

---

# Nested Counterfactual Identification from Arbitrary Surrogate Experiments

---

**Juan D. Correa**  
Columbia University  
jdcorrea@cs.columbia.edu

**Sanghack Lee**  
Seoul National University  
sanghack@snu.ac.kr

**Elias Bareinboim**  
Columbia University  
eb@cs.columbia.edu

## Abstract

The *Ladder of Causation* describes three qualitatively different types of activities an agent may be interested in engaging in, namely, seeing (observational), doing (interventional), and imagining (counterfactual) (Pearl and Mackenzie, 2018). The inferential challenge imposed by the causal hierarchy is that data is collected by an agent observing or intervening in a system (layers 1 and 2), while its goal may be to understand what would have happened had it taken a different course of action, contrary to what factually ended up happening (layer 3). While there exists a solid understanding of the conditions under which cross-layer inferences are allowed from observations to interventions, the results are somewhat scarcer when targeting counterfactual quantities. In this paper, we study the identification of nested counterfactuals from an arbitrary combination of observations and experiments. Specifically, building on a more explicit definition of nested counterfactuals, we prove the counterfactual unnesting theorem (CUT), which allows one to map arbitrary nested counterfactuals to unnested ones. For instance, applications in mediation and fairness analysis usually evoke notions of direct, indirect, and spurious effects, which naturally require nesting. Second, we introduce a sufficient and necessary graphical condition for counterfactual identification from an arbitrary combination of observational and experimental distributions. Lastly, we develop an efficient and complete algorithm for identifying nested counterfactuals; failure of the algorithm returning an expression for a query implies it is not identifiable.

## 1 Introduction

Counterfactuals provide the basis for notions pervasive throughout human affairs, such as credit assignment, blame and responsibility, and regret. One of the most powerful constructs in human reasoning —“what if?” questions— evokes hypothetical conditions usually contradicting the factual evidence. Judgment and understanding of critical situations found from medicine to psychology to business involve counterfactual reasoning, e.g.: “Joe received the treatment and died, would he be alive had he not received it?” “Had the candidate been male instead of female, would the decision from the admissions committee be more favorable?” or “Would the profit this quarter remain within 5% of its value had we increased the price by 2%?”. By and large, counterfactuals are key ingredients that go in the construction of explanations about why things happened as they did [17, 19].

The structural interpretation of causality provides proper semantics for representing counterfactuals [17, Ch. 7]. Specifically, each structural causal model (SCM)  $\mathcal{M}$  induces a collection of distributions related to the activities of seeing (called observational), doing (interventional), and imagining (counterfactual). The collection of these distributions is known as the *Ladder of causation* [19], and has also been called the *Pearl’s Causal Hierarchy* (PCH, for short) [2]. The PCH is a containment hierarchy; each type of distribution can be put in increasingly refined layers: observational content goes in layer 1; experimental in layer 2; counterfactual in layer 3; see Fig. 1.

It is understood that if we have all the information in the world about layer 1, there are still questions about layers 2 and 3 that are unanswerable, or technically undetermined; further, if we have data from layers 1 and 2, there are still questions in the world about layer 3 that are underdetermined [17, 2].

The inferential challenge in these settings arises because the generating model  $\mathcal{M}$  is not fully observed, nor data from all of the layers are necessarily available, perhaps due to the cost or the infeasibility of performing certain interventions. One common task found in the literature is to determine the effect of an intervention of a variable  $X$  on an outcome  $Y$ , say  $P(Y|do(X))$  (layer 2), using data from observations  $P(\mathbf{V})$  (layer 1), where  $\mathbf{V}$  is the set of observed variables, and possibly other interventions, e.g.,  $P(\mathbf{V}|do(Z))$ . Also, qualitative assumptions about the system are usually articulated in the form of a causal diagram  $\mathcal{G}$ . This setting has been studied in the literature under the rubric of non-parametric identification from a combination of observations and experiments. Multiple solutions exist, including Pearl’s celebrated do-calculus [16], and other increasingly refined methods that are computationally efficient, sufficient, and necessary [25, 8, 20, 26, 21, 10, 3, 13, 12].<sup>1</sup>

There is a growing literature about cross-layer inferences from data in layers 1 and 2 to quantities in layer 3. For example, a data scientist may be interested in evaluating the effect of an intervention on the group of subjects who receive the treatment instead of those randomly assigned to it. This measure is known as the *effect of treatment on the treated* [9, 17], and there exists a graphical condition for mapping it to a (layer 2) causal effect [23]. Further, there are also results on the identification of *path-specific effects*, which are counterfactuals that isolate specific paths in the graph [18, 1]. In particular, [24] provides a sufficient and necessary algorithm for identification of these effects from observational data, and [28] provides identification conditions from observational and experimental data in general canonical models. Further, [22] studied counterfactual identification under the assumption that all experimental distributions (i.e., over every subset of the observed variables) are available.<sup>2</sup>

In this paper, our goal is to identify the probability distribution of (possibly nested) counterfactual events from an *arbitrary* combination of user-specified observational and experimental distributions. To the best of our knowledge, this provides the first general treatment of nested counterfactual identification from arbitrary data collections. Moreover, it also provides the first, graphical and algorithmic, sufficient and necessary conditions for the identifications of counterfactuals from observational data alone (when no experimental data is available) and arbitrary causal diagrams. Moving up the PCH, our results allow for arbitrary quantities as inferential targets and for the addition of arbitrary experimental distributions to the input, increasing the flexibility of the solution.

For concreteness, consider the causal diagram shown in Fig. 2 and a counterfactual query called *direct effect*. This quantity represents the sensitivity of a variable  $Y$  to changes in another variable  $X$  while all other factors in the analysis remain fixed. Suppose  $X$  is level of exercise,  $M$  cholesterol levels, and  $Y$  cardiovascular disease. Exercising can improve cholesterol levels, which in turn affect the chances of developing cardiovascular disease. An interesting question is how much exercise prevents the disease by means other than regulating cholesterol. In counterfactual notation, this is to compare  $Y_{x, M_x}$  and  $Y_{x', M_x}$ , where  $x$  and  $x'$  are different values. The first quantity represents the value of  $Y$  when  $X=x$  and  $M$  varies accordingly. The second expression is the value  $Y$  attains if  $X$  is held constant at  $x'$  while  $M$  still follows  $X=x$ . The difference  $E[Y_{x', M_x} - Y_{x, M_x}]$  — known as the natural direct effect (NDE) — is non-zero if there is some direct effect of  $X$  on  $Y$ . In this instance, this nested counterfactual is identifiable only if observational data and experiments on  $X$  are available.

<sup>1</sup>In fact, this is a classic task in a larger family of problems known as data fusion, which include other challenges such as selection bias, transportability, to cite a few. For more details, see [4].

<sup>2</sup>For the sake of context, the work proposed here can be seen as a generalization of two tasks, counterfactual identification under the assumptions discussed earlier [22] and interventional identification from arbitrary experiments [13]. As discussed later on, we will be able to show, based on the machinery developed here, that the individual methods for those tasks can be combined and also be shown complete.

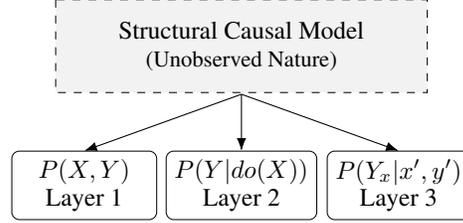


Figure 1: Every SCM induces different quantities in each layer of the PCH.

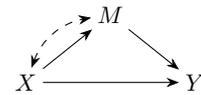


Figure 2: Causal diagram with treatment  $X$ , outcome  $Y$ , and mediator  $M$ .

After all, there is no general identification method for this particular counterfactual family (which also includes indirect and spurious effects) and, more broadly, other arbitrary nested counterfactuals that are well-defined in layer 3. Our goal is to understand the non-parametric identification of arbitrary nested and conditional counterfactuals when the input consists of any combination of observational and interventional distributions, whatever is available for the data scientist. More specifically, our contributions are as follows.

1. We look at nested counterfactuals from an SCM perspective and introduce machinery that supports counterfactual reasoning. In particular, we prove the counterfactual unnesting theorem (CUT), which allows one to map any nested counterfactual to an unnested one (Section 2).
2. Building on this new machinery, we derive sufficient and necessary graphical conditions and an algorithm to determine the identifiability of marginal nested counterfactuals from an arbitrary combination of observational and experimental distributions (Section 3).
3. We prove a reduction from conditional counterfactuals to marginal ones, and use it to derive a complete algorithm for their identification (Section 4).

Due to space constraints, all the proofs in this paper can be found in the full technical report [5].

## 1.1 Preliminaries

We denote variables by capital letters,  $X$ , and values by small letters,  $x$ . Bold letters,  $\mathbf{X}$  represent sets of variables and  $\mathbf{x}$  sets of values. The domain of a variable  $X$  is denoted by  $\mathfrak{X}_X$ . Two values  $\mathbf{x}$  and  $\mathbf{z}$  are said to be consistent if they share the common values for  $\mathbf{X} \cap \mathbf{Z}$ . We also denote by  $\mathbf{x} \setminus \mathbf{Z}$  the value of  $\mathbf{X} \setminus \mathbf{Z}$  consistent with  $\mathbf{x}$  and by  $\mathbf{x} \cap \mathbf{Z}$  the subset of  $\mathbf{x}$  corresponding to variables in  $\mathbf{Z}$ . We assume the domain of every variable is finite.

We rely on causal graphs and denote them with a calligraphic letter, e.g.,  $\mathcal{G}$ . We denote the set of vertices (i.e., variables) in  $\mathcal{G}$  as  $\mathbf{V}(\mathcal{G})$ . Given a graph  $\mathcal{G}$ ,  $\mathcal{G}_{\overline{\mathbf{W}\mathbf{X}}}$  is the result of removing edges coming into variables in  $\mathbf{W}$  and going out from variables in  $\mathbf{X}$ .  $\mathcal{G}[\mathbf{W}]$  denotes a vertex-induced subgraph including  $\mathbf{W}$  and the edges among its elements. We use kinship notation for graphical relationships such as parents, children, descendants, and ancestors of a set of variables. For example, the set of parents of  $\mathbf{X}$  in  $\mathcal{G}$  is  $Pa(\mathbf{X})_{\mathcal{G}} := \mathbf{X} \cup \bigcup_{X \in \mathbf{X}} Pa(X)_{\mathcal{G}}$ . Similarly, we define  $Ch()$ ,  $De()$ , and  $An()$ .

To articulate and formalize counterfactual questions, we require a framework that allows us to reason simultaneously about events from *alternative worlds*. Accordingly, we employ the Structural Causal Model (SCM) paradigm [17, Ch. 7]. An SCM  $\mathcal{M}$  is a 4-tuple  $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$ , where  $\mathbf{U}$  is a set of exogenous (latent) variables;  $\mathbf{V}$  is a set of endogenous (observable) variables;  $\mathcal{F}$  is a collection of functions such that each variable  $V_i \in \mathbf{V}$  is determined by a function  $f_i \in \mathcal{F}$ . Each  $f_i$  is a mapping from a set of exogenous variables  $\mathbf{U}_i \subseteq \mathbf{U}$  and a set of endogenous variables  $\mathbf{Pa}_i \subseteq \mathbf{V} \setminus \{V_i\}$  to the domain of  $V_i$ . Uncertainty is encoded through a probability distribution over the exogenous variables,  $P(\mathbf{U})$ . An SCM  $\mathcal{M}$  induces a *causal diagram*  $\mathcal{G}$  where  $\mathbf{V}$  is the set of vertices, there is a directed edge  $(V_j \rightarrow V_i)$  for every  $V_i \in \mathbf{V}$  and  $V_j \in \mathbf{Pa}_i$ , and a bidirected edge  $(V_i \leftrightarrow V_j)$  for every pair  $V_i, V_j \in \mathbf{V}$  such that  $U_i \cap U_j \neq \emptyset$  ( $V_i$  and  $V_j$  have a common exogenous parent).

We assume that the underlying model is recursive. That is, there are no cyclic dependencies among the variables. Equivalently, that is to say, that the corresponding causal diagram is acyclic. The set  $\mathbf{V}(\mathcal{G})$  can be partitioned into subsets called *c-components* [27] such that two variables belong to the same c-component if they are connected in  $\mathcal{G}$  by a path made entirely of bidirected edges.

## 2 SCMs and Nested Counterfactuals

Intervening on a system represented by an SCM  $\mathcal{M}$  results in a new model differing only on the mechanisms associated with the intervened variables [15, 6, 7]. If the intervention consists on fixing the value of a variable  $X$  to a constant  $x \in \mathfrak{X}_X$ , it induces a *submodel*, denoted  $\mathcal{M}_x$  [17, Def. 7.1.2]. To formally study nested counterfactuals, we extend this notion to models derived from interventions that replace functions from the original SCM with other, not necessarily constant, functions.

**Definition 1** (Derived Model). Let  $\mathcal{M}$  be an SCM,  $\widehat{\mathbf{U}} \subseteq \mathbf{U}$ ,  $X \in \mathbf{V}$ , and  $\widehat{X} : \widehat{\mathbf{U}} \rightarrow \mathfrak{X}_X$  a function. Then,  $\mathcal{M}_{\widehat{X}}$ , called the *derived model* of  $\mathcal{M}$  according to  $\widehat{X}$ , is identical to  $\mathcal{M}$ , except that the function  $f_X$  is replaced with a function  $\widehat{f}_X$  identical to  $\widehat{X}$ .

This definition is easily extendable to models derived from an intervention on a set  $\mathbf{X}$  instead of a singleton. When  $\widehat{\mathbf{X}}$  is a collection of functions  $\{\widehat{X} : \widehat{\mathbf{U}}_X \rightarrow \mathfrak{X}_X\}_{X \in \mathbf{X}}$ , the derived model  $\mathcal{M}_{\widehat{\mathbf{X}}}$  is obtained by replacing each  $f_X$  with  $\widehat{X}$  for  $X \in \mathbf{X}$ . Next, we discuss the concept of *potential response* [17, Def. 7.4.1] with respect to the derived models.

**Definition 2** (Potential Response). Let  $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$  be subsets of observable variables, let  $\mathbf{u}$  be a unit, and let  $\widehat{\mathbf{X}}(\mathbf{u})$  be a set of functions from  $\widehat{\mathbf{U}}_X \rightarrow \mathfrak{X}_X$ , for  $X \in \mathbf{X}$  where  $\widehat{\mathbf{U}}_X \subseteq \mathbf{U}$ . Then,  $\mathbf{Y}_{\mathbf{X}=\widehat{\mathbf{X}}}(\mathbf{u})$  (or  $\mathbf{Y}_{\widehat{\mathbf{X}}}(\mathbf{u})$ , for short) is called the *potential response* of  $\mathbf{Y}$  to  $\mathbf{X} = \widehat{\mathbf{X}}$ , and is defined as the solution of  $\mathbf{Y}$ , for a particular  $\mathbf{u}$ , in the derived model  $\mathcal{M}_{\widehat{\mathbf{X}}}$ .

A potential response  $Y_{\widehat{\mathbf{X}}}(\mathbf{u})$  describes the value that variable  $Y$  would attain for a unit (or individual)  $\mathbf{u}$  if the intervention  $\widehat{\mathbf{X}}$  is performed. This concept is tightly related to that of *potential outcome*, but the former explicitly allows for interventions that do not necessarily fix the variables in  $\mathbf{X}$  to a constant value. Averaging over the space of  $\mathbf{U}$ , a potential response  $Y_{\widehat{\mathbf{X}}}(\mathbf{u})$  induces a random variable that we will denote simply as  $Y_{\widehat{\mathbf{X}}}$ . If the intervention replaces a function  $f_X$  with a potential response of  $X$  in  $\mathcal{M}$ , we say the intervention is *natural*.

When variables are enumerated as  $W_1, W_2, \dots$ , we may add square brackets around the part of the subscript denoting interventions. We use  $\mathbf{W}_*$  to denote sets of arbitrary counterfactual variables. Let  $\mathbf{W}_* = \{W_{1[\widehat{\mathbf{T}}_1]}, W_{2[\widehat{\mathbf{T}}_2]}, \dots\}$  represent a set of counterfactual variables such that  $W_i \in \mathbf{V}$  and  $\mathbf{T}_i \subseteq \mathbf{V}$  for  $i = 1, \dots, l$ . Define  $\mathbf{V}(\mathbf{W}_*) = \{W \in \mathbf{V} \mid W_{\widehat{\mathbf{T}}} \in \mathbf{W}_*\}$ , that is, the set of observables that appear in  $\mathbf{W}_*$ . Let  $\mathbf{w}_*$  represent a vector of values, one for each variable in  $\mathbf{W}_*$  and define  $\mathbf{w}_*(\mathbf{X}_*)$  as the subset of  $\mathbf{w}_*$  corresponding to  $\mathbf{X}_*$  for any  $\mathbf{X}_* \subseteq \mathbf{W}_*$ .

The probability of any counterfactual event is given by

$$P(\mathbf{Y}_* = \mathbf{y}_*) = \sum_{\{\mathbf{u} \mid \mathbf{Y}_*(\mathbf{u}) = \mathbf{y}_*\}} P(\mathbf{u}), \quad (1)$$

where the predicate  $\mathbf{Y}_*(\mathbf{u}) = \mathbf{y}_*$  means  $\bigwedge_{\{Y_{\widehat{\mathbf{T}}} \in \mathbf{Y}_*\}} Y_{\widehat{\mathbf{T}}}(\mathbf{u}) = y$ .

When all variables in the expression have the same subscript, that is, they belong to the same submodel; we will often denote it as  $P_{\mathbf{X}}(W_1, W_2, \dots)$ .

For most real-world scenarios, having access to a fully specified SCM of the underlying system is unfeasible. Nevertheless, our analysis does not rely on such privileged access but the aspects of the model captured by the causal graph and data samples generated by the unobserved model.

## 2.1 Nested Counterfactuals

Potential responses can be compounded based on natural interventions. For instance, the counterfactual  $Y_{Z_x}(\mathbf{u})$  ( $Y_{Z=Z_x}(\mathbf{u})$ ) can be seen as the potential response of  $Y$  to an intervention that makes  $\widehat{Z}$  equal to  $Z_x$ . Notice that  $Z_x(\mathbf{u})$  is in itself a potential response, but from a different (nested) model. Hence we call  $Y_{Z_x}$  a *nested counterfactual*.

Recall the causal diagram in Fig. 2 and consider once again the NDE as

$$NDE_{x \rightarrow x', Z}(Y) = E[Y_{x'Z_x}] - E[Y_x]. \quad (2)$$

The second term is also equal to  $Y_{xZ_x}$  as  $Z_x$  is consistent with  $X = x$ , so it is the value  $Y$  listens to in  $\mathcal{M}_x$ . Meanwhile, the first one is related to  $P(Y_{x'Z_x})$ , the probability of a nested counterfactual.

## 2.2 Tools for Counterfactual Reasoning

Before characterizing the identification of counterfactuals from observational and experimental data, we develop from first principles a canonical representation of any such query. First, we extend the notion of ancestors for counterfactual variables, which subsumes the usual one described before.

**Definition 3** (Ancestors, of a counterfactual). Let  $Y_{\mathbf{x}}$  be such that  $Y \in \mathbf{V}, \mathbf{x} \subseteq \mathbf{V}$ . Then, the set of (counterfactual) ancestors of  $Y_{\mathbf{x}}$ , denoted  $An(Y_{\mathbf{x}})$ , consist of each  $W_{\mathbf{z}}$ , such that  $W \in An(Y)_{\mathcal{G}_{\widehat{\mathbf{X}}}}$  (which includes  $Y$  itself), and  $\mathbf{z} = \mathbf{x} \cap An(W)_{\mathcal{G}_{\widehat{\mathbf{X}}}}$ .

