M} \leftarrow$ MORALIZE($\mathcal{G}''$) then remove $\mathbf{X}$
  $\mathbf{Z}' \leftarrow \emptyset, \mathbf{Q} \leftarrow \mathbf{T}$ and mark all $v \in \mathbf{T}$ as visited
  while $\mathbf{Q} \neq \emptyset$ **do**
    $u \leftarrow \mathbf{Q}$.POP()
    if $u \in \mathbf{Y}$ **then: return** $\perp$
    $\mathbf{NR} \leftarrow$ GETNEIGHBORS($u, \mathcal{M}$) $\cap \mathbf{R}'$ that are not visited
    $\mathcal{G}'' \leftarrow \mathcal{G}'_{\underline{\mathbf{T} \cup \mathbf{Z}' \cup \mathbf{NR}}}$
    $\mathcal{M} \leftarrow$ MORALIZE($\mathcal{G}''$) then remove $\mathbf{X}$
    $\mathbf{N}' \leftarrow$ GETNEIGHBORS($u, \mathcal{M}$) that are not visited
    $\mathbf{NR}' \leftarrow \{w \in \mathbf{NR} |$ there exists an incoming arrow into w in $\mathcal{G}\}$
    $\mathbf{N} \leftarrow \mathbf{N}' \cup \mathbf{NR}', \mathbf{Z}' \leftarrow \mathbf{Z}' \cup \mathbf{NR}$
    $\mathbf{Q}$.INSERT($\mathbf{N}$) and mark all $w \in \mathbf{N}$ as visited
  end while
  return $\mathbf{Z}'$
end function

Figure 4: A function that facilitates the construction of a set that satisfies the third condition of the FD criterion.

$\mathbf{Y}$ where Π are not disconnected in $\mathcal{G}_{\underline{\mathbf{T} \cup \mathbf{Z}'}}$. After adding $\mathbf{NR}$ to $\mathbf{Z}'$, $\mathcal{M}$ must be reconstructed in a way that reflects the setting where all outgoing edges of $\mathbf{NR}$ are removed. BFS will be performed on such modified $\mathcal{M}$.

GETDEP checks if there exists any set of nodes $\mathbf{N}$ to be visited further. $\mathbf{N}$ consists of two sets: 1) $\mathbf{N}'$, all observed neighbors of u that are still reachable from u, even after removing all outgoing edges of $\mathbf{NR}$, and 2) $\mathbf{NR}' \subseteq \mathbf{NR}$ where for every node $w \in \mathbf{NR}$, there exists an incoming arrow into w in $\mathcal{G}$. All nodes in $\mathbf{NR}'$ must be checked because there might exist some BD path π from w to $y \in \mathbf{Y}$ that cannot be blocked by $\mathbf{X}$ in $\mathcal{G}$. If π cannot be disconnected from w to y, then the set $\mathbf{Z}$ will violate the third condition of the FD criterion relative to $(\mathbf{X}, \mathbf{Y})$.

The BFS continues until either a node $y \in \mathbf{Y}$ is visited, or no more node can be visited. If GETDEP returns a set $\mathbf{Z}'$, then we have that all BD paths from $\mathbf{T}$ to $\mathbf{Y}$ that cannot be blocked by $\mathbf{X}$ in $\mathcal{G}$ have been disconnected in $\mathcal{G}_{\underline{\mathbf{Z}}}$ while ensuring that there exists no BD path from $\mathbf{Z}$ to $\mathbf{Y}$ that cannot be blocked by $\mathbf{X}$ in $\mathcal{G}$. Therefore, $\mathbf{Z}$ satisfies the third condition of the FD criterion relative to $(\mathbf{X}, \mathbf{Y})$. Otherwise, if GETDEP returns $\perp$ (i.e., y is visited), then there does not exist any $\mathbf{Z}$ containing $\mathbf{T}$ that satisfies the third condition of the FD criterion relative to $(\mathbf{X}, \mathbf{Y})$. This is because there exists a BD path π from $t \in \mathbf{T}$ to y that cannot be blocked by $\mathbf{X}$ in $\mathcal{G}$; removing outgoing edges of all $w \in \mathbf{R}'$ that intersect π cannot disconnect π from t to y.

