Causal Effect Identification from Multiple Incomplete Data Sources: A General Search-based Approach

Santtu Tikka, santtu.tikka@jyu.fi
Department of Mathematics and Statistics
University of Jyvaskyla, Finland

Antti Hyttinen, antti.hyttinen@helsinki.fi
HIIT, Department of Computer Science
University of Helsinki, Finland

Juha Karvanen, juha.t.karvanen@jyu.fi
Department of Mathematics and Statistics
University of Jyvaskyla, Finland

Abstract

Causal effect identification considers whether an interventional probability distribution can be uniquely determined without parametric assumptions from measured source distributions and structural knowledge on the generating system. While complete graphical criteria and procedures exist for many identification problems, there are still challenging but important extensions that have not been considered in the literature. To tackle these new settings, we present a search algorithm directly over the rules of do-calculus. Due to generality of do-calculus, the search is capable of taking more advanced data-generating mechanisms into account along with an arbitrary type of both observational and experimental source distributions. The search is enhanced via a heuristic and search space reduction techniques. The approach, called `do-search`, is provably sound, and it is complete with respect to identifiability problems that have been shown to be completely characterized by do-calculus. When extended with additional rules, the search is capable of handling missing data problems as well. With the versatile search, we are able to approach new problems such as combined transportability and selection bias, or multiple sources of selection bias. We also perform a systematic analysis of bivariate missing data problems and study causal inference under case-control design.

A modification of (Tikka et al., 2019).

1 Introduction

A causal effect is defined as the distribution $P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{Z})$ where variables $\mathbf{Y}$ are observed, variables $\mathbf{X}$ are intervened upon (forced to values irrespective of their natural causes) and variables $\mathbf{Z}$ are conditioned on. Instead of placing various parametric restrictions based on background knowledge, we are interested in this paper in the question of identifiability: can the causal effect be uniquely determined from the distributions (data) we have and a graph representing our structural knowledge on the generating causal system.

In the most basic setting we are identifying causal effects from a single observational input distribution, corresponding to passively observed data. To solve such problems more generally than what is possible with the back-door adjustment (Spirtes et al., 1993; Pearl, 2009; Greenland et al., 1999), Pearl (1995) introduced *do-calculus*, a set of three rules that together with probability theory enable the manipulation of interventional distributions. Shpitser and Pearl (2006a) and Huang and Valtorta (2006) showed that do-calculus is complete by presenting polynomial-time algorithms whose each step can be seen as a rule of do-calculus or as an operation based on basic probability theory. The algorithms have a high practical value because the rules of do-calculus do not by themselves provide an indication on the order in which they should be applied. The algorithms save us from manual application of do-calculus, which is a tedious task in all but the simplest problems.

Since then many extensions of the basic identifiability problem have appeared. In identifiability using surrogate experiments (Bareinboim and Pearl, 2012b), or z-identifiability, an experimental distribution is available in addition to the observed probability distribution. For data observed in the presence of selection bias, both algorithmic and graphical identifiability results have been derived (Bareinboim and Tian, 2015; Correa et al., 2018). More generally, the presence of missing data necessitates the representation of the missingness mechanism, which poses additional challenges (Mohan et al., 2013; Shpitser et al., 2015). Another dimension of complexity is the number of available data sources. Identification from a mixture of observational and interventional distributions that originate from multiple conceptual domains is known as transportability for which complete solutions exist in a specific setting (Bareinboim and Pearl, 2014). Most of these algorithms are implemented in the R package `causaleffect` (Tikka and Karvanen, 2017a).

While completeness has been accomplished for a number of basic identifiability problems, there are still many challenging but important extensions to the identifiability problem that have not been studied so far. Table 1 recaps the current state of the art identifiability results; it also describes generalizations that we aim to investigate in this paper. To find solutions to the more complicated identifiability problems, we present a unified approach to the identification of observational and interventional causal queries by constructing a search algorithm that directly applies the rules of do-calculus. We impose no restrictions to the number or type of known input distributions: we thus provide a solution to problems for which no algorithmic solutions exist (row 7 in Table 1). We also extend to identifiability under missing data together with mechanisms related to selection bias and transportability (row 10 in Table 1).

To combat the inherent computational complexity of such a search-based approach, we derive rules and techniques that avoid unnecessary steps. We also present a search heuristic that considerably speeds up the search in the cases where the effect is indeed identifiable. The approach, called `do-search`, is provably sound and it retains the completeness in the cases previously proven to be solved by do-calculus rules. We can easily scale up to the problems sizes commonly reported in the literature. An R package (R Core Team, 2018) implementing `do-search` is also available on CRAN at:

`https://CRAN.R-project.org/package=dosearch`

	Problem (Reference)	**Target**	**Input (assumptions)**	**Missing data pattern**	**Solution (complete)**
1	Causal effect identifiability (Shpitser and Pearl, 2006a)	$P(\mathbf{Y} \mid \text{do}(\mathbf{X}))$	$P(\mathbf{V})$	None	ID (Yes)
2	Causal effect identifiability (Shpitser and Pearl, 2006b)	$P(\mathbf{Y} \mid \text{do}(\mathbf{X}), \mathbf{Z})$	$P(\mathbf{V})$	None	IDC (Yes)
3	z-identifiability (Bareinboim and Pearl, 2012b)	$P(\mathbf{Y} \mid \text{do}(\mathbf{X}), \mathbf{Z})$	$P(\mathbf{V})$, $P(\mathbf{V} \setminus \mathbf{B}_i \mid \text{do}(\mathbf{B}))$ (NE, ED)	None	zID (Yes)
4	mz-transportability (Bareinboim and Pearl, 2014)	$P(\mathbf{Y} \mid \text{do}(\mathbf{X}), \mathbf{Z})$	$\{P(\mathbf{V} \setminus (\mathbf{B}_i \cup \mathbf{T}_i) \mid \text{do}(\mathbf{B}_i), \mathbf{T}_i)\}$ (NEDD, ED)	None	TR^{mz} (Yes)
5	Surrogate outcome identifiability (Tikka and Karvanen, 2018b)	$P(\mathbf{Y} \mid \text{do}(\mathbf{X}), \mathbf{Z})$	$\{P(\mathbf{A}_i \mid \text{do}(\mathbf{B}_i), \mathbf{C}_i)\}$ (NE, SO)	None	TRSO (No)
6	Selection bias recoverability (Bareinboim and Tian, 2015)	$P(\mathbf{Y} \mid \text{do}(\mathbf{X}), \mathbf{Z})$	$P(\mathbf{V} \mid S)$	Selection	RC (?)
7	***Generalized identifiability***	$P(\mathbf{Y} \mid \text{do}(\mathbf{X}), \mathbf{Z})$	$\{P(\mathbf{A}_i \mid \text{do}(\mathbf{B}_i), \mathbf{C}_i)\}$	***None***	***None***
8	Missing data recoverability (Mohan et al., 2013)	$P(\mathbf{V})$	$P(\mathbf{V}^*)$	Restricted	Thm. 2 (Yes)
9	Missing data recoverability (Shpitser et al., 2015)	$P(\mathbf{V})$	$P(\mathbf{V}^*)$	Arbitrary	MID (?)
10	***Generalized identifiability with missing data***	$P(\mathbf{Y} \mid \text{do}(\mathbf{X}), \mathbf{Z})$	$\{P(\mathbf{A}_i^* \mid \text{do}(\mathbf{B}_i), \mathbf{C}_i^*)\}$	***Arbitrary***	***None***

Table 1: Solved and unsolved problems (in bold italic) in causal identification. Input $P(\mathbf{V})$ stands for passively observed joint distribution of all variables. Input $P(\mathbf{V}^*)$ is the joint distribution with missing data (see Section 5). Input $P(\mathbf{V} \mid S)$ means the joint distribution under selection bias. Input $P(\mathbf{V} \setminus \mathbf{B} \mid \text{do}(\mathbf{B}))$ stands for an experiment where all variables are measured and input $P(\mathbf{A} \mid \text{do}(\mathbf{B}))$ stands for an experiment where only a subset $\mathbf{A} \subset \mathbf{V}$ of the variables is measured. Notation $\{\cdot\}$ denotes a set of inputs enumerated by the index i. The variable sets present in the same distribution are disjoint. The assumptions of nested experiments (NE), entire distributions (ED) and nested experiments in different domains (NEDD) are explained in Section 2. Assumptions related to surrogate outcomes (SO) can be found in (Tikka and Karvanen, 2018b). The last column tells the algorithm or result that can be used to solve the problem and whether it provides a complete solution to the problem, or whether the completeness status is not known (?). An algorithm is complete if it returns a formula when the target query is identifiable. Problems 1–6 are special cases of problem 7 and problems 1–9 are special cases of problem 10.

The paper is structured as follows. Section 2 formulates our general search problem and explains the scenarios in Table 1 and previous research in detail. Section 3 presents the search algorithm, including the rules we use, search space reduction techniques, heuristics, theoretical properties, and finally simulations that demonstrate the efficacy of the search. Section 4 shows a number of new problems for which we can find solutions by using the search. These problems include combined transportability and selection bias, multiple sources of selection bias, and causal effect identification from arbitrary (experimental) distributions. Section 5 shows how the search can be extended to problems that involve missing data. This section also includes a systematic analysis of missing data problems and case-control designs. Section 6 discusses the merits and limitations of the approach. Section 7 offers concluding remarks.

2 The General Causal Effect Identification Problem

Our presentation is based on Structural Causal Models (SCM) and the language of directed graphs. We assume the reader to be familiar with these concepts and refer them to detailed works on these topics for extended discussion and descriptions, such as (Pearl, 2009) and (Koller and Friedman, 2009). Following the standard set-up of do-calculus (Pearl, 1995), we assume that the causal structure can be represented by a *semi-Markovian causal graph* G over a set of vertices $\mathbf{V}$ (see Fig 1(a) for example). The directed edges correspond to direct causal relations between the variables (relative to $\mathbf{V}$); directed edges do not form any cycles. Confounding of any two observed variables in $\mathbf{V}$ by some unobserved common cause is represented by a bidirected edge between the variables.

In a non-parametric setting, the problem of expressing a causal quantity of interest in terms of available information has been be described in various ways depending on the context. When available data are affected by selection bias or missing data, a typical goal is to "recover" some joint or marginal distributions. If data are available from multiple conceptual domains, a distribution is "transported" from the source domains, from which a combination of both observational and experimental data are available, to a target domain. The aforementioned can be expressed in the SCM framework by equipping the graph of the model with special vertices. However, on a fundamental level these problems are simply variations of the original identifiability problem of causal effects and as such, our goal is to represent them as a single generalized identifiability problem. Formally, identifiability can be defined as follows (Pearl, 2009; Shpitser and Pearl, 2008).

Definition 1 (Identifiability)**.** *Let* $\mathbf{M}$ *be a set of models with a description* T *and two objects* ϕ *and* θ *computable from each model. Then* ϕ *is* identifiable *from* θ *in* T *if* ϕ *is uniquely computable from* θ *in any model* $M \in \mathbf{M}$*. In other words, all models in* $\mathbf{M}$ *which agree on* θ *also agree on* ϕ*.*

In the simplest case, the description T refers to the graph induced by causal model, θ is the joint distribution of the observed variables $P(\mathbf{V})$ and the query ϕ is a causal effect $P(Y \,|\, \mathrm{do}(X))$. On the other hand, proving non-identifiability of ϕ from θ can be obtained by describing two models $M^1, M^2 \in \mathbf{M}$ such that θ is the same in M^1 and M^2, but object ϕ in M^1 is different from ϕ in M^2.

The general form for a causal identifiability problem that we consider in this paper is formulated as follows.

Input: A set of input distributions of the form $P(\mathbf{A}_i \,|\, \mathrm{do}(\mathbf{B}_i), \mathbf{C}_i)$, a query $P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{Z})$ and a semi-Markovian causal graph G over $\mathbf{V}$.

Task: Output a formula for the query $P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{Z})$ over the input distributions, or decide that it is not identifiable.

Here $\mathbf{A}_i, \mathbf{B}_i, \mathbf{C}_i$ are disjoint subsets of $\mathbf{V}$ for all i, and $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are disjoint subsets of $\mathbf{V}$. The causal graph G may contain vertices which describe mechanisms related to transportability and selection bias. In the following subsections we explain several important special cases of this problem definition, some that have been considered in the literature and some which have not been.

2.1 Previously Considered Scenarios as Special Cases

We restate the concepts of transportability and selection bias under the causal inference framework, and show that identifiability in the scenarios of rows 1–6 of Table 1 falls under the general form on row 7. We return to problems that involve missing data on rows 8–10 later in Section 5.