For a set of variables  $\mathbf{W}_*$ , we define  $An(\mathbf{W}_*)$  as the union of the ancestors of each variable in the set. That is,  $An(\mathbf{W}_*) = \bigcup_{W_{\mathbf{t}} \in \mathbf{W}_*} An(W_{\mathbf{t}})$ . For instance, in Fig. 3(a),  $An(Y_x) = \{Y_x, Z\}$ ,  $An(X_{yz}) =$

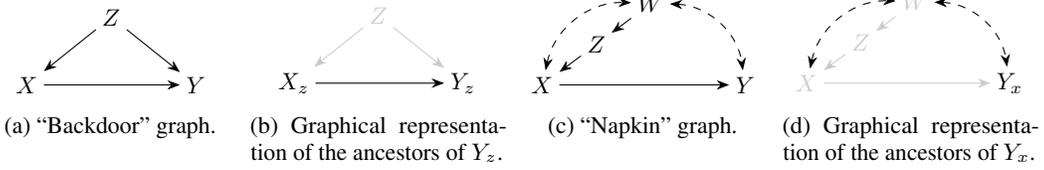


Figure 3: Two causal diagrams and the subgraphs considered when finding sets of ancestors for a counterfactual variable.

$\{X_z\}$  and  $An(Y_z) = \{Y_z, X_z\}$  (depicted in Fig. 3(b)). In Fig. 3(c)  $An(Z, Y_z) = \{Y_z, X_z, Z, W\}$  and  $An(Y_x) = \{Y_x\}$  (represented in Fig. 3(d)).

Probabilistic and causal inference with graphical models exploits the local structure among variables, specifically parent-child relationships, to infer and even estimate probabilities. In particular, Tian [27] introduced *c-factors* which have proven instrumental in solving many problems in causal inference. We naturally generalize this notion to the counterfactual setting with the following definition.

**Definition 4** (Counterfactual Factor (ctf-factor)). A counterfactual factor is a distribution of the form

$$P(W_{1[\mathbf{pa}_1]} = w_1, W_{2[\mathbf{pa}_2]} = w_2, \dots, W_{l[\mathbf{pa}_l]} = w_l), \quad (3)$$

where each  $W_i \in \mathbf{V}$  and there could be  $W_i = W_j$  for some  $i, j \in \{1, \dots, l\}$ .

For example, for Fig. 3(c)  $P(Y_x = y, Y_{x'} = y')$ ,  $P(Y_x = y, X_z = x)$  are ctf-factors but  $P(Y_z = y, Z_w = z)$  is not. Using the notion of ancestry introduced in Definition 3, we can factorize counterfactual probabilities as ctf-factors.

**Theorem 1** (Ancestral set factorization). Let  $\mathbf{W}_*$  be an ancestral set, that is,  $An(\mathbf{W}_*) = \mathbf{W}_*$ , and let  $\mathbf{w}_*$  be a vector with a value for each variable in  $\mathbf{W}_*$ . Then,

$$P(\mathbf{W}_* = \mathbf{w}_*) = P\left(\bigwedge_{W_t \in \mathbf{W}_*} W_{\mathbf{pa}_w} = w\right), \quad (4)$$

where each  $w$  is  $w_t$  and  $\mathbf{pa}_w$  is determined for each  $W_t \in \mathbf{W}_*$  as follows:

- (i) the values for variables in  $\mathbf{Pa}_w \cap \mathbf{T}$  are the same as in  $\mathbf{t}$ , and
- (ii) the values for variables in  $\mathbf{Pa}_w \setminus \mathbf{T}$  are taken from  $\mathbf{w}_*$  corresponding to the parents of  $W$ .

*Proof outline.* Following a reverse topological order in  $\mathcal{G}$ , look at each  $W_{it_i} \in \mathbf{W}_*$ . Since any parent of  $W_i$  not in  $\mathbf{T}_i$  must appear in  $\mathbf{W}_*$ , the composition axiom [17, 7.3.1] licenses adding them to the subscript. Then, by exclusion restrictions [16], any intervention not involving  $Pa(W_i)$  can be removed to obtain the form in Eq. (4).  $\square$

For example, consider the diagram in Fig. 3(c) and the counterfactual  $P(Y_x = y \mid X = x')$  known as the *effect of the treatment on the treated* (ETT) [9, 17]. First note that  $P(Y_x = y \mid X = x') = P(Y_x = y, X = x')/P(X = x')$  and that  $An(Y_x, X) = \{Y_x, X, Z, W\}$ , then

$$P(Y_x = y, X = x') = \sum_{z,w} P(Y_x = y, X = x', Z = z, W = w). \quad (5)$$

Then, by Theorem 1 we can write

$$P(Y_x = y, X = x') = \sum_{z,w} P(Y_x = y, X_z = x', Z_w = z, W = w). \quad (6)$$

Moreover, the following result describes a factorization of ctf-factors based on the c-component structure of the graph, which will prove instrumental in the next section.

**Theorem 2** (Counterfactual factorization). Let  $P(\mathbf{W}_* = \mathbf{w}_*)$  be a ctf-factor, let  $W_1 < W_2 < \dots$  be a topological order over the variables in  $\mathcal{G}[\mathbf{V}(\mathbf{W}_*)]$ , and let  $\mathbf{C}_1, \dots, \mathbf{C}_k$  be the c-components of the same graph. Define  $\mathbf{C}_{j*} = \{W_{\mathbf{pa}_w} \in \mathbf{W}_* \mid W \in \mathbf{C}_j\}$  and  $\mathbf{c}_{j*}$  as the values in  $\mathbf{w}_*$  corresponding to  $\mathbf{C}_{j*}$ , then  $P(\mathbf{W}_* = \mathbf{w}_*)$  decomposes as

$$P(\mathbf{W}_* = \mathbf{w}_*) = \prod_j P(\mathbf{C}_{j*} = \mathbf{c}_{j*}). \quad (7)$$

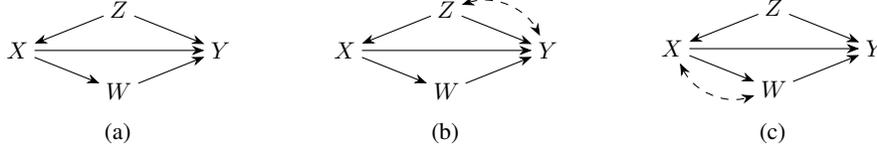


Figure 4: Three causal diagrams representing plausible structures in mediation analysis.

Furthermore, each factor can be computed from  $P(\mathbf{W}_* = \mathbf{w}_*)$  as

$$P(\mathbf{C}_{j_*} = \mathbf{c}_{j_*}) = \prod_{\{W_i \in \mathbf{C}_{j_*}\}} \frac{\sum_{\{w | W_{\text{pa}_w} \in \mathbf{W}_*, W_i < W\}} P(\mathbf{W}_* = \mathbf{w}_*)}{\sum_{\{w | W_{\text{pa}_w} \in \mathbf{W}_*, W_{i-1} < W\}} P(\mathbf{W}_* = \mathbf{w}_*)}. \quad (8)$$

Armed with these results, we consider the identification problem in the next section.

### 3 Counterfactual Identification from Observations and Experiments

In this section, we consider the identification of a counterfactual probability from a collection of observational and experimental distributions. This task can be seen as a generalization of that in [13] where the available data is the same, but the query is a causal effect  $P_{\mathbf{x}}(\mathbf{Y})$ . Let  $\mathbb{Z} = \{\mathbf{Z}_1, \mathbf{Z}_2, \dots\}$ ,  $\mathbf{Z}_j \subseteq \mathbf{V}$ , and assume that all of  $\{P_{\mathbf{z}_j}(\mathbf{V})\}_{\mathbf{z}_j \in \mathcal{X}_{\mathbf{Z}_j}, \mathbf{Z}_j \in \mathbb{Z}}$  are available. Notice that  $\mathbf{Z}_j = \emptyset$  is a valid choice corresponding to  $P(\mathbf{V})$  the observational (non-interventional) distribution.

**Definition 5** (Counterfactual Identification). A query  $P(\mathbf{Y}_* = \mathbf{y}_*)$  is said to be identifiable from  $\mathbb{Z}$  in  $\mathcal{G}$ , if  $P(\mathbf{Y}_* = \mathbf{y}_*)$  is uniquely computable from the distributions  $\{P_{\mathbf{z}_j}(\mathbf{V})\}_{\mathbf{z}_j \in \mathcal{X}_{\mathbf{Z}_j}, \mathbf{Z}_j \in \mathbb{Z}}$  in any causal model which induces  $\mathcal{G}$ .

Given an arbitrary query  $P(\mathbf{Y}_* = \mathbf{y}_*)$ , we could express it in terms of ctf-factors by writing  $P(\mathbf{Y}_* = \mathbf{y}_*) = \sum_{\mathbf{d}_* \setminus \mathbf{y}_*} P(\mathbf{D}_* = \mathbf{d}_*)$  where  $\mathbf{D}_* = \text{An}(\mathbf{Y}_*)$  and then using Theorem 1 to write  $P(\mathbf{D}_* = \mathbf{d}_*)$  as a ctf-factor. For instance, the ancestral set  $\mathbf{W}_* = \{Y_x, X, Z, W\}$  with  $\mathbf{w} = \{y, x', z, w\}$  in Eq. (6) can be written in terms of ctf-factors as

$$P(Y_x = y, X_z = x', Z_w = z, W = w) = P(Y_x = y, X_z = x', W = w)P(Z_w = z). \quad (9)$$

The following lemma characterizes the relationship between the identifiability of  $P(\mathbf{Y}_* = \mathbf{y}_*)$  and  $P(\mathbf{D}_* = \mathbf{d}_*)$ .

**Lemma 1.** Let  $P(\mathbf{W}_* = \mathbf{w}_*)$  be a ctf-factor and let  $\mathbf{Y}_* \subseteq \mathbf{W}_*$  be such that  $\mathbf{W}_* = \text{An}(\mathbf{Y}_*)$ . Then,  $\sum_{\mathbf{w}_* \setminus \mathbf{y}_*} P(\mathbf{W}_* = \mathbf{w}_*)$  is identifiable from  $\mathbb{Z}$  if and only if  $P(\mathbf{W}_* = \mathbf{w}_*)$  is identifiable from  $\mathbb{Z}$ .

Once the query of interest is in ctf-factor-form, the identification task reduces to identifying smaller ctf-factors according to the c-components of  $\mathcal{G}$ . In this respect, Theorem 2 implies the following

**Corollary 1.** Let  $P(\mathbf{W}_* = \mathbf{w}_*)$  be a ctf-factor and  $\mathbf{C}_j$  be a c-component of  $\mathcal{G}[\mathbf{V}(\mathbf{W}_*)]$ . Then, if  $P(\mathbf{C}_{j_*} = \mathbf{c}_{j_*})$  is not identifiable,  $P(\mathbf{W}_* = \mathbf{w}_*)$  is also not identifiable.

*Proof.* Assume for the sake of contradiction that  $P(\mathbf{C}_{j_*} = \mathbf{c}_{j_*})$  is not identifiable but  $P(\mathbf{W}_* = \mathbf{w}_*)$  is. Then, by Theorem 2, the former is identifiable from the latter, a contradiction.  $\square$

Let us consider the causal diagrams in Fig. 4 and the counterfactual  $Y_{x_1, W_{x_0}} = y, X = x$ , with  $x_0, x_1, x \in \mathcal{X}_X$ , used to define quantities for fairness analysis in [28] (e.g.,  $Y_{x_1, W_{x_0}} = y | X = x$ ):

$$\begin{aligned} P(Y_{x_1, W_{x_0}} = y, X = x) &= \sum_w P(Y_{x_1, w} = y, W_{x_0} = w, X = x) && \text{Unnesting} \quad (10) \end{aligned}$$

$$= \sum_{w, z} P(Y_{x_1, w} = y, W_{x_0} = w, X = x, Z = z) \quad \text{Complete ancestral set} \quad (11)$$

$$= \sum_{w, z} P(Y_{x_1, w, z} = y, W_{x_0} = w, X_z = x, Z = z) \quad \text{Write in ctf-factor-form} \quad (12)$$

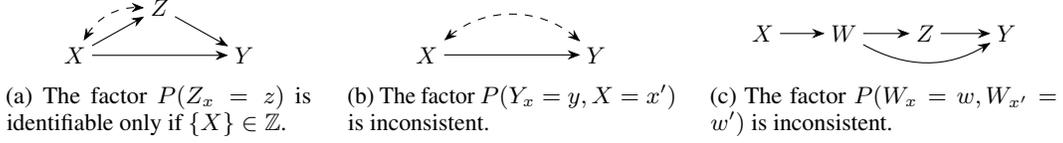


Figure 5: Examples of causal diagrams and inconsistent ctf-factors derived from them.

Due to the particular c-component structure of each model, we can factorize  $P(Y_{x_1, w, z} = y, W_{x_0} = w, X_z = x, Z = z)$  according to each model as:

$$P(Y_{x_1, w, z} = y)P(W_{x_0} = w)P(X_z = x)P(Z = z), \quad (13)$$

$$P(Y_{x_1, w, z} = y, Z = z)P(W_{x_0} = w)P(X_z = x), \text{ and} \quad (14)$$

$$P(Y_{x_1, w, z} = y)P(W_{x_0} = w, X_z = x)P(Z = z). \quad (15)$$

The question then becomes, whether ctf-factors corresponding to individual c-components can be identified from the available input. In this example, all factors in Eq. (13) and Eq. (14) are identifiable from  $P(\mathbf{V})$ . For Eq. (14) in particular, they are given by

$$P(Y = y, Z = z \mid W = w, X = x_1)P(W = w \mid X = x_0)P(X = x \mid Z = z). \quad (16)$$

In contrast, the factor  $P(W_{x_0} = w, X_z = x)$  in Eq. (15) (model Fig. 4(c)) is only identifiable if  $x = x_0$ . The following definition and theorem characterize the factors that can be identified from  $\mathbb{Z}$  and  $\mathcal{G}$ .

**Definition 6** (Inconsistent ctf-factor).  $P(\mathbf{W}_* = \mathbf{w}_*)$  is an inconsistent ctf-factor if it is a ctf-factor,  $\mathcal{G}[\mathbf{V}(\mathbf{W}_*)]$  has a single c-component, and one of the following situations hold:

- (i) there exist  $W_{\mathbf{t}} \in \mathbf{W}_*, Z \in \mathbf{T} \cap \mathbf{V}(\mathbf{W}_*)$  such that  $z \in \mathbf{t}, z' \in \mathbf{w}_*$  and  $z \neq z'$ , or
- (ii) there exists  $W_{i[\mathbf{t}_i]}, W_{j[\mathbf{t}_j]} \in \mathbf{W}_*$  and  $T \in \mathbf{T}_i \cap \mathbf{T}_j$  such that  $t \in \mathbf{t}_1, t' \in \mathbf{t}_2$  and  $t \neq t'$ .

**Theorem 3** (Ctf-factor identifiability). *A ctf-factor  $P(\mathbf{W}_* = \mathbf{w})$  is identifiable from  $\mathbb{Z}$  if and only if it is consistent. If consistent, let  $\mathbf{W} = \mathbf{V}(\mathbf{W}_*)$  and  $\mathbf{W}' = \mathbf{V} \setminus \mathbf{W}$ ; then  $P(\mathbf{W}_* = \mathbf{w}_*)$  is equal to  $P_{\mathbf{w}'}(\mathbf{w})$  where  $\mathbf{w}$  and  $\mathbf{w}'$  are consistent with  $\mathbf{w}_* \cup \bigcup_{\{W_{\mathbf{p}\mathbf{a}_w} \in \mathbf{W}_*\}} \mathbf{p}\mathbf{a}_w$ .*

Consider the  $NDE_{x \rightarrow x', Z}(Y)$  in Fig. 5(a), we can write

$$P(Y_{x'Z_x} = y) = \sum_z P(Y_{x'z} = y, Z_x = z) = \sum_z P(Y_{x'z} = y)P(Z_x = z). \quad (17)$$

While the factor  $P(Y_{x'z} = y)$  is identifiable from  $P(\mathbf{V})$  as  $P(Y = y \mid X = x', Z = z)$ , the second factor is identifiable only if experimental data on  $X$  is available, as  $P_x(z)$ .

We can also verify that the factor  $P(Y_x = y, X = x')$  in Fig. 5(b) is inconsistent. For another example consider the ETT-like expression  $P(Y_{xz} = y, X = x', Z = z')$  in Fig. 5(c), we have

$$\begin{aligned} P(Y_{xz} = y, X = x', Z = z') \\ = \sum_{w, w'} P(Y_{xz} = y, X = x', Z = z', W_x = w, W = w') \end{aligned} \quad (18)$$

$$= \sum_{w, w'} P(Y_{xz} = y, X = x', Z_{w'} = z', W_x = w, W_{x'} = w') \quad (19)$$

$$= \sum_{w, w'} P(Y_{xz} = y)P(X = x')P(Z_{w'} = z')P(W_x = w, W_{x'} = w'), \quad (20)$$

where the factor  $P(W_x = w, W_{x'} = w')$  is inconsistent.

Given a counterfactual variable  $\mathbf{Y}_{\mathbf{x}}$ , it could be the case that some values in  $\mathbf{x}$  become causally irrelevant to  $Y$  after the rest of  $\mathbf{x}$  has been fixed. Formally,

**Lemma 2** (Interventional Minimization). *Let  $\|\mathbf{Y}_*\| := \bigcup_{Y_{\widehat{\mathbf{x}}} \in \mathbf{Y}_*} \|Y_{\widehat{\mathbf{x}}}\|$  such that  $\|Y_{\widehat{\mathbf{x}}}\| := Y_{\widehat{\mathbf{z}}}$  where  $\mathbf{Z} = \mathbf{X} \cap \text{An}(Y)_{\mathcal{G}_{\widehat{\mathbf{x}}}}$  and  $\widehat{\mathbf{Z}}$  is consistent with  $\|\widehat{\mathbf{X}}\|, \|\mathbf{x}\| = \mathbf{x}$  and  $\|\emptyset\| = \emptyset$ . Then,  $\mathbf{Y}_* = \|\mathbf{Y}_*\|$ .*

Moreover, such simplification may reveal counterfactual expressions with equivalent or contradicting events. In Fig. 3(c),  $(Y_{xz} = y, Y_{xz'} = y') = (Y_x = y, Y_x = y')$  which has probability 0 if  $y \neq y'$ , or  $(Y_{xz} = y, Y_{xz'} = y)$  that is simply  $(Y_x = y)$ . Similarly, the probabilities of counterfactual events of the form  $P(X_x = x'), x \neq x'$ , and  $P(X_x = x)$  are trivially 0 and 1 respectively.

The following result shows how nested counterfactuals can be written in terms of non-nested ones.

---

**Algorithm 1** CTFIDU( $\mathbf{Y}_*, \mathbf{y}_*, \mathbb{Z}, \mathcal{G}$ )

---

**Input:**  $\mathcal{G}$  causal diagram over variables  $\mathbf{V}$ ;  $\mathbf{Y}_*$  a set of counterfactual variables in  $\mathbf{V}$ ;  $\mathbf{y}_*$  a set of values for  $\mathbf{Y}_*$ ; and available distribution specification  $\mathbb{Z}$ .

**Output:**  $P(\mathbf{Y}_* = \mathbf{y}_*)$  in terms of available distributions or FAIL if not identifiable from  $\langle \mathcal{G}, \mathbb{Z} \rangle$ .