Lemma 2 (Correctness of GETCAND3RDFDC)**.** GETCAND3RDFDC($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R}'$) *in Step 2 of Alg. 1 generates a set of variables* $\mathbf{R}''$ *where* $\mathbf{I} \subseteq \mathbf{R}'' \subseteq \mathbf{R}'$. $\mathbf{R}''$ *consists of all and only variables* v *such that there exists a subset* $\mathbf{Z}$ *with* $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}'$ *and* $v \in \mathbf{Z}$ *that satisfies the third condition of the FD criterion relative to* $(\mathbf{X}, \mathbf{Y})$. *Further, every set* $\mathbf{Z}$ *with* $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$ *that satisfies both the second and the third conditions of the FD criterion must be a subset of* $\mathbf{R}''$.

Remark: Even though every set $\mathbf{Z}$ with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}'$ that satisfies the third condition of the FD criterion must be a subset of $\mathbf{R}''$, *not* every subset $\mathbf{Z} \subseteq \mathbf{R}''$ satisfies the third condition of the FD criterion, as illustrated by the following example.

Example 4. In Example 3, GETCAND3RDFDC outputs $\mathbf{R}'' = \{A, B, C\}$. However, for $\mathbf{Z} = \{B\}$, the BD path $\{B \leftarrow A \rightarrow D \rightarrow Y\}$ is not blocked by $\{X\}$; for $\mathbf{Z} = \{C\}$, the BD path $\{C \leftarrow A \rightarrow D \rightarrow Y\}$ is not blocked by $\{X\}$.

On the other hand, we show that $\mathbf{Z} = \mathbf{R}''$ itself satisfies the third condition of the FD criterion, as shown in the following.

Lemma 3. *$\mathbf{R}''$ generated by* GETCAND3RDFDC *(in Step 2 of Alg. 1) satisfies the third condition of the FD criterion, that is, all BD paths from $\mathbf{R}''$ to $\mathbf{Y}$ are blocked by $\mathbf{X}$.*

Step 3 of FINDFDSET

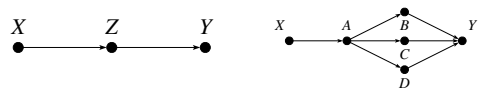
Finally, in Step 3, FINDFDSET looks for a set $\mathbf{Z} \subseteq \mathbf{R}''$ that satisfies the first condition of the FD criterion relative to $(\mathbf{X}, \mathbf{Y})$, that is, $\mathbf{Z}$ intercepts all causal paths from $\mathbf{X}$ to $\mathbf{Y}$. To facilitate checking whether a set $\mathbf{Z}$ intercepts all causal paths from $\mathbf{X}$ to $\mathbf{Y}$, we introduce the concept of causal path graph defined as follows.

Definition 2. (Causal Path Graph) Let $\mathcal{G}$ be a causal graph and $\mathbf{X}, \mathbf{Y}$ disjoint sets of variables. A *causal path graph* $\mathcal{G}'$ relative to $(\mathcal{G}, \mathbf{X}, \mathbf{Y})$ is a graph over $\mathbf{X} \cup \mathbf{Y} \cup PCP(\mathbf{X}, \mathbf{Y})$, where $PCP(\mathbf{X}, \mathbf{Y}) = (De(\mathbf{X})_{\mathcal{G}_{\overline{\mathbf{X}}}} \setminus \mathbf{X}) \cap An(\mathbf{Y})_{\mathcal{G}_{\underline{\mathbf{X}}}}$[1], constructed as follows:

1. Construct a subgraph $\mathcal{G}'' = \mathcal{G}_{\mathbf{X} \cup \mathbf{Y} \cup PCP(\mathbf{X}, \mathbf{Y})}$.
2. Construct a graph $\mathcal{G}' = \mathcal{G}''_{\overline{\mathbf{X}}\underline{\mathbf{Y}}}$, then remove all bidirected edges from $\mathcal{G}'$.

