Causal Effect Identification with Context-specific Independence Relations of Control Variables

Ehsan Mokhtarian
EPFL, Switzerland

Fateme Jamshidi
EPFL, Switzerland

Jalal Etesami
EPFL, Switzerland

Negar Kiyavash
EPFL, Switzerland

Abstract

We study the problem of causal effect identification from observational distribution given the causal graph and some context-specific independence (CSI) relations. It was recently shown that this problem is NP-hard, and while a sound algorithm to learn the causal effects is proposed in Tikka et al. (2019), no provably complete algorithm for the task exists. In this work, we propose a sound and complete algorithm for the setting when the CSI relations are limited to observed nodes with no parents in the causal graph. One limitation of the state of the art in terms of its applicability is that the CSI relations among all variables, even unobserved ones, must be given (as opposed to learned). Instead, We introduce a set of graphical constraints under which the CSI relations can be learned from mere observational distribution. This expands the set of identifiable causal effects beyond the state of the art.

1 INTRODUCTION

Data-driven approaches to identify a causal effect from a combination of observations, experiments, and side information about the problem of interest is central to science. Causal effect identification considers whether an interventional probability distribution can be uniquely determined from available information (Pearl (1995); Tian and Pearl (2003)).

In the absence of unobserved variables in the system, Pearl (1995) showed that the causal graph along with the observational distribution suffices to uniquely identify all interventional distributions. On the other

hand, when there are hidden variables, causal identifications become more challenging. Pearl (1995) introduced three rules, known as do-calculus, for identifying causal effect in the presence of unobserved variables. Shpitser and Pearl (2006) later showed that applying these rules along with probabilistic manipulations is *complete* to determine whether an interventional distribution is identifiable from only observational distribution given the causal graph. Interestingly, when further side information about the underlying generative model is available, the completeness results of do-calculus is no longer valid. That is, although more causal effects become identifiable, do-calculus based methods fail to identify them. For example, consider the causal graph in Figure 1. The causal effect of X on Y is not identifiable from the graph, while given a set of context-specific independence (CSI) relations; it becomes identifiable (see Example 2).

Statistical independence, specifically conditional independence (CI) relations between a set of random variables, plays an important role in causal inference (Mokhtarian et al., 2021). An important generalization of this concept is CSI (Boutilier et al. (1996); Shimony (1991)) which refers to a conditional independence relation that is true only in a specific context (See Section 3 for more details). Consider the following example. Smoking normally has a causal effect on blood pressure. But, when a person has a ratio of beta and alpha lipoproteins larger than a threshold, whether or not he smokes is unlikely to affect his blood pressure (Edwards and Toma (1985)). Thus, the blood pressure is independent of smoking in the context of this ratio. CSI relations have been used to analyze, e.g., gene expression data (Barash and Friedman (2002)), parliament elections, prognosis of heart disease (Nyman et al. (2014)), etc. They have also been used to improve probabilistic inference (Chavira and Darwiche (2008); Dal et al. (2018)) and structure learning (Chickering et al. (1997); Hyttinen et al. (2018)). Similar to Tikka et al. (2019), in this work, we study the causal effect identification problem with extra information in the form of CSI relations. Our contributions are as follows.
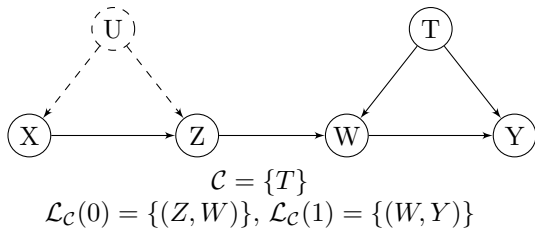
$$\mathcal{C} = \{T\}$$
$$\mathcal{L}_{\mathcal{C}}(0) = \{(Z, W)\}, \mathcal{L}_{\mathcal{C}}(1) = \{(W, Y)\}$$

Figure 1: Causal graph and $\mathcal{L}_{\mathcal{C}}$ in Example 2

- We study the causal identification problem in the presence of CSI relations of a subset of observed variables, called *control variables*, as well as the causal graph. We show that this problem is equivalent to a series of causal effect identifications only from causal graphs (Theorem 1). Consequently, we propose the first sound and complete algorithm (Algorithm 1) for this problem.

- We introduce a graphical constraint under which the CSI relations of control variables can be inferred only from the observational distribution (Algorithm 2). This expands the set of identifiable causal effects beyond the state-of-the-art approaches. More precisely, do-calculus-based methods determine the identifiability of a causal effect only from the causal graph without utilizing the other available source of knowledge, that is the observational distribution. Algorithm 2 uses both the causal graph and the observational distribution to determine the identifiability.

To prove the completeness result of our proposed algorithm in Theorem 1, we introduce a novel proof technique. In this technique, in order to show the identifiablity of a causal effect $P_{\mathbf{t}}(\mathbf{s})$ (see Section 2 for formal definition) from the observational distribution $P(\mathbf{N})$, instead of finding an exact formula for $P_{\mathbf{t}}(\mathbf{s})$ from $P(\mathbf{N})$, we show that it can be approximated by a functional of $P(\mathbf{N})$ with arbitrary accuracy.

1.1 Related Work

The causal effect identification problem has been extensively studied in the literature. Given the causal graph and the observational distribution, there are several sound and complete algorithms in the literature for identifying causal effects.

Pearl (1995) proposed three rules, known as do-calculus, which, along with probabilistic manipulations, suffice to derive a formula for a causal effect $P_{\mathbf{t}}(\mathbf{s})$ based on observational distribution $P(\mathbf{N})$. Later, Shpitser and Pearl (2006) showed that these rules are complete for the identification of causal effect from $\mathcal{G}$. That is, if do-calculus rules fail to obtain

a formula, the causal effect is non-identifiable.

Tian and Pearl (2003) proposed another algorithm for this problem and Huang and Valtorta (2008) later proved its completeness. In this algorithm, interventional distributions are expressed by Q functions using $Q[\mathbf{s}](\mathbf{n}) := P_{\mathbf{n} \backslash \mathbf{s}}(\mathbf{s})$. Subsequently, they show that the identifiablity of $P_{\mathbf{t}}(\mathbf{s})$ from $P(\mathbf{N})$ is equivalent to the identifiablity of a particular Q function from $P(\mathbf{N})$ and propose an algorithm for the identifiablity of the Q function from $P(\mathbf{N})$.

There are other variants of causal identification problem in the literature. In these variants, causal effect identification problem is studied under different information sets. For instance, Bareinboim and Pearl (2012) assumed there exists a subset of variables $\mathbf{Z}$ that could be intervened upon. That is, we have access to $P_{\mathbf{z}'}(\mathbf{v} \backslash \mathbf{z}')$ for all $\mathbf{Z}' \subseteq \mathbf{Z}$, the observational distribution, and the causal graph. They presented a sound and complete algorithm for identification of a causal effect. However, in practice, we might not have access to $P_{\mathbf{z}'}(\mathbf{v} \backslash \mathbf{z}')$ for all $\mathbf{Z}' \subseteq \mathbf{Z}$. Lee et al. (2019) generalized Bareinboim and Pearl (2012)'s result by restricting the set of interventional distributions $P_{\mathbf{z}'}(\mathbf{v} \backslash \mathbf{z}')$ to some $\mathbf{Z}' \subseteq \mathbf{V}$ by proposing a sound and complete algorithm for this setting. Lee and Bareinboim (2020) studied the problem of causal effect identification when the available distributions are only *partially* observable and proposed a sound algorithm to compute a causal effect in terms of the available distributions. They did not show the completeness of their algorithm. Zhang and Bareinboim (2021) considered the problem of bounding causal effects from experiments when the assignment of treatment is randomized while the subject compliance is imperfect. Due to unobserved variables, the causal effects are not identifiable. Thus, they propose bounds over the causal effects.

Tikka et al. (2019) studied the causal effect identification in the presence of CSI relations, observational distribution, and the causal graph. They showed that the problem in general is NP-hard and proposed a sound algorithm for this setting, but they did not provide completeness results.

Robins et al. (2020) considered the problem of identifiability for Controlled Direct Effect (CDE) that is an average causal effect when CSI relations are available. They showed that classical do-calculus based methods cannot determine the identifiability of CDE and proposed an algorithm that calculates CDE when additional CSI relations were available.