- 1: let  $\mathbf{Y}_* \leftarrow \|\mathbf{Y}_*\|$ .
  - 2: **if** there exists  $Y_x \in \mathbf{Y}_*$  with two or more different values in  $\mathbf{y}_*(Y_x)$  or  $Y_y \in \mathbf{Y}_*$  with  $\mathbf{y}_*(Y_y) \neq y$  **then return 0**.
  - 3: **if** there exists  $Y_x \in \mathbf{Y}_*$  with two consistent values in  $\mathbf{y}_*(Y_x)$  or  $Y_y \in \mathbf{Y}_*$  with  $\mathbf{y}_*(Y_y) = y$  **then** remove repeated variables from  $\mathbf{Y}_*$  and values  $\mathbf{y}_*$ .
  - 4: let  $\mathbf{W}_* \leftarrow An(\mathbf{Y}_*)$ , and let  $\mathbf{C}_{1*}, \dots, \mathbf{C}_{k*}$  be corresponding ctf-factors in  $\mathcal{G}[\mathbf{V}(\mathbf{W}_*)]$ .
  - 5: **for each**  $\mathbf{C}_{i*}$  s.t.  $(\mathbf{C}_{i*} = \mathbf{c}_{i*})$  is not inconsistent,  $\mathbf{Z} \in \mathbb{Z}$  s.t.  $\mathbf{C}_i \cap \mathbf{Z} = \emptyset$  **do**
  - 6:   let  $\mathbf{B}_i$  be the c-component of  $\mathcal{G}_{\mathbf{Z}}$  such that  $\mathbf{C}_i \subseteq \mathbf{B}_i$ , compute  $P_{\mathbf{V} \setminus \mathbf{B}_i}(\mathbf{B}_i)$  from  $P_{\mathbf{Z}}(\mathbf{V})$ .
  - 7:   **if** IDENTIFY( $\mathbf{C}_i, \mathbf{B}_i, P_{\mathbf{V} \setminus \mathbf{B}_i}(\mathbf{B}_i), \mathcal{G}$ ) does not FAIL **then**
  - 8:     let  $P_{\mathbf{V} \setminus \mathbf{C}_i}(\mathbf{C}_i) \leftarrow$  IDENTIFY( $\mathbf{C}_i, \mathbf{B}_i, P_{\mathbf{V} \setminus \mathbf{B}_i}(\mathbf{B}_i), \mathcal{G}$ ).
  - 9:     let  $P(\mathbf{C}_{i*} = \mathbf{c}_{i*}) \leftarrow$   $P_{\mathbf{V} \setminus \mathbf{C}_i}(\mathbf{C}_i)$  evaluated with values  $(\mathbf{c}_{i*} \cup \bigcup_{\mathbf{C}_t \in \mathbf{C}_{i*}} \mathbf{pa}_{\mathbf{C}_t})$ .
  - 10:     move to the next  $\mathbf{C}_i$ .
  - 11:   **end if**
  - 12: **end for**
  - 13: **if** any  $P(\mathbf{C}_{i*} = \mathbf{c}_{i*})$  is inconsistent or was not identified from  $\mathbb{Z}$  **then return FAIL**.
  - 14: **return**  $P(\mathbf{Y}_* = \mathbf{y}_*) \leftarrow \sum_{\mathbf{w}_* \setminus \mathbf{y}_*} \prod_i P(\mathbf{C}_{i*} = \mathbf{c}_{i*})$ .
- 

**Theorem 4** (Counterfactual Unnesting Theorem (CUT)). *Let  $\widehat{\mathbf{X}}, \widehat{\mathbf{Z}}$  be any natural interventions on disjoint sets  $\mathbf{X}, \mathbf{Z} \subseteq \mathbf{V}$ . Then, for  $\mathbf{Y} \subseteq \mathbf{V}$  disjoint from  $\mathbf{X}$  and  $\mathbf{Z}$  such that  $\mathbf{X} \subseteq An(\mathbf{Y})_{\mathcal{G}_{\widehat{\mathbf{Z}}}}$ ,  $P(\mathbf{Y}_{\widehat{\mathbf{Z}}, \widehat{\mathbf{X}}} = \mathbf{y})$  is identifiable iff  $P(\mathbf{Y}_{\widehat{\mathbf{Z}}, \mathbf{x}} = \mathbf{y}, \widehat{\mathbf{X}} = \mathbf{x})$  is identifiable for every  $\mathbf{x}$ , and given by*

$$P(\mathbf{Y}_{\widehat{\mathbf{Z}}, \widehat{\mathbf{X}}} = \mathbf{y}) = \sum_{\mathbf{x} \in \mathcal{X}_{\mathbf{X}}} P(\mathbf{Y}_{\widehat{\mathbf{Z}}, \mathbf{x}} = \mathbf{y}, \widehat{\mathbf{X}} = \mathbf{x}). \quad (21)$$

For instance, for the model in Fig. 2 we can write

$$P(Y_{x'Z_x} = y) = \sum_z P(Y_{x'z} = y, Z_x = z). \quad (22)$$

As Theorem 4 allows us to rewrite any nested counterfactual in terms of a non-nested one, we focus on the latter and assume that any given counterfactual is already unnested.

Using the results in this section, we propose the algorithm CTFIDU (Algorithm 1) which given a set of counterfactual variables  $\mathbf{Y}_*$ , values  $\mathbf{y}_*$ , a collection of observational and experimental distributions  $\mathbb{Z}$ , and a causal diagram  $\mathcal{G}$ ; outputs an expression for  $P(\mathbf{Y}_* = \mathbf{y}_*)$  in terms of the specified distributions or FAIL if the query is not identifiable from such input in  $\mathcal{G}$ . Line 1 removes irrelevant subscripts from the query by virtue of Lemma 2. Then, lines 2 and 3 look for inconsistent events and redundant events, respectively. Line 4 finds the relevant ctf-factors consisting of a single c-component, as licensed by Theorem 1 and Theorem 2. As long as the factors are consistent, and allowed by Theorem 3, lines 6-11 carry out identification of the causal effect  $P_{\mathbf{V} \setminus \mathbf{C}_i}(\mathbf{C}_i)$  from the available distributions employing the algorithm IDENTIFY [26] as a subroutine (see an example in the next section). The procedure fails if any of the factors  $P(\mathbf{C}_{i*} = \mathbf{c}_{i*})$  is inconsistent or not identifiable from  $\mathbb{Z}$ . Otherwise, it returns the corresponding expression.

**Theorem 5** (CTFIDU completeness). *A counterfactual probability  $P(\mathbf{Y}_* = \mathbf{y}_*)$  is identifiable from  $\mathbb{Z}$  and  $\mathcal{G}$  if and only if CTFIDU returns an expression for it.*

This result ascertains that Algorithm 1 solves the identification task for counterfactuals in the form  $P(\mathbf{Y}_* = \mathbf{y}_*)$ .<sup>3</sup> Still, there exist other quantities that take into account evidence from events observed in the system, and cannot be written in this particular form. Since these queries represent important aspects of the underlying system, we will discuss them in the next section.

---

<sup>3</sup>One corollary of this result is that two identification algorithms ([22, 13]) can be combined and be shown complete; for details, see Appendix F. Even though our results were developed from first principles using an entirely new approach and machinery (which was the key to derive Algorithm 1 and prove its completeness), practitioners familiar with these algorithms may find the completeness of this combined approach attractive.



(a) While  $P(Y_x=y \mid Z_x=z, X=x')$  is identifiable from  $\mathbb{Z}$ ,  $P(Y_x=y, Z_x=z, X=x')$  is not. (b)  $P(Y_x=y \mid Z_x=z, X=x')$  is not identifiable from  $\mathbb{Z}$  because of the factor  $P(Y_{xz}=y, X=x')$ .

Figure 6: Examples of conditional queries.

---

**Algorithm 2** CTFID( $\mathbf{Y}_*, \mathbf{y}_*, \mathbf{X}_*, \mathbf{x}_*, \mathbb{Z}, \mathcal{G}$ )

---

**Input:**  $\mathcal{G}$  causal diagram over variables  $\mathbf{V}$ ;  $\mathbf{Y}_*, \mathbf{X}_*$  a set of counterfactual variables in  $\mathbf{V}$ ;  $\mathbf{y}_*, \mathbf{x}_*$  a set of values for  $\mathbf{Y}_*$  and  $\mathbf{X}_*$ ; and available distribution specification  $\mathbb{Z}$ .

**Output:**  $P(\mathbf{Y}_*=\mathbf{y}_* \mid \mathbf{X}_*=\mathbf{x}_*)$  in terms of available distributions or FAIL if non-ID from  $\langle \mathcal{G}, \mathbb{Z} \rangle$ .

- 1: Let  $\mathbf{A}_{1*}, \mathbf{A}_{2*}, \dots$  be the ancestral components of  $\mathbf{Y}_* \cup \mathbf{X}_*$  given  $\mathbf{X}_*$ .
  - 2: Let  $\mathbf{D}_*$  be the union of the ancestral components containing a variable in  $\mathbf{Y}_*$  and  $\mathbf{d}_*$  the corresponding set of values.
  - 3: let  $Q \leftarrow \text{CTFIDU}(\bigcup_{D_t \in \mathbf{D}_*} \mathbf{D}_{\text{pa}_d}, \mathbf{d}_*, \mathbb{Z}, \mathcal{G})$ .
  - 4: **return**  $\sum_{\mathbf{d}_* \setminus (\mathbf{y}_* \cup \mathbf{x}_*)} Q / \sum_{\mathbf{d}_* \setminus \mathbf{x}_*} Q$ .
- 

## 4 Identification of Conditional Counterfactuals

In this section, we consider counterfactual quantities of the form  $P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*)$ . It is immediate to write such a query as  $P(\mathbf{W}_* = \mathbf{w}_*) / \sum_{\mathbf{y}_*} P(\mathbf{W}_* = \mathbf{w}_*)$  with  $\mathbf{W}_* = \mathbf{Y}_* \cup \mathbf{X}_*$ , and try to identify it using CTFIDU. Nevertheless, depending on the graphical structure, the original query may be identifiable even if the latter is not. To witness, consider the causal diagram in Fig. 6(a) and the counterfactual  $P(Y_x = y \mid Z_x = z, X = x')$ , which can be written as  $P(Y_x = y, Z_x = z, X = x') / \sum_y P(Y_x = y, Z_x = z, X = x')$ . Following the strategy explained so far, the numerator is equal to  $P(Y_z = y)P(Z_x = z, X = x')$ , where the second ctf-factor is inconsistent, and therefore not identifiable from  $\mathbb{Z}$ . Nevertheless, the conditional query is identifiable as

$$\frac{P(Y_{xz} = y)P(Z_x = z, X = x')}{P(Z_x = z, X = x') \sum_y P(Y_{xz} = y)} = P(Y_{xz} = y) = P(Y = y \mid Z = z, X = x). \quad (23)$$

To characterize such simplifications of the query, we look at the causal diagram paying special attention to variables after the conditioning bar that are also ancestors of those before. Let  $\mathbf{X}_*(W_t) = \mathbf{V}(\|\mathbf{X}_* \parallel \cap \text{An}(W_t))$ , that is, the primitive variables in  $\mathbf{X}_*$  that are also ancestors of  $\mathbf{W}_t$ .

**Definition 7** (Ancestral components). Let  $\mathbf{W}_*$  be a set of counterfactual variables,  $\mathbf{X}_* \subseteq \mathbf{W}_*$ , and  $\mathcal{G}$  be a causal diagram. Then the ancestral components induced by  $\mathbf{W}_*$ , given  $\mathbf{X}_*$ , are sets  $\mathbf{A}_{1*}, \mathbf{A}_{2*}, \dots$  that form a partition over  $\text{An}(\mathbf{W}_*)$ , made of unions of ancestral sets  $\text{An}(W_t)_{\mathcal{G}_{\mathbf{X}_*(W_t)}}$ ,  $W_t \in \mathbf{W}_*$ . Sets  $\text{An}(W_{1[t_1]})_{\mathcal{G}_{\mathbf{X}_*(W_{1[t_1]})}}$  and  $\text{An}(W_{2[t_2]})_{\mathcal{G}_{\mathbf{X}_*(W_{2[t_2]})}}$  are put together if they are not disjoint or there exists a bidirected arrow in  $\mathcal{G}$  connecting variables in those sets.

**Lemma 3** (Conditional Query Reduction). *Let  $\mathbf{Y}_*, \mathbf{X}_*$  be two sets of counterfactual variables and let  $\mathbf{D}_*$  be the set of variables in the same ancestral component, given  $\mathbf{X}_*$ , as any variable in  $\mathbf{Y}_*$ , then*

$$P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*) = \frac{\sum_{\mathbf{d}_* \setminus (\mathbf{y}_* \cup \mathbf{x}_*)} P(\bigwedge_{D_t \in \mathbf{D}_*} \mathbf{D}_{\text{pa}_d} = d)}{\sum_{\mathbf{d}_* \setminus \mathbf{x}_*} P(\bigwedge_{D_t \in \mathbf{D}_*} \mathbf{D}_{\text{pa}_d} = d)}, \quad (24)$$

where  $\text{pa}_d$  is consistent with  $\mathbf{t}$  and  $\mathbf{d}_*$ , for each  $D_t \in \mathbf{D}_*$ . Moreover,  $P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*)$  is identifiable from  $\mathbb{Z}$  if and only if  $P(\bigwedge_{D_t \in \mathbf{D}_*} \mathbf{D}_{\text{pa}_d} = d)$  is identifiable from  $\mathbb{Z}$ .

Using the notion of ancestral components and Lemma 3, we propose a conditional version of CTFIDU (Algorithm 2). Due to Lemma 3, it is easy to see that CTFID is complete. In case that  $\mathbf{W}_* = \mathbf{Y}_* \cup \mathbf{X}_*$  contains nested counterfactuals, Theorem 4 can be used first and the new variables in the expression are added to  $\mathbf{W}_*$ , then the new sum indexes added to  $\mathbf{d}_*$  in Eq. (24).

**Theorem 6** (CTFID completeness). *A counterfactual probability  $P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*)$  is identifiable from  $\mathbb{Z}$  and  $\mathcal{G}$  if and only if CTFID returns an expression for it.*

To illustrate the mechanics of our algorithm, consider the causal diagram  $\mathcal{G}$  in Fig. 7(a) and the identification of the query  $P(Y_{x_1 z_{x_0}} = y_1 \mid X_{w_0} = x_1, T_{r_1} = t_1, R = r_0)$  from  $\mathbb{Z} = \{\{\}, \{W, T\}\}$ . First, we can use Theorem 4 to write the query as:

$$\sum_z P(Y_{x_1 z} = y_1, Z_{x_0} = z \mid X_{w_0} = x_1, T_{r_1} = t_1, R = r_0). \quad (25)$$

The query within the sum can be processed using  $\text{CTFID}(\mathbf{Y}_* = \{Y_{x_1 z}, Z_{x_0}\}, \mathbf{y}_* = \{y_1, z\}, \mathbf{X}_* = \{X_{w_0}, T_{r_1}, R\}, \mathbf{x}_* = \{x_1, t_1, r_0\}, \mathbb{Z}, \mathcal{G})$ .

At line 1, the algorithm looks for the ancestral components of  $\mathbf{Y}_* \cup \mathbf{X}_*$  given  $\mathbf{X}_*$ . This results in three ancestral components, which can be labeled as  $\mathbf{A}_{1*} = \{Y_z, X_{w_0}, T, R\}$ ,  $\mathbf{A}_{2*} = \{Z_{x_0}\}$  and  $\mathbf{A}_{3*} = \{T_{r_1}\}$ . Then, we let  $\mathbf{D}$  be the union of the ancestral sets that contain variables in  $\mathbf{Y}_*$ , that is  $\mathbf{D}_* = \{Y_z, X_{w_0}, T, R, Z_{x_0}\}$ . We also gather the values  $\mathbf{d}_* = \{y_1, x_1, t, r_0, z\}$ , where  $t$  is a new value introduced for  $T$  that will appear as the index in a sum on line 3.

At this point,  $\text{CTFIDU}$  is invoked with arguments  $(\mathbf{Y}_* = \{Y_z, X_{w_0 t}, T_{r_0}, R, Z_{x_0 r_0}\}, \mathbf{y}_* = \mathbf{d}_*, \mathbb{Z}, \mathcal{G})$ . In this case lines 1-3 will not make any change. Note that  $An(D_{\mathbf{pa}_d}) = \{D_{\mathbf{pa}_d}\}$ , hence their union is also an ancestral set, and  $\mathbf{W}_*$  will be equal to the  $\mathbf{Y}_*$  given to  $\text{CTFIDU}$ . The c-components in the graph  $\mathcal{G}[\mathbf{V}(\mathbf{W}_*)]$  are  $\{X, R, Y\}$ ,  $\{T\}$  and  $\{Z\}$ , then we need to consider the ctf-factors  $P(Y_z = y_1, X_{w_0 t} = x_1, R = r_0)$ ,  $P(T_{r_0} = t)$  and  $P(Z_{x_0 r_0} = z)$ . In this case all of the ctf-factors are consistent, so we can try to identify each of them from  $\mathbb{Z}$ . Specifically, we will obtain<sup>4</sup>

$$P(Y_z = y_1, X_{w_0 t} = x_1, R = r_0) = P_{w_0, t}(r_0) P_{w_0, t}(x_1 \mid r_0) P_{w_0, t}(y_1 \mid r_0, x_1, z), \quad (26)$$

$$P(T_{r_0} = t) = P(t \mid r_0), \text{ and} \quad (27)$$

$$P(Z_{x_0 r_0} = z) = P_{w, t'}(z \mid r_0, x_0); \quad (28)$$

where  $w$  and  $t'$  could be any values in  $\mathfrak{X}_W$  and  $\mathfrak{X}_T$ , respectively. We could just take  $w_0$  and  $t$  for simplicity. Then, considering the sum over  $z$  introduced for unnesting, the final result is:

$$P(Y_{x_1 z_{x_0}} = y_1 \mid X_{w_0} = x_1, T_{r_1} = t_1) = \quad (29)$$

$$\sum_{z, t} P_{w_0, t}(r_0) P_{w_0, t}(x_1 \mid r_0) P_{w_0, t}(y_1 \mid r_0, x_1, z) P(t \mid r_0) P_{w_0, t}(z \mid r_0, x_0). \quad (30)$$

## 5 Conclusions

We investigated in this paper the problem of nested and non-nested counterfactual identification from an arbitrary combination of observational and experimental distributions. First, we introduced fundamental building blocks for counterfactual reasoning, which allowed us to prove several properties of nested counterfactual distributions, including the counterfactual unnesting theorem (Theorem 4), and the ancestral and counterfactual factorization theorems (Theorem 1, Theorem 2). Moreover, we introduced a graphical condition (Definition 6, Theorem 3) and developed an efficient algorithm (Algorithm 1) for identifying marginal counterfactuals. We then proved their sufficiency and necessity for the task of nested counterfactual identification (Theorem 5). Lastly, we reduced the identification of conditional counterfactuals to that of marginal ones (Lemma 3) and provided a corresponding complete algorithm (Algorithm 2) for this task (Theorem 6). These results advance the state of the art regarding the distributions the inference engine expects as input, and the query it can generate as the output. In terms of the input, it accepts any combination of observational and experimental distributions, and in terms of the output, it considers an arbitrary nested counterfactual distribution. This work closes the long-standing inferential gap within the layers of the Pearl's Causal Hierarchy, and now one can move within all layers of the hierarchy in a very flexible and general way.