A function GETCAUSALPATHGRAPH$(\mathcal{G}, \mathbf{X}, \mathbf{Y})$ for constructing a causal path graph is presented in Fig. 9 in the Appendix.

Example 5. Consider the causal graph $\mathcal{G}'$ shown in Fig. 1b with $\mathbf{X} = \{X\}$, and $\mathbf{Y} = \{Y\}$. The causal path graph $\mathcal{G}''$ relative to $(\mathcal{G}', \{X\}, \{Y\})$ is shown in Fig. 5b. All causal paths from $\{X\}$ to $\{Y\}$ in $\mathcal{G}'$ are present in $\mathcal{G}''$.



(a) $\mathcal{G}$ (b) $\mathcal{G}''$

Figure 5: Two causal path graphs generated from (a) the causal graph in Fig. 1a, and (b) the causal graph in Fig. 1b. Both preserve all and only causal paths from $\{X\}$ to $\{Y\}$ in the original graphs.

After constructing a causal path graph $\mathcal{G}'$ relative to $(\mathcal{G}, \mathbf{X}, \mathbf{Y})$, we use the function TESTSEP$(\mathcal{G}', \mathbf{X}, \mathbf{Y}, \mathbf{Z})$ to check if $\mathbf{Z}$ is a separator of $\mathbf{X}$ and $\mathbf{Y}$ in $\mathcal{G}'$. Based on the following lemma, $\mathbf{Z}$ satisfies the first condition of the FD criterion relative to $(\mathbf{X}, \mathbf{Y})$ if and only if TESTSEP returns True.

Lemma 4. *Let $\mathcal{G}$ be a causal graph and $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ disjoint sets of variables. Let $\mathcal{G}'$ be the causal path graph relative to $(\mathcal{G}, \mathbf{X}, \mathbf{Y})$. Then, $\mathbf{Z}$ satisfies the first condition of the FD criterion relative to $(\mathbf{X}, \mathbf{Y})$ if and only if $\mathbf{Z}$ is a separator of $\mathbf{X}$ and $\mathbf{Y}$ in $\mathcal{G}'$.*

Given the set $\mathbf{R}''$ that contains every set $\mathbf{Z}$ with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$ that satisfies both the second and the third conditions of the FD criterion (Lemma 2), it may appear that we need to search for a set $\mathbf{Z} \subseteq \mathbf{R}''$ that satisfies the first condition of the FD criterion. We show instead that all we need is to check whether the set $\mathbf{R}''$ itself satisfies the first condition which has been shown to satisfy the second and third conditions by Lemma 3. This result is summarized in the following lemma.

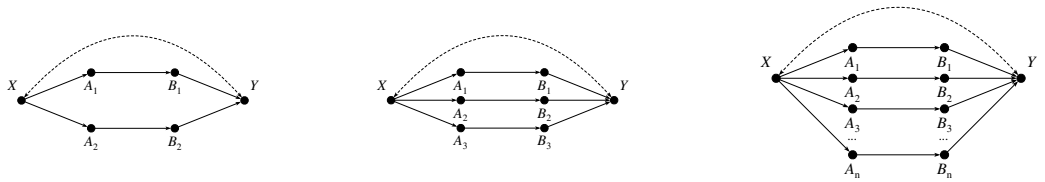
Lemma 5. *There exists a set $\mathbf{Z}_0$ satisfying the FD criterion relative to $(\mathbf{X}, \mathbf{Y})$ with $\mathbf{I} \subseteq \mathbf{Z}_0 \subseteq \mathbf{R}$ if and only if $\mathbf{R}''$ generated by* GETCAND3RDFDC *(in Step 2 of Alg. 1) satisfies the FD criterion relative to $(\mathbf{X}, \mathbf{Y})$.*

Example 6. Continuing Example 3. In Step 3, FINDFDSET outputs $\mathbf{Z} = \mathbf{R}'' = \{A, B, C\}$ since $\mathbf{Z}$ is a separator of $\{X\}$ and $\{Y\}$ in the causal path graph $\mathcal{G}''$ in Fig. 5b.