CSI relations have also been used for structure learning. As an example, Ramanan and Natarajan (2020) used CSI relations to identify the candidate set of causal relationships for learning structural causal mod-

Table 1: Table of notations

Notation	Description
Pa_X	Parents of X
Anc_X	Ancestors of X
$\Lambda_{\mathcal{G}}$	Set of observed roots in $\mathcal{G}$
$\mathcal{C}$	Set of control variables
$\mathcal{C}^X$	$Pa_X \cap \mathcal{C}$
$D_{\mathcal{C}}$	Domain of $\mathcal{C}$

els from observational data. Hyttinen et al. (2018) considered the problem of structure learning for Bayesian networks with CSI relations. They proposed orientation rules that utilizes CSI relations to orient edges.

2 PRELIMINARIES

In this paper, we use capital and small letters to denote random variables and their realizations, respectively. Similarly, sets of random variables are denoted by bold capital letters and sets of their realizations by small bold letters. Observable and unobservable variables are represented in solid and dashed circles in the figures, respectively. We assume all the variables are discrete with finite domain.

Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a directed acyclic graph (DAG) with a finite set of vertices $\mathbf{V}$ and a set of edges $\mathbf{E}$. $X \in \mathbf{V}$ is called a *parent* of Y if $(X, Y) \in \mathbf{E}$. X is called an *ancestor* of Y if there exists a directed path from X to Y. For $X \in \mathbf{V}$, Pa_X and Anc_X denote the set of parents and the set of ancestors of X, respectively. Note that $X \in Anc_X$. For $\mathbf{X} \subseteq \mathbf{V}$, $Anc_{\mathbf{X}}$ denotes the union of ancestors of the variables in $\mathbf{X}$. We use $D_{\mathbf{X}}$ to denote the domain of $\mathbf{X}$. A vertex $X \in \mathbf{V}$ is called a *root* if it has no parents. $\Lambda_{\mathcal{G}}$ denotes the set of observed roots in $\mathcal{G}$. Table 1 summarizes some of the notations used in this paper.

To model the environment, we assume that the variables are generated by a Structural Equation Model (SEM) (Pearl, 2009). In such models, each variable $X \in \mathbf{V}$ is generated as $f_X(Pa_X, \epsilon_X)$, where f_X is a deterministic function and ϵ_X is the exogenous noise corresponding to X such that the noise variables are jointly independent. Suppose $\mathcal{M}$ is a SEM with the causal DAG $\mathcal{G}$ and the joint distribution $P^{\mathcal{M}}(\mathbf{V})$ over $\mathbf{V} = \mathbf{N} \cup \mathbf{U}$, where $\mathbf{N}$ and $\mathbf{U}$ are the set of observable and unobservable variables, respectively. An intervention $do(\mathbf{T} = \mathbf{t})$ in $\mathcal{M}$ on a set $\mathbf{T}$ is defined as forcing $\mathbf{T}$ to be $\mathbf{t}$ and eliminating the impact of other variables on those in $\mathbf{T}$. $P_{\mathbf{t}}(\mathbf{s}) := P(\mathbf{S} = \mathbf{s}|do(\mathbf{T} = \mathbf{t}))$ denotes the post-interventional distribution of $\mathbf{S}$ after the intervention $do(\mathbf{T} = \mathbf{t})$ (Pearl (2009)). In the classic problem of causal effect identification, the goal is to compute $P_{\mathbf{t}}(\mathbf{s})$ from the observational distribution.

Definition 1 (ID from $\mathcal{G}$ (Pearl, 2009)). *The causal effect of $\mathbf{T}$ on $\mathbf{S}$ is said to be identifiable from $\mathcal{G}$ if for any $\mathbf{t} \in D_{\mathbf{T}}$ and $\mathbf{s} \in D_{\mathbf{S}}$, $P_{\mathbf{t}}^{\mathcal{M}}(\mathbf{s})$ is uniquely computable from $P^{\mathcal{M}}(\mathbf{N})$ in any SEM $\mathcal{M}$ with causal graph $\mathcal{G}$ such that $P^{\mathcal{M}}(\mathbf{n}) > 0$ for any $\mathbf{n} \in D_{\mathbf{N}}$.*

When an interventional distribution is identifiable, it can be uniquely computed from the joint observational distribution by a series of continuous operations (e.g., marginalization, Bayes rule, and the law of total probability). Formally, *uniquely computable* in Definition 1 means that there exists a continuous operator $\mathcal{F}$ such that $\mathcal{F}(P^{\mathcal{M}}(\mathbf{N})) = P_{\mathbf{t}}^{\mathcal{M}}(\mathbf{s})$.

Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{S}$ be three disjoint subsets of vertices in $\mathcal{G}$. $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{S}$, denoted by $\mathbf{X} \perp_{\mathcal{G}} \mathbf{Y}|\mathbf{S}$, if every path between $\mathbf{X}$ and $\mathbf{Y}$ is blocked[1] by $\mathbf{S}$. As we mentioned in the related work section, Pearl (1995) proposed following do-calculus rules.

Rule 1 (Insertion/deletion of observations)

$$P_{\mathbf{x}}(\mathbf{y} \mid \mathbf{z}, \mathbf{w}) = P_{\mathbf{x}}(\mathbf{y} \mid \mathbf{w}) \text{ , if } \mathbf{Y} \perp_{\mathcal{G}_{\overline{\mathbf{X}}}} \mathbf{Z}|\mathbf{X}, \mathbf{W}.$$

Rule 2 (Action/observation exchange)

$$P_{\mathbf{x}, \mathbf{z}}(\mathbf{y} \mid \mathbf{w}) = P_{\mathbf{x}}(\mathbf{y} \mid \mathbf{w}, \mathbf{z}) \text{ , if } \mathbf{Y} \perp_{\mathcal{G}_{\overline{\mathbf{X}}, \underline{\mathbf{Z}}}} \mathbf{Z}|\mathbf{X}, \mathbf{W}.$$

Rule 3 (Insertion/deletion of actions)

$$P_{\mathbf{x}, \mathbf{z}}(\mathbf{y} \mid \mathbf{w}) = P_{\mathbf{x}}(\mathbf{y} \mid \mathbf{w}) \text{ , if } \mathbf{Y} \perp_{\mathcal{G}_{\overline{\mathbf{X}}, \overline{\mathbf{Z}(\mathbf{W})}}} \mathbf{Z}|\mathbf{X}, \mathbf{W}.$$

In above, $\mathcal{G}_{\overline{\mathbf{X}}}$ (or $\mathcal{G}_{\underline{\mathbf{X}}}$) denotes the graph obtained by deleting all the in-going (or out-going) edges of the variables in $\mathbf{X}$ from $\mathcal{G}$, and $\mathbf{Z}(\mathbf{W})$ is the set of nodes in $\mathbf{Z}$ that are not ancestors of any node in $\mathbf{W}$.

3 CONTROL VARIABLES AND IDENTIFIABILITY FROM $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$

Let $\mathbf{X}, \mathbf{Y}, \mathbf{S}$ be three disjoint subsets of $\mathbf{V}$. $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{S}$ denotes the Conditional Independence (CI) of $\mathbf{X}$ and $\mathbf{Y}$ given $\mathbf{S}$. A context-specific independence (CSI) relation is of the form $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{S}_1 = \mathbf{s}_1, \mathbf{S}_2$, where $\mathbf{X}, \mathbf{Y}, \mathbf{S}_1$, and $\mathbf{S}_2$ are disjoint subsets of $\mathbf{V}$ and $\mathbf{s}_1 \in D_{\mathbf{S}_1}$.

Example 1. *Consider the DAG in Figure 2 with the following SEM.*

$$\begin{aligned} T &= Be(0.5), & Z &= Be(0.5), \\ Y &= TX + Z, & X &= T \oplus Z, \end{aligned}$$

where $Be(q)$ and $\oplus$ denote Bernoulli distribution with parameter q and logical xor, respectively. In this example, Y is a function of X and there is no CI relation between them. However, the following CSI relation holds: $X \perp\!\!\!\perp Y|T = 0, Z$.
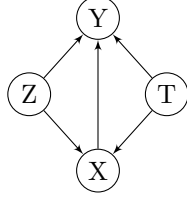
Figure 2: Causal graph of the SEM in Example 1.