<sup>4</sup>For details on this example see Appendix E.

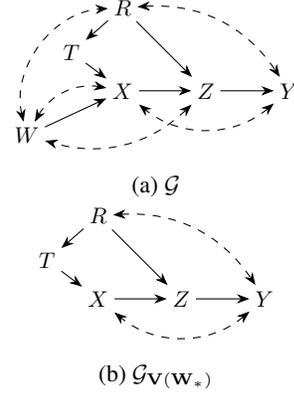


Figure 7: Graphs used for the running example of the algorithms.

## Acknowledgments and Disclosure of Funding

Elias Bareinboim and Juan D. Correa were supported in part by funding from the NSF (IIS-1704352, IIS-1750807), Amazon, JP Morgan, and The Alfred P. Sloan Foundation. Sanghack Lee was supported by the New Faculty Startup Fund from Seoul National University.

## References

- [1] Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of Path-Specific Effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence {IJCAI-05}*, pages 357–363, Edinburgh, UK. Morgan-Kaufmann Publishers.
- [2] Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. (2020). On Pearl’s Hierarchy and the Foundations of Causal Inference. Technical Report R-60, Causal Artificial Intelligence Lab, Columbia University. In: “Probabilistic and Causal Inference: The Works of Judea Pearl”, ACM Books, in press.
- [3] Bareinboim, E. and Pearl, J. (2012). Causal Inference by Surrogate Experiments: z-Identifiability. In Murphy, N. d. F. and Kevin, editors, *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 113–120. AUAI Press.
- [4] Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. In Shiffrin, R. M., editor, *Proceedings of the National Academy of Sciences*, volume 113, pages 7345–7352. National Academy of Sciences.
- [5] Correa, J. D., Lee, S., and Bareinboim, E. (2021). Nested Counterfactual Identification from Arbitrary Surrogate Experiments. Technical Report R-79, Causal AI Lab, Columbia University.
- [6] Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, 70:161–189.
- [7] Dawid, A. P. (2015). Statistical Causality from a Decision-Theoretic Perspective. *Annual Review of Statistics and Its Application*, 2(1):273–303.
- [8] Galles, D. and Pearl, J. (1995). Testing identifiability of causal effects. In Besnard, P. and Hanks, S., editors, *Uncertainty in Artificial Intelligence 11*, pages 185–195. Morgan Kaufmann, San Francisco.
- [9] Heckman, J. J. (1992). Randomization and Social Policy Evaluation. In Manski, C. and Garfinkle, I., editors, *Evaluations: Welfare and Training Programs*, pages 201–230. Harvard University Press, Cambridge, MA.
- [10] Huang, Y. and Valorta, M. (2006). Identifiability in Causal Bayesian Networks: A Sound and Complete Algorithm. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*, pages 1149–1156. AAAI Press, Menlo Park, CA.
- [11] Huang, Y. and Valorta, M. (2008). On the completeness of an identifiability algorithm for semi-Markovian models. *Annals of Mathematics and Artificial Intelligence*, 54(4):363–408.
- [12] Lee, S. and Bareinboim, E. (2020). Causal effect identifiability under partial-observability. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119. PMLR.
- [13] Lee, S., Correa, J. D., and Bareinboim, E. (2019). General Identifiability with Arbitrary Surrogate Experiments. In *Proceedings of the Thirty-Fifth Conference Annual Conference on Uncertainty in Artificial Intelligence*, Corvallis, OR. AUAI Press.
- [14] Lee, S., Correa, J. D., and Bareinboim, E. (2020). Generalized Transportability: Synthesis of Experiments from Heterogeneous Domains. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, Menlo Park, CA. AAAI Press.
- [15] Pearl, J. (1994). A probabilistic calculus of actions. In de Mantaras, R. L. and D. Poole, editors, *Uncertainty in Artificial Intelligence 10*, pages 454–462. Morgan Kaufmann, San Mateo, CA.
- [16] Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.
- [17] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition.
- [18] Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420. Morgan Kaufmann, San Francisco, CA.

- [19] Pearl, J. and Mackenzie, D. (2018). *The Book of Why*. Basic Books, New York.
- [20] Pearl, J. and Robins, J. M. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In Besnard, P. and Hanks, S., editors, *Uncertainty in Artificial Intelligence 11*, pages 444–453. Morgan Kaufmann, San Francisco.
- [21] Shpitser, I. and Pearl, J. (2006). Identification of Joint Interventional Distributions in Recursive semi-Markovian Causal Models. In *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*, volume 2, pages 1219–1226.
- [22] Shpitser, I. and Pearl, J. (2007). What Counterfactuals Can Be Tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pages 352–359. AUAI Press, Vancouver, BC, Canada.
- [23] Shpitser, I. and Pearl, J. (2009). Effects of Treatment on the Treated: Identification and Generalization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, Montreal, Quebec. AUAI Press.
- [24] Shpitser, I. and Sherman, E. (2018). Identification of Personalized Effects Associated With Causal Pathways. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, pages 530–539.
- [25] Spirtes, P., Glymour, C. N., and Scheines, R. (2001). *Causation, Prediction, and Search*. MIT Press, 2nd edition.
- [26] Tian, J. and Pearl, J. (2002a). A General Identification Condition for Causal Effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI 2002)*, pages 567–573, Menlo Park, CA. AAAI Press/The MIT Press.
- [27] Tian, J. and Pearl, J. (2002b). On the Testable Implications of Causal Models with Hidden Variables. *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI-02)*, pages 519–527.
- [28] Zhang, J. and Bareinboim, E. (2018). Fairness in Decision-Making-The Causal Explanation Formula. In *AAAI Conference on Artificial Intelligence*.

# Nested Counterfactual Identification from Arbitrary Surrogate Experiments

## Supplemental Material

### A Counterfactual Unnesting Theorem

**Theorem 4** (Counterfactual Unnesting Theorem (CUT)). *Let  $\widehat{\mathbf{X}}, \widehat{\mathbf{Z}}$  be any natural interventions on disjoint sets  $\mathbf{X}, \mathbf{Z} \subseteq \mathbf{V}$ . Then, for  $\mathbf{Y} \subseteq \mathbf{V}$  disjoint from  $\mathbf{X}$  and  $\mathbf{Z}$  such that  $\mathbf{X} \subseteq \text{An}(\mathbf{Y})_{\mathcal{G}_{\widehat{\mathbf{Z}}}}$ ,  $P(\mathbf{Y}_{\widehat{\mathbf{Z}}, \widehat{\mathbf{X}}} = \mathbf{y})$  is identifiable iff  $P(\mathbf{Y}_{\widehat{\mathbf{Z}}, \mathbf{x}} = \mathbf{y}, \widehat{\mathbf{X}} = \mathbf{x})$  is identifiable for every  $\mathbf{x}$ , and given by*

$$P(\mathbf{Y}_{\widehat{\mathbf{Z}}, \widehat{\mathbf{X}}} = \mathbf{y}) = \sum_{\mathbf{x} \in \mathfrak{X}_{\mathbf{X}}} P(\mathbf{Y}_{\widehat{\mathbf{Z}}, \mathbf{x}} = \mathbf{y}, \widehat{\mathbf{X}} = \mathbf{x}). \quad (21)$$

*Proof.* (soundness) From Eq. (1) we have:  $P(\mathbf{Y}_{\widehat{\mathbf{Z}}, \widehat{\mathbf{X}}} = \mathbf{y}) = \sum_{\{\mathbf{u} \mid \mathbf{Y}_{\widehat{\mathbf{Z}}, \widehat{\mathbf{X}}}(\mathbf{u}) = \mathbf{y}\}} P(\mathbf{u})$ . In  $\mathcal{M}_{\widehat{\mathbf{Z}}, \widehat{\mathbf{X}}}$ , for any given  $\mathbf{u}$ , the variables in  $\mathbf{X}$  take a particular value-vector  $\mathbf{x} = \widehat{\mathbf{X}}(\mathbf{u})$ . Then, we can partition  $\{\mathbf{u} \mid \mathbf{Y}_{\widehat{\mathbf{Z}}, \widehat{\mathbf{X}}}(\mathbf{u}) = \mathbf{y}\}$  as  $\bigcup_{\mathbf{x} \in \mathfrak{X}_{\mathbf{X}}} \{\mathbf{u} \mid \mathbf{Y}_{\widehat{\mathbf{Z}}, \mathbf{x}}(\mathbf{u}) = \mathbf{y}, \widehat{\mathbf{X}}(\mathbf{u}) = \mathbf{x}\}$ . Hence

$$P(\mathbf{Y}_{\widehat{\mathbf{Z}}, \widehat{\mathbf{X}}} = \mathbf{y}) = \sum_{\mathbf{x} \in \mathfrak{X}_{\mathbf{X}}} \sum_{\{\mathbf{u} \mid \mathbf{Y}_{\widehat{\mathbf{Z}}, \mathbf{x}}(\mathbf{u}) = \mathbf{y}, \widehat{\mathbf{X}}(\mathbf{u}) = \mathbf{x}\}} P(\mathbf{u}) = \sum_{\mathbf{x} \in \mathfrak{X}_{\mathbf{X}}} P(\mathbf{Y}_{\widehat{\mathbf{Z}}, \mathbf{x}} = \mathbf{y}, \widehat{\mathbf{X}} = \mathbf{x}). \quad (31)$$

(necessity) Suppose  $P(\mathbf{Y}_{\widehat{\mathbf{Z}}, \mathbf{x}} = \mathbf{y}, \widehat{\mathbf{X}} = \mathbf{x})$  is not identifiable from  $\mathbb{Z}$ . Since  $\mathbf{X} \subseteq \text{An}(\mathbf{Y})_{\mathcal{G}_{\widehat{\mathbf{Z}}}}$  there exist  $X_r \in \widehat{\mathbf{X}}$  and a set of variables  $\mathbf{D}_{\widehat{\mathbf{Z}}}$  such that a path  $X \rightarrow D_1 \rightarrow D_2 \rightarrow \dots \rightarrow D_k \rightarrow Y$  exists in  $\mathcal{G}$  for some  $Y_{\widehat{\mathbf{Z}}} \in \mathbf{Y}_*$  with  $k \geq 0$ . This implies that  $D_{1[\widehat{\mathbf{Z}}]}$  is a function of  $X_r$ ,  $D_{i+1[\widehat{\mathbf{Z}}]}$  is a function of  $D_{i[\widehat{\mathbf{Z}}]}$  and  $Y_{\widehat{\mathbf{Z}}}$  is a function of  $D_{k[\widehat{\mathbf{Z}}]}$ . Then, we can use Lemma 4 from  $X$  to  $Y$  following this path to prove the result. □

### B Proofs on ctf-factors and factorization results

**Lemma 2** (Interventional Minimization). *Let  $\|\mathbf{Y}_*\| := \bigcup_{Y_{\widehat{\mathbf{X}}} \in \mathbf{Y}_*} \|Y_{\widehat{\mathbf{X}}}\|$  such that  $\|Y_{\widehat{\mathbf{X}}}\| := Y_{\widehat{\mathbf{Z}}}$  where  $\mathbf{Z} = \mathbf{X} \cap \text{An}(Y)_{\mathcal{G}_{\widehat{\mathbf{X}}}}$  and  $\widehat{\mathbf{Z}}$  is consistent with  $\|\widehat{\mathbf{X}}\|$ ,  $\|\mathbf{x}\| = \mathbf{x}$  and  $\|\emptyset\| = \emptyset$ . Then,  $\mathbf{Y}_* = \|\mathbf{Y}_*\|$ .*

*Proof.* Let us fix a  $\mathbf{u}$ , so a counterfactual variable  $Y_{\widehat{\mathbf{X}}}$  takes the value of the potential response  $Y_{\widehat{\mathbf{X}}}(\mathbf{u})$ . Note that this potential response is equal to  $f_y((\mathbf{Pa}_y)_{\widehat{\mathbf{X}}}(\mathbf{u}), \mathbf{u})$ . Next, for each  $W \in \mathbf{Pa}_y$ , we consider the response  $W_{\widehat{\mathbf{X}}}(\mathbf{u})$ . If  $W \in \mathbf{X}$ ,  $W_{\widehat{\mathbf{X}}}(\mathbf{u}) = \widehat{W}(\mathbf{u})$  where  $\widehat{W} \in \widehat{\mathbf{X}}$ . Else, we can evaluate  $W_{\widehat{\mathbf{X}}}(\mathbf{u})$  as  $f_w((\widehat{\mathbf{Pa}}_w)_{\widehat{\mathbf{X}}}(\mathbf{u}), \mathbf{u})$ . Then, by following this procedure recursively, we evaluate every function in  $\text{An}(Y)_{\mathcal{G}_{\widehat{\mathbf{X}}}}$  and use the variables/values of  $\widehat{\mathbf{X}}$  only for those variables in  $\mathbf{Z} = \mathbf{X} \cap \text{Pa}(\text{An}(Y)_{\mathcal{G}_{\widehat{\mathbf{X}}}})_{\mathcal{G}} = \mathbf{X} \cap \text{An}(Y)_{\mathcal{G}_{\widehat{\mathbf{X}}}}$ .

It follows that evaluating  $Y_{\widehat{\mathbf{Z}}}(\mathbf{u})$ , for  $\widehat{\mathbf{X}} \cap \mathbf{Z}$  and the same  $\mathbf{u}$  will result in exactly the same value as  $Y_{\widehat{\mathbf{X}}}(\mathbf{u})$ , and the result follows. If this equality holds for any  $\mathbf{u}$ , it will hold for any averaging over them as well. □

**Theorem 1** (Ancestral set factorization). *Let  $\mathbf{W}_*$  be an ancestral set, that is,  $\text{An}(\mathbf{W}_*) = \mathbf{W}_*$ , and let  $\mathbf{w}_*$  be a vector with a value for each variable in  $\mathbf{W}_*$ . Then,*

$$P(\mathbf{W}_* = \mathbf{w}_*) = P\left(\bigwedge_{W_t \in \mathbf{W}_*} W_{\mathbf{pa}_w} = w\right), \quad (4)$$

where each  $w$  is  $w_t$  and  $\mathbf{pa}_w$  is determined for each  $W_t \in \mathbf{W}_*$  as follows:

- (i) the values for variables in  $\mathbf{Pa}_w \cap \mathbf{T}$  are the same as in  $\mathbf{t}$ , and
- (ii) the values for variables in  $\mathbf{Pa}_w \setminus \mathbf{T}$  are taken from  $\mathbf{w}_*$  corresponding to the parents of  $W$ .

*Proof.* Let  $W_1 < W_2 < \dots < W_k$  be a topological order of the variables in  $\mathbf{W}$  according to  $\mathcal{G}$ . We transform the expression considering the instances of each  $W_i$  at a time, in reverse topological order. Let  $W_{i\mathbf{t}_i}$  be any instance of  $W_i$  in  $\mathbf{W}_*$ . For every  $W \in \mathbf{Pa}_{W_i} \setminus \mathbf{T}_i$ , the variable  $W_{\mathbf{t}}$  where and  $\mathbf{T} = \mathbf{T}_i \cap \text{An}(W)_{\mathcal{G}_{\overline{\mathbf{T}_i}}}$  must be in  $\mathbf{W}_*$  as  $W < W_i$ . Moreover,  $W_{\mathbf{t}} = W_{\mathbf{t}_i}$  because interventions on  $\mathbf{T}_i \setminus \mathbf{T}$  have no effect on  $W_{\mathbf{t}}$ . Then, using composition for  $W_i$  and all such parent  $W$  we get  $W_{i[\mathbf{t}_i]} = W_{i[\mathbf{t}_i \cup (\mathbf{pa}_{w_i} \setminus \mathbf{t}_i)]}$ , which is equal to  $W_{i[\mathbf{pa}_{w_i}]}$ . Repeating this process for  $i = k, k-1, \dots, 1$  ends with Eq. (4).  $\square$

**Theorem 2** (Counterfactual factorization). *Let  $P(\mathbf{W}_* = \mathbf{w}_*)$  be a ctf-factor, let  $W_1 < W_2 < \dots$  be a topological order over the variables in  $\mathcal{G}[\mathbf{V}(\mathbf{W}_*)]$ , and let  $\mathbf{C}_1, \dots, \mathbf{C}_k$  be the  $c$ -components of the same graph. Define  $\mathbf{C}_{j*} = \{W_{\mathbf{pa}_w} \in \mathbf{W}_* \mid W \in \mathbf{C}_j\}$  and  $\mathbf{c}_{j*}$  as the values in  $\mathbf{w}_*$  corresponding to  $\mathbf{C}_{j*}$ , then  $P(\mathbf{W}_* = \mathbf{w}_*)$  decomposes as*

$$P(\mathbf{W}_* = \mathbf{w}_*) = \prod_j P(\mathbf{C}_{j*} = \mathbf{c}_{j*}). \quad (7)$$

Furthermore, each factor can be computed from  $P(\mathbf{W}_* = \mathbf{w}_*)$  as

$$P(\mathbf{C}_{j*} = \mathbf{c}_{j*}) = \prod_{\{W_i \in \mathbf{C}_j\}} \frac{\sum_{\{w \mid W_{\mathbf{pa}_w} \in \mathbf{W}_*, W_i < W\}} P(\mathbf{W}_* = \mathbf{w}_*)}{\sum_{\{w \mid W_{\mathbf{pa}_w} \in \mathbf{W}_*, W_{i-1} < W\}} P(\mathbf{W}_* = \mathbf{w}_*)}. \quad (8)$$

*Proof.* Equation (7) follows immediately from independence restrictions. To show Eq. (8) let  $W_k$  be such that  $W_i \leq W_k$  for all  $W_i \in \mathbf{C}_j$ , then rewrite the factor as follows:

$$P(\mathbf{C}_{j*} = \mathbf{c}_{j*}) = P(\mathbf{C}_{j*} = \mathbf{c}_{j*}) \prod_{\{W_i \in \mathbf{C}_j, W_i < W_k\}} \frac{P(\bigwedge_{\{W_{\mathbf{pa}_w} \in \mathbf{C}_{j*}, W \leq W_i\}} W_{\mathbf{pa}_w} = w)}{P(\bigwedge_{\{W_{\mathbf{pa}_w} \in \mathbf{C}_{j*}, W \leq W_i\}} W_{\mathbf{pa}_w} = w)}. \quad (32)$$

Rearranging factors we get

$$P(\mathbf{C}_{j*} = \mathbf{c}_{j*}) = \prod_{\{W_i \in \mathbf{C}_j\}} \frac{P(\bigwedge_{\{W_{\mathbf{pa}_w} \in \mathbf{C}_{j*}, W \leq W_i\}} W_{\mathbf{pa}_w} = w)}{P(\bigwedge_{\{W_{\mathbf{pa}_w} \in \mathbf{C}_{j*}, W < W_i\}} W_{\mathbf{pa}_w} = w)}, \quad (33)$$

where  $P(\bigwedge_{\{W_{\mathbf{pa}_w} \in \mathbf{C}_{j*}, W < W_i\}} W_{\mathbf{pa}_w} = w) = 1$  if  $W_i = W_1$ . Next, we can multiply each factor in Eq. (33) by  $1 = P(\bigwedge_{\{W_{\mathbf{pa}_w} \in \mathbf{W}_* \setminus \mathbf{C}_{j*}, W < W_i\}} W_{\mathbf{pa}_w} = w) / P(\bigwedge_{\{W_{\mathbf{pa}_w} \in \mathbf{W}_* \setminus \mathbf{C}_{j*}, W < W_i\}} W_{\mathbf{pa}_w} = w)$ . Then, by Eq. (7)

$$\begin{aligned} & P(\bigwedge_{\{W_{\mathbf{pa}_w} \in \mathbf{W}_* \setminus \mathbf{C}_{j*}, W < W_i\}} W_{\mathbf{pa}_w} = w) P(\bigwedge_{\{W_{\mathbf{pa}_w} \in \mathbf{C}_{j*}, W \leq W_i\}} W_{\mathbf{pa}_w} = w) \\ &= P(\bigwedge_{\{W_{\mathbf{pa}_w} \in \mathbf{W}_*, W \leq W_i\}} W_{\mathbf{pa}_w} = w) \end{aligned} \quad (34)$$

and we have

$$P(\mathbf{C}_{j*} = \mathbf{c}_{j*}) = \prod_{\{W_i \in \mathbf{C}_j\}} \frac{P(\bigwedge_{\{W_{\mathbf{pa}_w} \in \mathbf{C}_{j*}, W \leq W_i\}} W_{\mathbf{pa}_w} = w)}{P(\bigwedge_{\{W_{\mathbf{pa}_w} \in \mathbf{C}_{j*}, W < W_i\}} W_{\mathbf{pa}_w} = w)}. \quad (35)$$