Causal Effect Identification Input is restricted to a passive observational distribution $P(\mathbf{V})$. The target is either a causal effect $P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}))$ for row 1 of Table 1 or a conditional causal effect $P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{Z})$ for row 2 of Table 1 (Shpitser and Pearl, 2006a,b).

z-identifiability Similarly to ordinary causal effect identification, the input consists of the passive observational distribution $P(\mathbf{V})$ but also of experimental distributions known as surrogate experiments on a set $\mathbf{B}$ (Bareinboim and Pearl, 2012b). Two restricting assumptions, called here nested experiments and entire distributions, apply to surrogate experiments. Experiments are called nested experiments (NE) when for each experiment intervening a set of variables $\mathbf{B}$, experiments intervening on all subsets of $\mathbf{B}$ are available as well. Entire distributions (ED) denote the assumption that the union of observed and intervened variables is always the set of all variables $\mathbf{V}$.

Surrogate Outcome Identifiability Surrogate outcomes generalize the notion of surrogate experiments from z-identifiability. For surrogate outcomes, the assumption of nested experiments still holds, but the assumption of entire distributions can be dropped. Some less strict assumptions (SO) still apply (Tikka and Karvanen, 2018b). The idea of surrogate outcomes is that data from previous experiments are available, but the target $\mathbf{Y}$ was at most only partially measured in these experiments and the experiments do not have to be disjoint from $\mathbf{X}$.

Transportability The problem of incorporating data from multiple causal domains is known as transportability (Bareinboim and Pearl, 2013). Formally, the goal is to identify a query in a target domain π^* using data from source domains $\pi_1, \ldots, \pi_n$. The domains are represented in the causal graph using a special set of transportability nodes $\mathbf{T}$ which is partitioned into disjoint subsets $\mathbf{T}_1, \ldots, \mathbf{T}_n$ corresponding to each domain π_i. The causal graph contains an extra edge $T_{ij} \to V_j$ whenever a functional discrepancy in f_{V_j} or in $P(u_{V_j})$ exists between the target domain π^* and the source domain π_i. The discrepancy is active if $T_{ij} = 1$ and inactive otherwise. A distribution associated with a domain π_i is of the form $P(\mathbf{A} \,|\, \mathrm{do}(\mathbf{B}), \mathbf{C}, \mathbf{T}_i = 1, \mathbf{T}_{-i} = 0)$. In other words, only the discrepancies between the π_i and π^* are active. A distribution corresponding to the target domain has no active discrepancies meaning that it is of the form $P(\mathbf{A} \,|\, \mathrm{do}(\mathbf{B}), \mathbf{C}, \mathbf{T} = 0)$. Any variable is conditionally independent from inactive transportability nodes since their respective edges vanish. Furthermore, since transportability nodes set to 0 vanish, we can assume any present transportability node to have the value 1. Thus an input distribution from a domain π_i takes the form $P(\mathbf{A} \,|\, \mathrm{do}(\mathbf{B}), \mathbf{C}, \mathbf{T}_i)$. In the specific case of mz-transportability, the assumptions of entire distributions (ED) and nested experiments in different domains (NEDD) apply, which means that $P(\mathbf{V} \setminus (\mathbf{B}'_i \cup \mathbf{T}_i) \,|\, \mathrm{do}(\mathbf{B}'_i), \mathbf{T}_i)$ is available for every subset $\mathbf{B}'_i$ of $\mathbf{B}_i$ in each domain π_i.

Selection Bias Recoverability Selection bias can be seen as a special case of missing data, where the mechanism responsible for the preferential selection is represented in the causal graph by a special sink vertex S (Bareinboim and Pearl, 2012a). Typical input for the recoverability problem is $P(\mathbf{V} \,|\, S = 1)$, the joint distribution observed under selection bias. Just as in the case of transportability nodes, selection bias nodes only appear when the mechanism has been enabled. Thus we may assume that the input is of form $P(\mathbf{V} \,|\, S)$. More generally, we can consider input distributions of the form $P(\mathbf{A} \,|\, \mathrm{do}(\mathbf{B}), \mathbf{C}, S)$.

2.2 New Scenarios as Special Cases

The following settings are special cases of the general identifiability problem of row 7 in Table 1, that do not fall under any of the problems of rows 1–6. They serve as interesting additions to the cases considered in the literature. Concrete examples on these new scenarios are presented in Section 4. Section 5 extends the general problem of row 7 in Table 1 to the general problem with missing data on row 10 while also showcasing the special cases of rows 8 and 9.

Multiple Data Sources with Partially Overlapping Variable Sets The scenario where only subsets of variables are ever observed together has been extensively considered in the causal discovery literature (Danks et al., 2009; Tillman and Spirtes, 2011; Triantafillou et al., 2010), but not in the context of causal effect identification. In the basic setting the input consists of passively observed distributions $P(\mathbf{A}_i)$ such that $\mathbf{A}_i \subset \mathbf{V}$. We may also observe experimental distributions $P(\mathbf{A}_i \,|\, \mathrm{do}(\mathbf{B}_i))$ (Hyttinen et al., 2012; Triantafillou and Tsamardinos, 2015) or even conditionals $P(\mathbf{A}_i \,|\, \mathrm{do}(\mathbf{B}_i), \mathbf{C}_i)$. Our approach sets no limitations for the number and types of input distributions.

Combining Transportability and Selection Bias To the best of our knowledge, the frameworks of transportability and selection bias have not been considered simultaneously. The combination of these scenarios fits into the general problem formulation. For example, we may have access to two observational distributions originating from different source domains, but affected by the same biasing mechanism: $P(\mathbf{A}_1 \,|\, \mathbf{C}_1, T_1, S)$ and $P(\mathbf{A}_2 \,|\, \mathbf{C}_2, T_2, S)$, where T_1 and T_2 are the transportability nodes corresponding to the two source domains and S is the selection bias node.

Recovering from Multiple Sources of Selection Bias In recent literature on selection bias as a causal inference problem, the focus has been on settings where only a single selection bias node is present (e.g. Bareinboim et al., 2014; Correa and Bareinboim, 2017;

Algorithm 1 An outline of a search for causal effect identification.

```
Input: Target $Q = P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W})$, a semi-Markovian graph $G$ and a set of known input
    distributions $\mathbf{P} = \{P_1, \ldots, P_n\}$.
Output: A formula for $Q$ or NA if the effect is not identifiable.
for each $P_i \in \mathbf{P}$ do
    Derive new distributions computable from $P_i$ such that:
        • The required d-separation criteria are satisfied by $G$.
        • For multiple inputs, both inputs must be in $\mathbf{P}$.
    Add the new identified distributions to $\mathbf{P}$.
    If $Q$ was derived, return a formula for it.
Return NA.
```

Correa et al., 2018). However, multiple sources of selection bias are typical in longitudinal studies where dropout occurs at different stages of the study. Our approach is applicable for an arbitrary number of selection bias mechanisms and input distributions affected by arbitrary combinations of these mechanisms. In other words, if $\mathbf{S}$ is the set of all selection bias nodes present in the graph, the inputs can take the form $P(\mathbf{A} \,|\, \mathrm{do}(\mathbf{B}), \mathbf{C}, \mathbf{S}')$, where $\mathbf{S}'$ is an arbitrary subset of $\mathbf{S}$.

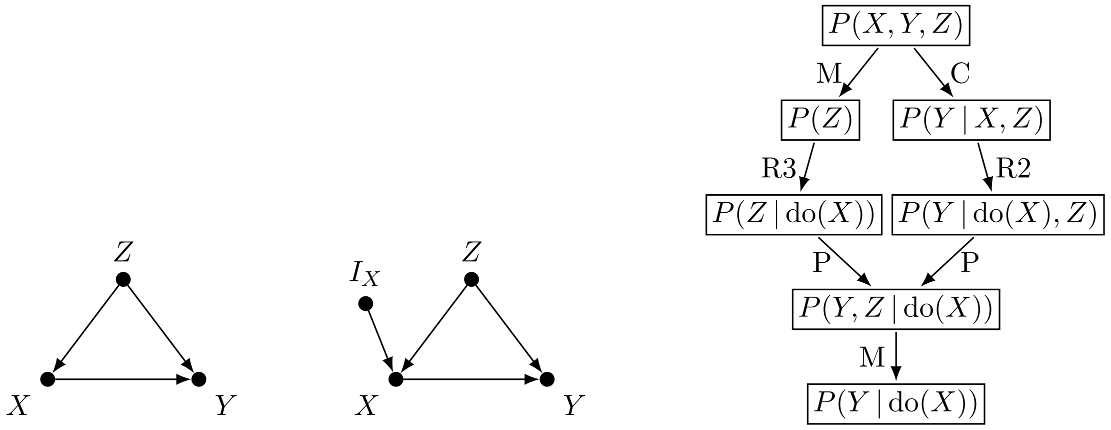
3 A Search Based Approach for Causal Effect Identification

The key to identification of causal effects is that interventional expressions can be manipulated using the rules of do-calculus. We present these rules for augmented DAGs where an additional intervention variable I_X such that $I_X \to X$ is added to the induced graph for each variable X (Spirtes et al., 1993; Pearl, 2009; Lauritzen, 2000) (see Figure 1(b)). Now a d-separation condition of the form $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \,|\, \mathbf{X}, \mathbf{W} \,||\, \mathbf{X}$ means that $\mathbf{Y}$ and $\mathbf{Z}$ are d-separated by $\mathbf{X}$ and $\mathbf{W}$ in a graph where edges incoming to (intervened) $\mathbf{X}$ have been removed (Hyttinen et al., 2015; Dawid, 2002). The three rules of do-calculus Pearl (1995) can be expressed as follows:

$$\begin{aligned}
P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{Z}, \mathbf{W}) &= P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W}), \text{ if } \mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \,|\, \mathbf{X}, \mathbf{W} \,||\, \mathbf{X} \\
P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}, \mathbf{Z}), \mathbf{W}) &= P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{Z}, \mathbf{W}), \text{ if } \mathbf{Y} \perp\!\!\!\perp \mathbf{I_Z} \,|\, \mathbf{X}, \mathbf{Z}, \mathbf{W} \,||\, \mathbf{X} \\
P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}, \mathbf{Z}), \mathbf{W}) &= P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W}), \text{ if } \mathbf{Y} \perp\!\!\!\perp \mathbf{I_Z} \,|\, \mathbf{X}, \mathbf{W} \,||\, \mathbf{X}
\end{aligned}$$

The rules are often referred to as insertion/deletion of observations, exchange of actions and observations, and insertion/deletion of actions respectively. Each rule of do-calculus is only applicable if the accompanying d-separation criterion (on the right-hand side) holds in the underlying graph. In addition to these rules, most derivations require basic probability calculus.

Do-calculus directly motivates a forwards search over its rules. The outline of this type of search is given in Algorithm 1. The algorithm derives new identifiable distributions based on what has been given as the input or identified in the previous steps. For each identified distribution every rule of do-calculus and standard probability manipulations of marginalization and conditioning are applied in succession, until the target distribution is


Z
X
Y
I_X
Z
X
Y
$P(X,Y,Z)$
M
C
$P(Z)$
$P(Y \mid X,Z)$
R3
R2
$P(Z \mid \mathrm{do}(X))$
$P(Y \mid \mathrm{do}(X),Z)$
P
P
$P(Y,Z \mid \mathrm{do}(X))$
M
$P(Y \mid \mathrm{do}(X))$

(a) Example graph.

(b) Augmented graph.

(c) A derivation for $P(Y \mid \mathrm{do}(X))$.

Figure 1: The back-door criterion holds in the example graph (a) for Z. The augmented graph (b) includes the intervention node I_X for X explicitly. The labels M, C, P, R2 and R3 in the derivation of (c) refer to marginalization, conditioning, product rule and rules 2 and 3 of do-calculus respectively (see Table 2). The required d-separation conditions $Y \perp\!\!\!\perp I_X \mid Z, X$ for R2 and $Z \perp\!\!\!\perp I_X$ for R3 hold in the augmented graph (b).

found, or no new distributions can be found to be identifiable. A preliminary version of this kind of search is used by Hyttinen et al. (2015) as a part of an algorithmic solution to causal effect identifiability when the underlying graph is unavailable.

The formulas produced by Algorithm 1 correspond to short derivations and unnecessarily complicated expressions are avoided. Also, only distributions guaranteed to be identifiable are derived and used during the search. Formulas for intermediary queries that were identified during the search are also available as a result. Alternatively, one could also start with the target and search towards the input distributions; a search in this direction will spend time deriving a number expressions that are anyway non-identifiable based on the input. A depth-first search would produce unnecessarily complicated expressions.