The results in this section are summarized as follows.

Theorem 1 (Correctness of FINDFDSET)**.** *Let $\mathcal{G}$ be a causal graph, $\mathbf{X}, \mathbf{Y}$ disjoint sets of variables, and $\mathbf{I}, \mathbf{R}$ sets of variables such that $\mathbf{I} \subseteq \mathbf{R}$. Then,* FINDFDSET$(\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R})$ *outputs a set $\mathbf{Z}$ with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$ that satisfies the FD criterion relative to $(\mathbf{X}, \mathbf{Y})$, or outputs $\perp$ if none exists, in $O(n^3(n+m))$ time, where n and m represent the number of nodes and edges in $\mathcal{G}$.*

[1] A notation introduced by van der Zander et al. [29] to denote the set of variables on proper causal paths from $\mathbf{X}$ to $\mathbf{Y}$.



(a) $\mathcal{G}$ (b) $\mathcal{G}'$ (c) $\mathcal{G}''$

Figure 6: Three examples of the FD criterion to demonstrate that total number of FD adjustment sets may be exponential with respect to the number of nodes in a graph.

4 Enumerating Front-door Adjustment Sets

Our goal in this section is to develop an algorithm that lists *all* FD adjustment sets in a causal diagram. In general, there may exist exponential number of such sets, which means that any listing algorithm will take exponential time to list them all. We will instead look for an algorithm that has an interesting property known as *polynomial delay* [26]. In words, poly-delay algorithms output the first answer (or indicate none is available) in polynomial time, and take polynomial time to output each consecutive answer as well. Consider the following example.

Algorithm 2 LISTFDSETS $(\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R})$

1: **Input:** $\mathcal{G}$ a causal diagram; $\mathbf{X}, \mathbf{Y}$ disjoint sets of variables; $\mathbf{I}, \mathbf{R}$ sets of variables.
2: **Output:** Listing front-door adjustment set $\mathbf{Z}$ relative to $(\mathbf{X}, \mathbf{Y})$ where $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$.
3: **if** FINDFDSET$(\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R}) \neq \perp$ **then:**
4: **if** $\mathbf{I} = \mathbf{R}$ **then:** Output $\mathbf{I}$
5: **else:**
6: $v \leftarrow$ any variable from $\mathbf{R} \setminus \mathbf{I}$
7: LISTFDSETS$(\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I} \cup \{v\}, \mathbf{R})$
8: LISTFDSETS$(\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R} \setminus \{v\})$

Example 7. Consider the three causal graphs in Fig. 6. In $\mathcal{G}$ shown in Fig. 6a, there exists 9 valid FD adjustment sets relative to $(\{X\}, \{Y\})$. In $\mathcal{G}'$, presented in Fig. 6b, two variables A_3 and B_3 are added from $\mathcal{G}$, forming an additional causal path from X to Y. 27 FD adjustment sets relative to $(\{X\}, \{Y\})$ are available in $\mathcal{G}'$. If another causal path $X \to A_4 \to B_4 \to Y$ is added to $\mathcal{G}'$, then there are 81 FD adjustment sets relative to $(\{X\}, \{Y\})$. As shown in Fig. 6c, in a graph $\mathcal{G}''$ with similar pattern with causal path $X \to A_i \to B_i \to Y, i = 1, \ldots n$, there are at least 2^n number of FD adjustment sets.

We have developed an algorithm named LISTFDSETS, shown in Alg. 2, that lists all FD adjustment sets $\mathbf{Z}$ relative to $(\mathbf{X}, \mathbf{Y})$ satisfying $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$ with polynomial delay, given a causal diagram $\mathcal{G}$, disjoint sets of variables $\mathbf{X}$ and $\mathbf{Y}$, and two sets of variables $\mathbf{I}$ and $\mathbf{R}$.