Finally, note that no value for  $W$  such that  $W_i < W$  appears in the numerator or denominator of the factor associated with  $W_i$  in Eq. (35), so we can write

$$P(\bigwedge_{\{W_{\mathbf{pa}_w} \in \mathbf{C}_{j*}, W \leq W_i\}} W_{\mathbf{pa}_w} = w) = \sum_{\{w \mid W_{\mathbf{pa}_w} \in \mathbf{W}_*, W_{i-1} < W\}} P(\mathbf{W}_* = \mathbf{w}_*) \quad (36)$$

to obtain Eq. (8).  $\square$

## C Proofs of the results on Identification

**Lemma 4.** *Suppose  $P(\mathbf{W}_* = \mathbf{w}_*)$  is not identifiable from a set of available distributions in a causal diagram  $\mathcal{G}$ . Let  $A_{1[t_1]}, A_{2[t_2]} \in \mathbf{W}_*$  such that  $A_{2[t_2]}$  is a function of  $A_{1[t_1]}$ . Then  $\sum_{a_1} P(\mathbf{W}_* = \mathbf{w}_*)$  is not identifiable from the same input either.*

*Proof.* Let  $\mathcal{M}^{(1)}$  and  $\mathcal{M}^{(2)}$  be the two models witnessing the non-identifiability of  $P(\mathbf{W}_* = \mathbf{w}_*)$ , they agree on available distributions, but for  $\mathbf{w}_*$  we have  $P^{(1)}(\mathbf{W}_* = \mathbf{w}_*) = \alpha$ ,  $P^{(2)}(\mathbf{W}_* = \mathbf{w}_*) = \beta$  with  $\alpha \neq \beta$ . Assume, without loss of generality that  $\alpha > \beta$ . We will extend the strategy used by [11] to construct two models  $\mathcal{M}^{(1)'}$  and  $\mathcal{M}^{(2)'}$  where the domain of  $A_2$  is  $\mathfrak{X}_{A_2} \times \{0, 1\}$ , where  $\mathfrak{X}_{A_2}$  is the domain of  $A_2$  in  $\mathcal{M}_1, \mathcal{M}_2$ . Let  $F(A_1)$  be a probability function from  $\mathfrak{X}_{A_1}$  to  $\{0, 1\}$ , such that  $P(F(a_1) = k) > 0, k = 0, 1$  and  $P(F(a_1) = 0) = 1 - P(F(a_1) = 1)$ . In  $\mathcal{M}^{(i)'}$ ,  $i = 1, 2$  we define the new function for  $A_2$  such that:

$$P^{(i)'}((a_2, k) \mid \mathbf{pa}_{a_2}, \mathbf{u}_{a_2}) = P^{(i)}(a_2 \mid \mathbf{pa}_{a_2}, \mathbf{u}_{a_2})P(F(a_1)=k). \quad (37)$$

For any other  $V_j \in \mathbf{V} \setminus \{A_2\}$  we keep the same function such that  $P^{(i)'}(v_j \mid \mathbf{pa}_j, u_j) = P^{(i)}(v_j \mid \mathbf{pa}_j, u_j)$ . In particular, the functions of variables that are children of  $A_2$  simply ignore the new bit in  $A_2$ 's domain. We can verify that for any counterfactual  $\mathbf{Z}_*$  on which  $\mathcal{M}^{(1)}$  and  $\mathcal{M}^{(2)}$  agree, we have

$$P^{(i)'}(\mathbf{Z}_* = \mathbf{z}_*) = \sum_{\mathbf{d}_* \setminus \mathbf{z}_*} P^{(i)'}(\mathbf{D}_* = \mathbf{d}_*), \quad (38)$$

where  $\mathbf{D}_* = An(\mathbf{Z}_*)$  and  $\mathbf{d}_*$  is the union of  $\mathbf{z}_*$  and the indexing values of the sum. By Theorem 1 this can be written in the form of a ctf-factor as

$$P^{(i)'}(\mathbf{Z}_* = \mathbf{z}_*) = \sum_{\mathbf{d}_* \setminus \mathbf{z}_*} P^{(i)'} \left( \bigwedge_{D_t \in \mathbf{D}_*} D_{\mathbf{pa}_d} = d \right), \quad (39)$$

and then sum over  $\mathbf{U}$

$$P^{(i)'}(\mathbf{Z}_* = \mathbf{z}_*) = \sum_{\mathbf{u}, \mathbf{d}_* \setminus \mathbf{z}_*} P^{(i)'} \left( \bigwedge_{D_t \in \mathbf{D}_*} D_{\mathbf{pa}_d} = d \mid \mathbf{u} \right) P^{(i)'}(\mathbf{u}). \quad (40)$$

Note that  $P^{(i)'}(\mathbf{u}) = P^{(i)}(\mathbf{u})$ . Also, once  $\mathbf{U}$  has been fixed, any  $D_{\mathbf{pa}_d}$  becomes deterministic and independent of any other variable, as it only depends on  $\mathbf{pa}_d$  and  $\mathbf{U}_d$ . Then we can write

$$P^{(i)'} \left( \bigwedge_{D_t \in \mathbf{D}_*} D_{\mathbf{pa}_d} = d \mid \mathbf{u} \right) = \prod_{D_t \in \mathbf{D}_*} P^{(i)'}(D_{\mathbf{pa}_d} = d \mid \mathbf{u}) = \prod_{D_t \in \mathbf{D}_*} P^{(i)'}(d \mid \mathbf{pa}_d, \mathbf{u}), \quad (41)$$

For any  $D \neq A_2$  we have  $P^{(i)'}(d \mid \mathbf{pa}_d, \mathbf{u}) = P^{(i)}(d \mid \mathbf{pa}_d, \mathbf{u})$ . Then  $P^{(i)'}(\mathbf{Z}_* = \mathbf{z}_*) = P^{\mathcal{M}_i}(\mathbf{Z}_* = \mathbf{z}_*)$ , if  $\mathbf{D}_*$  does not contain any  $A_2$ .

By construction, the factor for  $A_2$  is given by Eq. (37). Hence if  $\mathbf{D}_*$  does contain some instance of  $A_2$  and  $\mathbf{d}_*$  has some  $(a_2, k)$ ,

$$P^{(i)'}(\mathbf{Z}_* = \mathbf{z}_*) = \sum_{\mathbf{d}_* \setminus \mathbf{z}_*} P^{(i)} \left( \bigwedge_{D_t \in \mathbf{D}_*} D_{\mathbf{pa}_d} = d \right) P(F(a_1) = k). \quad (42)$$

If  $(a_2, k) \notin \mathbf{z}_*$ , we can sum out  $k$  to obtain  $P^{(i)'}(\mathbf{Z}_* = \mathbf{z}_*) = P^{(i)}(\mathbf{Z}_* = \mathbf{z}_*)$ . If  $a_1 \notin (\mathbf{d}_* \setminus \mathbf{z}_*)$  then the expression becomes

$$P^{(i)'}(\mathbf{Z}_* = \mathbf{z}_*) = P(F(a_1) = k) \sum_{\mathbf{d}_* \setminus \mathbf{z}_*} P^{(i)} \left( \bigwedge_{D_t \in \mathbf{D}_*} D_{\mathbf{pa}_d} = d \right) = P(F(a_1) = k)P^{(i)}(\mathbf{Z}_* = \mathbf{z}_*). \quad (43)$$

For any input  $P_t(\mathbf{V} \setminus \mathbf{T}) = P(\bigwedge_{V_i \in \mathbf{V} \setminus \mathbf{T}} V_{i[t]})$ , either  $A_{1[t]}$  appears in the expression and does not have an index in  $\mathbf{d}_* \setminus \mathbf{z}_*$ , or  $A_1 \in \mathbf{T}$  so that no instance of  $A_1$  is an ancestor of any other variable in

the expression. Hence any experimental distribution of the form  $P_{\mathbf{t}}(\mathbf{V} \setminus \mathbf{T})$ , given as input ( $\mathbf{T} \in \mathbb{Z}$ ), will match in  $\mathcal{M}^{(1)'}$  and  $\mathcal{M}^{(2)'}$  as it matches in  $\mathcal{M}^{(1)}$  and  $\mathcal{M}^{(2)}$ .

Consider the assignment  $\mathbf{w}_* \setminus \{a_2\}, (a'_2, 0)$ , by construction we have

$$P^{(i)'}(\mathbf{W}_* = (\mathbf{w}_* \setminus \{a_2\}) \cup (a_2, 0)) = P^{(i)}(\mathbf{W}_* = \mathbf{w}_*)P(F(a_1) = 0) \quad (44)$$

For  $a_1 \in \mathbf{w}_*, a'_1 \neq a_1$ , let  $P(F(a_1) = 0) = 1/2$  and  $P(F(a'_1) = 0) = (\alpha - \beta)/4$ . Then

$$\begin{aligned} & \sum_{a'_1} P^{(i)}(\mathbf{W}_* = (\mathbf{w}_* \setminus \{a_1, a_2\}) \cup \{a'_1, (a_2, 0)\}) \\ &= \sum_{a'_1} P^{(i)}(\mathbf{W}_* = (\mathbf{w}_* \setminus \{a_1, a_2\}) \cup \{a'_1, a_2\})P(F(a'_1) = 0) \end{aligned} \quad (45)$$

For  $\mathcal{M}^{(1)'}$  this means

$$\begin{aligned} & \sum_{a'_1} P^{(1)'}(\mathbf{W}_* = (\mathbf{w}_* \setminus \{a_1, a_2\}) \cup \{a'_1, (a_2, 0)\}) \\ &= \frac{1}{2}\alpha + \left(\frac{\alpha-\beta}{4}\right) \sum_{a_1 \neq a'_1} P^{(1)}(\mathbf{W}_* = (\mathbf{w}_* \setminus \{a_1, a_2\}) \cup \{a'_1, a_2\}) \end{aligned} \quad (46)$$

$$> \frac{1}{2}\alpha \quad (47)$$

As for  $\mathcal{M}'_2$ :

$$\begin{aligned} & \sum_{a'_1} P^{(2)'}(\mathbf{W}_* = (\mathbf{w}_* \setminus \{a_1, a_2\}) \cup \{a'_1, (a_2, 0)\}) \\ &= \frac{1}{2}\beta + \left(\frac{\alpha-\beta}{4}\right) \sum_{a_1 \neq a'_1} P^{(2)}(\mathbf{W}_* = (\mathbf{w}_* \setminus \{a_1, a_2\}) \cup \{a'_1, a_2\}) \end{aligned} \quad (48)$$

$$< \frac{1}{2}\beta + \frac{\alpha-\beta}{4} \quad (49)$$

$$< \frac{1}{2}\alpha. \quad (50)$$

Then,  $\mathcal{M}^{(1)'}$  and  $\mathcal{M}^{(2)'}$  are compatible with  $\mathcal{G}$ , match in the available distributions and yield different  $\sum_{a_1} P(\mathbf{W}_* = \mathbf{w}_*)$ .  $\square$

**Lemma 1.** *Let  $P(\mathbf{W}_* = \mathbf{w}_*)$  be a ctf-factor and let  $\mathbf{Y}_* \subseteq \mathbf{W}_*$  be such that  $\mathbf{W}_* = An(\mathbf{Y}_*)$ . Then,  $\sum_{\mathbf{w}_* \setminus \mathbf{y}_*} P(\mathbf{W}_* = \mathbf{w}_*)$  is identifiable from  $\mathbb{Z}$  if and only if  $P(\mathbf{W}_* = \mathbf{w}_*)$  is identifiable from  $\mathbb{Z}$ .*

*Proof.* If  $P(\mathbf{W}_* = \mathbf{w}_*)$  is identifiable, then  $\sum_{\mathbf{w}_* \setminus \mathbf{y}_*} P(\mathbf{W}_* = \mathbf{w}_*)$  is clearly identifiable. Since  $\mathbf{W}_*$  is ancestral with respect of  $\mathbf{Y}_*$ , every variable in  $\mathbf{W}_* \setminus \mathbf{Y}_*$  must have a children in  $\mathbf{W}_*$ , that is, another variable that is a function of it. Then, we can use Lemma 4 in a topological order for the variables in  $\mathbf{W}_* \setminus \mathbf{Y}_*$  to claim that if  $P(\mathbf{W}_* = \mathbf{w}_*)$  is not identifiable then  $\sum_{\mathbf{w}_* \setminus \mathbf{y}_*} P(\mathbf{W}_* = \mathbf{w}_*)$  is also not identifiable.  $\square$

**Lemma 5** (Inconsistent factor non-identifiability). *Let  $P(\mathbf{W}_* = \mathbf{w}_*)$  be an inconsistent ctf-factor. Then,  $P(\mathbf{W}_* = \mathbf{w}_*)$  is not identifiable from  $\mathbf{P}_*$ .*

*Proof.* We construct two models  $\mathcal{M}^{(1)}$  and  $\mathcal{M}^{(2)}$  agreeing on  $\mathbf{P}_*^{(1)}$  and  $\mathbf{P}_*^{(2)}$  disagreeing on the inconsistent factor. Without loss of generality, let us ignore parents (intervened) that are consistent. Further, let  $W_{1[t_1]}$  and  $W_{l[t_l]}$  be inconsistent ones such that (i)  $z = \mathbf{w}_*(W_{1[t_1]})$  and  $z' \in t_l = \{z'\}$  are different and (ii)  $t \in t_1 = \{t\}$  and  $t' \in t_l = \{t'\}$  are different. We let  $\mathbf{U} = \{U_1, \dots, U_{l-1}\}$  be the unobserved variables where  $W_i \leftarrow U_i \rightarrow W_{i+1}$  unless stated otherwise.

**(Case i)** Let  $W_{1[t_1]} = W_1$  and  $W_{l[t_l]} = W_{l[w'_l]}$ . We investigate the case with regard to  $l = 2$  and  $l > 2$ .

(Case  $l = 2$ ) Let  $U_1$  and  $W_1$  be ternary and let  $U = U_1$  for simplicity.

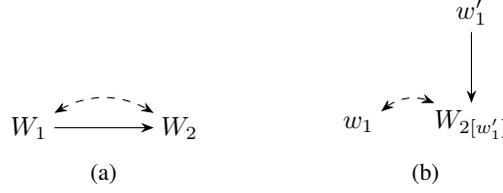


Figure 8: (Case i,  $l = 2$ )

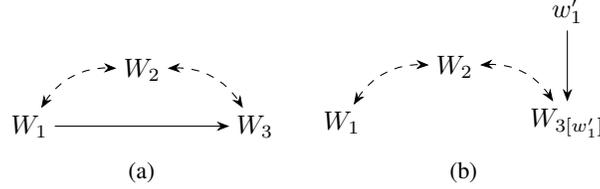


Figure 9: (Case i,  $l > 1$ )

$$W_1(u) = u \quad (51)$$

$$W_l^{(1)}(w_1, u) = 1_{w_1=1, u=2} \quad W_l^{(2)}(w_1, u) = 1_{w_1=1, u=0} \quad (52)$$

Then, two models agree on  $\mathbf{P}_*$ . In the observational distribution,  $W_1$  values are determined by the common values (i.e.,  $W_l^{(1)}$  and  $W_l^{(2)}$  behave the same when  $w_1 = u$ ). For interventional distributions,  $W_1$  behaves the same marginally over  $U$ . However,  $w_2$  value as in  $(w_1 = 0, W_{2[w_1=1]})$  is determined differently in the two models since we only look at the case where  $U = 0$  and  $W_1 = 1$  (0 for  $\mathcal{M}^{(1)}$  and 1 for  $\mathcal{M}^{(2)}$ ).

(Case  $l > 2$ ) Let  $W_1$  and  $U_1$  be ternaries and others be binaries with  $P(\mathbf{U})$  a uniform distribution. Let

$$\begin{aligned} W_1(u_1) &= u_1 \\ W_2(u_1, u_2) &= 1_{u_1=0, u_2=1} + 1_{u_1=2, u_2=0} \\ \forall_{3 \leq i \leq l-1} W_i(u_{i-1}, u_i) &= u_{i-1} \oplus u_i \\ W_l^{(1)}(w_1, u_{l-1}) &= 1_{w_1=1, u_{l-1}=1} & W_l^{(2)} &= 1_{w_1=1, u_{l-1}=0}. \end{aligned}$$

We first show that they agree on the observation for each  $U_1$  value. Let  $\mathbf{C}_{a:b} = \{C_a, \dots, C_b\}$  for  $a \leq b$  and  $\emptyset$  otherwise.

(Subcase  $U_1 = 0$ )  $W_1 = 0$  and  $W_l = 0$ .  $\mathbf{W}_{2:l-1}$  behaves the same by construction.

(Subcase  $U_1 = 1$ )  $W_1 = 1$  and  $W_2 = 0$ . Two models behave the same for  $\mathbf{W}_{3:l}$  in a complementing  $\mathbf{u}_{2:l-1}$ :

$$\mathbf{W}_{3:l}^{(1)}(\mathbf{u}_{2:l-1}) = \mathbf{W}_{3:l}^{(2)}(\overline{\mathbf{u}_{2:l-1}}). \quad (53)$$

Combined with the uniformness of  $P(\mathbf{U}_{2:l-1})$ ,  $P(\mathbf{W}|U_1 = 1)$  are the same for both models.

(Subcase  $U_1 = 2$ )  $W_1 = 2$  and  $W_l = 0$  are fixed and  $\mathbf{W}_{2:l-1}$  behave the same.

From the three subcases, both models agree on the observation.

Now we show they agree on  $do(W_1 = w_1)$  (intervening on other variables is trivial, and intervening other variables together with  $W_1$  becomes trivial once the two models agree on  $do(W_1 = w_1)$ ). In the case of  $do(W_1 = 0)$  or  $do(W_1 = 2)$ ,  $W_l = 0$  and  $\mathbf{W}_{2:l-1}$  behave the same. Hence, done. Now we focus on  $do(W_1 = 1)$ . We reuse a similar argument as shown in Equation (53). Let  $\bar{u}_1 = 2 - u_1$ . Then,  $(\mathbf{W}_{2:l})^{(1)}(\mathbf{u}) = (\mathbf{W}_{2:l})^{(2)}(\bar{\mathbf{u}})$ . Adding independent noises excluding  $Y$  ensures that the distributions are positive for  $\mathbf{P}_*$ .

Finally, we show that they disagree on  $P(W_1 = 0, \mathbf{W}_{2:l-1}, W_{l,[w_1=1]})$ . With  $W_1 = 0$  observed,  $U_1 = 0$ . Then,  $\bigoplus \mathbf{W} = \bigoplus \mathbf{W}_{2:l} = U_{l-1} \oplus W_l$ , which is 0 for  $\mathcal{M}^{(1)}$  and 1 for  $\mathcal{M}^{(2)}$ . Hence, they disagree.