The search can easily derive for example the back-door criterion in the graph of Figure 1(a) as shown by the derivation in Figure 1(c). The target is $Q = P(Y \mid \mathrm{do}(X))$ and input is $\mathbf{P} = \{P(X,Y,Z)\}$. From $P(X,Y,Z)$ the search first derives the marginal $P(Z)$ and the conditional $P(Y \mid X,Z)$. Then $P(Z \mid \mathrm{do}(X))$ is derived by the third rule of do-calculus because $Z \perp\!\!\!\perp I_X$. The second rule derives $P(Y \mid \mathrm{do}(X),Z)$ from $P(Y \mid X,Z)$ as $Y \perp\!\!\!\perp I_X \mid Z, X$. The two terms can be combined via the product rule of probability calculus to get $P(Y,Z \mid \mathrm{do}(X))$ and finally the target is $P(Y \mid \mathrm{do}(X))$ is just a marginalization of this. The familiar formula $\sum_Z P(Y \mid X,Z)P(Z)$ is thus obtained.

However, it is not straightforward to make a search over do-calculus computationally feasible. The search space in Figure 1(c) shows only the parts that resulted in the identifying formula: for example all passively observed marginals and conditionals over $\mathbf{V}$ can be derived from the input $P(\mathbf{V})$. Especially in a non-identifiable case a naive search may go through a huge space before it can return the non-identifiable verdict. The choice of rules is also not obvious: a redundant rule may make the search faster or slower; false non-

Rule	Additional Input	Output	Description
1+		$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{Z}, \mathbf{W})$	Insertion of observations
1−		$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W} \setminus \mathbf{Z})$	Deletion of observations
2+		$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}, \mathbf{Z}), \mathbf{W} \setminus \mathbf{Z})$	Observation to action exchange
2−		$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X} \setminus \mathbf{Z}), \mathbf{Z}, \mathbf{W})$	Action to observation exchange
3+		$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}, \mathbf{Z}), \mathbf{W})$	Insertion of actions
3−		$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X} \setminus \mathbf{Z}), \mathbf{W})$	Deletion of actions
4		$P(\mathbf{Y} \setminus \mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$	Marginalization
5		$P(\mathbf{Y} \setminus \mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{Z}, \mathbf{W})$	Conditioning
6+	$P(\mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{W} \setminus \mathbf{Z})$	$P(\mathbf{Y}, \mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$	Chain rule multiplication
6−	$P(\mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{Y}, \mathbf{W})$	$P(\mathbf{Y}, \mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$	Chain rule multiplication

Table 2: The rules used to manipulate input distributions of the form $P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$. The output distribution is identified if the input is identified and if the corresponding d-separation criteria hold in the graph (for rules $1\pm, 2\pm$ and $3\pm$) or if the additional input has also been identified (rules $6\pm$). The sets $\mathbf{Y}, \mathbf{X}$ and $\mathbf{W}$ are disjoint. The role of the set $\mathbf{Z}$ depends on the rule being applied.

identifiability may be concluded if a necessary rule is missing. Also the order in which the rules are applied can have a large impact on the performance of the search. In the following sections we will provide highly non-trivial solutions to these challenges.

3.1 Rules

Table 2 lists the full set of rules used to manipulate distributions during the search, generalizing Hyttinen et al. (2015).

Do-calculus Rules $1\pm, 2\pm$ and $3\pm$ correspond to the rules of do-calculus such that rules $1+, 2+, 3+$ are used to add conditional variables and interventions and rules $1-, 2-, 3-$ are used to remove them. Each rule is only valid if the corresponding d-separation criterion given in the beginning of Section 3 hold.

Probability theory Rule 4 performs marginalization over $\mathbf{Z} \subset \mathbf{Y}$, and produces a summation at the formula level:

$$P(\mathbf{Y} \setminus \mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{W}) = \sum_{\mathbf{Z}} P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W}).$$

Similarly, rule 5 conditions on a subset $\mathbf{Z} \subset \mathbf{Y}$ to obtain the following formula:

$$P(\mathbf{Y} \setminus \mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{Z}, \mathbf{W}) = \frac{P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})}{\sum_{\mathbf{Y} \setminus \mathbf{Z}} P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})}.$$

Rules 6+ and 6− perform multiplication using the chain rule of probability which requires two known distributions. When rule 6+ is applied, the distribution $P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$ is

Rule	Validity condition	Termination condition
1+	$\mathbf{Z} \cap (\mathbf{Y} \cup \mathbf{X} \cup \mathbf{W}) = \emptyset$	
1−	$\mathbf{Z} \subseteq \mathbf{W}$	$\mathbf{W} = \emptyset$
2+	$\mathbf{Z} \subseteq \mathbf{W}$	$\mathbf{W} = \emptyset$
2−	$\mathbf{Z} \subseteq \mathbf{X}$	$\mathbf{X} = \emptyset$
3+	$\mathbf{Z} \cap (\mathbf{Y} \cup \mathbf{X} \cup \mathbf{W}) = \emptyset$	
3−	$\mathbf{Z} \subseteq \mathbf{X}$	$\mathbf{X} = \emptyset$
4	$\mathbf{Z} \subset \mathbf{Y}$	$\|\mathbf{Y}\| = 1$
5	$\mathbf{Z} \subset \mathbf{Y}$	$\|\mathbf{Y}\| = 1$
6+	$\mathbf{Z} \subseteq \mathbf{W}$	$\mathbf{W} = \emptyset$
6−	$\mathbf{Z} \cap (\mathbf{Y} \cup \mathbf{X} \cup \mathbf{W}) = \emptyset$	

Table 3: The conditions for the enumerated subset $\mathbf{Z}$ for applying the rules of Table 2 to a term $P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W})$. For rules 6+ and 6−, the conditions specify valid variables of the second required term.

known and we check whether $P(\mathbf{Z} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W} \setminus \mathbf{Z})$ is known as well. For rule 6−, the roles of the distributions are reversed. In the case of rule 6+, $\mathbf{Z}$ is a subset of $\mathbf{W}$ and we obtain

$$P(\mathbf{Y}, \mathbf{Z} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W}) = P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W}) P(\mathbf{Z} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W} \setminus \mathbf{Z}).$$

The two version of the chain rule are needed: it may be the case that when expanding $P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W})$ with rule 6+ the additional input $P(\mathbf{Z} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W} \setminus \mathbf{Z})$ is only identified later in the search. Then, $P(\mathbf{Y}, \mathbf{Z} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W})$ is identified when rule 6− is applied to $P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W})$.

3.2 Improving the Efficacy of the Search

In this section, we present various techniques that improved the efficiency of the search. These findings are implemented in a search algorithm in Section 3.3.

3.2.1 Term Expansion

Term expansion refers to the process of deriving new distributions from an input distribution using the rules of Table 2. By *term* we mean a single identified distribution. A term is considered *expanded* if the rules of Table 2 have been applied to it in every possible way when the term is in the role of the input. Note that an expanded distribution may still take the role of an additional input when another term is being expanded. Consider the step of expanding the input term in Table 2 to all possible outputs with any rule. This can be done by enumerating every non-empty subset $\mathbf{Z}$ of $\mathbf{V}$, and applying the rule with regard to it.

Table 3 outlines the requirements for $\mathbf{Z}$ for each rule of the search. Table 3 tells us that when an observation $\mathbf{Z}$ is added using rule 1+, it cannot be contained in any of the sets $\mathbf{Y}, \mathbf{X}$ or $\mathbf{W}$ since they are already present in the term. Only observations that are present can be removed, which is why $\mathbf{Z}$ has to a subset of $\mathbf{W}$ when applying rule 1−. We may skip the application of this rule if the set of observations is empty for the current term. The exchange of observations to experiments using rule 2+ has similar requirements for set $\mathbf{Z}$ as rule 1−. Exchanging experiments to observations using rule 2− works in a similar fashion. Only experiments that are present can be exchanged which means that $\mathbf{Z} \subseteq \mathbf{X}$. This rule can be skipped if the set of experiments is empty. New experiments are added using rule 3+ with similar requirements as rule 1+. Well-defined subsets for using rule 3− are the same as for rule 2−. For rules 4 and 5, the only requirement is that $\mathbf{Z}$ is a proper subset of $\mathbf{Y}$. When the chain rule is applied with rule 6+, we require that the variables of the second product term is observed in the first term. When applied in reverse with rule 6−, the variables of the second term must not be present in the first term.

3.2.2 Termination Conditions

Additionally, Table 3 lists the termination condition: if it is satisfied by the current term to be expanded we know that the rule cannot be applied to it. The following simple lemma shows that when any of the termination conditions hold, no new distributions can be derived from it using the respective rule, which allows the search to directly proceed to the next rule.

Lemma 1. *Let G be a semi-Markovian graph and let $\mathbf{Y}, \mathbf{X}$ and $\mathbf{W}$ be disjoint subsets of $\mathbf{V}$. Then all of the following are true:*

(i) If $\mathbf{W} = \emptyset$, then rule 1− of Table 2 cannot be used.

(ii) If $\mathbf{W} = \emptyset$, then rule 2+ of Table 2 cannot be used.

(iii) If $\mathbf{X} = \emptyset$, then rule 2− of Table 2 cannot be used.

(iv) If $\mathbf{X} = \emptyset$, then rule 3− of Table 2 cannot be used.

(v) If $|\mathbf{Y}| = 1$, then rule 4 of Table 2 cannot be used.

(vi) If $|\mathbf{Y}| = 1$, then rule 5 of Table 2 cannot be used.

(vii) If $\mathbf{W} = \emptyset$, then rule 6+ of Table 2 cannot be used.

Proof. For (i), the set $\mathbf{W}$ is empty so the application of rule 1− using any subset $\mathbf{Z}$ would result in $P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W} \setminus \mathbf{Z}) = P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W})$ which is already identified. For (ii), the set $\mathbf{W}$ is empty so no observation can be exchanged for an action using the second rule of do-calculus. For (iii), the set $\mathbf{X}$ is empty so no action can be exchanged for an observation using the second rule of do-calculus. For (iv), the set $\mathbf{X}$ is empty so the application of rule 3− using any subset $\mathbf{Z}$ would result in $P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X} \setminus \mathbf{Z}), \mathbf{W}) = P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W})$ which is already identified. For (v) and (vi), the set $\mathbf{Y}$ only has a single vertex, so it cannot have a non-empty subset. For (vii), the set $\mathbf{W}$ is empty so no subset $\mathbf{Z} \subset \mathbf{W}$ can exist for the second input. □

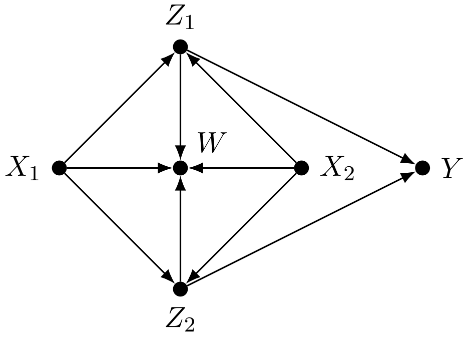


Figure 2: A graph for the example where all rules of Table 2 are required for identifying the target quantity.

3.2.3 Rule Necessity

The rule 1 of do-calculus can be omitted as shown by Huang and Valtorta (2006, Lemma 4). Instead of inserting an observation using rule 1, we can insert an intervention and then exchange it for an observation. Similarly, an observation can be removed by first exchanging it for an intervention and then deleting the intervention. It follows that rules 1+ and 1− of Table 2 are unnecessary for the search.

The following example shows that the remaining rules of Table 2 are all necessary. In the graph of Figure 2, the causal effect $P(Y, X_1 \,|\, \mathrm{do}(X_2), W)$ can be identified from the inputs $P(W \,|\, \mathrm{do}(X_2), Y, X_1)$, $P(Y \,|\, \mathrm{do}(X_2), Z_1, Z_2, X_1)$, $P(X_1 \,|\, \mathrm{do}(X_2), W)$, $P(Z_2, X_2 \,|\, \mathrm{do}(X_1))$ and $P(Z_1 \,|\, \mathrm{do}(X_1, Y), X_2)$ when all rules are available, but not when any individual rule is omitted. This can be verified by running the search algorithm presented at the beginning of Section 3 or the more advanced algorithm of Section 3.3 with each rule turned off individually.

3.2.4 Early Detection of Non-identifiable Instances

Worst-case performance of the search can be improved by detecting non-identifiable quantities directly based on the set of inputs before launching the search. The following theorem provides a sufficient criterion for non-identifiability.