Example 1 shows that, in general, CSI relations cannot be inferred merely from the causal graph. They have to be either provided as side information or inferred via additional assumptions. In order to encode CSI relations succinctly into the causal graph, analogous to Tikka et al. (2019); Pensar et al. (2015), we define a set of labels as follows.

Definition 2 (Label set $\mathcal{L}_{\mathbf{C}}$). *Suppose $\mathbf{C} \subseteq \mathbf{V}$ and denote by $\mathcal{L}_{\mathbf{C}}(\mathbf{c})$ a set of edges $(Y, X) \in \mathbf{E}$ such that $X \perp\!\!\!\perp Y | \mathbf{C} = \mathbf{c}, \mathbf{S}$ for some $\mathbf{S} \subseteq \mathbf{V} \setminus (\{X, Y\} \cup \mathbf{C})$. We define the label set $\mathcal{L}_{\mathbf{C}} := (\mathcal{L}_{\mathbf{C}}(\mathbf{c}) : \mathbf{c} \in D_{\mathbf{C}})$.*

Accordingly, a SEM $\mathcal{M}$ is said to be *compatible* with $\mathcal{L}_{\mathbf{C}}$, if for any $(Y, X) \in \mathcal{L}_{\mathbf{C}}(\mathbf{c})$, X is no longer a function of Y when $\mathbf{C} = \mathbf{c}$. In Example 1, $\mathcal{L}_T(0) = \{(X, Y)\}$ and $\mathcal{L}_T(1) = \varnothing$.

It is important to emphasize that Tikka et al. (2019) assumed all CSI relations $X \perp\!\!\!\perp Y | \mathbf{C} = \mathbf{c}$ are given as side information, where $\mathbf{C} = Pa_X \setminus \{Y\}$. This includes CSI relations with conditioning set $\mathbf{C}$ that could include unobserved variables. Therefore, it is impossible to obtain all such CSI relations from merely observational distribution. In this paper, we relax this assumption and restrict $\mathbf{C}$ to be a particular subset of observed variables called *control variables*.

Definition 3. *The set of control variables, denoted by $\mathcal{C}$, is a subset of the observed roots, i.e., $\mathcal{C} \subseteq \Lambda_{\mathcal{G}}$.*

Knowing the label set of control variables, i.e., $\mathcal{L}_{\mathcal{C}}$, is equivalent to knowing the edges that will be omitted from $\mathcal{G}$ for different realizations of $\mathcal{C}$. For instance, in Example 1, $\mathcal{L}_T(0) = \{(X, Y)\}$ implies that the edge (X, Y) will be deleted from the graph when $T = 0$. We will show that such side information can be utilized to identify some causal effects that would have been non-identifiable from merely do-calculus.

Next, we formally define identifiablity from $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$

Definition 4 (ID from $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$). *Suppose $\mathbf{T}, \mathbf{S} \subseteq \mathbf{N} \setminus \mathcal{C}$ are disjoint subsets of variables that are not control variables. The causal effect of $\mathbf{T}$ on $\mathbf{S}$ is said to be identifiable from $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$ if for any $\mathbf{t} \in D_{\mathbf{T}}$ and $\mathbf{s} \in D_{\mathbf{S}}$, $P_{\mathbf{t}}^{\mathcal{M}}(\mathbf{s})$ is uniquely computable from $P^{\mathcal{M}}(\mathbf{N})$ in any SEM $\mathcal{M}$ with causal graph $\mathcal{G}$ and compatible with*

---

[1]See Pearl (2009) for definition of blocking.

$\mathcal{L}_{\mathcal{C}}$ *such that $P^{\mathcal{M}}(\mathbf{n}) > 0$ for any $\mathbf{n} \in D_{\mathbf{N}}$.*

Next example demonstrates a scenario in which a non-identifiable causal effect from $\mathcal{G}$ becomes identifiable from $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$.

Example 2. *Consider the DAG in Figure 1 in which $D_T = \{0, 1\}$. It is known that $P_x(y)$ is not identifiable from $\mathcal{G}$. However, given $\mathcal{C} = \{T\}$, $\mathcal{L}_{\mathcal{C}}(0) = \{(Z, W)\}$, and $\mathcal{L}_{\mathcal{C}}(1) = \{(W, Y)\}$, $P_x(y)$ becomes identifiable and is equal to $P(y)$.*

Problem description: Suppose we are given the DAG $\mathcal{G}$ and the label set $\mathcal{L}_{\mathcal{C}}$, where $\mathcal{C} \subseteq \Lambda_{\mathcal{G}}$ is the set of control variables. This paper seeks to determine when $P_{\mathbf{t}}(\mathbf{S})$ is identifiable from $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$.

4 MAXIMAL-REGULAR LABELS

Pensar et al. (2015) first introduced the notion of *maximality* and *regularity* for a certain class of graphs called labeled DAGs. Herein, we extend these notions for $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$ and show how to modify $\mathcal{G}$ and $\mathcal{L}_{\mathcal{C}}$ such that $\mathcal{L}_{\mathcal{C}}$ becomes *maximal-regular* with respect to $\mathcal{G}$.

Definition 5. *$\mathcal{L}_{\mathcal{C}}$ is regular w.r.t. DAG $\mathcal{G}$ if for any $(Y, X) \in \mathcal{L}_{\mathcal{C}}$, $\mathcal{C}^X$ is not empty, where $\mathcal{C}^X$ denotes $Pa_X \cap \mathcal{C}$.*

We now show that *redundant* edges from $\mathcal{G}$ can be removed such that $\mathcal{L}_{\mathcal{C}}$ will be regular w.r.t. the new graph.

Lemma 1. *If $(Y, X) \in \mathcal{L}_{\mathcal{C}}(\mathbf{c})$ and $\mathcal{C}^X = \varnothing$, then $X \perp\!\!\!\perp Y | Pa_X \setminus \{Y\}$ and (Y, X) can be deleted from $\mathcal{G}$.*

To prove Lemma 1, we first prove the following lemma.

Lemma 2. *If $(Y, X) \in \mathcal{L}_{\mathcal{C}}(\mathbf{c})$, then*

$$X \perp\!\!\!\perp Y | \mathcal{C} = \mathbf{c}, Pa_X \setminus (\{Y\} \cup \mathcal{C}).$$

Proof. Let $\mathcal{H}$ be the DAG obtained by removing (Y, X) from $\mathcal{G}$, and $\mathcal{M}_{\mathbf{c}}$ be the SEM over $\mathbf{V} \setminus \mathcal{C}$ when $do(\mathcal{C} = \mathbf{c})$. As $(Y, X) \in \mathcal{L}_{\mathcal{C}}(\mathbf{c})$, $\mathcal{H}$ is a causal graph for $\mathcal{M}_{\mathbf{c}}$. Moreover, since $\mathcal{C} \subseteq \Lambda_{\mathcal{G}}$, we have

$$P^{\mathcal{M}_{\mathbf{c}}}(\mathbf{V} \setminus \mathcal{C}) = P^{\mathcal{M}}(\mathbf{V} \setminus \mathcal{C} | \mathcal{C} = \mathbf{c}). \qquad (1)$$

Because $(Y, X) \in \mathbf{E}$ and $\mathcal{G}$ is a DAG, $X \notin Anc_Y$. Thus, local Markov property (Pearl (2009)) for $\mathcal{M}_{\mathbf{c}}$ implies that $X \perp\!\!\!\perp_{P^{\mathcal{M}_{\mathbf{c}}}} Y | Pa_X \setminus (\{Y\} \cup \mathcal{C})$. Combining this with Equation (1) implies that $X \perp\!\!\!\perp Y | \mathcal{C} = \mathbf{c}, Pa_X \setminus (\{Y\} \cup \mathcal{C})$. $\qquad \square$

Next, we prove Lemma 1.