Example 8. Consider the causal graph $\mathcal{G}'$ shown in Fig. 1b with $\mathbf{X} = \{X\}$, $\mathbf{Y} = \{Y\}$, $\mathbf{I} = \emptyset$ and $\mathbf{R} = \{A, B, C, D\}$. LISTFDSETS outputs $\{A, B, C\}$, $\{A, B\}$, $\{A, C\}$, $\{A\}$ one by one, and finally stops as no more adjustment sets exist.

The algorithm LISTFDSETS takes the same search strategy as the listing algorithm LISTSEP [29] that enumerates all BD adjustment sets with polynomial delay. LISTFDSETS implicitly constructs a binary search tree where each tree node $\mathcal{N}(\mathbf{I}', \mathbf{R}')$ represents the collection of all FD adjustment sets $\mathbf{Z}$ relative to $(\mathbf{X}, \mathbf{Y})$ with $\mathbf{I}' \subseteq \mathbf{Z} \subseteq \mathbf{R}'$. The search starts from the root tree node $\mathcal{N}(\mathbf{I}, \mathbf{R})$, indicating that LISTFDSETS will list all FD adjustment sets $\mathbf{Z}$ relative to $(\mathbf{X}, \mathbf{Y})$ with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$.

Upon visiting a node $\mathcal{N}(\mathbf{I}', \mathbf{R}')$, LISTFDSETS first calls the function FINDFDSET (line 3) to decide whether it is necessary to search further from $\mathcal{N}$. If FINDFDSET outputs $\perp$, then there does not exist any FD adjustment set $\mathbf{Z}_0$ with $\mathbf{I}' \subseteq \mathbf{Z}_0 \subseteq \mathbf{R}'$ and there is no need to search further. Otherwise, $\mathcal{N}$ spawns two children, $\mathcal{N}_1$ and $\mathcal{N}_2$, and LISTFDSETS continues the search over each child separately. $\mathcal{N}_1$ in line 7 represents the collection of all FD adjustment sets $\mathbf{Z}_1$ relative to $(\mathbf{X}, \mathbf{Y})$ where $\mathbf{I}' \cup \{v\} \subseteq \mathbf{Z}_1 \subseteq \mathbf{R}'$. On the other hand, $\mathcal{N}_2$ in line 8 represents the collection of all FD adjustment sets $\mathbf{Z}_2$ where $\mathbf{I}' \subseteq \mathbf{Z}_2 \subseteq \mathbf{R}' \setminus \{v\}$. $\mathcal{N}_1$ and $\mathcal{N}_2$ are disjoint and thus the search never

overlaps, which is crucial to guaranteeing that LISTFDSETS runs in polynomial delay. Finally, a leaf tree node $\mathcal{L}$ is reached when $\mathbf{I}' = \mathbf{R}'$, and LISTFDSETS outputs a valid FD adjustment set $\mathbf{I}'$.

Example 9. Continuing from Example 8. Fig. 7 shows a search tree generated by running LISTFDSETS$(\mathcal{G}', \{X\}, \{Y\}, \emptyset, \{A, B, C, D\})$. Initially, the search starts from the root tree node $\mathcal{N}(\emptyset, \{A, B, C, D\})$. Since FINDFDSET returns a set $\{A, B, C\}$, $\mathcal{N}$ branches out into two children $\mathcal{N}'(\{A\}, \{A, B, C, D\})$ and $\mathcal{N}''(\emptyset, \{B, C, D\})$. The search continues from the left child $\mathcal{N}'$ until reaching the leaf tree node $\mathcal{L}_1(\{A, B, C, D\}, \{A, B, C, D\})$ where FINDFDSET returns $\perp$. LISTFDSETS backtracks to the parent tree node $\mathcal{N}_1(\{A, B, C\}, \{A, B, C, D\})$ and then checks the next leaf $\mathcal{L}_2(\{A, B, C\}, \{A, B, C\})$ where FINDFDSET returns a set $\{A, B, C\}$, a valid FD admissible set relative to $(\{X\}, \{Y\})$. LISTFDSETS outputs $\{A, B, C\}$. Next, LISTFDSETS backtracks to the tree node $\mathcal{N}_2(\{A, B\}, \{A, B, C, D\})$ and reaches the leaf $\mathcal{L}_3(\{A, B\}, \{A, B\})$ where FINDFDSET outputs $\{A, B\}$, and thus LISTFDSETS outputs $\{A, B\}$. LISTFDSETS continues and outputs two sets $\{A, C\}$ and $\{A\}$ in order. Finally, LISTFDSETS backtracks to the root $\mathcal{N}$ and checks the right child $\mathcal{N}''$ where FINDFDSET returns $\perp$. LISTFDSETS does not search further from $\mathcal{N}''$ and stops as no more tree node is left to be visited.