Theorem 1. *Let G be a semi-Markovian graph, let $Q = P(\mathbf{Y} \,|\, do(\mathbf{X}), \mathbf{W})$ and let*

$$\mathbf{P} = \{P(\mathbf{A}_1 \,|\, do(\mathbf{B}_1), \mathbf{C}_1), \ldots, P(\mathbf{A}_n \,|\, do(\mathbf{B}_n), \mathbf{C}_n)\}.$$

Then Q is not identifiable from $\mathbf{P}$ in G if

$$\mathbf{Y} \not\subseteq \bigcup_{i=1}^{n} \mathbf{A}_i,$$

Proof. Since $\mathbf{Y} \not\subseteq \bigcup_{i=1}^{n} \mathbf{A}_i$, there exists a variable $Y' \in \mathbf{Y}$ such that none of the sets $\mathbf{A}_i$ contain it. We construct two models, M^1 and M^2, such that $P^1(Y' \,|\, \mathrm{do}(\mathrm{Pa}(Y')_G)) = P^1(Y') = P^2(Y' + c \,|\, \mathrm{do}(\mathrm{Pa}(Y')_G)) = P^2(Y' + c)$ where $c \neq 0$ is a constant. For any child

V_i of Y', we define the structural equations so that $f_i^2(\mathrm{Pa}(V_i)_G \setminus Y', Y', U_i) = f_i^1(\mathrm{Pa}(V_i)_G \setminus Y', Y' - c, U_i)$. For all other variables, the structural equations are the same for the models M^1 and M^2. We have that $P^1(Y' \mid \mathrm{do}(\mathbf{X}), \mathbf{W}) \neq P^2(Y' \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$ while all inputs $\mathbf{P}$ are the same for the models M^1 and M^2. It follows that $P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$ is not identifiable. □

In other words, Theorem 1 can be used to verify that the entire set $\mathbf{Y}$ of a target distribution $P(\mathbf{Y} \mid \cdot)$ cannot be constructed from the inputs. If this is the case, the target quantity is not identifiable.

3.2.5 Heuristics

During the search, we always expand one term at a time through the rules and store the newly identified distributions. In order for the search to perform fast, we need to decide which branches are the most promising and should therefore be expanded first. We can do this by defining a proximity function relating the source terms and the target query, and by always expanding the closest term first. Our suggestion here is motivated by the way an educated person might apply do-calculus in a manual derivation. Our chosen proximity function h links the target distribution $P^t = P(\mathbf{A}_t \mid \mathrm{do}(\mathbf{B}_t), \mathbf{C}_t)$ and a source distribution $P^s = P(\mathbf{A}_s \mid \mathrm{do}(\mathbf{B}_s), \mathbf{C}_s)$ in the following way:

$$\begin{aligned} h(P^t, P^s) = 10|\mathbf{A}_t \cap \mathbf{A}_s| + 5|\mathbf{B}_t \cap \mathbf{B}_s| + 3|\mathbf{C}_t \cap \mathbf{C}_s| - 2|\mathbf{A}_t \setminus \mathbf{A}_s| - 2|\mathbf{B}_t \setminus \mathbf{B}_s| \\ - 2|\mathbf{B}_s \setminus \mathbf{B}_t| - |\mathbf{C}_t \setminus \mathbf{C}_s| - |\mathbf{C}_s \setminus \mathbf{C}_t|. \end{aligned}$$

Each input distribution and terms derived using the search are assigned into a priority queue, where the priority is determined by the value given by h. Distributions closer to the target are prioritized over other terms. The weight 10 for the term $|\mathbf{A}_t \cap \mathbf{A}_s|$ indicates that having the correct response variables is considered as the first priority. Having the correct intervention is considered as the second priority (weight 5) and having the correct condition as the third priority (weight 3). The remaining terms in h penalize variables that are in the target distribution but not in the source distribution or vice versa. Again, variables that are intervened on are considered to be more important than conditioning variables.

3.3 The Search Algorithm

We take Algorithm 1 as our starting point and compile the results of Section 3.2 into a new search algorithm called `do-search`. This algorithm is capable of solving generalized identifiability problems (row 7 in Table 1) while streamlining the search process through a heuristic search order and elimination of redundant rules and subsets. The pseudo-code for `do-search` is shown in Algorithm 2.

The algorithm begins by checking whether the query can be solved trivially without performing the search. This can happen if the target Q is a member of the set of inputs or if Theorem 1 applies. Next, we note that each input distribution in the set $\mathbf{P}$ is marked as unexpanded at the beginning of the search. Distributions in $\mathbf{P}$ are expanded one at a time by applying every rule of Table 2 in every possible way.

The iteration over the unexpanded distributions $\mathbf{U}$ proceeds as follows (lines 4–5). Each input distribution and terms derived from it using the search are assigned into a priority

Algorithm 2 `do-search`

```
Input: Target $Q = P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W})$, a semi-Markovian graph $G$ and a set of known
    distributions $\mathbf{P} = \{P_1, \ldots, P_n\}$.
Output: A formula $F$ for $Q$ in terms of $\mathbf{P}$ or NA
 1: if $Q \in \mathbf{P}$, return $Q$
 2: if target is non-identifiable by Theorem 1, then return NA
 3: let $\mathbf{U}$ be the set of unexpanded distributions, initially $\mathbf{U} := \mathbf{P}$
 4: while $\mathbf{U} \neq \emptyset$, do
 5:   let $P'$ be the unexpanded distribution closest to the target: $P' = \operatorname*{argmax}_{P_i \in \mathbf{U}} h(Q, P_i)$
 6:   let $\mathbf{M}$ be the set of rules of Table 2
 7:   Remove rules 1+ and 1− from $\mathbf{M}$
 8:   Remove those rules from $\mathbf{M}$ where termination criteria of Table 3 hold for $P'$
 9:   let $\mathbf{P}^*$ be the set of all distributions derived from $P'$ using the rules in $\mathbf{M}$
10:   for each new candidate distribution $P^* \in \mathbf{P}^*$, do
11:     if $P^*$ is already in $\mathbf{P}$, then continue
12:     if conditions of Table 3 are not satisfied by $P^*$, then continue
13:     if an additional input is required that is not in $\mathbf{P}$, then continue
14:     if d-separation criteria of Table 2 are not satisfied by $G$, then continue
15:     if $P^* = Q$, then
16:       Derive a formula $F$ for $Q$ by backtracking.
17:       return $F$
18:     Add $P^*$ to $\mathbf{P}$, add $P^*$ to $\mathbf{U}$
19:   Mark $P'$ as expanded: remove $P'$ from $\mathbf{U}$
20: return NA
```

queue, where the priority is determined by the value given by the proximity function h. Distributions closer to the target are expanded first. In the implementation, only the actual memory addresses of the distribution objects are placed into the queue. The set $\mathbf{P}$ is implemented as a hash table that serves as a container for all input distributions and those derived from them. Each new distribution is assigned a unique index that also serves the hash function for this table. The distribution objects contained in the table are represented uniquely by three integers corresponding to the sets $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$ of the general form $P(\mathbf{A} \,|\, \mathrm{do}(\mathbf{B}), \mathbf{C})$. The distribution objects also contain additional auxiliary information such as which rule was used to derive it, whether it is expanded or not and from which distribution it was obtained. This information is used to construct the derivation if the target is found to be identifiable.

Multiple distributions can share the same value of the proximity function h. In the case that multiple candidates share the maximal value, the one that was derived the earliest takes precedence. When the unexpanded distribution currently closest to the target is determined, the rules of Table 2 are applied sequentially for all valid subsets dictated by Table 3. When rules one, two and three of do-calculus are considered the necessary d-separation criteria is checked in G (line 14). For the chain rule, the presence of the required second input is also verified. The reverse lookup is implemented by using another hash table, where the hash function is based on the unique representation of each distribution object. The values contained in the table are the indices of the derived distributions. The same hash table is also used to verify that we do not derive again distributions that have been previously found to be identifiable from the inputs.

We construct a set $\mathbf{M}$ of applicable rules for each unexpanded distribution P' using the termination criteria of Table 3 (lines 6–8). If all the necessary criteria have been found to hold for an applicable rule and a subset, the newly derived distribution P^* is added to the set of known distributions and placed into the priority queue as an unexpanded distribution. When the applicable rules and subsets have been exhausted for the current distribution P', it is marked as expanded and removed from the queue (line 19). If the target distribution is found at any point (line 15), a formula is returned for it in terms of the original inputs. Alternatively, we can also continue deriving distributions to obtain different search paths to the target that can possibly produce different formulas for it. If instead we exhaust the set of unexpanded distributions by emptying the queue, the target is deemed non-identifiable by the search (line 20).

We keep track of the rules that were used to derive each new distribution in the search. This allows us to construct a graph of the derivation where each root node is a member of the original input set $\mathbf{P}$ and their descendants are the distributions derived from them during the search. Each edge represents a manipulation of the parent node(s) to obtain the child node. For an identifiable target quantity, the formula F is obtained by backtracking the chain of manipulations recursively until the roots are reached (line 16). The derivation of the example in the beginning of Section 3 depicted in Figure 1(c) can be efficiently found by applying this procedure.

3.4 Soundness and Completeness Properties

We are ready to establish some key theoretical properties of `do-search`. The first theorem considers the correctness of the search.

Theorem 2 (Soundness). `do-search` *always terminates: if it returns an expression for the target Q, it is correct, if it returns NA then Q is not identifiable with respect to the rules of do-calculus and standard probability manipulations (in Table 2).*

Proof. Each new distribution is derived by using only well-defined manipulations as outlined by Table 3 and by ensuring that the required d-separation criteria hold in G when rules of do-calculus are concerned. It follows that if the search terminates and returns a formula for the target distribution, it was reached from the set input distributions through a chain of valid manipulations. If `do-search` terminates as a result of Theorem 1, we are done. Suppose now that Theorem 1 does not apply. By definition, `do-search` enumerates every rule of Table 2 for every well-defined subset of Table 3. By Lemma 1, no distributions are left out by applying the termination criteria of Table 3. We know that if rules $1-$ and $1+$ of Table 3 are omitted, the distributions generated by these rules can be obtained by a combination of rules $2-, 2+, 3-$ and $3+$. Furthermore, the order in which the distributions are expanded has no effect, as every possible manipulation is still carried out. The search will eventually terminate, since distributions that have already been derived are not added again to the set of unexpanded distributions and there are only finitely many ways to apply the rules of Table 2. □

The following theorem provides a completeness result in connection to existing identifiability results. Since do-calculus has been shown to be complete with respect to (conditional) causal effect identifiability, z-identifiability and transportability, it follows that `do-search` is complete for these problems as well.

Theorem 3 (Completeness)**.** *If* `do-search` *returns NA in the settings in rows 1–4 in Table 1, then the query is non-identifiable.*

Proof. Do-calculus has been shown complete in these settings. The rules of probability calculus encode what is used in the algorithms as can be seen for example from the proofs of Theorem 7 and Lemmas 4–8 of (Shpitser and Pearl, 2006a). □

It is not known whether the rules implemented in `do-search` are sufficient for other more general identifiability problems since it is conceivable that some additional rules might exist that would be required to achieve completeness. One such generalization is the inclusion of missing data in the causal model, which we present in Section 5. However, if one were to show that do-calculus (or any other set of rules included in `do-search`) is complete for some special case of the generalized identifiability problem, then `do-search` would be complete for this problem as well. In the following sections we will use the term "identifiable by `do-search`" to refer to causal queries that can be indentified by `do-search`.

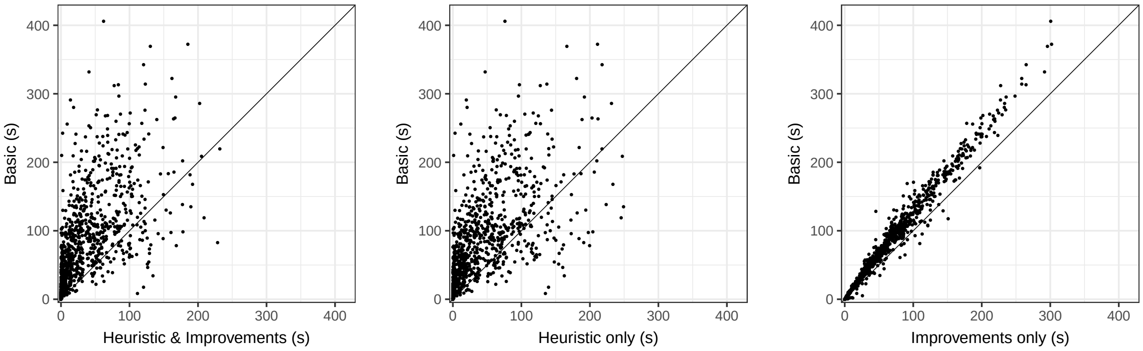
3.5 Simulations

We implemented `do-search` (Algorithm 2) in C++. Here we report the findings of a simulation study to assess the running time performance of `do-search` and the impact of the improvements outlined in Section 3.2 as well as the search heuristic described in Section 3.2.5.