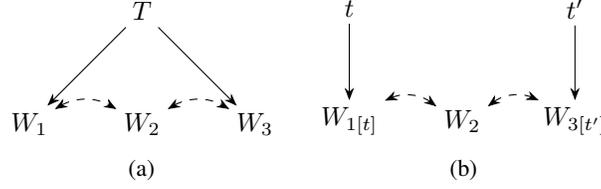


Figure 10

**(Case ii)** Let  $T$  and  $U$  be fair coins. In case of  $l = 1$ , let  $U_1$  be the exogenous variable specific to  $W_1 = W_l$ . Let

$$\begin{aligned}
 W_1^{(1)}(u_1, t) &= u_1 & W_1^{(2)}(u_1, t) &= u_1 \oplus t \\
 \forall 2 \leq i \leq l-1 W_i(u_{i-1}, u_i) &= u_{i-1} \oplus u_i \\
 W_l^{(1)}(u_{l-1}, t) &= u_{l-1} & W_l^{(2)}(u_{l-1}, t) &= u_{l-1} \oplus t.
 \end{aligned}$$

First, both models share  $P(\mathbf{W}|t)$ . For any  $t$  value and model,  $\mathbf{U}$  maps to  $\mathbf{W}$  uniformly where their bit-parity equals to zero. However, the bit-parity of  $(W_{1[t=0]}, \mathbf{W}_{2:l-1}, W_{l[t=1]})$  differ for the models since  $T$  is counted only once in  $\mathcal{M}^{(2)}$ , and, the bit-parity becomes odd.  $\square$

**Theorem 3** (Ctf-factor identifiability). *A ctf-factor  $P(\mathbf{W}_* = \mathbf{w})$  is identifiable from  $\mathbb{Z}$  if and only if it is consistent. If consistent, let  $\mathbf{W} = \mathbf{V}(\mathbf{W}_*)$  and  $\mathbf{W}' = \mathbf{V} \setminus \mathbf{W}$ ; then  $P(\mathbf{W}_* = \mathbf{w}_*)$  is equal to  $P_{\mathbf{w}'}(\mathbf{w})$  where  $\mathbf{w}$  and  $\mathbf{w}'$  are consistent with  $\mathbf{w}_* \cup \bigcup_{\{W_{\mathbf{pa}_w} \in \mathbf{W}_*\}} \mathbf{pa}_w$ .*

*Proof.* Let  $\mathbf{z} = \mathbf{w}_* \cup \bigcup_{\{W_{\mathbf{pa}_w} \in \mathbf{W}_*\}} \mathbf{pa}_w$ . Since the factor is consistent,  $\mathbf{z}$  contains at most one value per observable. We can write  $P(\mathbf{W}_* = \mathbf{w}) = P(\bigwedge_{\{W_{\mathbf{pa}_w} \in \mathbf{W}_*\}} W_{\mathbf{pa}_w} = w)$  where every  $\mathbf{pa}_w$  is consistent with  $\mathbf{z}$ .

Let  $W \in \mathbf{W}$  be such that  $\mathbf{Pa}_w \cap \mathbf{W} = \emptyset$ , then  $\mathbf{Pa}_w \subseteq \mathbf{W}'$  and  $W_{\mathbf{pa}_w} = W_{\mathbf{w}'}$  where  $\mathbf{w}'$  has any valid value for  $\mathbf{W}' \setminus \mathbf{Z}$ . For a  $W$  with parents in  $\mathbf{W}$ , suppose every  $T \in \mathbf{Pa}_w \setminus \mathbf{W}'$  already appears as  $T_{\mathbf{w}'}$ . Then, we can use composition to make  $W_{[\mathbf{pa}_w \cup \mathbf{w}']} = W_{\mathbf{w}'}$ . Applying this reasoning from parent to children in  $\mathbf{W}$  we get

$$P(\mathbf{W}_* = \mathbf{w}) = P\left(\bigwedge_{\{W_{\mathbf{pa}_w} \in \mathbf{W}_*\}} W_{\mathbf{w}'} = w\right) = P_{\mathbf{w}'}(\mathbf{w}). \quad (54)$$

The ‘only if’ portion of the statement follows immediately from Lemma 5.  $\square$

### C.1 Algorithm Completeness

**Theorem 5** (CTFIDU completeness). *A counterfactual probability  $P(\mathbf{Y}_* = \mathbf{y}_*)$  is identifiable from  $\mathbb{Z}$  and  $\mathcal{G}$  if and only if CTFIDU returns an expression for it.*

*Proof.* (soundness) The algorithm starts by removing irrelevant subscripts from the query by virtue of Lemma 2 (line 1). Then, lines 2 and 3 look for inconsistent events and redundant events, respectively. Line 4 finds the relevant ctf-factors consisting of a single c-component, as licensed by Theorem 1 and Theorem 2. As long as the factors are consistent, and allowed by Theorem 3, lines 6-11 carry out identification of the causal effect  $P_{\mathbf{V} \setminus \mathbf{C}_i}(\mathbf{C}_i)$  from the available distributions employing the algorithm IDENTIFY [26] as a subroutine. Then, the factor  $P(\mathbf{C}_{i*} = \mathbf{c}_{i*})$  is identified by evaluating  $P_{\mathbf{V} \setminus \mathbf{C}_i}(\mathbf{C}_i)$  with the proper values according to Theorem 3. If all the ctf-factors are consistent and identifiable, the algorithm returns the query as in line 14. This expression is equal to the original query due to Theorem 1 and Theorem 2.

(necessity) The procedure fails if any of the factors  $P(\mathbf{C}_{i*} = \mathbf{c}_{i*})$  is inconsistent or not identifiable from  $\mathbb{Z}$ . If inconsistent, Theorem 3 ascertains the non-identifiability of the factor from  $\mathbb{Z}$ . On the other hand, IDENTIFY (in line 7) fails to obtain  $P_{\mathbf{V} \setminus \mathbf{C}_i}(\mathbf{C}_i)$  for every  $\mathbf{Z} \in \mathbb{Z}$  such that  $\mathbf{C}_i \cap \mathbf{Z} = \emptyset$ , and the corresponding  $P_{\mathbf{V} \setminus \mathbf{B}_i}(\mathbf{B}_i)$  in  $\mathcal{G}$  only if there exists some  $\mathbf{T}_i$  s.t.  $\mathbf{C}_i \subset \mathbf{T}_i \subseteq \mathbf{B}_i$  [11, Thm. 3]. Let  $\mathbf{T} = \bigcup_{\{\mathbf{z} \in \mathbb{Z} | \mathbf{z} \cap \mathbf{C}_i = \emptyset\}} \mathbf{T}_i$  and note that  $\mathbf{C}_i$  and  $\mathbf{T}_i$  both form a single c-component in each  $\mathcal{G}_{\mathbf{z}}$ .

As  $\mathcal{G}[\mathbf{T}]$  contains all edges in each  $\mathcal{G}_{\mathbf{Z}}[\mathbf{T}_i]$ ,  $\mathcal{G}_{\mathbf{T}}$  must also be a single c-component as every variable has a bidirected path to  $\mathbf{C}_i$ .

Let  $\mathcal{H}$  be a minimal subgraph of  $\mathcal{G}_{\mathbf{T}}$  such that every edge (directed or bidirected) that can be removed without changing the c-component structure and the ancestral relationships between the variables, has been removed. Then,  $\mathcal{H}$  is a Thicket [13, Def. 6] for  $P_{\mathbf{V} \setminus \mathbf{C}_i}(\mathbf{C}_i)$  relative to  $\mathcal{G}$  and  $\mathbb{Z}$  as it is made of a single c-component, for every  $\mathbf{Z}$  there exists  $\langle \mathcal{H}[\mathbf{T}_i], \mathcal{H}[\mathbf{C}_i] \rangle$ , a hedge [14] rooted in  $\mathbf{C}_i$ .

The existence of Thicket  $\mathcal{H}$  implies the non-identifiability of  $P_{\mathbf{V} \setminus \mathbf{C}_i}(\mathbf{C}_i)$  from  $\mathbb{Z}$ . That is, there exist two models that agree on all distributions given by  $\mathbb{Z}$  but differ on  $P_{\mathbf{V} \setminus \mathbf{C}_i}(\mathbf{C}_i)$  for a particular set of values  $\mathbf{v}$ . With a proper encoding of the values, consistent with  $(\mathbf{c}_{i*} \cup \bigcup_{\mathbf{C}_t \in \mathbf{C}_{i*}} \mathbf{pa}_c)$ , this is also a counterexample for  $P(\mathbf{C}_{i*} = \mathbf{c}_{i*})$ .

If any  $P(\mathbf{C}_{i*} = \mathbf{c}_{i*})$  is not identifiable Corollary 1 implies that  $P(\mathbf{W}_* = \mathbf{w}_*)$  is also not identifiable, and in turn Lemma 1 then implies that  $P(\mathbf{Y}_* = \mathbf{y}_*)$  is not identifiable.  $\square$

## D Conditional Counterfactuals

**Lemma 6.** *Let  $A, B$  and  $C$  be binary random variables causally related as given by the chain  $A \rightarrow B \rightarrow C$ . Suppose  $P(B = 1 \mid A = 1) = \alpha$  and  $P(B = 1 \mid A = 0) = 1 - \alpha$ , for some  $\alpha \in [0, 1]$ . Then, for any  $\beta$  such that  $|1/2 - \beta| \leq |1/2 - \alpha|$  there is always a function  $f_C$  such that  $P(C = 1 \mid A = 1) = \beta$ ,  $P(C = 1 \mid A = 0) = 1 - \beta$  and  $P(A, B, C)$  is a positive distribution.*

*Proof.* Let  $P(C = 1 \mid B = 1) = x$  and  $P(C = 1 \mid B = 0) = 1 - x$ , then

$$\beta = P(C = 1 \mid A = 1) = \sum_b P(C = 1 \mid b)P(b \mid A = 1) = 1 - \alpha + x(2\alpha - 1) \quad (55)$$

$$x = \frac{\alpha + \beta - 1}{2\alpha - 1}. \quad (56)$$

Since  $x$  must belong to the interval  $(0, 1)$ , we can bound  $\beta \in (1 - \alpha, \alpha)$  if  $\alpha \leq 1/2$  and  $\beta \in (\alpha, 1 - \alpha)$  if  $\alpha > 1/2$ . Both conditions are satisfied when  $|1/2 - \beta| \leq |1/2 - \alpha|$  as assumed, so any solution  $x$  is a valid probability.

Then, we can define the function for  $C$  as

$$f_C = B \oplus U_c, \quad (57)$$

where  $U_c$  is a binary unobservable,  $P(U_c = 0) = x$  and  $\oplus$  is the logical xor operator.  $\square$

**Lemma 7.** *Let  $A$  and  $B$  be two variables in a causal graph where  $A \leftarrow A_n \leftarrow A_{n-1} \leftarrow \dots \leftarrow A_1 \leftarrow C \rightarrow B_1 \rightarrow \dots \rightarrow B_{m-1} \rightarrow B_m \rightarrow B$ . The variables  $A_1, \dots, A_n, B_1, \dots, B_m$  are observable,  $C$  could be observable or unobservable and  $m, n$  are nonnegative integers. Then we can define functions for all variables involved such that they are binary and*

$$P(a, b) = \begin{cases} \frac{1}{2}\gamma & \text{if } a = b \\ \frac{1}{2}(1 - \gamma) & \text{otherwise} \end{cases}, \quad (58)$$

for any  $\gamma \in (0, 1)$ .

*Proof.* First, if  $C$  is unobservable set  $P(C = 1) = 1/2$ , else define an unobservable  $U_c$  with  $P(U_c = 1) = 1/2$  and let  $f_C = U_c$ . Let  $\alpha, \beta \in (0, 1)$  be parameters to decide later.

If  $n = 0$  define  $f_A$  such that  $P(A = 1 \mid C = 1) = \alpha$ ,  $P(A = 1 \mid C = 0) = 1 - \alpha$ . Similarly, if  $m = 0$  define  $f_B$  such that  $P(B = 1 \mid C = 1) = \alpha$ ,  $P(B = 1 \mid C = 0) = 1 - \alpha$ .

Suppose  $n > 0$ , then we will define the functions for  $A_1, A_2, \dots, A_n, A$  such that  $P(A_i = 1 \mid C = 1)$  gets closer to  $\alpha$  as  $i$  increases. If  $\alpha < 1/2$ , set  $f_{A_1}$  such that  $P(A_1 = 1 \mid C = 1) = \alpha/(n + 1)$ ,  $P(A_1 = 1 \mid C = 0) = 1 - \alpha/(n + 1)$ . Then use lemma 6 to define  $f_{A_i}$ ,  $i = 1, \dots, n$  and  $f_A$  such that  $P(A_i = 1 \mid C = 1) = i\alpha/(n + 1)$ ,  $P(A_i = 1 \mid C = 0) = 1 - i\alpha/(n + 1)$  and finally  $P(A = 1 \mid C = 1) = (n + 1)\alpha/(n + 1) = \alpha$ ,  $P(A = 1 \mid C = 0) = 1 - \alpha$ .

If  $\alpha > 1/2$  use the same strategy but starting from  $P(A_1 = 1 \mid C = 1) = 1 - (1 - \alpha)/(n + 1)$  and decreasing as  $P(A_i = 1 \mid C = 1) = 1 - i(1 - \alpha)/(n + 1)$ , to obtain  $P(A = 1 \mid C = 1) = 1 - (n + 1)(1 - \alpha)/(n + 1) = \alpha$ .

The same procedure is applied for  $B_1, \dots, B_m, B$  to obtain  $P(B = 1 \mid C = 1) = \beta$ ,  $P(B = 1 \mid C = 0) = 1 - \beta$ .

Finally,

$$P(A = 1, B = 1) = \sum_c P(A = 1 \mid c)P(B = 1 \mid c)P(c) \quad (59)$$

$$= \frac{1}{2}[\alpha\beta + (1 - \alpha)(1 - \beta)] \quad (60)$$

$$P(A = 0, B = 0) = \frac{1}{2}[(1 - \alpha)(1 - \beta) + \alpha\beta] \quad (61)$$

$$P(A = 0, B = 1) = \frac{1}{2}[(1 - \alpha)\beta + \alpha(1 - \beta)] \quad (62)$$

$$P(A = 1, B = 0) = \frac{1}{2}[\alpha(1 - \beta) + (1 - \alpha)\beta]. \quad (63)$$

If  $\gamma < 1/2$  make  $\beta = 1 - \alpha$  and  $\alpha = \frac{1 \pm \sqrt{1 - 2\gamma}}{2}$ . If  $\gamma \geq 1/2$ , let  $\beta = \alpha$  and  $\alpha = \frac{1 \pm \sqrt{2\gamma - 1}}{2}$ . It is just a matter of algebra to verify that  $P(A, B)$  results in the intended distribution.  $\square$

**Lemma 8.** *Let  $\mathbf{W}_*$  be a set of counterfactual variables,  $\mathbf{X}_*$ , and let  $\mathbf{A}_{1*}, \mathbf{A}_{2*}, \dots$  be the ancestral components of  $\mathbf{W}_*$  given  $\mathbf{X}_*$ . Then*

$$P(\mathbf{W}_* = \mathbf{w}_*) = \prod_j \sum_{\mathbf{a}_{j*} \setminus \mathbf{w}_*} P\left(\bigwedge_{A_t \in \mathbf{A}_{j*}} A_{\mathbf{pa}_a} = a\right), \quad (64)$$

where  $\mathbf{pa}_a$  is consistent with  $\mathbf{t}$  and  $\mathbf{a}_{j*}$ , for each  $A_t \in \mathbf{A}_{j*}$ .

*Proof.* For the sake of simplicity let  $\mathbf{X}_* = \|\mathbf{X}_*\|$ . First, note that by definition  $\mathbf{A}_{1*}, \mathbf{A}_{2*}, \dots$  must be mutually disjoint and their union is equal to  $\mathbf{A}_* = An(\mathbf{W}_*)$ . Then, by Theorem 1 we can write

$$P(\mathbf{W}_* = \mathbf{w}_*) = \sum_{\mathbf{a}_* \setminus \mathbf{w}_*} P\left(\bigwedge_j \mathbf{A}_{j*} = \mathbf{a}_{j*}\right). \quad (65)$$

Furthermore, no variable in two distinct ancestral components could be in the same c-component in  $\mathcal{G}[\mathbf{V}(\mathbf{A}_*)]$ . Hence, by Theorem 2 we have

$$P(\mathbf{W}_* = \mathbf{w}_*) = \sum_{\mathbf{a}_* \setminus \mathbf{w}_*} \prod_j P\left(\bigwedge_{A_t \in \mathbf{A}_{j*}} A_{\mathbf{pa}_a} = a\right). \quad (66)$$

To prove the result, it suffices to show that this expression can be factorized such that the sum over each  $\mathbf{a}_{j*} \setminus \mathbf{w}_*$  only affects factor  $j$ . For this, we argue that for  $A_t \in \mathbf{A}_{j*}$  no value of  $\mathbf{pa}_a$  or  $a$  appears in  $\mathbf{a}_{k*} \setminus \mathbf{w}_*$  for any  $k \neq j$ . To see why this is the case, assume for the sake of contradiction that (i) there is some value in  $\mathbf{pa}_a$  that appears in such a  $\mathbf{a}_{k*} \setminus \mathbf{w}_*$  or (ii)  $a$  appears in the same set. For case (i), this implies that there is an ancestor of  $A_t$  (not in  $\mathbf{X}_*$ ) which is also an ancestor of some variable in component  $\mathbf{A}_{k*}$ . For case (ii), this means that  $A_t \notin \mathbf{X}_*$  is an ancestor of some variable in component  $\mathbf{A}_{k*}$ . In both cases would imply that  $\mathbf{A}_{j*}$  and  $\mathbf{A}_{k*}$  are not disjoint, a contradiction.  $\square$

**Lemma 3** (Conditional Query Reduction). *Let  $\mathbf{Y}_*, \mathbf{X}_*$  be two sets of counterfactual variables and let  $\mathbf{D}_*$  be the set of variables in the same ancestral component, given  $\mathbf{X}_*$ , as any variable in  $\mathbf{Y}_*$ , then*

$$P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*) = \frac{\sum_{\mathbf{d}_* \setminus (\mathbf{y}_* \cup \mathbf{x}_*)} P(\bigwedge_{D_t \in \mathbf{D}_*} D_{\mathbf{pa}_d} = d)}{\sum_{\mathbf{d}_* \setminus \mathbf{x}_*} P(\bigwedge_{D_t \in \mathbf{D}_*} D_{\mathbf{pa}_d} = d)}, \quad (24)$$

where  $\mathbf{pa}_d$  is consistent with  $\mathbf{t}$  and  $\mathbf{d}_*$ , for each  $D_t \in \mathbf{D}_*$ . Moreover,  $P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*)$  is identifiable from  $\mathbb{Z}$  if and only if  $P(\bigwedge_{D_t \in \mathbf{D}_*} D_{\mathbf{pa}_d} = d)$  is identifiable from  $\mathbb{Z}$ .