Our results are summarized in the following theorem, which provides both the correctness, completeness, and poly-delay complexity of the proposed algorithm.

Theorem 2 (Correctness of LISTFDSETS)**.** *Let $\mathcal{G}$ be a causal graph, $\mathbf{X}, \mathbf{Y}$ disjoint sets of variables, and $\mathbf{I}, \mathbf{R}$ sets of variables.* LISTFDSETS$(\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R})$ *enumerates all and only sets $\mathbf{Z}$ with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$ that satisfy the FD criterion relative to $(\mathbf{X}, \mathbf{Y})$ in $O(n^4(n+m))$ delay where n and m represent the number of nodes and edges in $\mathcal{G}$.*

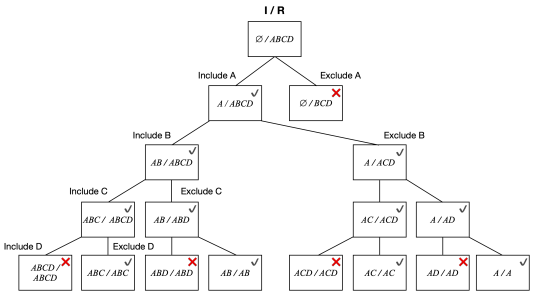


Figure 7: A search tree illustrating the running of LISTFDSETS in Example 9.

5 Discussion and Conclusions

This work has some limitations and can be extended in several directions. First, Pearl's FD criterion is not complete in the sense that there might exist a causal effect that can be computed by the FD adjustment formula (Eq. (3)) but the set $\mathbf{Z}$ does not satisfy the FD criterion. While the BD criterion has been generalized to a complete criterion for BD adjustment [23], it is an interesting open problem to come up with a complete criterion for sets satisfying the FD adjustment. Second, this work assumes that the causal diagram is given (or inferred based on scientists' domain knowledge and/or data). Although this assumption is quite common throughout the causal inference literature, more recent work has moved to finding BD admissible sets given incomplete or partially specified causal diagrams, e.g., maximal ancestral graphs (MAGs) [29], partial ancestral graphs (PAGs) [18], and completed partially directed acyclic graphs (CPDAGs) [18]. It is an interesting and certainly challenging future work to develop algorithms for finding FD admissible sets in these types of graphs. Other interesting future research topics include developing algorithms for finding minimal, minimum, and minimum cost FD adjustment sets, which are available for the BD adjustment sets [30], as well as algorithms for finding conditional FD adjustment sets [8, 4]. Having said all of that, we believe that the results developed in this paper is a necessary step towards solving these more challenging problems.

After all, we started from the observation that identification is not restricted to BD adjustment, and Pearl's FD criterion provides a classic strategy for estimating causal effects from observational data and qualitative knowledge encoded in the form of a causal diagram. The criterion is drawing more attention in recent years and statistically efficient and doubly robust estimators have been developed for estimating the FD estimand from finite samples. In this paper, we develop algorithms that given a causal diagram $\mathcal{G}$, find an admissible FD set (Alg. 1 FINDFDSET, Thm. 1) and enumerate all admissible FD sets with polynomial delay (Alg. 2 LISTFDSETS, Thm. 2). We hope that the methods and algorithms proposed in this work will help scientists to use the FD strategy for causal effects estimation in the practical applications and are useful for scientists in study design to select covariates based on desired properties, including cost, feasibility, and statistical power.