Our synthetic simulation scenario consisted of 1071 semi-Markovian causal graphs of 10 vertices that were generated at random by first generating a random topological order of the vertices followed by a random lower triangular adjacency matrices for both directed and bidirected edges. Graphs without a directed path from X to Y were discarded. We sampled sequentially input distributions of the form $P(\mathbf{A} \mid \mathrm{do}(\mathbf{B}), \mathbf{C})$ at random by generating disjoint subsets such that $\mathbf{A}$ is always non-empty. This was continued until the target quantity $P(Y \mid \mathrm{do}(X))$ was found to be identifiable by the search. Then for each graph, we recorded the search times for set of inputs that first resulted in the query to be identified and for the last set such that the target was non-identifiable. In other words, each graph generates two simulation instances, one for an identifiable query and one for a non-identifiable query. This setting directly corresponds to the setting of partially overlapping experimental data sets discussed in Section 2.2 for which no other algorithmic solutions exist.

To understand the impact of the search heuristic and the various improvements, we compare four different search configurations: the basic `do-search` without the search heuristic or improvements[1], one that only uses the search heuristic, one that only uses the improvements of Section 3.2 and one that uses them both.

[1] In this configuration, terms are expanded in the order they were identified; the conditions in Table 3 are not checked.



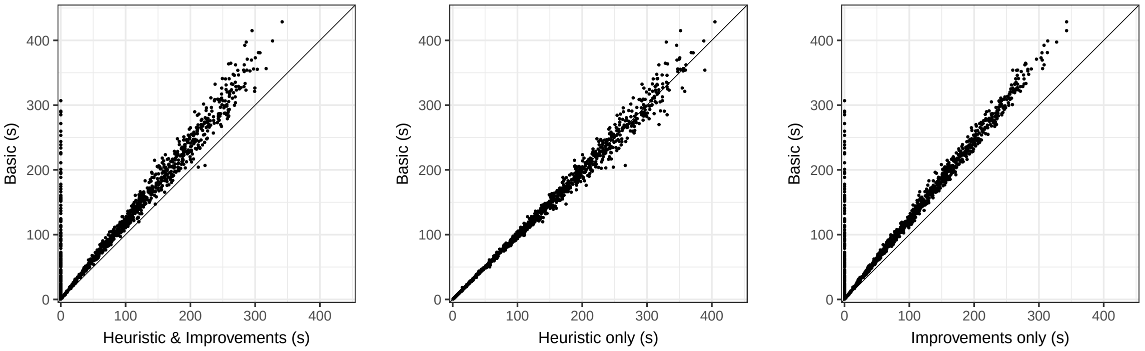
(a) (b) (c)

Figure 3: Scatter plots of the search times from identifiable instances under different search configurations compared to the basic `do-search` without a heuristic or improvements.

Figure 3 shows the search times of the configurations compared to the basic configuration for identifiable instances. Most importantly, a vast majority of instances (93 %) are solved faster than the basic configuration when both heuristics and improvements are used. The average search time with both heuristics and improvements enabled was 32.7 seconds and 75.2 seconds for the basic configuration. The search heuristic provides the greatest benefit for these instances as can be seen from Figure 3(b). Using a heuristic can also hinder performance by leading the search astray and by causing additional computational steps through the evaluation of the proximity function. For example, there is a small number of instances where the search time is over ten times slower than the basic configuration when using a heuristic. Fortunately, there are several instances in the opposite direction, where the heuristic provides over one hundred fold reduction in search time. Curiously, even using the improvements sometimes results in slower search times. This is most likely due to the elimination of rule 1 of do-calculus, since it may be the case that the basic search is able to use this rule to reach the target distribution faster. More importantly, Figure 3(c) shows that the improvements clearly benefit the search. Furthermore, the benefit tends to increase as the instances get harder.

Figure 4 shows the search times of the configurations for non-identifiable instances. Relying only on a search heuristic provides no benefit here, as expected. The improvements to the search are most valuable for these instances, and in this scenario every non-identifiable instance was solved faster than baseline using the improvements, and when applied with the heuristic only three instance were slower than baseline. The average search time with both heuristic and improvements enabled was 105.2 seconds and 139.7 seconds for the basic configuration. The almost zero second instances are a result of Theorem 1 when no search has to be performed in order to determine the instance to be non-identifiable. The benefit of the improvements tends to increase as the instances get harder also for these instances.

Finally we examined the average run time performance of `do-search`, with all improvements and heuristics enabled. We replicated the previously described simulation scenario with the same number of instances (1071) for graphs up to 10 vertices. Figure 5 shows the



(a)

(b)

(c)

Figure 4: Scatter plots of the search times from non-identifiable instances under different search configurations compared to the baseline configuration.

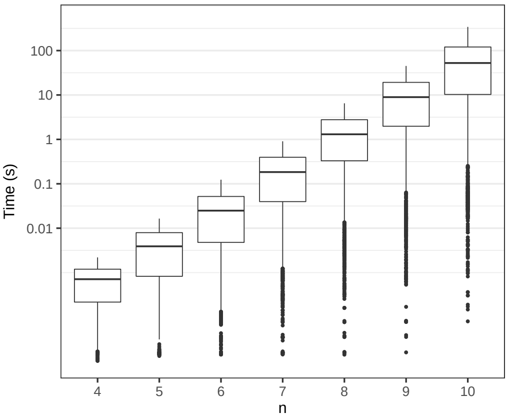


Figure 5: Boxplots of search times for both identifiable and non-identifiable instances in graphs of $n = 4, \ldots, 10$ vertices. The vertical axis uses a logarithmic scaling. Instances where the search time was less than 10^{-5} seconds were omitted for clarity.

boxplots of search times on a log-scale for graphs of different size, including both identifiable and non-identifiable instances. Note that for every graph size there are a number of easily solvable instances that show up as outliers in this plot. 10-node instances are solved routinely under 100 seconds. In this plot, the running times increase exponentially with increasing graph size (or number of variables).

4 New Causal Effect Identification Results

We present a number of results for various identifiability problems to showcase the versatility of `do-search`.

4.1 Multiple Data Sources with Partially Overlapping Variable Sets

Earlier generalizations of the identifiability problem assume nested experiments or entire distributions with the exception of surrogate outcome identifiability (Tikka and Karvanen, 2018b) which also has its own intricate assumptions regarding the available distributions. None of these assumptions are needed in `do-search` and it can be used to solve identifiability problems from completely arbitrary collections of input distributions.

We showcase identifiability from multiple experimental distributions by two examples. In the first example we consider identifiability of $P(Y_1, Y_2 \,|\, \mathrm{do}(X_1, X_2))$ in the graph of Figure 6(a) from $P(\mathbf{V})$, $P(Y_1, Y_2 \,|\, \mathrm{do}(X_1), Z, W, X_2)$, $P(W \,|\, \mathrm{do}(X_1, X_2))$, $P(Z \,|\, \mathrm{do}(X_2))$ and $P(Y_2 \,|\, \mathrm{do}(X_1), Y_1, Z, W, X_2)$. The target quantity is identifiable and `do-search` produces the following formula for it

$$\sum_{Z} \left(P(Z' \,|\, \mathrm{do}(X_2)) \sum_{W} P(W \,|\, \mathrm{do}(X_1, X_2)) \left(\sum_{Y_2} P(Y_1, Y_2 \,|\, \mathrm{do}(X_1), Z, W, X_2) \right) \times \right.$$
$$\left. \frac{P(Y_1, Y_2 \,|\, \mathrm{do}(X_1), Z, W, X_2)}{\sum_{Y_2'} P(Y_1, Y_2' \,|\, \mathrm{do}(X_1), Z, W, X_2)} \right).$$

In the second example we consider identifiability of $P(Y_1, Y_2 \,|\, \mathrm{do}(X_1, X_2))$ in the graph of Figure 6(b) from $P(\mathbf{V})$, $P(Y_1 \,|\, \mathrm{do}(X_1), Y_2, W, Z, X_2)$, $P(X_2, W \,|\, \mathrm{do}(X_1))$, $P(X_2 \,|\, \mathrm{do}(X_1, W))$, $P(Y_2 \,|\, \mathrm{do}(X_1), Z, W, X_2)$, $P(Y_2 \,|\, \mathrm{do}(Z), X_1, W, X_2)$, and $P(Y_1, Y_2 \,|\, \mathrm{do}(Z), W, X_1, X_2)$. Again, the target quantity is identifiable and `do-search` outputs the following formula

$$\sum_{W} \left(P(W \,|\, \mathrm{do}(X_1), X_2) \sum_{X_2} P(X_2 \,|\, \mathrm{do}(X_1, W)) \times \right.$$
$$\left. \frac{\sum_Z P(X_2, W, Z \,|\, X_1) P(Y_1, Y_2 \,|\, \mathrm{do}(X_1), X_2, W, Z)}{\sum_{Y_1', Y_2', Z} P(X_2, W, Z \,|\, X_1) P(Y_1', Y_2' \,|\, \mathrm{do}(X_1), X_2, W, Z)} \right).$$

This example shows that a heuristic approach can also help us to find shorter formulas. If we run `do-search` again without the heuristic in this instance, the output formula is instead

$$\sum_{W,Z} \left(P(Z) P(W \,|\, X_2, X_1, Z) \sum_{X_2} P(X_2 \,|\, X_1, Z) \sum_{Y_2} P(Y_2 \,|\, \mathrm{do}(X_1), X_2, W, Z) \times \right.$$
$$\left. P(Y_1 \,|\, \mathrm{do}(X_1), X_2, Y_2, W, Z) \frac{P(Y_2 \,|\, \mathrm{do}(X_1), X_2, W, Z) P(Y_1 \,|\, \mathrm{do}(X_1), X_2, Y_2, W, Z)}{\sum_{Y_2'} P(Y_2' \,|\, \mathrm{do}(X_1), X_2, W, Z) P(Y_1 \,|\, \mathrm{do}(X_1), X_2, Y_2', W, Z)} \right).$$

4.2 Combining Transportability and Selection Bias

Input distributions that originate from multiple sources while being simultaneously affected by selection bias can be considered with `do-search`. This kind of problem cannot be solved with algorithms RC or TR$^{\mathrm{mz}}$ of Table 1. As an example we consider one source domain and a target domain with two input data sets: a biased distribution $P(X, Y, Z \,|\, S)$ from

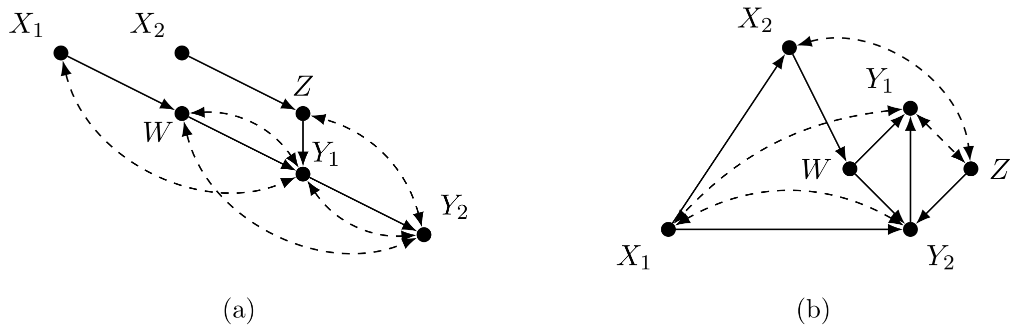


Figure 6: Graphs for the examples on identifiability problems combining both observational and experimental distributions.

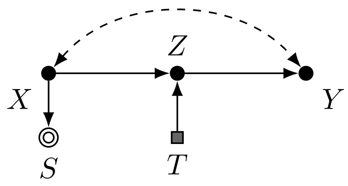


Figure 7: Graph that contains both selection bias and transportability nodes.

the target domain and an unbiased experimental distribution $P(Y, Z \,|\, \mathrm{do}(X), T)$ from the source domain. We evaluate the query $P(Y \,|\, \mathrm{do}(X))$ in the graph of Figure 7 using these inputs. In the figure transportability node T is depicted as a gray square and selection bias node S is depicted as an open double circle. The query is identifiable and `do-search` outputs the following formula for it

$$P(Y \,|\, \mathrm{do}(X)) = \sum_{Z} P(Y \,|\, \mathrm{do}(X), Z, T) \sum_{Y'} P(Z, Y' \,|\, X, S).$$

4.3 Recovering from Multiple Sources of Selection Bias

We present an example where bias originates from two sources with two input data sets: a distribution affected by both biasing mechanisms $P(X, Y, Z, W_1, W_2 \,|\, S_1, S_2)$ and a distribution affected only by a single bias source $P(Z \,|\, S_1)$. We evaluate the query $P(Y \,|\, \mathrm{do}(X))$ in the graph of Figure 8 using the inputs. The query is identifiable and the following formula is obtained by `do-search`

$$\sum_{Z} P(Z \,|\, S_1) P(Y \,|\, X, Z, W_1, W_2, S_1, S_2).$$

5 Extension to Missing Data Problems

The SCM framework can be extended to describe missing data mechanisms. For each variable V_i that is not fully observed, two special vertices are added to the causal graph.