*Proof.*

$$P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*) = \frac{P(\mathbf{Y}_* = \mathbf{y}_*, \mathbf{X}_* = \mathbf{x}_*)}{\sum_{\mathbf{y}_*} P(\mathbf{Y}_* = \mathbf{y}_*, \mathbf{X}_* = \mathbf{x}_*)} \quad (67)$$

Let  $\mathbf{A}_{1*}, \mathbf{A}_{2*}, \dots$  be the ancestral components of  $\mathbf{Y}_* \cup \mathbf{X}_*$ , given  $\mathbf{X}_*$ . Then, by Lemma 8

$$P(\mathbf{Y}_* = \mathbf{y}_*, \mathbf{X}_* = \mathbf{x}_*) = \prod_j \sum_{\mathbf{a}_{j*} \setminus (\mathbf{y}_* \cup \mathbf{x}_*)} P\left(\bigwedge_{A_t \in \mathbf{A}_{j*}} A_{\text{pa}_a} = a\right). \quad (68)$$

Let  $Y_t \in \mathbf{Y}_*$  and  $y$  its corresponding value in  $\mathbf{y}_*$ . If there exists any variable  $A_t \in \mathbf{A}_*$  such that  $y \in \mathbf{t}$ ,  $Y_t$  and  $A_t$  must be in the same  $\mathbf{A}_{j*}$ , as the former is then an ancestor of the latter. Let  $\mathbf{A}_{\mathbf{y}_*}$  be the union of ancestral components that contain ancestors of  $\mathbf{Y}_*$ , and let  $\mathbf{A}_{\bar{\mathbf{y}}_*}$  be the rest. Then, if we sum over  $\mathbf{y}_*$  in the right-hand-side of Eq. (68), the factors with variables in  $\mathbf{A}_{\bar{\mathbf{y}}_*}$  can be taken out of the sum.

$$\begin{aligned} & P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*) \\ &= \frac{\left(\sum_{\mathbf{a}_{\bar{\mathbf{y}}_*} \setminus \mathbf{x}_*} P\left(\bigwedge_{A_t \in \mathbf{A}_{\bar{\mathbf{y}}_*}} A_{\text{pa}_a} = a\right)\right) \left(\sum_{\mathbf{a}_{\mathbf{y}_*} \setminus (\mathbf{y}_* \cup \mathbf{x}_*)} P\left(\bigwedge_{A_t \in \mathbf{A}_{\mathbf{y}_*}} A_{\text{pa}_a} = a\right)\right)}{\left(\sum_{\mathbf{a}_{\bar{\mathbf{y}}_*} \setminus \mathbf{x}_*} P\left(\bigwedge_{A_t \in \mathbf{A}_{\bar{\mathbf{y}}_*}} A_{\text{pa}_a} = a\right)\right) \sum_{\mathbf{y}_*} \left(\sum_{\mathbf{a}_{\mathbf{y}_*} \setminus (\mathbf{y}_* \cup \mathbf{x}_*)} P\left(\bigwedge_{A_t \in \mathbf{A}_{\mathbf{y}_*}} A_{\text{pa}_a} = a\right)\right)} \end{aligned} \quad (69)$$

After cancelling out the first factor, simplifying the expression, and defining  $\mathbf{D}_* = \mathbf{A}_{\mathbf{y}_*}$ , we obtain Eq. (24).

For simplicity, let  $\mathcal{G}'$  be the same as  $\mathcal{G}$  after removing all edges out of  $\mathbf{V}(\mathbf{X}_* \cap \mathbf{D}_*)$  and any edge out of  $\mathbf{V}(\mathbf{Y}_* \cap \text{An}(\mathbf{X}_*))$ .  $\mathcal{G}'$  and  $\mathcal{G}$  have the same c-component structure and the same ancestral relationships; therefore and by the same reasoning, the query is not identifiable from  $\langle \mathcal{G}', \mathbb{Z} \rangle$  either. For the rest of the proof, let  $\mathcal{G}$  denote  $\mathcal{G}'$  instead. Clearly, if  $P(\bigwedge_{D_t \in \mathbf{D}_*} D_{\text{pa}_d} = d)$  is identifiable, so it is  $P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*)$ . For the second part of the statement, suppose this query is not identifiable and let  $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$  be two models witnessing this fact.

For simplicity, let

$$\rho(\mathbf{y}, \mathbf{x}) = \sum_{\mathbf{d}_* \setminus (\mathbf{y}_* \cup \mathbf{x}_*)} P\left(\bigwedge_{D_t \in \mathbf{D}_*} D_{\text{pa}_d} = d\right) \quad (70)$$

then

$$P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*) = \frac{\rho(\mathbf{y}_*, \mathbf{x}_*)}{\sum_{\mathbf{y}_*} \rho(\mathbf{y}_*, \mathbf{x}_*)}. \quad (71)$$

If  $\rho(\mathbf{x}) = \sum_{\mathbf{y}} \rho(\mathbf{y}, \mathbf{x})$  is identifiable then  $P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*)$  must be non-identifiable, else

$$\rho(\mathbf{y}, \mathbf{x}) = P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*) \rho(\mathbf{x}), \quad (72)$$

contradicting the assumption that the  $\rho(\mathbf{y}, \mathbf{x})$  is not identifiable. Therefore, we can further assume  $\rho(\mathbf{x})$  is not identifiable, for the rest of the argument.

Let  $(\mathbf{y}', \mathbf{x}')$  be an assignment such that  $\rho(\mathbf{y}', \mathbf{x}')$  differs between  $\mathcal{M}^{(1)}$  and  $\mathcal{M}^{(2)}$ , then let

$\mathbf{X}_*$	$\mathbf{Y}_*$	$\rho^{(1)}(\mathbf{X}, \mathbf{Y})$	$\rho^{(2)}(\mathbf{X}, \mathbf{Y})$
$\mathbf{x}'$	$\mathbf{y}'$	$a$	$b$
$\mathbf{x}'$	$\neq \mathbf{y}'$	$c$	$d$
$\neq \mathbf{x}'$	$\mathbf{y}'$	$e$	$f$
$\neq \mathbf{x}'$	$\neq \mathbf{y}'$	$g$	$h$

Due to the non-identifiability of  $\rho(\mathbf{y}, \mathbf{x})$  we have  $a \neq b$  and without loss of generality we can assume  $a > b$ . Similarly, due to the non-identifiability of  $\rho(\mathbf{x})$  we have  $a + c \neq b + d$ . For  $\mathcal{M}^{(1)}$  and  $\mathcal{M}^{(2)}$ :

$$\rho^{(1)}(\mathbf{y}' \mid \mathbf{x}') = \frac{a}{a + c} \quad (73)$$

$$\rho^{(2)}(\mathbf{y}' \mid \mathbf{x}') = \frac{b}{b + d}. \quad (74)$$

These probabilities are equal if and only if  $ad = bc$ . Hence, if they are not equal, we are done as  $\rho^{(1)}(\mathbf{y}' \mid \mathbf{x}') \neq \rho^{(2)}(\mathbf{y}' \mid \mathbf{x}')$ . If they are equal, let  $X_t \in \mathbf{X}_* \cap \mathbf{D}_*$  be such that  $X_t$  and some

$Y_{\mathbf{r}} \in \mathbf{Y}_*$  have a common ancestor or they have ancestors sharing a bidirected edge. Assume  $X_{\mathbf{t}}$  is such that no other element in  $\mathbf{X}_*$  is closer to  $Y_{\mathbf{r}}$  in the corresponding path in  $\mathcal{G}$ . Let  $\bar{p}$  be the path in  $\mathcal{G}$  that goes from  $X$  to  $Y$ , passing to the common ancestor or bidirected alluded before. Add a bit to every variable in  $\bar{p}$  and denote them with subscript  $p$ . Define independent functions for the bits which we will parametrize later.

We define two new models,  $\mathcal{M}^{(1)'}$  and  $\mathcal{M}^{(2)'}$ , based on  $\mathcal{M}^{(1)}$  and  $\mathcal{M}^{(2)}$ . For every variable in  $\bar{p}$  except for  $X$  and  $Y$ , append the corresponding extra bit, defined in  $\bar{p}$ , to the original variables in  $\mathcal{M}^{(1)}$  and  $\mathcal{M}^{(2)}$ . Rename  $X$  and  $Y$  as  $\tilde{X}, \tilde{Y}$  and make them unobservable, then define  $X$  in the new models with functions:

$$f'_x = \begin{cases} x' & \text{if } X_p = 1 \\ \tilde{X} & \text{otherwise,} \end{cases} \quad (75)$$

where  $X_p$  is unobservable too and  $x' = \mathbf{x}'(X)$ , for which the query disagrees in  $\mathcal{M}^{(1)}$  and  $\mathcal{M}^{(2)}$ .

Analogously, define

$$f'_y = \begin{cases} y' & \text{if } Y_p = 1 \\ \tilde{Y} & \text{otherwise.} \end{cases} \quad (76)$$

According to our choice of  $X$ , there exists  $Z \in \bar{p} \cap An(X) \cap An(Y)$  (possibly  $Y$  itself) or there exists  $Z_1, Z_2 \in \bar{p}$  with  $Z_1 \in An(X), Z_2 \in An(Y)$  and there is an edge  $Z_1 \leftarrow \dots \rightarrow Z_2$  in  $\bar{p}$ . For the parametrization of the extra bits in  $\bar{p}$  define a new unobservable  $U$  and let  $Z_p = U$  if the common ancestor is observable or let  $U$  be the unobservable parent of  $Z_{1p}$  and  $Z_{2p}$  in the second. Notice that  $Z_1$  and  $Z_2$  may be equal to  $X$  and  $Y$  themselves.

Using lemma 7 we will parametrize  $\bar{p}$  such that

$X_0$	$Y_0$	$P(X_0, Y_0)$
1	1	$\frac{1}{2}\gamma$
1	0	$\frac{1}{2}(1 - \gamma)$
0	1	$\frac{1}{2}(1 - \gamma)$
0	0	$\frac{1}{2}\gamma$

for some  $\gamma \in (0, 1)$  that we will pick later.

**Claim 1** (Disagreement on the query).  $\mathcal{M}^{(1)'}$  and  $\mathcal{M}^{(2)'}$  disagree on the query for any  $\gamma$  such that  $c - d \neq [(a + c + 1)h - (b + d + 1)g](1 - \gamma)$ .

*Proof.* For  $\mathcal{M}^{(1)'}$  and  $\mathcal{M}^{(2)'}$  we have

$$\rho'(\mathbf{x}', \mathbf{y}') = \sum_{x_p, y_p, \tilde{x}, \tilde{y}} \rho'(\mathbf{x}', \mathbf{y}', \tilde{x}, \tilde{y} \mid x_p, y_p) P(x_p, y_p). \quad (77)$$

Going over each possible combination of  $X_p$  and  $Y_p$ , we get

$$\rho'(\mathbf{x}', \mathbf{y}') = \rho(\mathbf{X} = \mathbf{x}', \mathbf{Y} = \mathbf{y}') P(X_p = 0, Y_p = 0) \quad (78)$$

$$+ \rho(\mathbf{Y} = \mathbf{y}') P(X_p = 1, Y_p = 0) \quad (79)$$

$$+ \rho(\mathbf{X} = \mathbf{x}') P(X_p = 0, Y_p = 1) \quad (80)$$

$$+ P(X_p = 1, Y_p = 1). \quad (81)$$

Similarly,

$$\rho'(\mathbf{x}') = \rho(\mathbf{X} = \mathbf{x}') P(X_p = 0) + P(X_p = 1). \quad (82)$$

For  $\mathcal{M}^{(1)'}$

$$\rho^{(1)'(\mathbf{y}' | \mathbf{x}') = \frac{\frac{1}{2}a\gamma + \frac{1}{2}(a+e)(1-\gamma) + \frac{1}{2}(a+c)(1-\gamma) + \frac{1}{2}\gamma}{\frac{1}{2}(a+c) + \frac{1}{2}} \quad (83)$$

$$= \frac{a\gamma + (2a+c+e)(1-\gamma) + \gamma}{a+c+1} \quad (84)$$

$$= \frac{a - (a+c+e)(\gamma-1) + \gamma}{a+c+1} \quad (85)$$

$$= \frac{a + (a+c+e)(1-\gamma) - (1-\gamma) + 1}{a+c+1} \quad (86)$$

$$= \frac{a - (1 - (a+c+e))(1-\gamma) + 1}{a+c+1} \quad (87)$$

$$= \frac{a - g(1-\gamma) + 1}{a+c+1} \quad (88)$$

Analogously for  $\mathcal{M}^{(2)'}$ :

$$\rho^{(2)'(\mathbf{y}' | \mathbf{x}') = \frac{b - h(1-\gamma) + 1}{b+d+1} \quad (89)$$

Those two are equal if and only if

$$\begin{aligned} & ab - bg(1-\gamma) + b + ad - dg(1-\gamma) + d + a - g(1-\gamma) + 1 \\ &= ab - ah(1-\gamma) + a + bc - ch(1-\gamma) + c + b - h(1-\gamma) + 1 \end{aligned} \quad (90)$$

$$\begin{aligned} & \iff \\ & -bg(1-\gamma) + ad - dg(1-\gamma) + d - g(1-\gamma) \\ &= -ah(1-\gamma) + bc - ch(1-\gamma) + c - h(1-\gamma) \end{aligned} \quad (91)$$

Recall that we have  $ad = bc$ , which also implies that  $c \neq d$ , else  $a = b$  which is a contradiction. Then, the condition for equality can be further simplified as

$$c - d = [(a+c+1)h - (b+d+1)g](1-\gamma). \quad (92)$$

The left hand side is nonzero, all  $a, b, c, d, g$  and  $h$  are fixed, and  $\gamma$  is a free parameter. Therefore, as long as we pick a  $\gamma$  such that the equality doesn't hold, we get that  $\rho^{(1)'(\mathbf{y}' | \mathbf{x}') = P^{(1)'(\mathbf{Y}_* = \mathbf{y}' | \mathbf{X}_* = \mathbf{x}')}$  does not match in  $\mathcal{M}^{(1)'}$  and  $\mathcal{M}^{(2)'}$ .  $\square$

**Claim 2** (Agreement on given distributions). *Let  $\mathbf{Z} \subset \mathbf{V}$  be any subset of observable variables and let  $\mathbf{Z} \in \mathbb{Z}$ . If  $\mathcal{M}^{(1)}$  and  $\mathcal{M}^{(2)}$  agree on  $P_{\mathbf{z}}(\mathbf{V})$ , then  $\mathcal{M}^{(1)'}$  and  $\mathcal{M}^{(2)'}$  also agree on  $P_{\mathbf{z}'}(\mathbf{V})$ .*

*Proof.* As the input distributions are in layer 2, let us use c-factor notation, where each c-factor corresponds to a causal effect. For any  $\mathbf{C} \subseteq \mathbf{V}$ , the quantity  $Q[\mathbf{C}](\mathbf{v})$  is the c-factor of  $\mathbf{C}$  and denotes the following function

$$Q[\mathbf{C}](\mathbf{v}) = \sum_{\mathbf{u}(\mathbf{C})} \prod_{\{i | V_i \in \mathbf{C}\}} P^k(v_i | \mathbf{pa}_i, \mathbf{u}_i) P^k(\mathbf{u}(\mathbf{C})), \quad (93)$$

where  $\mathbf{U}(\mathbf{C}) = \bigcup_{V_i \in \mathbf{C}} \mathbf{U}_i$ .

Let  $\mathbf{C}_1, \mathbf{C}_2, \dots$  be the c-components of  $\mathcal{G}_{\bar{\mathbf{z}}}$ . By assumption we have  $P_{\mathbf{z}}^{(1)}(\mathbf{V} \setminus \mathbf{Z}) = Q^{(1)}[\mathbf{V} \setminus \mathbf{Z}] = Q^{(2)}[\mathbf{V} \setminus \mathbf{Z}] = P_{\mathbf{z}}^{(2)}(\mathbf{V} \setminus \mathbf{Z})$ , and since any  $Q[\mathbf{C}_j]$  is identifiable from  $Q[\mathbf{V} \setminus \mathbf{Z}]$  [27], it follows  $Q^{(1)}[\mathbf{C}_j] = Q^{(2)}[\mathbf{C}_j]$  for any  $\mathbf{C}_j$ .

$\mathcal{M}^{(k)'}$  is identical to  $\mathcal{M}^{(k)}$ ,  $k = 1, 2$ , except for the functions of the observables in the path  $\bar{p}$ . For any variable  $T$  not in  $\bar{p}$ , but with a parent on it, the function  $f_T$  remains the same and it simply ignores the extra bit that its parent has in  $\mathcal{M}^{(k)'}$ .

Let  $\mathbf{C}_j$  be a c-component containing some set of variables  $\mathbf{R}$  in  $\bar{p}$  different to  $X$  and  $Y$  (the endpoints of  $\bar{p}$ ). First, by definition

$$Q^{(k)'[\mathbf{C}_j](\mathbf{v}) = \sum_{\mathbf{u}(\mathbf{C}_j)} \prod_{V_i \in \mathbf{C}_j} P^{(k)'(v_i | \mathbf{pa}_i, \mathbf{u}_i) P^{(k)'(\mathbf{u}(\mathbf{C}_j))}. \quad (94)$$

For any  $S \notin \bar{p}$ ,  $R \in \mathbf{R}$ ,  $X$  and  $Y$  that could be in  $\mathbf{C}_j$  in  $\mathcal{M}^{(k)'}$ , their corresponding factors in the previous expression can be re-written in terms of probabilities of  $\mathcal{M}^{(k)}$ , as follows:

$$P^{(k)'}(s \mid \mathbf{pa}_s, \mathbf{u}_s) = P^{(k)}(s \mid \mathbf{pa}_s, \mathbf{u}_s), \quad (95)$$

$$P^{(k)'}(r \mid \mathbf{pa}_r, \mathbf{u}_r) = P^{(k)}(r \mid \mathbf{pa}_r, \mathbf{u}_r)P(r_p \mid (\mathbf{pa}_r)_p), \quad (96)$$

$$P^{(k)'}(y \mid \mathbf{pa}_y, \mathbf{u}_y) = P^{(k)}(y \mid \mathbf{pa}_y, \mathbf{u}_y)P(Y_p = 0 \mid (\mathbf{pa}_y)_p) \\ + 1[y = y']P(Y_p = 1 \mid (\mathbf{pa}_y)_p), \quad (97)$$

$$P^{(k)'}(x \mid \mathbf{pa}_x, \mathbf{u}_x) = P^{(k)}(x \mid \mathbf{pa}_x, \mathbf{u}_x)P(X_p = 0 \mid (\mathbf{pa}_x)_p) \\ + 1[x = x']P(X_p = 1 \mid (\mathbf{pa}_x)_p). \quad (98)$$

It follows that

$$Q^{(k)'}[\mathbf{C}_j](\mathbf{v}) = \left( \prod_{R \in \mathbf{R}} P(r_p \mid (\mathbf{pa}_r)_p) \right) \left[ \begin{aligned} & Q^{(k)}[\mathbf{C}_j](\mathbf{v})P(Y_p = 0 \mid (\mathbf{pa}_y)_p)P(X_p = 0 \mid (\mathbf{pa}_x)_p) \\ & + Q^{(k)}[\mathbf{C}_j \setminus \{Y\}](\mathbf{v})P(Y_p = 1 \mid (\mathbf{pa}_y)_p)P(X_p = 0 \mid (\mathbf{pa}_x)_p)1[y = y'] \\ & + Q^{(k)}[\mathbf{C}_j \setminus \{X\}](\mathbf{v})P(Y_p = 0 \mid (\mathbf{pa}_y)_p)P(X_p = 1 \mid (\mathbf{pa}_x)_p)1[x = x'] \\ & + Q^{(k)}[\mathbf{C}_j \setminus \{X, Y\}](\mathbf{v})P(Y_p = 1 \mid (\mathbf{pa}_y)_p)P(X_p = 1 \mid (\mathbf{pa}_x)_p)1[x = x', y = y'] \end{aligned} \right]. \quad (99)$$