Proof. Let $\mathbf{A} = Pa_X \setminus \{Y\}$. Since $(Y, X) \in \mathcal{L}_{\mathcal{C}}(\mathbf{c})$, Lemma 2 implies that for any $\mathbf{a} \in D_{\mathbf{A}}$ and $y_1, y_2 \in D_Y$,

$$P(X | \mathbf{c}, \mathbf{a}, y_1) = P(X | \mathbf{c}, \mathbf{a}, y_2).$$

Since $X \perp_{\mathcal{G}} C | Pa_X$, Rule 1 of do-calculus implies that $P(X|\mathbf{c},\mathbf{a},y_i) = P(X|\mathbf{a},y_i)$, for $i \in \{1,2\}$. Thus,

$$P(X|\mathbf{a},y_1) = P(X|\mathbf{a},y_2).$$

Hence, $X \perp\!\!\!\perp Y | \mathbf{A}$. $\qquad\square$

Lemma 1 implies that if an edge (Y,X) belongs to $\mathcal{L}_{\mathcal{C}}$ and $\mathcal{C}^X = \varnothing$, then (Y,X) can be removed from the causal graph, i.e., $\mathcal{G}' = (\mathbf{V}, \mathbf{E} \setminus \{(Y,X)\})$ is still a causal graph for the joint distribution $P(\mathbf{V})$. Therefore, by removing all such edges from $\mathcal{G}$, we obtain a DAG such that for all $(Y,X) \in \mathcal{L}_{\mathcal{C}}$, $\mathcal{C}^X$ is not empty.

Example 3. *Consider the causal DAG in Figure 1 with control variable $\mathcal{C} = \{T\}$, where $D_T = \{0,1\}$ and label set $\mathcal{L}_{\mathcal{C}} = (\mathcal{L}_{\mathcal{C}}(0), \mathcal{L}_{\mathcal{C}}(1))$, where $\mathcal{L}_{\mathcal{C}}(0) = \{(Z,W)\}$ and $\mathcal{L}_{\mathcal{C}}(1) = \{(W,Y),(X,Z)\}$. This label set is not regular w.r.t. $\mathcal{G}$ because $(X,Z) \in \mathcal{L}_{\mathcal{C}}(1)$ but $\mathcal{C}^Z = \varnothing$. However, by removing edge (X,Z) from the DAG, $\mathcal{L}_{\mathcal{C}}$ becomes regular w.r.t. the new DAG.*

The following lemma shows that only the realizations of $\mathcal{C}^X$, e.g., $\mathbf{c}^X$, are relevant to determine whether an edge (Y,X) belongs to $\mathcal{L}_{\mathcal{C}}$ or not.

Lemma 3. *If $(Y,X) \in \mathcal{L}_{\mathcal{C}}(\mathbf{c})$ and $\mathcal{C}^X \neq \varnothing$, then $(Y,X) \in \mathcal{L}_{\mathcal{C}^X}(\mathbf{c}^X)$.*

Proof. Suppose $\mathbf{A} = Pa_X \setminus (\mathcal{C}^X \cup \{Y\})$. Since $(Y,X) \in \mathcal{L}_{\mathcal{C}}(\mathbf{c})$, Lemma 2 implies that for any $\mathbf{a} \in D_{\mathbf{A}}$ and $y_1, y_2 \in D_Y$,

$$P_{\mathbf{c}}(X|\mathbf{a},y_1) = P_{\mathbf{c}}(X|\mathbf{a},y_2).$$

Because variables in $\mathcal{C}$ have no parents then $\mathcal{G}_{\overline{\mathcal{C}^X},\overline{\mathcal{C} \setminus \mathcal{C}^X}(Pa_X \setminus \mathcal{C}^X)} = \mathcal{G}$ and $X \perp_{\mathcal{G}} \mathcal{C} \setminus \mathcal{C}^X | Pa_X$. Hence, rule 3 of do-calculus implies that $P_{\mathbf{c}}(X|Pa_X \setminus \mathcal{C}^X) = P_{\mathbf{c}^X}(X|Pa_X \setminus \mathcal{C}^X)$. By combining the above equations, we obtain

$$P_{\mathbf{c}^x}(X|\mathbf{a},y_1) = P_{\mathbf{c}^x}(X|\mathbf{a},y_2).$$

This concludes that $(Y,X) \in \mathcal{L}_{\mathcal{C}^x}(\mathbf{c}^x)$. $\qquad\square$

Corollary 1. *Suppose $(Y,X) \in \mathcal{L}_{\mathcal{C}}(\mathbf{c}_1)$ for some $\mathbf{c}_1 \in D_{\mathcal{C}}$. Let $\mathbf{c}_2 \in D_{\mathcal{C}}$ such that $\mathbf{c}_1 \neq \mathbf{c}_2$ but $\mathbf{c}_1^{\mathbf{X}} = \mathbf{c}_2^{\mathbf{X}}$, then we can add (Y,X) to $\mathcal{L}_{\mathcal{C}}(\mathbf{c}_2)$.*

Definition 6 (Maximal-regular)**.** *We say $\mathcal{L}_{\mathcal{C}}$ is maximal w.r.t. $\mathcal{G}$ if we cannot add any new label to $\mathcal{L}_{\mathcal{C}}$ using Corollary 1. Label set $\mathcal{L}_{\mathcal{C}}$ is called maximal-regular w.r.t. $\mathcal{G}$ if it is both maximal and regular w.r.t. $\mathcal{G}$.*

Example 4. *Consider the causal graph in Figure 1 with $D_T = \{0,1\}$, $\mathcal{C} = \{T\}$, $\mathcal{L}_{\mathcal{C}}(0) = \{(Z,W)\}$ and $\mathcal{L}_{\mathcal{C}}(1) = \{(W,Y)\}$. In this case, $\mathcal{L}_{\mathcal{C}}$ is maximal-regular w.r.t. this causal graph.*

5 MAIN RESULT

Next theorem states that the identifiablity of $P_{\mathbf{t}}(\mathbf{s})$ given $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$ is equivalent to the identifiability of $P_{\mathbf{t}}(\mathbf{s})$ from a list of DAGs. Note that Tian and Pearl (2003) introduced a sound and complete algorithm for checking whether $P_{\mathbf{t}}(\mathbf{s})$ is identifiable from a DAG. Therefore, using the next theorem, we can develop a sound and complete algorithm for identifiability of $P_{\mathbf{t}}(\mathbf{s})$ from $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$.

Theorem 1. *Suppose $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a DAG with observable variables $\mathbf{N} \subseteq \mathbf{V}$, and let $\mathcal{C} \subseteq \Lambda_{\mathcal{G}}$ and $\mathbf{T}, \mathbf{S} \subseteq \mathbf{N} \setminus \mathcal{C}$ be three disjoint subsets. Furthermore, suppose the set of labels $\mathcal{L}_{\mathcal{C}}$ is maximal-regular w.r.t. $\mathcal{G}$. Causal effect of $\mathbf{T}$ on $\mathbf{S}$ is identifiable from $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$ if and only if the causal effect of $\mathbf{T}$ on $\mathbf{S}$ is identifiable from $\mathcal{G}_{\mathbf{c}} = (\mathbf{V}', \mathbf{E}' \setminus \mathcal{L}_{\mathcal{C}}(\mathbf{c}))$ for every $\mathbf{c} \in D_{\mathcal{C}}$, where $\mathbf{V}' = \mathbf{V} \setminus \mathcal{C}$ and $\mathbf{E}' = \mathbf{E} \cap (\mathbf{V}' \times \mathbf{V}')$.*

Proof. Sufficiency: Suppose $\mathcal{M}$ is a SEM with causal graph $\mathcal{G}$ and compatible with label set $\mathcal{L}_{\mathcal{C}}$. For every $\mathbf{c} \in D_{\mathcal{C}}$, we construct a SEM $\mathcal{M}_{\mathbf{c}}$ over $\mathbf{V}'$ by setting $\mathcal{C}$ to be $\mathbf{c}$ in the equations of $\mathcal{M}$. Hence, $\mathcal{G}_{\mathbf{c}} = (\mathbf{V}', \mathbf{E}' \setminus \mathcal{L}_{\mathcal{C}}(\mathbf{c}))$ is a causal graph for $\mathcal{M}_{\mathbf{c}}$ and

$$P^{\mathcal{M}_{\mathbf{c}}}(\mathbf{N} \setminus \mathcal{C}) = P_{\mathbf{c}}^{\mathcal{M}}(\mathbf{N} \setminus \mathcal{C}) = P^{\mathcal{M}}(\mathbf{N} \setminus \mathcal{C} \mid \mathbf{c}). \quad (2)$$