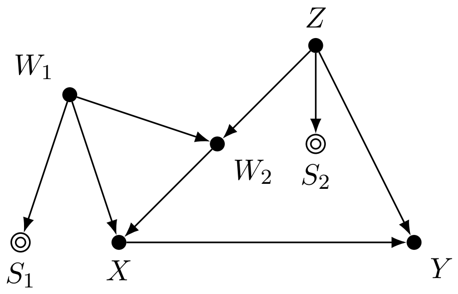


Figure 8: Graph where two selection bias nodes are present.

The vertex V_i^* is the observed proxy variable which is linked to the true variable V_i via the missingness mechanism (Little and Rubin, 1986; Mohan et al., 2013):

$$V_i^* = \begin{cases} V_i, & \text{if } R_{V_i} = 1, \\ \text{NA}, & \text{if } R_{V_i} = 0, \end{cases} \tag{1}$$

where NA denotes a missing value and R_{V_i} is called the response indicator (of V_i). In other words, the variable V_i^* that is actually observed matches the true value V_i if it is not missing ($R_{V_i} = 1$). Figure 10 depicts some examples of graphs containing missing data mechanisms.

The observed vertices of the causal diagram are partitioned into four categories

$$\mathbf{V} = \mathbf{V}_o \cup \mathbf{V}_m \cup \mathbf{V}^* \cup \mathbf{R},$$

where $\mathbf{V}_o$ is the set of fully observed variables, $\mathbf{V}_m$ is the set of partially observed variables, $\mathbf{V}^*$ is the set of all proxy variables and $\mathbf{R}$ is the set of response indicators. Our method is also capable of processing queries when the causal graph contains missing data mechanisms where the sets $\mathbf{A}_i, \mathbf{B}_i$ and $\mathbf{C}_i$ of the input distributions are restricted to contain observed variables in $\mathbf{V}^* \cup \mathbf{V}_o \cup \mathbf{R}$. An active response indicator $R_{V_i} = 1$ is denoted by $R^1_{V_i}$. Proxy variables are not explicitly shown in the graphs of this section for clarity.

Determining identifiability is more challenging under missing data. As evidence of this, even some non-interventional queries require the application of do-calculus (Mohan and Pearl, 2018). Furthermore, the rules used in the search of Table 2 are no longer sufficient and deriving the desired quantity necessitates the use of additional rules that stem from the definition of the proxy variables and the response indicator. Each new partially observed variable also has a higher impact on computational complexity, since the corresponding response indicator and proxy variable are always added to the graph as well.

Table 4 extends the set of rules of Table 2 to missing data problems by providing manipulations related to the missingness mechanism. The missing data column lists extended requirements for the valid subset if missing data mechanisms are present in the graph. The following notation is used in the table: $\mathbf{R}^a$ is the set of active response indicators for the current term, $\mathbf{V}^t$ denotes the set of partially observed variables corresponding to the proxy variables present in the current term. and $\mathbf{V}^p$ denotes the set of proxy variables corresponding to the partially observed variables present in the current term. For example, if the current term is $P(Y, Z^* \,|\, X^*)$, the aforementioned sets would be $\mathbf{V}^t = \{Z, X\}$ and $\mathbf{V}^p = \{Y^*\}$. The sets $\mathbf{Z}^t$ and $\mathbf{Z}^p$ are defined accordingly with respect to the set $\mathbf{Z}$.

Rule	Input	Additional Input	Output	Description
1+	$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$		$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{Z}, \mathbf{W})$	Insertion of observations
1−	$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$		$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W} \setminus \mathbf{Z})$	Deletion of observations
2+	$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$		$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}, \mathbf{Z}), \mathbf{W} \setminus \mathbf{Z})$	Obs. to action exchange
2−	$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$		$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X} \setminus \mathbf{Z}), \mathbf{Z}, \mathbf{W})$	Action to obs. exchange
3+	$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$		$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}, \mathbf{Z}), \mathbf{W})$	Insertion of actions
3−	$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$		$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X} \setminus \mathbf{Z}), \mathbf{W})$	Deletion of actions
4	$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$		$P(\mathbf{Y} \setminus \mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$	Marginalization
5	$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$		$P(\mathbf{Y} \setminus \mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{Z}, \mathbf{W})$	Conditioning
6+	$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$	$P(\mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{W} \setminus \mathbf{Z})$	$P(\mathbf{Y}, \mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$	Chain rule multiplication
6−	$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$	$P(\mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{Y}, \mathbf{W})$	$P(\mathbf{Y}, \mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$	Chain rule multiplication
7+	$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$	$P(\mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$	$P(\mathbf{Y} \setminus \mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{Z}, \mathbf{W})$	Chain rule conditioning
7−	$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$	$P(\mathbf{Y}, \mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{W} \setminus \mathbf{Z})$	$P(\mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$	Chain rule conditioning
8.1	$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$		$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W} \setminus \mathbf{Z}, \mathbf{Z}^1)$	Enable response indicators
8.2	$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$		$P(\mathbf{Y} \setminus \mathbf{Z}, \mathbf{Z}^1 \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$	Enable response indicators
9.1	$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W}, \mathbf{R}^1_{\mathbf{Z}})$		$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W} \setminus \mathbf{Z}^*, \mathbf{Z}, \mathbf{R}^1_{\mathbf{Z}})$	Proxy variable exchange
9.2	$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{W}, \mathbf{R}^1_{\mathbf{Z}})$		$P(\mathbf{Y} \setminus \mathbf{Z}^*, \mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{W}, \mathbf{R}^1_{\mathbf{Z}})$	Proxy variable exchange
9.3	$P(\mathbf{Y}, \mathbf{R}^1_{\mathbf{Z}} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$		$P(\mathbf{Y} \setminus \mathbf{Z}^*, \mathbf{Z}, \mathbf{R}^1_{\mathbf{Z}} \mid \mathrm{do}(\mathbf{X}), \mathbf{W})$	Proxy variable exchange

Table 4: Extended set of rules for missing data problems used to manipulate input distributions. Rules $1\pm, 2\pm, 3\pm, 4, 5$ and $6\pm$ are the same as in Table 2. For rules $7\pm$, the additional input has also been identified. The sets $\mathbf{Y}, \mathbf{X}, \mathbf{W}$ and $\mathbf{R}_{\mathbf{Z}}$ are disjoint. The roles of the sets $\mathbf{Z}$ and $\mathbf{R}_{\mathbf{Z}}$ depend on the rule being applied.

Rules 7+ and 7− perform conditioning using the chain rule. These rules are necessary in the case that set $\mathbf{Z}$ contains missing data mechanisms that have been enabled and thus cannot be marginalized over when attempting to use rule 5. The input is then of the following form, for example in the case of rule 7+:

$$P(\mathbf{Y}, \mathbf{Z} \mid \mathrm{do}(\mathbf{X}), \mathbf{W}) = P(\mathbf{Y}, \mathbf{Z}', \mathbf{R}' \mid \mathrm{do}(\mathbf{X}), \mathbf{W}),$$

where $\mathbf{Z}'$ does not contain any missing data mechanisms (or is possibly empty), $\mathbf{R}'$ contains only active missing data mechanisms and $\mathbf{Z} = \mathbf{Z}' \cup \mathbf{R}'$.

Rules 8.1 and 8.2 are used to enable response indicators, which then facilitates the use of rules $9.1, 9.2$ and 9.3. These three rules exchange proxy variables to their true counterparts when the corresponding response indicators are enabled. For example, rule 8.1 can be used on $P(Y, X^* \mid R_X)$ to first obtain $P(Y, X^* \mid R^1_X)$ by enabling R_X. Then, rule 9.2 can applied to this distribution to obtain $P(Y, X \mid R^1_X)$ by exchanging X^* for X.

Similarly to Table 3, Table 5 outlines the valid subsets $\mathbf{Z}$ for applying the extended rules of Table 4. A major difference to the original validity and termination conditions is the addition of the missing data condition that outlines the additional requirements that must be satisfied when missingness mechanisms are present. For the rules that are shared by Tables 2 and 4, the missing data condition ensures that a true variable and its proxy counterpart never appear in the same term at the same time. For example, we cannot add an intervention on X to $P(X^*)$. It also ensures that we do not carry out summation over enabled response indicators in the case of rules 4 and 5. When applying rules 8.1 or 8.2, the condition also ensures that we do not attempt to enable a response indicator that is already enabled. For rules $9.1, 9.2$ and 9.3, the conditions guarantees that a proxy can only be exchanged to a true variable if its corresponding response indicator is enabled in the term.

Rule	Validity condition	Missing data condition	Termination condition
1+	$\mathbf{Z} \cap (\mathbf{X} \cup \mathbf{Y} \cup \mathbf{W}) = \emptyset$	$\mathbf{Z} \cap (\mathbf{V}^t \cup \mathbf{V}^p \cup \mathbf{Z}^t \cup \mathbf{Z}^p) = \emptyset$	
1−	$\mathbf{Z} \subseteq \mathbf{W}$		$\mathbf{W} = \emptyset$
2+	$\mathbf{Z} \subseteq \mathbf{W}$	$\mathbf{Z} \cap \mathbf{V}^* = \emptyset$	$\mathbf{W} = \emptyset$
2−	$\mathbf{Z} \subseteq \mathbf{X}$		$\mathbf{X} = \emptyset$
3+	$\mathbf{Z} \cap (\mathbf{Y} \cup \mathbf{X} \cup \mathbf{W}) = \emptyset$	$\mathbf{Z} \cap (\mathbf{V}^* \cup \mathbf{V}^t) = \emptyset$	
3−	$\mathbf{Z} \subseteq \mathbf{X}$		$\mathbf{X} = \emptyset$
4	$\mathbf{Z} \subset \mathbf{Y}$	$\mathbf{Z} \cap (\mathbf{R}^a \cap \mathbf{Y}) = \emptyset$	$\lvert\mathbf{Y}\rvert = 1$
5	$\mathbf{Z} \subset \mathbf{Y}$	$(\mathbf{Y} \setminus \mathbf{Z}) \cap (\mathbf{R}^a \cap \mathbf{Y}) = \emptyset$	$\lvert\mathbf{Y}\rvert = 1$
6+	$\mathbf{Z} \subseteq \mathbf{W}$		$\mathbf{W} = \emptyset$
6−	$\mathbf{Z} \cap (\mathbf{Y} \cup \mathbf{X} \cup \mathbf{W}) = \emptyset$	$\mathbf{Z} \cap (\mathbf{V}^t \cup \mathbf{V}^p \cup \mathbf{Z}^t \cup \mathbf{Z}^p) = \emptyset$	
7+		$\mathbf{Z} \subset \mathbf{Y}$	$\lvert\mathbf{Y}\rvert = 1$
7−		$\mathbf{Z} \subseteq \mathbf{W}$	$\mathbf{W} = \emptyset$
8.1		$\mathbf{Z} \subseteq \mathbf{R} \cap \mathbf{W}, \mathbf{Z} \cap \mathbf{R}^a = \emptyset$	$\mathbf{R} \cap \mathbf{W} = \emptyset$
8.2		$\mathbf{Z} \subseteq \mathbf{R} \cap \mathbf{Y}, \mathbf{Z} \cap \mathbf{R}^a = \emptyset$	$\mathbf{R} \cap \mathbf{Y} = \emptyset$
9.1		$\mathbf{Z} \subseteq \mathbf{V}^* \cap \mathbf{W}, \mathbf{R}_{\mathbf{Z}} \subseteq \mathbf{R}^a$	$\mathbf{R}^a = \emptyset$
9.2		$\mathbf{Z} \subseteq \mathbf{V}^* \cap \mathbf{Y}, \mathbf{R}_{\mathbf{Z}} \subseteq \mathbf{R}^a$	$\mathbf{R}^a = \emptyset$
9.3		$\mathbf{Z} \subseteq \mathbf{V}^* \cap \mathbf{Y}, \mathbf{R}_{\mathbf{Z}} \subseteq \mathbf{R}^a$	$\mathbf{R}^a = \emptyset$

Table 5: The conditions for the enumerated subset $\mathbf{Z}$ for applying the rules of Table 2 to a term in the input column. For rules 6± and 7±, the conditions specify valid variables of the second required term. Validity conditions for rules $1\pm, 2\pm, 3\pm, 4, 5$ and 6± are the same as in Table 3.