References

[1] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

[2] Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint arXiv:2003.12659*, 2020.

[3] Ronald A Fisher. Design of Experiments. *British Medical Journal*, 1(3923):554, 1936.

[4] Isabel R Fulcher, Ilya Shpitser, Stella Marealle, and Eric J Tchetgen Tchetgen. Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):199–214, 2020.

[5] Adam N. Glynn and Konstantin Kashin. Front-door versus back-door adjustment with unmeasured confounding: Bias formulas for front-door and hybrid adjustments with application to a job training program. *Journal of the American Statistical Association*, 113(523):1040–1049, 2018.

[6] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, Mar 1998.

[7] Yimin Huang and Marco Valtorta. Identifiability in Causal Bayesian Networks: A Sound and Complete Algorithm. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*, pages 1149–1156. AAAI Press, Menlo Park, CA, 2006.

[8] Paul Hünermund and Elias Bareinboim. Causal inference and data-fusion in econometrics. Technical Report R-51, Causal Artificial Intelligence Lab, Columbia University, Dec 2019.

[9] Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating causal effects using weighting-based estimators. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 10186–10193, New York, NY, 2020. AAAI Press.

[10] Yonghan Jung, Jin Tian, and Elias Bareinboim. Learning causal effects via weighted empirical risk minimization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12697–12709, Vancouver, Canada, Jun 2020. Curran Associates, Inc.

[11] Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating identifiable causal effects through double machine learning. *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021.

[12] Steffen L Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996.

[13] Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems(with discussion). *Journal of the Royal Statistical Society*, 50(2):157–224, 1988.

[14] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.

[15] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[16] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2000.

[17] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, 2018.

[18] Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H. Maathuis. Complete Graphical Characterization and Construction of Adjustment Sets in Markov Equivalence Classes of Ancestral Graphs. *Journal of Machine Learning Research*, 18, 2018.

[19] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2017.

[20] James Robins and Andrea Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In Nicholas Jewell, Klaus Dietz, and Vernon Farewell, editors, *AIDS Epidemiology: Methodological Issues*, pages 297–331, Boston, MA, 1992. Birkhäuser Boston.

[21] Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *The Journal of Machine Learning Research*, 21:1–86, Sep 2020.

[22] Ilya Shpitser and Judea Pearl. Identification of Joint Interventional Distributions in Recursive semi-Markovian Causal Models. In *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*, volume 2, pages 1219–1226, 2006.

[23] Ilya Shpitser, Tyler J VanderWeele, and James M Robins. On the validity of covariate adjustment for estimating causal effects. *Proceedings of the Twenty Sixth Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 527–536, 2010.

[24] Ezequiel Smucler, Facundo Sapienza, and Andrea Rotnitzky. Efficient adjustment sets in causal graphical models with hidden variables. *Biometrika*, 109(1):49–65, Mar 2022.

[25] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer-Verlag, New York, 1993.

[26] Ken Takata. Space-optimal, backtracking algorithms to list the minimal vertex separators of a graph. *Discrete Applied Mathematics*, 158(15):1660–1667, 2010.

[27] Johannes Textor and Maciej Liskiewicz. Adjustment Criteria in Causal Diagrams: An Algorithmic Perspective. In Avi Pfeffer and Fabio Cozman, editors, *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 681–688. AUAI Press, 2011.

[28] Jin Tian and Judea Pearl. A General Identification Condition for Causal Effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI 2002)*, pages 567–573, Menlo Park, CA, 2002. AAAI Press/The MIT Press.

[29] Benito van der Zander, Maciej Liskiewicz, and Johannes Textor. Constructing separators and adjustment sets in ancestral graphs. In *Proceedings of UAI 2014*, pages 907–916, 2014.

[30] Benito van der Zander, Maciej Liskiewicz, and Johannes Textor. Separators and adjustment sets in causal graphs: Complete criteria and an algorithmic framework. *Artificial Intelligence*, 270: 1–40, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Check Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We do not expect this work has negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Check the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [N/A]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]