The above relations hold due to the fact that $\mathcal{C} \subseteq \Lambda_{\mathcal{G}}$, that is, intervened variables do not have parents. As $P_{\mathbf{t}}(\mathbf{S})$ is identifiable from $\mathcal{G}_{\mathbf{c}}$, $P_{\mathbf{t}}^{\mathcal{M}_{\mathbf{c}}}(\mathbf{S})$ is uniquely computable from $P^{\mathcal{M}_{\mathbf{c}}}(\mathbf{N} \setminus \mathcal{C})$, and therefore, from $P^{\mathcal{M}}(\mathbf{N})$. From the definition of $\mathcal{M}_{\mathbf{c}}$, we have

$$P_{\mathbf{t}}^{\mathcal{M}_{\mathbf{c}}}(\mathbf{S}) = P_{\mathbf{c},\mathbf{t}}^{\mathcal{M}}(\mathbf{S}) = P_{\mathbf{t}}^{\mathcal{M}}(\mathbf{S} \mid \mathbf{c}).$$

Hence, $P_{\mathbf{t}}^{\mathcal{M}}(\mathbf{S} \mid \mathbf{c})$ is uniquely computable from $P^{\mathcal{M}}(\mathbf{N})$ for every $\mathbf{c} \in D_{\mathcal{C}}$. On the other hand, for any $\mathbf{t} \in D_{\mathbf{T}}$, we have

$$\begin{aligned} P_{\mathbf{t}}^{\mathcal{M}}(\mathbf{S}) &= \sum_{\mathbf{c} \in D_{\mathcal{C}}} P_{\mathbf{t}}^{\mathcal{M}}(\mathbf{S} \mid \mathbf{c}) P_{\mathbf{t}}^{\mathcal{M}}(\mathbf{c}) \\ &= \sum_{\mathbf{c} \in D_{\mathcal{C}}} P_{\mathbf{t}}^{\mathcal{M}}(\mathbf{S} \mid \mathbf{c}) P^{\mathcal{M}}(\mathbf{c}). \end{aligned} \quad (3)$$

Equation (3) implies that $P_{\mathbf{t}}^{\mathcal{M}}(\mathbf{S})$ is uniquely computable from $P^{\mathcal{M}}(\mathbf{N})$ and the causal effect of $\mathbf{T}$ on $\mathbf{S}$ is identifiable from $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$. Note that the functional that maps $P^{\mathcal{M}}(\mathbf{N})$ to $P_{\mathbf{t}}^{\mathcal{M}}(\mathbf{S})$ in (3) is continuous.

Necessity: For any $\mathbf{c}_0 \in D_{\mathcal{C}}$, we need to show that $P_{\mathbf{t}}(\mathbf{S})$ is identifiable from $\mathcal{G}_{\mathbf{c}_0}$ knowing that $P_{\mathbf{t}}(\mathbf{S})$ is identifiable from $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$. We use proof by contradiction. Suppose there exists two SEMs $\mathcal{M}_1$ and $\mathcal{M}_2$ over $\mathbf{V}'$ with causal graph $\mathcal{G}_{\mathbf{c}_0}$ such that $P^{\mathcal{M}_1}(\mathbf{N} \setminus \mathcal{C}) = P^{\mathcal{M}_2}(\mathbf{N} \setminus \mathcal{C}) > 0$, but there exists $\mathbf{t} \in D_{\mathbf{T}}$ and $\mathbf{s} \in D_{\mathbf{S}}$ such that $P_{\mathbf{t}}^{\mathcal{M}_1}(\mathbf{s}) \neq P_{\mathbf{t}}^{\mathcal{M}_2}(\mathbf{s})$. Let $\delta = |P_{\mathbf{t}}^{\mathcal{M}_1}(\mathbf{s}) -$

$P_{\mathbf{t}}^{\mathcal{M}_2}(\mathbf{s})| > 0$. We now construct two SEMs $\mathcal{M}_i^\epsilon$ for $i \in \{1,2\}$ over $\mathbf{V}$, where $\epsilon > 0$ is a small number.

First, we define $P^{\mathcal{M}_i^\epsilon}(\mathcal{C}) := \prod_{C \in \mathcal{C}} P^{\mathcal{M}_i^\epsilon}(C)$. $\{P^{\mathcal{M}_i^\epsilon}(C) : C \in \mathcal{C}\}$ are selected such that $P^{\mathcal{M}_i^\epsilon}(C = c)$ is positive for all $c \in D_C$ and $P^{\mathcal{M}_i^\epsilon}(\mathbf{c_0}) = 1 - \epsilon$. Next, for each $X \in \mathbf{V} \backslash \mathcal{C}$ we define the equation of X in $\mathcal{M}_i^\epsilon$ as follows.

$$X = \begin{cases} X^{\mathcal{M}_i} & \text{if } \mathcal{C}^X = \mathbf{c_0}^X \\ \sim \text{uniform in } D_X & \text{otherwise,} \end{cases} \quad (4)$$

where $X^{\mathcal{M}_i}$ indicates the random variable X in SEM $\mathcal{M}_i$. Note that $\mathcal{M}_i^\epsilon$ is compatible with $\mathcal{G}$ since X is a function of $\mathcal{C}^X$ and parents of X in $\mathcal{G}_{\mathbf{c_0}}$. Next, we show that $\mathcal{M}_i^\epsilon$ is also compatible with $\mathcal{L}_{\mathcal{C}}(\mathbf{c})$ for each $\mathbf{c} \in D_{\mathcal{C}}$, and therefore, compatible with $\mathcal{L}_{\mathcal{C}}$. For each $X \in \mathbf{V} \backslash \mathcal{C}$, if $\mathbf{c}^X \neq \mathbf{c}_0^X$, then X is not a function of any other variables in $\mathcal{M}_i^\epsilon$. If $\mathbf{c}^X = \mathbf{c}_0^X$, then $X = X^{\mathcal{M}_i}$. In this case, for each edge (Y, X) in $\mathcal{G}_{\mathbf{c_0}}$, $(Y, X) \notin \mathcal{L}_{\mathcal{C}}(\mathbf{c})$ since $\mathcal{L}_{\mathcal{C}}$ is maximal and $(Y, X) \notin \mathcal{L}_{\mathcal{C}}(\mathbf{c_0})$. Hence, $\mathcal{M}_i^\epsilon$ is compatible with $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$.

From the definition of $\mathcal{M}_i^\epsilon$ and the fact that $\mathcal{C} \subseteq \Lambda_{\mathcal{G}}$, we obtain

$$P^{\mathcal{M}_i}(\mathbf{N} \backslash \mathcal{C}) = P^{\mathcal{M}_i^\epsilon}(\mathbf{N} \backslash \mathcal{C} \mid \mathbf{c_0}), \quad (5)$$

$$P_{\mathbf{t}}^{\mathcal{M}_i}(\mathbf{S}) = P_{\mathbf{t}}^{\mathcal{M}_i^\epsilon}(\mathbf{S} \mid \mathbf{c_0}). \quad (6)$$

We can write $P^{\mathcal{M}_i^\epsilon}(\mathbf{N})$ as

$$P^{\mathcal{M}_i^\epsilon}(\mathbf{N} \backslash \mathcal{C}, \mathbf{c_0}) \mathbb{1}_{\{\mathcal{C} = \mathbf{c_0}\}} + P^{\mathcal{M}_i^\epsilon}(\mathbf{N}) \mathbb{1}_{\{\mathcal{C} \neq \mathbf{c_0}\}}.$$

The second term is not larger than $P^{\mathcal{M}_i^\epsilon}(\mathcal{C} \neq \mathbf{c_0})$ which is equal to ϵ. On the other hand, from (5), we obtain

$$P^{\mathcal{M}_i^\epsilon}(\mathbf{N} \backslash \mathcal{C}, \mathbf{c_0}) = P^{\mathcal{M}_i^\epsilon}(\mathbf{c_0}) P^{\mathcal{M}_i^\epsilon}(\mathbf{N} \backslash \mathcal{C} \mid \mathbf{c_0})$$
$$= (1 - \epsilon) P^{\mathcal{M}_i}(\mathbf{N} \backslash \mathcal{C}). \quad (7)$$

Therefore, we have

$$0 \leq P^{\mathcal{M}_i^\epsilon}(\mathbf{N}) - (1 - \epsilon) P^{\mathcal{M}_i}(\mathbf{N} \backslash \mathcal{C}) \mathbb{1}_{\{\mathcal{C} = \mathbf{c_0}\}} \leq \epsilon. \quad (8)$$