Additional terminations conditions also apply to the new rules and their correctness is easily verified.

Lemma 2. *Let G be a semi-Markovian graph and let $\mathbf{Y}, \mathbf{X}$ and $\mathbf{W}$ be disjoint subsets of $\mathbf{V}$. Then all of the following are true:*

(i) If $\mathbf{W} = \emptyset$, then rule $7-$ of Table 4 cannot be used.

(ii) If $\mathbf{R} \cap \mathbf{W} = \emptyset$ then rule 8.1 of Table 4 cannot be used.

(iii) If $\mathbf{R} \cap \mathbf{Y} = \emptyset$ then rule 8.2 of Table 4 cannot be used.

(iv) If $\mathbf{R}^a = \emptyset$, then rules $9.1, 9.2$ or 9.3 of Table 4 cannot be used.

Proof. For (i), the set $\mathbf{W}$ is empty so no subset $\mathbf{Z} \subset \mathbf{W}$ can exist for the second input. For (ii), and the set $\mathbf{R} \cap \mathbf{W}$ is empty so no assignment $(\mathbf{W} \cap \mathbf{R}) = 1$ can be performed. Similarly for (iii), the set $\mathbf{R} \cap \mathbf{Y}$ is empty so no assignment $(\mathbf{Y} \cap \mathbf{R}) = 1$ can be performed. For (iv) the set of active response indicators $\mathbf{R}^a$ is empty, so no transformation from proxy variables to true variables via the missingness mechanism in (1) can be made. □

The task of selecting a suitable heuristic becomes more difficult when missing data are involved with the identifiability problem. The approach of Section 3.2.5 is no longer directly applicable due to the relation between proxy variables, response indicators and partially observed variables. The proximity function considers X and X^* as entirely different variables despite their connection and does not prefer the inclusion of response indicators.

If the heuristic is applies as such, the search path will often involve a large number of manipulations which in turn leads to complicated expressions. For these reasons we do not apply a heuristic to missing data problems, but expand terms in the order in which they were identified. The improvements described in Section 3.2 still apply.

It is straightforward to adapt `do-search` to the new extended set of rules. In the pseudocode shown in Algorithm 2, we simply replace all references to Tables 2 and 3 by references to Tables 4 and Tables 5, respectively. When the validity condition is checked, we also verify that the missing data condition holds. Lemma 2 guarantees the correctness of the new termination criteria. Theorem 1 is also valid when the sets $\mathbf{A}_i$ are replaced by $\mathbf{A}_i \cup \mathbf{A}_i^t$, since it may be possible to exchange some proxy variable to a true variable that is present in the set $\mathbf{Y}$ of the target $P(\mathbf{Y} \,|\, \mathrm{do}(\mathbf{X}), \mathbf{W})$.

5.1 Systematic Analysis of Bivariate Missing Data Problems

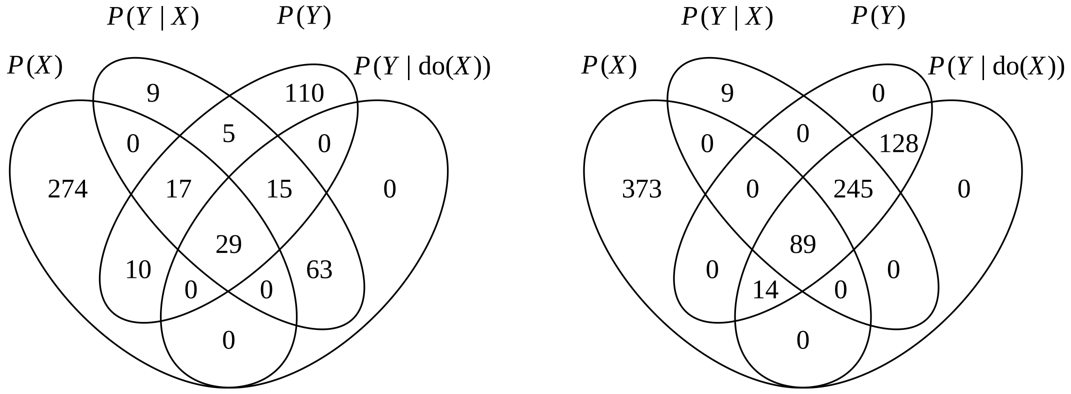
We apply `do-search` using the extended rule set of Table 4 for all identifiability problems in bivariate missingness graphs. By bivariate missingness graphs we mean semi-Markovian graphs for two variables, X and Y, and their missingness indicators, R_X and R_Y. Noting that edges from $\{R_X, R_Y\}$ to $\{X, Y\}$ are not allowed, there are 9216 such graphs. We consider only 6144 graphs of which 3072 have the edge $X \rightarrow Y$ and 3072 do not have an edge between X and Y. Graphs with the edge $Y \rightarrow X$ are obtained from the studied graphs by swapping the roles of X and Y. The maximum number of edges in a bivariate missingness graph is 12 (when a bidirected edge is counted as a single edge).

The available theoretical results for missing data problems include a theorem by Mohan et al. (2013) that gives a sufficient and necessary condition for the identifiability of the joint distribution $P(\mathbf{V})$ but is restricted to graphs that do not have edges between the missingness indicators (row 8 of Table 1). In our example, 5120 graphs out of 6144 have such edges. The algorithm by Shpitser et al. (2015) does not have this restriction but it is not known if the algorithm is complete (row 9 of Table 1). Similarly, it is not known if rules of Table 4 are complete for missing data problems or if some additional rules or tools are needed for identification in general. Differently from the theorem by Mohan et al. (2013) and the algorithm by Shpitser et al. (2015), `do-search` can also address missing data problems where we consider identification of a marginal or conditional distribution.

The queries $P(X, Y)$, $P(X)$, $P(Y)$, $P(Y \,|\, X)$ and $P(Y \,|\, \mathrm{do}(X))$ were evaluated using `do-search` in these 6144 graphs with the input distribution $P(X^*, Y^*, R_X, R_Y)$. The results are summarized by Venn diagrams in Figure 9. The results are also available as a data set `bivariate_missingness` in the R package implementing `do-search`. Using this data set we are able to prove some non-identifiability results and find interesting special cases. The following theorem gives sufficient conditions for non-identifiability in terms of the number of edges.

Theorem 4. *Let K denote the number of edges in a bivariate missingness graph that does not have edge $Y \rightarrow X$. The joint distribution $P(X, Y)$ is not identifiable by `do-search` if $K > 5$, marginal distribution $P(X)$ is not identifiable by `do-search` if $K > 9$, marginal distribution $P(Y)$ and conditional distribution $P(Y \,|\, X)$ are not identifiable by `do-search` if $K > 8$.*

Proof. By direct evaluation in every possible bivariate missingness graph. □



(a) Graphs with arrow $X \to Y$.

(b) Graphs without arrow $X \to Y$ or $Y \to X$.

Figure 9: Venn diagrams indicating the number of graphs were different distributions can be identified `do-search`. The intersection of $P(X)$ and $P(Y \mid X)$ shows the number of graphs were $P(X, Y)$ can be identified. The total number of possible graphs is 3072 in both cases.

The next result specifies the graph with the largest number of edges where both the joint distribution of X and Y and the causal effect of X on Y can be identified.

Theorem 5. *The graph in Figure 10(a) is the only bivariate missingness graph that (i) has edge $X \to Y$, (ii) has five edges, and (iii) allows for the identification of $P(X, Y)$ and $P(Y \mid do(X))$ by `do-search`.*

Proof. By direct evaluation in every possible bivariate missingness graph. □

The third result specifies the graph with the largest number of edges where the marginal distributions are identifiable while the joint distribution and the causal effect of X on Y are non-identifiable.

Theorem 6. *The graph in Figure 10(b) is the only bivariate missingness graph that (i) has five edges, and (ii) allows for the identification of $P(X)$ and $P(Y)$, and (iii) does not allow for the identification of $P(X, Y)$ or $P(Y \mid do(X))$ by `do-search`. No bivariate missingness graph that has more than five edges fulfills the conditions (ii) and (iii).*

Proof. By direct evaluation in every possible bivariate missingness graph. □

Some interesting examples are shown in Figure 10. Graphs (a) and (b) are the unique graphs that fulfill the conditions specified in Theorems 5 and 6, respectively. Graph (c) is the smallest graph were marginals $P(X)$ and $P(Y)$ can be identified but the joint distribution $P(X, Y)$ or causal effect $P(Y \mid \mathrm{do}(X))$ cannot be identified by `do-search`. In graph

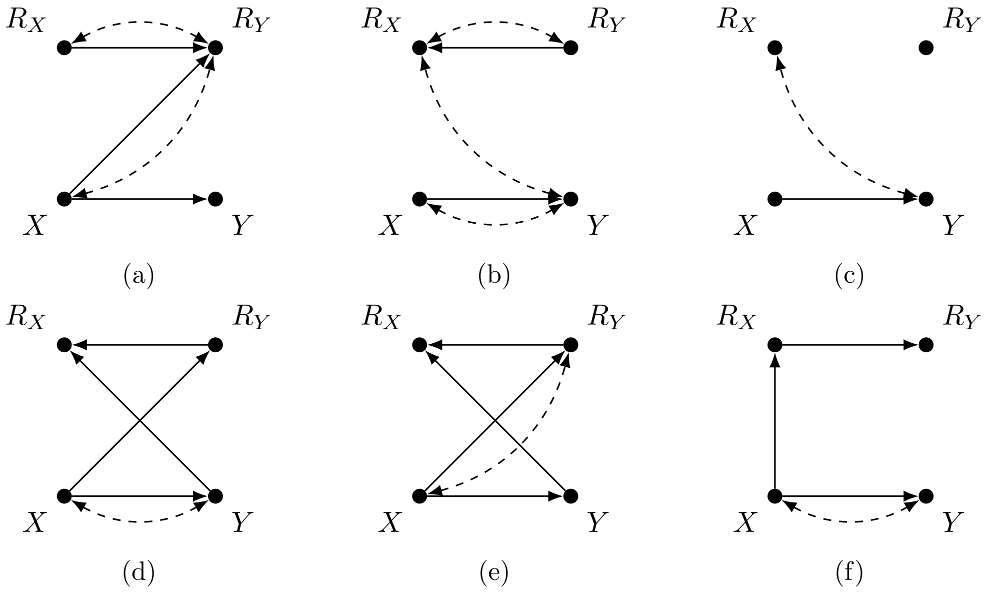


Figure 10: Missingness graphs used as example cases. Proxy variables are omitted for clarity.

(d), $P(X)$, $P(Y)$, $P(X,Y)$ and $P(Y \mid \mathrm{do}(X))$ are not identifiable by `do-search` but the conditional distribution $P(Y \mid X)$ can be identified as follows

$$P(Y \mid X) = \frac{P(Y \mid R_Y = 1)P(X \mid Y, R_X = 1, R_Y = 1)}{\sum_{Y'} P(Y' \mid R_Y = 1)P(X \mid Y', R_X = 1, R_Y = 1)}. \tag{2}$$

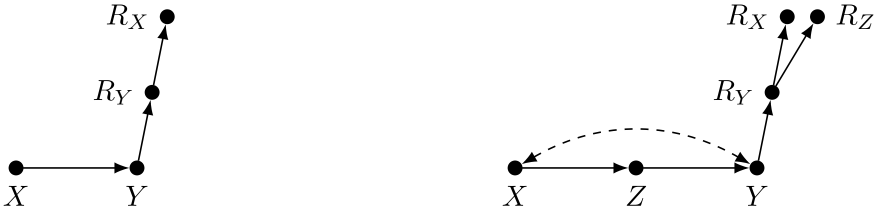
In equation (2), the numerator resembles the joint distribution $P(X,Y \mid R_X = 1, R_Y = 1)$ but is different because Y and R_X are not independent. The denominator is the marginal of this pseudo joint distribution. In graph (e), $P(X)$, $P(Y)$ and $P(X,Y)$ are not identifiable by `do-search` but $P(Y \mid X)$ and $P(Y \mid \mathrm{do}(X))$ are identifiable and can be both estimated with equation (2). In graph (f), $P(X,Y)$, $P(X)$ and $P(Y \mid \mathrm{do}(X))$ are not identifiable by `do-search` but $P(Y)$ and $P(Y \mid X)$ can be identified as follows

$$\begin{aligned} P(Y) &= \sum_{R_X, X^*} P(Y \mid X^*, R_X, R_Y = 1)P(R_X, X^*), \qquad (3) \\ P(Y \mid X) &= P(Y \mid X, R_X = 1, R_Y = 1) \end{aligned}$$

In equation (3), the summation also goes over the cases where $X^* = \mathrm{NA}$ and the distribution of Y must be estimated also on the condition that X is not observed.

5.2 Causal Inference under Case-control Design

Case-control design (Breslow, 1996) is commonly used in epidemiology to study risk factors of rare diseases. In the basic setup, a fixed number of disease cases and a fixed number of controls are selected for the risk factor measurements. When the disease is rare, this design leads to substantial savings in the sample size compared to simple random sampling.



(a) Basic case-control design.

(b) Case-control design for the front-door situation.

Figure 11: Missingness graph for the case-control examples.

Figure 11(a) shows the missingness graph for a situation where the inclusion to the study (indicator R_Y) depends on the disease endpoint Y. The risk factors X are measured for the subset $R_Y = 1$ but occasionally the values are missing (indicator R_X). It is immediately seen that neither the causal effect $P(Y \mid \mathrm{do}(X))$ nor conditional distribution $P(Y \mid X)$ can be identified because of the arrow $Y \rightarrow R_Y$. However, if the prevalence of the disease in population, i.e., the marginal distribution $P(Y)$, is known, the causal effect $P(Y \mid \mathrm{do}(X))$ can be identified. The result is provided by `do-search`

$$P(Y \mid \mathrm{do}(X)) = \frac{P(Y)P(X \mid Y, R_Y = 1, R_X = 1)}{\sum_{Y'} P(Y')P(X \mid Y', R_Y = 1, R_X = 1)}. \tag{4}$$

In typical applications response Y is binary but in the non-parametric formula of equation (4) response can be discrete or continuous.

A more complicated example is shown in Figure 11(b) where the causal effect of risk factor X on disease endpoint Y fulfills the front-door criterion (Pearl, 1995) with respect to mediator Z and the data are collected from a case-control design where the selection depends Y and there is occasional item non-response in X and Z. We observe data $P(Y^*, X^*, Z^*, R_Y, R_X, R_Z)$ and know the marginal distribution $P(Y)$ from other sources. Applying `do-search` we obtain the result

$$\begin{aligned} &P(Y \mid \mathrm{do}(X)) = \\ &\sum_Z \Bigg[\frac{\sum_{Y'} P(Y')P(X, Z \mid Y', R_X = 1, R_Y = 1, R_Z = 1)}{\sum_{Z',Y'} P(Y')P(X, Z' \mid Y', R_X = 1, R_Y = 1, R_Z = 1)} \times \\ &\quad \sum_{X'} \Bigg(\sum_{Y',Z'} P(Y')P(X', Z' \mid Y', R_X = 1, R_Y = 1, R_Z = 1) \times \\ &\quad\quad \frac{P(Y)P(X', Z \mid Y, R_X = 1, R_Y = 1, R_Z = 1)}{\sum_{Y'} P(Y')P(X', Z \mid Y', R_X = 1, R_Y = 1, R_Z = 1)} \Bigg) \Bigg]. \end{aligned} \tag{5}$$

Expression (5) follows the general structure of the front-door adjustment

$$P(Y \mid \mathrm{do}(X)) = \sum_Z P(Z \mid X) \sum_{X'} P(X')P(Y \mid X', Z),$$

where

$$
\begin{aligned}
P(Z \,|\, X) &= \frac{\sum_{Y'} P(Y')P(X, Z \,|\, Y', R_X = 1, R_Y = 1, R_Z = 1)}{\sum_{Z',Y'} P(Y')P(X, Z' \,|\, Y', R_X = 1, R_Y = 1, R_Z = 1)}, \\
P(X) &= \sum_{Y',Z'} P(Y')P(X, Z' \,|\, Y', R_X = 1, R_Y = 1, R_Z = 1), \\
P(Y \,|\, X, Z) &= \frac{P(Y)P(X, Z \,|\, Y, R_X = 1, R_Y = 1, R_Z = 1)}{\sum_{Y'} P(Y')P(X, Z \,|\, Y', R_X = 1, R_Y = 1, R_Z = 1)}.
\end{aligned}
$$

Note that $P(X, Y, Z) = P(Y)P(X, Z \,|\, Y, R_X = 1, R_Y = 1, R_Z = 1)$. In (Karvanen, 2015), a similar example was studied assuming that X, Z and Y are binary but in expression (5) there are no such restrictions.

6 Discussion

The presented algorithm, `do-search`, removes the need for manual application of do-calculus, which is time-consuming and prone to errors. Systematic analyses such as the one in Section 5.1 are practically unreachable with manual application of do-calculus. Superiority of `do-search` over a simple forwards breadth-first search was attained through a combination of a search heuristic and a reduction of the search space. Some further approaches were attempted but later discarded as non-beneficial. These include caching d-separation criteria that hold in the graph after they are first evaluated, pre-computing valid subsets for each subset size and enumerating subsets in an order of increasing cardinality.

As the simulations showed, our intuitive heuristic yielded significant improvements in search performance. The proximity function defined in Section 3.2.5 uses only the information contained in the distributions themselves. One approach could be to also take the structure of the graph into account in the proximity function. Further study is needed for finding a heuristic that performs well when missing data mechanisms are present in the graph.

The scalability of `do-search` is limited due to vast search space of possibly identified causal effects. Currently, algorithms with polynomial complexity currently exist only for the simpler problems (see Table 1). However, based on the simulation results, `do-search` solves identifiability problems in graphs of ten vertices in under two minutes on average. Typically graphs analyzed in literature related to identifiability problems have fewer vertices. The theoretical computational complexity of the general form of the causal identifiability problem defined in Section 2 remains an important and interesting question.

The search could also be used to obtain formulas that are in some sense simpler than those produced by existing identifiability algorithms. A simplification algorithm by Tikka and Karvanen (2017b) functions as a post-processing step after the identifying formula has already been obtained by the ID algorithm, but has exponential complexity. Given a measure of simplicity, the search heuristic could be adjusted to find simple formulas directly without resorting to separate simplification procedures. In some specific contexts, such as the standard causal effect identifiability problem, an approach known as pruning (Tikka and Karvanen, 2018a) could be incorporated into the search. Pruning refers to the removal of vertices from the graph, that are not required for determining identifiability.

Finally we note that identifiability has also been studied under the assumption that the functional relationships depicted by the causal model are linear (Angrist et al., 1996; van der Zander and Liskiewicz, 2016; Chen et al., 2017) or non-parametric with additive error terms (Peters et al., 2014; Peña and Bendtsen, 2017) and when the causal graph is not completely known (Maathuis et al., 2009; Entner et al., 2013; Hyttinen et al., 2015; Perković et al., 2015; Malinsky and Spirtes, 2017; Jaber et al., 2018). Extending the search in these directions is an interesting line of future research.

7 Conclusion

We presented `do-search`: a do-calculus based search capable of solving identifiability problems for which no known solutions exist. This contribution is especially useful for researchers working in the field of causal inference to confirm theoretical results or to find counterexamples to identifiability claims. In practical terms, the search can also provide solutions to complicated problems such as combining transportability and selection bias, recovering from multiple bias sources or identifying causal quantities in the presence of missing data that cannot be solved by any other existing method. An R package providing an implementation of `do-search` is available on CRAN.

Acknowledgments

This work belongs to the thematic research area "Decision analytics utilizing causal models and multiobjective optimization" (DEMO) supported by Academy of Finland (grant number 311877). AH was supported by Academy of Finland through grant 295673. The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

E. Bareinboim and J. Pearl. Controlling selection bias in causal inference. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, volume 22, pages 100–108, 2012a.

E. Bareinboim and J. Pearl. Causal inference by surrogate experiments: z-identifiability. In N. de Freitas and K. Murphy, editors, *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 113–120. AUAI Press, 2012b.

E. Bareinboim and J. Pearl. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1:107–134, 2013.

E. Bareinboim and J. Pearl. Transportability from multiple environments with limited experiments: Completeness results. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pages 280–288, 2014.

E. Bareinboim and J. Tian. Recovering causal effects from selection bias. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 3475–3481, 2015.

E. Bareinboim, J. Tian, and J. Pearl. Recovering from selection bias in causal and statistical inference. In *Proceedings of the 28th AAAI Conference on Neural Information Processing Systems*, 2014.

N. E. Breslow. Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, 91(433):14–28, 1996.

B. Chen, D. Kumor, and E. Bareinboim. Identification and model testing in linear structural equation models using auxiliary variables. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 757–766, 2017.

J. Correa and E. Bareinboim. Causal effect identification by adjustment under confounding and selection biases. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017.

J. Correa, J. Tian, and E. Bareinboim. Generalized adjustment under confounding and selection biases. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

D. Danks, C. Glymour, and R. E. Tillman. Integrating locally learned causal structures with overlapping variables. In *Advances in Neural Information Processing Systems*, pages 1665–1672, 2009.

A. P. Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189, 2002.

D. Entner, P. Hoyer, and P. Spirtes. Data-driven covariate selection for nonparametric estimation of causal effects. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, volume 31, pages 256–264. PMLR, 2013.

S. Greenland, J. M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46, 1999.

Y. Huang and M. Valtorta. Pearl's calculus of intervention is complete. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 217–224. AUAI Press, 2006.

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Causal discovery of linear cyclic models from multiple experimental data sets with overlapping variables. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 387–396, 2012.

A. Hyttinen, F. Eberhardt, and M. Järvisalo. Do-calculus when the true graph is unknown. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 395–404. AUAI Press, 2015.

A. Jaber, J. Zhang, and E. Bareinboim. Causal identification under Markov equivalence. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, pages 978–987. AUAI Press, 2018.

J. Karvanen. Study design in causal models. *Scandinavian Journal of Statistics*, 42(2): 361–377, 2015.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, 2009.

S. L. Lauritzen. Causal inference from graphical models. In D. R. Cox O. E. Barndorff-Nielsen and C. Klüppelberg, editors, *Complex Stochastic Systems*, chapter 2, pages 67–107. CRC Press, 2000.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* John Wiley & Sons, Inc., New York, NY, USA, 1986.

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

D. Malinsky and P. Spirtes. Estimating bounds on causal effects in high-dimensional and possibly confounded systems. *International Journal of Approximate Reasoning*, 88:371–384, 2017.

K. Mohan and J. Pearl. Graphical models for processing missing data. 2018. Forthcoming, https://arxiv.org/abs/1801.03583.

K. Mohan, J. Pearl, and J. Tian. Graphical models for inference with missing data. In *Advances in Neural Information Systems*, volume 26, pages 1277–1285, 2013.

J. M. Peña and M. Bendtsen. Causal effect identification in acyclic directed mixed graphs and gated models. *International Journal of Approximate Reasoning*, 90:56–75, 2017.

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

J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, second edition, 2009.

E. Perković, J. Textor, M. Kalisch, and M. Maathuis. A complete generalized adjustment criterion. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 682–691. AUAI Press, 2015.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence – Volume 2*, pages 1219–1226. AAAI Press, 2006a.

I. Shpitser and J. Pearl. Identification of conditional interventional distributions. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 437–444. AUAI Press, 2006b.

I. Shpitser and J. Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

I. Shpitser, K. Mohan, and J. Pearl. Missing data as a causal and probabilistic problem. In Marina Meila and Tom Heskes, editors, *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 802–811. AUAI Press, 2015.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* Springer-Verlag, 1993. (2nd ed. MIT Press 2000).

S. Tikka and J. Karvanen. Identifying causal effects with the R package causaleffect. *Journal of Statistical Software*, 76(12):1–30, 2017a.

S. Tikka and J. Karvanen. Simplifying probabilistic expressions in causal inference. *Journal of Machine Learning Research*, 18(36):1–30, 2017b.

S. Tikka and J. Karvanen. Enhancing identification of causal effects by pruning. *Journal of Machine Learning Research*, 18(194):1–23, 2018a.

S. Tikka and J. Karvanen. Surrogate outcomes and transportability. Submitted manuscript, http://arxiv.org/abs/1806.07172, 2018b.

S. Tikka, A. Hyttinen, and J. Karvanen. Causal effect identification from multiple incomplete data sources: A general search-based approach. Submitted manuscript, https://arxiv.org/abs/1902.01073, 2019.

R. Tillman and P. Spirtes. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 3–15, 2011.

S. Triantafillou and I. Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16: 2147–2205, 2015.

S. Triantafillou, I. Tsamardinos, and I. Tollis. Learning causal structure from overlapping variable sets. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 860–867, 2010.

B. van der Zander and M. Liskiewicz. On searching for generalized instrumental variables. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016.