Since $P^{\mathcal{M}_1}(\mathbf{N} \backslash \mathcal{C}) = P^{\mathcal{M}_2}(\mathbf{N} \backslash \mathcal{C})$, Equation (8) implies

$$|P^{\mathcal{M}_1^\epsilon}(\mathbf{N}) - P^{\mathcal{M}_2^\epsilon}(\mathbf{N})| \leq \epsilon. \quad (9)$$

Note that $P_{\mathbf{t}}^{\mathcal{M}_i^\epsilon}(\mathbf{c}) = P^{\mathcal{M}_i^\epsilon}(\mathbf{c})$, since $\mathcal{C} \subseteq \Lambda_{\mathcal{G}}$. Hence,

$$P_{\mathbf{t}}^{\mathcal{M}_i^\epsilon}(\mathbf{S}) = \sum_{\mathbf{c}} P_{\mathbf{t}}^{\mathcal{M}_i^\epsilon}(\mathbf{S} \mid \mathbf{c}) P^{\mathcal{M}_i^\epsilon}(\mathbf{c})$$
$$\leq (1 - \epsilon) P_{\mathbf{t}}^{\mathcal{M}_i^\epsilon}(\mathbf{S} \mid \mathbf{c_0}) + \epsilon. \quad (10)$$

Equations (6) and (10) imply that

$$0 \leq P_{\mathbf{t}}^{\mathcal{M}_i^\epsilon}(\mathbf{S}) - (1 - \epsilon) P_{\mathbf{t}}^{\mathcal{M}_i}(\mathbf{S}) \leq \epsilon. \quad (11)$$

Because $\delta = |P_{\mathbf{t}}^{\mathcal{M}_1}(\mathbf{s}) - P_{\mathbf{t}}^{\mathcal{M}_2}(\mathbf{s})|$, Equation (11) implies

$$|P_{\mathbf{t}}^{\mathcal{M}_1^\epsilon}(\mathbf{S}) - P_{\mathbf{t}}^{\mathcal{M}_2^\epsilon}(\mathbf{S})| \geq (1 - \epsilon) \delta - 2\epsilon. \quad (12)$$

Recall that $\mathcal{M}_i^\epsilon$ is compatible with $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$ and $P^{\mathcal{M}_i^\epsilon}(\mathbf{n}) > 0$ for any $\mathbf{n} \in D_{\mathbf{N}}$. Hence, there exists a continuous operator $\mathcal{F}$ such that $\mathcal{F}(P^{\mathcal{M}_i^\epsilon}(\mathbf{N})) = P_{\mathbf{t}}^{\mathcal{M}_i^\epsilon}(\mathbf{S})$. Note that $\mathcal{F}$ does not depend on ϵ. Also, δ is a constant number. In this case, Equations (9) and (12) for sufficiently small ϵ contradict with the continuity of $\mathcal{F}$. $\qquad \square$

As the proof of the sufficiency in Theorem 1 is constructive, it allows us to develop an algorithm (Algorithm 1) for causal effect identification from $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$.


Algorithm 1: Causal effect ID from $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$.

1: **Input:** $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, $\mathcal{L}_{\mathcal{C}}$
2: **Output:** A formula for $P_{\mathbf{t}}(\mathbf{S})$ based on $P(\mathbf{N})$, or return non-identifiable.
   // To make regular:
3: **while** $\exists (Y, X) \in \mathcal{L}_{\mathcal{C}}$ such that $\mathcal{C}^X = \varnothing$ **do**
4:    Delete (Y, X) from $\mathcal{G}$
   // To make maximal:
5: **while** $\exists \mathbf{c}_1, \mathbf{c}_2 \in D_{\mathcal{C}}, (Y, X) \in \mathcal{L}_{\mathcal{C}}(\mathbf{c}_1)$ such that $\mathbf{c}_1^{\mathbf{X}} = \mathbf{c}_2^{\mathbf{X}}$ and $(Y, X) \notin \mathcal{L}_{\mathcal{C}}(\mathbf{c}_2)$ **do**
6:    Add (Y, X) to $\mathcal{L}_{\mathcal{C}}(\mathbf{c}_2)$
7: $\mathbf{V}' \leftarrow \mathbf{V} \backslash \mathcal{C}$
8: $\mathbf{N}' \leftarrow \mathbf{N} \backslash \mathcal{C}$
9: $\mathbf{E}' \leftarrow \mathbf{E} \cap (\mathbf{V}' \times \mathbf{V}')$
10: **for** $\mathbf{c} \in D_{\mathcal{C}}$ **do**
11:    $\mathcal{G}_{\mathbf{c}} \leftarrow (\mathbf{V}', \mathbf{E}' \backslash \mathcal{L}_{\mathcal{C}}(\mathbf{c}))$
12:    **if** $P_{\mathbf{t}}(\mathbf{S})$ is not ID from $\mathcal{G}_{\mathbf{c}}$ **then**
13:       **Return** Non-identifiable.
14:    **else**
15:       $P^{\mathbf{c}}(\mathbf{N}') \leftarrow P(\mathbf{N} \backslash \mathcal{C} \mid \mathbf{c})$
16:       $\mathbf{F}_{\mathbf{c}} \leftarrow$ A formula for $P_{\mathbf{t}}(\mathbf{S})$ based on $P^{\mathbf{c}}(\mathbf{N}')$ using graph $\mathcal{G}_{\mathbf{c}}$
17: **Return** $\sum_{\mathbf{c} \in D_{\mathcal{C}}} \mathbf{F}_{\mathbf{c}} P(\mathbf{c})$.

Algorithm 1 regularizes the labels in lines 3-4 and makes them maximal in lines 5-6 using Lemma 1 and Corollary 1, respectively. Then, based on Theorem 1, for each $\mathbf{c} \in D_{\mathcal{C}}$ it checks the identifiability of $P_{\mathbf{t}}(\mathbf{S})$ in DAG $\mathcal{G}_{\mathbf{c}} = (\mathbf{V}', \mathbf{E}' \backslash \mathcal{L}_{\mathcal{C}}(\mathbf{c}))$. If for any $\mathbf{c}$, this is non-identifiable, the algorithm terminates and outputs non-identifiable. Otherwise, it defines the distribution $P^{\mathbf{c}}$ over $\mathbf{N}'$ in line 15 based on Equation (2). Since $P_{\mathbf{t}}(\mathbf{S})$ is identifiable from $\mathcal{G}_{\mathbf{c}}$, the algorithm finds a formula for $P_{\mathbf{t}}(\mathbf{S})$ in terms of $P^{\mathbf{c}}(\mathbf{N}')$ in line 16. This step can be performed using any do-calculus based method. Finally, Algorithm 1 outputs a formula for $P_{\mathbf{t}}(\mathbf{S})$ in terms of $P(\mathbf{N})$ in line 17, using Equation (3).

Corollary 2. *Algorithm 1 for the problem of causal effect identification from $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$ is sound and complete.*

$$\mathcal{C} = \{T\}$$
$$\mathcal{L}_\mathcal{C}(0) = \{(Z_1, Y)\}$$
$$\mathcal{L}_\mathcal{C}(1) = \{(X_2, Z_2)\}$$
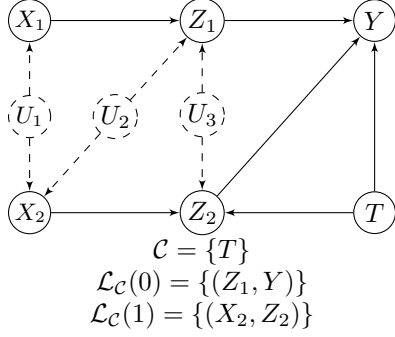
Figure 3: Causal graph and $\mathcal{L}_\mathcal{C}$ in Example 5.

Example 5. *Consider the causal graph with control variable T, $D_T = \{0, 1\}$, and the label set $\mathcal{L}_\mathcal{C}$ depicted in Figure 3. Note that $\mathcal{L}_\mathcal{C}$ is maximal-regular. In this case, the causal effect of $\{X_1, X_2\}$ on Y, i.e., $P_{x_1, x_2}(Y)$ is non-identifiable from the graph alone. However, it is identifiable when we have access to both the graph and the label set $\mathcal{L}_\mathcal{C}$. In this case, Algorithm 1 finds the following formula.*

$$P_{x_1, x_2}(Y) = \mathbf{F}_0 P(T = 0) + \mathbf{F}_1 P(T = 1),$$

where

$$\mathbf{F}_0 = \sum_{Z_2} P(Y \mid Z_2, T = 0) P(Z_2 \mid X_2, T = 0),$$

$$\mathbf{F}_1 = \sum_{Z_1, Z_2} P(Y \mid Z_1, Z_2, T = 1) P(Z_1, Z_2 \mid X_1, T = 1).$$

6 LEARNING $\mathcal{L}_\mathcal{C}$ FROM $P(\mathbf{N})$

In the previous section, we proposed an algorithm for causal effect identification, when a set of labels $\mathcal{L}_\mathcal{C}$ are given as side information. In this section, we propose Algorithm 2 for inferring these labels from observational distribution.

Algorithm 2: Learning $\mathcal{L}_\mathcal{C}$ from $P(\mathbf{N})$.

1: **Input:** $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, $P(\mathbf{N})$
2: **Output:** $\mathcal{L}_\mathcal{C}$
3: **for** $\mathbf{c} \in D_\mathcal{C}$ **do**
4:    **for** $X, Y \in \mathbf{N}$ such that $(Y, X) \in \mathbf{E}$ **do**
5:       **if** $X \perp\!\!\!\perp Y | \mathbf{c}, Anc_{\{X,Y\}} \cap \mathbf{N} \setminus \mathcal{C}$ **then**
6:          Add (Y, X) to $\mathcal{L}_\mathcal{C}(\mathbf{c})$
7: **Return** $\mathcal{L}_\mathcal{C}$

In order to add an edge $(Y, X) \in \mathbf{E}$ to $\mathcal{L}_\mathcal{C}(\mathbf{c})$, Algorithm 2 evaluates the following CSI in line 5.

$$X \perp\!\!\!\perp Y | \mathbf{c}, Anc_{\{X,Y\}} \cap \mathbf{N} \setminus \mathcal{C}. \qquad (13)$$

Note that the conditioning set in Equation (13) is a subset of observed variables. Hence, we can evaluate this CSI from $P(\mathbf{N})$ when X and Y are observed.

If Equation (13) holds for $(Y, X) \in \mathbf{E}$, by Definition 2, $(Y, X) \in \mathcal{L}_\mathcal{C}(\mathbf{c})$, which proves the soundness of the algorithm. Although the reverse does not hold in general, we present a graphical constraint in the following theorem, under which the reverse also holds true, and consequently, the algorithm is complete.

Theorem 2. *Suppose $(Y, X) \in \mathbf{E}$ such that X is not a function of Y when $\mathcal{C} = \mathbf{c}$. Let $\mathcal{H}$ be the DAG obtained by removing (Y, X) from $\mathcal{G}$. If I) $X, Y \in \mathbf{N}$ and II) X and Y are d-separable in $\mathcal{H}$ given subsets of $\mathbf{N}$, then Algorithm 2 correctly add (Y, X) to $\mathcal{L}_\mathcal{C}(\mathbf{c})$.*

To prove this theorem, we use the following definition and Lemma by Verma and Pearl (1991).

Definition 7 (Inducing path). *A path $\mathcal{P}$ between two observed variables X and Y is called an inducing path if I) every non-endpoint observed vertex in $\mathcal{P}$ is a collider[2], and II) every collider in $\mathcal{P}$ belongs to $Anc_{\{X,Y\}}$.*

Lemma 4 (Verma and Pearl (1991)). *There is no inducing path between X and Y in $\mathcal{H}$ if and only if $X \perp\!\!\!\perp_\mathcal{H} Y | (Anc_{\{X,Y\}}) \cap \mathbf{N}$.*

Now, we prove Theorem 2.

Proof. X and Y are d-separable in $\mathcal{H}$ by subsets of $\mathbf{N}$. This is equivalent to saying there is no inducing path between X and Y in $\mathcal{H}$ (Verma and Pearl (1991)). Also, Lemma 4 presents a necessary and sufficient graphical condition for this constraint. Hence, Equation (13) holds and Algorithm 2 correctly adds (Y, X) to $\mathcal{L}_\mathcal{C}(\mathbf{c})$ in line 6. □

The following example demonstrates a scenario where Algorithm 2 can infer a label set from the observational distribution such that a non-identifiable causal effect (using only the information from the causal graph) becomes identifiable (using both the graph and the inferred label set).

Example 6. *Consider again the example in Figure 3. Since I) $Z_1, Y \in \mathbf{N}$ and II) Z_2 d-separates Z_1 and Y in the resulting graph after removing (Z_1, Y), Theorem 2 implies that Algorithm 2 can infer $\mathcal{L}_\mathcal{C}(0)$. Similarly, I) $X_2, Z_2 \in \mathbf{N}$ and II) X_2 and Z_2 are d-separated by empty set in the resulting graph after removing (X_2, Z_2). Hence, Algorithm 2 can infer $\mathcal{L}_\mathcal{C}(1)$, and therefore, $\mathcal{L}_\mathcal{C}$. As we showed in Example 5, Algorithm 1 can then identify the causal effect of $\{X_1, X_2\}$ on Y while this causal effect is non-identifiable from the graph alone.*

---

[2]A vertex V_i is called a collider on a path $(V_1, \ldots V_k)$ if $1 < i < k$ and $V_{i-1} \rightarrow V_i \leftarrow V_{i+1}$.

$$\mathcal{C} = \{C\}$$
$$\mathcal{L}_{\mathcal{C}}(0) = \{(U, X)\}$$
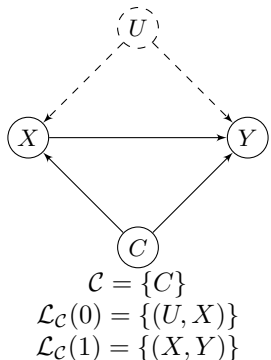$$\mathcal{L}_{\mathcal{C}}(1) = \{(X, Y)\}$$

Figure 4: Causal graph and $\mathcal{L}_{\mathcal{C}}$ in Example 7.

7 COMPLEXITY

Algorithm 1 determines the identifiability of $P_{\mathbf{t}}(\mathbf{s})$ from $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$ by checking the identifiability of $P_{\mathbf{t}}(\mathbf{s})$ in $\mathcal{G}_{\mathbf{c}} = (\mathbf{V}', \mathbf{E}' \setminus \mathcal{L}_{\mathcal{C}}(\mathbf{c}))$ for each $\mathbf{c} \in D_{\mathcal{C}}$. On the other hand, Tian and Pearl (2003) proposed a sound and complete algorithm to check the identifiability of a causal effect from a causal graph. The complexity of this algorithm is polynomial in the number of vertices of the graph. Therefore, Algorithm 1 determines the identifiability of $P_{\mathbf{t}}(\mathbf{s})$ from $(\mathcal{G}, \mathcal{L}_{\mathcal{C}})$ by calling Tian and Pearl (2003)'s algorithm at most $|D_{\mathcal{C}}|$ number of times. Consequently, when the number of control variables does not grow by the number of vertices in the causal graph (e.g., $|\mathcal{C}| = r$ for some fixed r), complexity of Algorithm 1 will also be polynomial in the number of vertices of the graph.

In the next example, we compare our algorithm with the method in Tikka et al. (2019).

Example 7. *Consider the causal graph with control variable C, $D_C = \{0, 1\}$, and the label set $\mathcal{L}_{\mathcal{C}}$ depicted in Figure 3. Herein, the goal is to identify the causal effect of X on Y. In this case, Algorithm 1 finds the following formula.*

$$P_x(Y) = \mathbf{F}_0 P(C = 0) + \mathbf{F}_1 P(C = 1), \quad (14)$$

where

$$\mathbf{F}_0 = P(Y \mid X, C = 0),$$
$$\mathbf{F}_1 = P(Y \mid C = 1).$$

Algorithm 1 finds $\mathbf{F}_0$ and $\mathbf{F}_1$ by solving only two identifiablity problems from two graphs with vertices $\{X, Y\}$. On the other hand, Tikka et al. (2019) provide eight rules $(R1 - R8)$ that can be applied recursively to compute $P_x(Y)$ from $P(X, Y, C)$. In this example, there are many ways that these rules can be applied. For example, $(R2 \rightarrow R1 \rightarrow R1 \rightarrow R7 \rightarrow R4 \rightarrow R6 \rightarrow R2)$ or $(R2 \rightarrow R7 \rightarrow R3 \rightarrow R1 \rightarrow R1 \rightarrow R4 \rightarrow R6 \rightarrow R2)$ (See Figure 4 in Tikka et al. (2019) for more details.)

Applying the eight rules with these specific orders will result in the same formula as in Equation (14). It is worthy to note that finding a correct order for applying these rules is challenging and can be computationally expensive. Therefore, our algorithm finds the solution faster than the method in Tikka et al. (2019).

8 EXPERIMENTS

The MATLAB implementation of Algorithm 1 along with our implementation of the algorithm by Tian and Pearl (2003) are publicly available[3]. In this section, we illustrate the performance of these methods over a set of random graphs in Figure 5. Next, we describe the settings of this experiment.

The variables are assumed to be binary. The skeleton of the graphs is generated from Erdos-Renyi model $G(n, p)$ (Erdős and Rényi (1960)), where n is the number of variables and p is the probability of an edge. n is chosen in a wide range from 30 to 100, and $p = \frac{\log(n)}{n}$. After constructing the skeleton, the edges are oriented based on a random ordering over the vertices. Each variable belongs to the set of observed variables with probability 0.7. Also, each variable in $\Lambda_{\mathcal{G}}$ is randomly selected as a control variable with probability 0.8. To make the model simpler, we replace the control variables $\mathcal{C}$ with one variable C, such that the children of C are the union of children of the variables in $\mathcal{C}$. To build the labels, for each edge $(Y, X) \in \mathbf{E}$ such that $C \in Pa_X$, either we add (Y, X) to no $\mathcal{L}_C(c)$ with probability 0.2, or otherwise, we add it to one of the labels $\mathcal{L}_C(c)$ uniformly at random. Each point on the plots is reported as the average of 1000 runs, and the shaded areas indicate the 80% confidence intervals. We randomly split $\mathbf{N}$ into two subsets $\mathbf{S}$ and $\mathbf{T}$. Then, we run both algorithms to identify the causal effect $P_{\mathbf{t}}(\mathbf{s})$.

Figure 5a shows the runtime of the algorithms in seconds. As this figure suggests, both of these algorithms are practically fast and are scalable to large graphs. Figure 5b shows the percentage of the runs in which the corresponding algorithm manages to identify the causal effect. This figure shows that given the label set $\mathcal{L}_{\mathcal{C}}$, more causal effects will be identifiable than the case where only the causal graph is given.

We did not report the results of the algorithm by Tikka et al. (2019) in this figure since their method is not scalable to large graphs. For instance, even for the graphs of size 10, their algorithm sometimes requires 30 minutes of runtime to terminate, while our method requires around 0.02 seconds to terminate for the graphs with 100 variables.

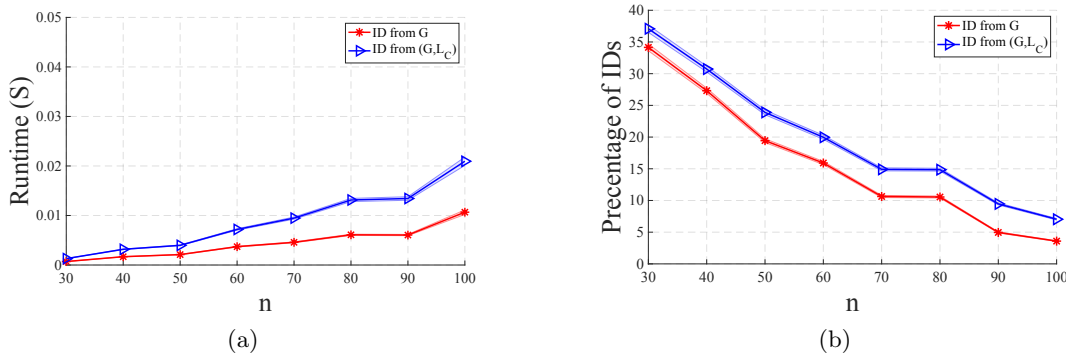
[3]https://github.com/Ehsan-Mokhtarian/causalID

Figure 5: Performance of Algorithm 1 and the identification algorithm by Tian and Pearl (2003) over random graphs generated from Erdos-Renyi models.

9 CONCLUSION

We studied the causal effect identification problem when extra side information about the underlying generative causal model in the form of CSI relations is available. To this end, we showed that when CSI relations of control variables are given, the identifiability of an interventional distribution from observational distribution is equivalent to a series of causal effect identifications only from causal graphs. Since there exist sound and complete algorithms for causal effect identification only from a causal graph, we could develop a sound and complete algorithm for causal effect identification in the presence of CSI relations. Although, such CSI relations in general cannot be inferred from observational distribution, we introduce a graphical constraint under which CSI relations of control variables can be inferred from the observation distribution.

References

Barash, Y. and Friedman, N. (2002). Context-specific bayesian clustering for gene expression data. *Journal of Computational Biology*, 9(2):169–191.

Bareinboim, E. and Pearl, J. (2012). Causal inference by surrogate experiments: Z-identifiability. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, page 113–120, Arlington, Virginia, USA. AUAI Press.

Boutilier, C., Friedman, N., Goldszmidt, M., and Koller, D. (1996). Context-specific independence in bayesian networks. *Uncertainty in Artificial Intelligence*.

Chavira, M. and Darwiche, A. (2008). On probabilistic inference by weighted model counting. *Artificial Intelligence*, 172(6-7):772–799.

Chickering, D. M., Heckerman, D., and Meek, C. (1997). A bayesian approach to learning bayesian networks with local structure. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc.

Dal, G. H., Laarman, A. W., and Lucas, P. J. (2018). Parallel probabilistic inference by weighted model counting. In *International Conference on Probabilistic Graphical Models*, pages 97–108. PMLR.

Edwards, D. and Toma, H. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika*, 72(2):339–351.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61.

Huang, Y. and Valtorta, M. (2008). On the completeness of an identifiability algorithm for semi-markovian models. *Annals of Mathematics and Artificial Intelligence*, 54(4):363–408.

Hyttinen, A., Pensar, J., Kontinen, J., and Corander, J. (2018). Structure learning for bayesian networks over labeled dags. In *International Conference on Probabilistic Graphical Models*, pages 133–144. PMLR.

Lee, S. and Bareinboim, E. (2020). Causal effect identifiability under partial-observability. In *International Conference on Machine Learning*, pages 5692–5701. PMLR.

Lee, S., Correa, J. D., and Bareinboim, E. (2019). General identifiability with arbitrary surrogate experiments. In *Uncertainty in Artificial Intelligence*, pages 389–398. PMLR.

Mokhtarian, E., Akbari, S., Ghassami, A., and Kiyavash, N. (2021). A recursive markov boundary-

based approach to causal structure learning. In *The KDD'21 Workshop on Causal Discovery*, pages 26–54. PMLR.

Nyman, H., Pensar, J., Koski, T., and Corander, J. (2014). Stratified graphical models-context-specific independence in graphical models. *Bayesian Analysis*, 9(4):883–908.

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

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

Pensar, J., Nyman, H., Koski, T., and Corander, J. (2015). Labeled directed acyclic graphs: a generalization of context-specific independence in directed graphical models. *Data mining and knowledge discovery*, 29(2):503–533.

Ramanan, N. and Natarajan, S. (2020). Causal learning from predictive modeling for observational data. *Frontiers in big Data*, 3:34.

Robins, J. M., Richardson, T. S., and Shpitser, I. (2020). An interventionist approach to mediation analysis. *arXiv preprint arXiv:2008.06019*.

Shimony, S. E. (1991). Explanation, irrelevance and statistical independence. In *Proceedings of the ninth National conference on Artificial intelligence-Volume 1*, pages 482–487.

Shpitser, I. and Pearl, J. (2006). *Identification of joint interventional distributions in recursive semi-Markovian causal models*. eScholarship, University of California.

Tian, J. and Pearl, J. (2003). On the identification of causal effects. Technical report, Department of Computer Science, University of California.

Tikka, S., Hyttinen, A., and Karvanen, J. (2019). Identifying causal effects via context-specific independence relations. *Advances in Neural Information Processing Systems 32*.

Verma, T. and Pearl, J. (1991). *Equivalence and synthesis of causal models*. UCLA, Computer Science Department Los Angeles, CA.

Zhang, J. and Bareinboim, E. (2021). Bounding causal effects on continuous outcome. In *Proceedings of the 35nd AAAI Conference on Artificial Intelligence*.