Since  $X$  and  $Y$  have no descendants in  $\mathcal{G}$ ,

$$Q^{(k)}[\mathbf{C}_j \setminus \{X, Y\}](\mathbf{v}) = \sum_{x, y} Q^{(k)}[\mathbf{C}_j](\mathbf{v}) \quad (100)$$

$$Q^{(k)}[\mathbf{C}_j \setminus \{X\}](\mathbf{v}) = \sum_x Q^{(k)}[\mathbf{C}_j](\mathbf{v}) \quad (101)$$

$$Q^{(k)}[\mathbf{C}_j \setminus \{Y\}](\mathbf{v}) = \sum_y Q^{(k)}[\mathbf{C}_j](\mathbf{v}), \quad (102)$$

all match between  $\mathcal{M}^{(1)}$  and  $\mathcal{M}^{(2)}$ . Consequently, every c-factor in the right-hand side of (99) is the same in those models, and since every other term is also the same in both  $\mathcal{M}^{(1)'}$  and  $\mathcal{M}^{(2)'}$ , we conclude that  $Q^{(1)'}[\mathbf{C}_j](\mathbf{v}) = Q^{(2)'}[\mathbf{C}_j](\mathbf{v})$ , which in turn implies our claim since

$$P_{\mathbf{z}}^{(k)'}(\mathbf{v}) = \prod_j Q^{(k)'}[\mathbf{C}_j](\mathbf{v}). \quad (103)$$

□

In summary,  $\mathcal{M}^{(1)'}$  and  $\mathcal{M}^{(2)'}$  induce  $\mathcal{G}$  and matching  $\mathbb{Z}$ , yet they differ on the value for  $P'(\mathbf{Y}_* = \mathbf{y}' \mid \mathbf{X}_* = \mathbf{x}')$ , proving the non-identifiability of the query. □

**Theorem 6** (CTFID completeness). *A counterfactual probability  $P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*)$  is identifiable from  $\mathbb{Z}$  and  $\mathcal{G}$  if and only if CTFID returns an expression for it.*

*Proof.* Lines 1 and 2 determine the set of variables and values  $\mathbf{D}_*$  and  $\mathbf{d}_*$  as specified by Lemma 3 to reduce the conditional query to a marginal one. Then, line 3 invokes CTFID to identify this quantity. If CTFID succeeds, the returned expression is equal to the conditional query. If it fails, Lemma 3 implies that the conditional query is also not identifiable. □

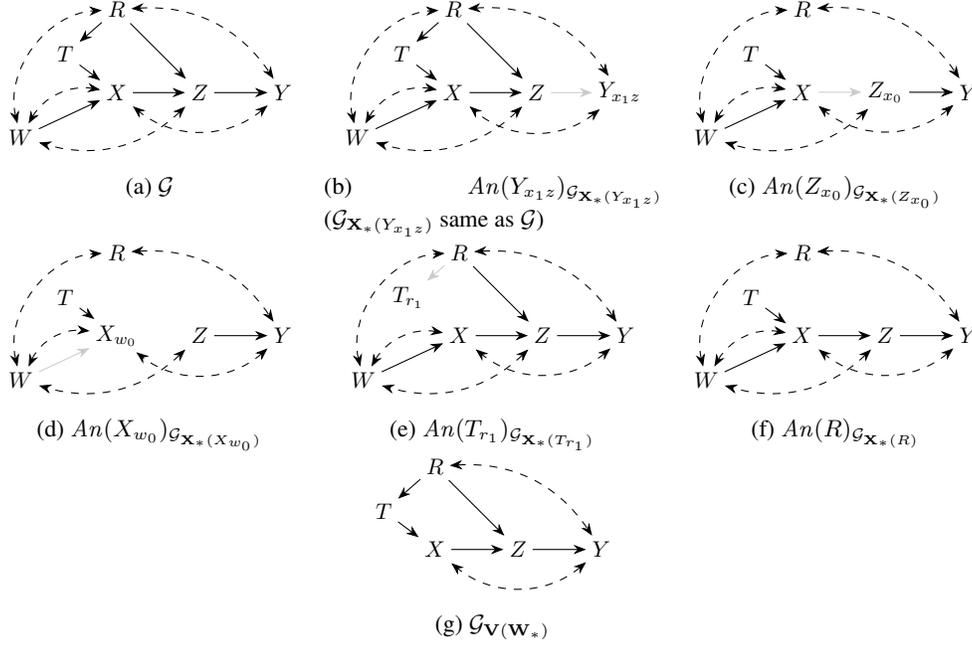


Figure 11: Graphs used for the running example of the algorithms.

## E Running example of the algorithm

Consider the causal diagram  $\mathcal{G}$  in Fig. 11(a) and the identification of the query  $P(Y_{x_1 z} = y_1 \mid X_{w_0} = x_1, T_{r_1} = t_1, R = r_0)$  from  $\mathcal{Z} = \{\{Y\}, \{W, T\}\}$ . First, we can use Theorem 4 to write the query as:

$$\sum_z P(Y_{x_1 z} = y_1, Z_{x_0} = z \mid X_{w_0} = x_1, T_{r_1} = t_1, R = r_0). \quad (104)$$

The query within the sum can be processed using  $\text{COND-CTFID}(\mathbf{Y}_* = \{Y_{x_1 z}, Z_{x_0}\}, \mathbf{y}_* = \{y_1, z\}, \mathbf{X}_* = \{X_{w_0}, T_{r_1}, R\}, \mathbf{x}_* = \{x_1, t_1, r_0\}, \mathcal{Z}, \mathcal{G})$ .

At line 1 the algorithm looks for the ancestral components of  $\mathbf{Y}_* \cup \mathbf{X}_*$  given  $\mathbf{X}_*$ . For this let us look at the ancestral sets of each  $\mathbf{W}_t \in \mathbf{Y}_* \cup \mathbf{X}_*$  in the graph  $An(W_t)G_{X_*(W_t)}$ . Recall the set  $\mathbf{X}_*(W_t)$  is defined as  $\mathbf{V}(\|\mathbf{X}_*\| \cap An(W_t))$ . Next, noting that  $\|\mathbf{X}_*\| = \{X_{w_0}, T_{r_1}, R = r_0\}$ , we compute each of this sets as

$$\mathbf{X}_*(Y_{x_1 z}) = \mathbf{V}(\|\mathbf{X}_*\| \cap \{Y_z\}) = \emptyset, \quad (105)$$

$$\mathbf{X}_*(Z_{x_0}) = \mathbf{V}(\|\mathbf{X}_*\| \cap \{Z_{x_0}, R\}) = \{R\}, \quad (106)$$

$$\mathbf{X}_*(X_{w_0}) = \mathbf{V}(\|\mathbf{X}_*\| \cap \{X_{w_0}, T, R\}) = \{X, R\}, \quad (107)$$

$$\mathbf{X}_*(T_{r_1}) = \mathbf{V}(\|\mathbf{X}_*\| \cap \{T_{r_1}\}) = \{T\}, \quad (108)$$

$$\mathbf{X}_*(R) = \mathbf{V}(\|\mathbf{X}_*\| \cap \{R\}) = \{R\}. \quad (109)$$

Then, we compute the ancestral sets as

$$An(Y_{x_1 z})G_{X_*(Y_{x_1 z})} = \{Y_z\}, \quad (\text{Fig. 11(b)}) \quad (110)$$

$$An(Z_{x_0})G_{X_*(Z_{x_0})} = \{Z_{x_0}\}, \quad (\text{Fig. 11(c)}) \quad (111)$$

$$An(X_{w_0})G_{X_*(X_{w_0})} = \{X_{w_0}, T\}, \quad (\text{Fig. 11(d)}) \quad (112)$$

$$An(T_{r_1})G_{X_*(T_{r_1})} = \{T_{r_1}\}, \quad (\text{Fig. 11(e)}) \quad (113)$$

$$An(R)G_{X_*(R)} = \{R\}. \quad (\text{Fig. 11(f)}) \quad (114)$$

Although the ancestral sets are disjoint, there are relevant bidirected arrows connecting the pairs  $(Y, X)$  and  $(Y, R)$ , resulting in three ancestral components. They can be labeled as  $\mathbf{A}_{1*} = \{Y_z, X_{w_0}, T, R\}$ ,

---

**Algorithm 3** IDENTIFY( $\mathbf{C}, \mathbf{T}, Q, \mathcal{G}$ )

---

**Input:**  $\mathbf{C} \subseteq \mathbf{T} \subseteq \mathbf{V}$ ,  $Q = Q[\mathbf{T}] = P_{\mathbf{V} \setminus \mathbf{T}}(\mathbf{T})$  and graph  $\mathcal{G}$ . Assuming  $\mathcal{G}_{[\mathbf{C}]}$  and  $\mathcal{G}_{[\mathbf{T}]}$  are composed of a single c-component.

**Output:** Expression for  $Q[\mathbf{C}] = P_{\mathbf{V} \setminus \mathbf{C}}(\mathbf{C})$  in terms of  $Q$  or FAIL.

- 1: Let  $\mathbf{A} \leftarrow An(\mathbf{C})_{\mathcal{G}_{[\mathbf{T}]}}$ .
  - 2: **if**  $\mathbf{A} = \mathbf{C}$  **then return**  $Q[\mathbf{C}] = \sum_{\mathbf{t} \setminus \mathbf{c}} Q$ .
  - 3: **if**  $\mathbf{A} = \mathbf{T}$  **then return** FAIL.
  - 4: **if**  $\mathbf{A} = \mathbf{C}$  **then**
  - 5:   Let  $\mathbf{T}'$  be the c-component containing  $\mathbf{C}$  in  $\mathcal{G}_{[\mathbf{A}]}$ .
  - 6:   Compute  $Q[\mathbf{T}']$  from  $Q[\mathbf{A}] = \sum_{\mathbf{t} \setminus \mathbf{a}} Q$ .
  - 7:   **return** IDENTIFY( $\mathbf{C}, \mathbf{T}', Q[\mathbf{T}'], \mathcal{G}$ ).
  - 8: **end if**
- 

$\mathbf{A}_{2*} = \{Z_{x_0}\}$  and  $\mathbf{A}_{3*} = \{T_{r_1}\}$ . Then, we let  $\mathbf{D}$  be the union of the ancestral sets that contain variables in  $\mathbf{Y}_*$ , that is  $\mathbf{D}_* = \{Y_z, X_{w_0}, T, R, Z_{x_0}\}$ . We also gather the corresponding values  $\mathbf{d}_* = \{y_1, x_1, t, r_0, z\}$ , where  $t$  is a new value introduced for  $T$  that will appear as the index in a sum on line 3.

At this point CTFID is invoked with arguments ( $\mathbf{Y}_* = \{Y_z, X_{w_0}, T_{r_0}, R, Z_{x_0}, r_0\}$ ,  $\mathbf{y}_* = \mathbf{d}_*, \mathbb{Z}, \mathcal{G}$ ). In this case lines 1-3 will not make any change. Note that  $An(D_{\mathbf{pa}_d}) = \{D_{\mathbf{pa}_d}\}$ , hence their union is also an ancestral set, and  $\mathbf{W}_*$  will be equal to the  $\mathbf{Y}_*$  given to CTFID. The c-components in the graph  $\mathcal{G}[\mathbf{V}(\mathbf{W}_*)]$  are  $\{X, R, Y\}$ ,  $\{T\}$  and  $\{Z\}$ , then we need to consider the ctf-factors  $P(Y_z = y_1, X_{w_0} = x_1, R = r_0)$ ,  $P(T_{r_0} = t)$  and  $P(Z_{x_0} = z)$ . In this case all of the ctf-factors are consistent, so we can try to identify each of them from  $\mathbb{Z}$ .

For this we use the IDENTIFY [26] algorithm (Algorithm 3). This algorithm is written in terms of c-factors, usually denoted as  $Q[\mathbf{C}]$ , for some set  $\mathbf{C} \subseteq \mathbf{V}$ . Such a factor is equivalent to  $P_{\mathbf{V} \setminus \mathbf{C}}(\mathbf{C})$ . When calling IDENTIFY with arguments ( $\mathbf{C}, \mathbf{T}, Q[\mathbf{T}] = P_{\mathbf{V} \setminus \mathbf{T}}(\mathbf{T}), \mathcal{G}$ ), the algorithm determines if  $P_{\mathbf{V} \setminus \mathbf{C}}(\mathbf{C})$  can be computed as a function of  $P_{\mathbf{V} \setminus \mathbf{T}}(\mathbf{T})$  in  $\mathcal{G}$ .

The ctf-factor  $P(Y_z = y_1, X_{w_0} = x_1, R = r_0)$  can be obtained as a function of  $P_{\mathbf{V} \setminus \{Y, X, R\}}(Y, X, R)$  evaluated with a set of values consistent with  $\{y_1, z, w_0, t, x_1, r_0\}$ . For  $\mathbf{Z} = \{z\}$ , the corresponding input distribution is  $P(\mathbf{V})$  factorizes according to the c-components of  $\mathcal{G}$  as  $P_T(\mathbf{V} \setminus \{T\})P_{\mathbf{V} \setminus \{T\}}(T)$ . This factorization follows from Lemma 2 in [27]. IDENTIFY will fail to obtain the factor from  $P_T(\mathbf{V} \setminus \{T\})$  and CTFID will try  $\mathbf{Z} = \{W, T\}$ . Now the corresponding input distribution is  $P_{W, T}(\mathbf{V} \setminus \{W, T\}) = P_{W, T, Z}(Y, X, R)P_{W, Y, X, R, T}(Z)$ , where the target factor  $P_{\mathbf{V} \setminus \{Y, X, R\}}(Y, X, R)$  is the first in the factorization.

The ctf-factor  $P(T_{r_0} = t)$  can be evaluated from  $P_{\mathbf{V} \setminus \{T\}}(T)$ , which can be obtained from  $P(\mathbf{V})$ . Finally, the ctf-factor  $P(Z_{x_0} = z)$  is related to  $P_{\mathbf{V} \setminus \{Z\}}(Z)$ , which can be obtained from  $P_{W, T}(\mathbf{V} \setminus \{W, T\})$ .

Specifically, we will obtain

$$P(Y_z = y_1, X_{w_0} = x_1, R = r_0) = P_{w_0, t}(r_0)P_{w_0, t}(x_1 | r_0)P_{w_0, t}(y_1 | r_0, x_1, z), \quad (115)$$

$$P(T_{r_0} = t) = P(t | r_0), \text{ and} \quad (116)$$

$$P(Z_{x_0} = z) = P_{w, t'}(z | r_0, x_0); \quad (117)$$

where  $w$  and  $t'$  could be any values in  $\mathfrak{X}_W$  and  $\mathfrak{X}_T$ , respectively. We could just take  $w_0$  and  $t$  for simplicity.

Then, the expression returned is

$$\sum_t P_{w_0, t}(r_0)P_{w_0, t}(x_1 | r_0)P_{w_0, t}(y_1 | r_0, x_1, z)P(t | r_0)P_{w_0, t}(z | r_0, x_0). \quad (118)$$

Considering again the sum over  $z$  introduced for the unnesting, the final result is:

$$P(Y_{x_1 Z_{x_0}} = y_1 | X_{w_0} = x_1, T_{r_1} = t_1, R = r_0) = \quad (119)$$

$$\sum_{z, t} P_{w_0, t}(r_0)P_{w_0, t}(x_1 | r_0)P_{w_0, t}(y_1 | r_0, x_1, z)P(t | r_0)P_{w_0, t}(z | r_0, x_0). \quad (120)$$

---

**Algorithm 4**  $\text{GID}^*(\mathbf{Y}_*, \mathbf{y}_*, \mathbb{Z}, \mathcal{G})$ 

---

**Input:**  $\mathcal{G}$  a causal diagram,  $(\mathbf{Y}_* = \mathbf{y}_*)$  a counterfactual event; and available distribution specification  $\mathbb{Z}$ .

**Output:**  $P(\mathbf{Y}_* = \mathbf{y}_*)$  in terms of available distributions specified by  $\mathbb{Z}$  or FAIL if not identifiable from  $\langle \mathcal{G}, \mathbb{Z} \rangle$ .

- 1: **if**  $\mathbf{Y}_* = \emptyset$  **then return** 1
  - 2: **if** there exists  $x \in \mathbf{y}_*$  for some  $X_{x'}$  with  $x \neq x'$  **then return** 0
  - 3: **if** there exists  $x \in \mathbf{y}_*$  for some  $X_x$  with  $x = x'$  **then return**  $\text{GID}^*(\mathbf{Y}_* \setminus \{X_x\}, \mathbf{y}_* \setminus \{x\}, \mathbb{Z}, \mathcal{G})$
  - 4:  $(\mathcal{G}', (\mathbf{Y}'_*, \mathbf{y}'_*)) \leftarrow \text{make-cg}(\mathcal{G}, (\mathbf{Y}_*, \mathbf{y}_*))$
  - 5: **if**  $\mathbf{Y}'_*$  is **INCONSISTENT** **then return** 0
  - 6: **if**  $\mathcal{G}'$  has more than one c-component  $\mathbf{S}_1, \dots, \mathbf{S}_k$  **then**
  - 7:     **return**  $\sum_{\mathbf{v}(\mathcal{G}') \setminus \mathbf{y}'_*} \prod_i \text{GID}^*((\mathbf{S}_i)_{\mathbf{v}(\mathcal{G}') \setminus \mathbf{s}_i}, \mathbf{y}'_*(\mathbf{S}_i), \mathbb{Z}, \mathcal{G}[\mathbf{S}_i])$
  - 8: **end if**
  - 9: **if**  $\mathcal{G}'$  has a single c-component  $\mathbf{S}$  **then**
  - 10:     **if**  $(\mathbf{S} = \mathbf{s})$  is inconsistent **then return** FAIL
  - 11:     let  $\mathbf{x} = \bigcup_{S_t \in \mathbf{S}} \mathbf{t}$
  - 12:     **return**  $\text{GID}(\mathbf{s}, \mathbf{x}, \mathbb{Z}, \mathcal{G})$
  - 13: **end if**
- 

## F Relation with Previous Algorithms

In Algorithm 4, we rewrite  $\text{ID}^*$  [22] with the notation used in this paper and modify it for this task. Specifically, instead of returning  $P_{\mathbf{x}}(\bigwedge_{S_t \in \mathbf{S}} S)$  in the last line, we invoke  $\text{GID}$  [13] to try to identify this effect from the available input distributions  $\mathbb{Z}$ .

In Lemma 9 we show that from the failure of  $\text{GID}^*$  we can establish the existence of a ctf-factor  $(\mathbf{C}_* = \mathbf{c}_*)$  corresponding to a c-component of  $\mathcal{G}[\mathbf{V}(An(\mathbf{Y}_*))]$  that is either inconsistent or not identifiable from  $\mathbb{Z}$  and  $\mathcal{G}$ . From this fact, the completeness of  $\text{GID}^*$  follows as a corollary of the completeness of  $\text{CTFID}$ , that fails only under the same condition. Specifically, we can use Theorem 3, Lemma 5, and Lemma 1 as we did for  $\text{CTFID}$  to prove the same result for  $\text{GID}^*$ .

**Lemma 9.** *If  $\text{GID}^*$  fails at line 10 for a counterfactual query  $P(\mathbf{Y}_* = \mathbf{y}_*)$  given  $\mathcal{G}$  and  $\mathbb{Z}$ , then there exists a c-component of  $\mathcal{G}[\mathbf{V}(An(\mathbf{Y}_*))]$  with ctf-factor  $(\mathbf{C}_* = \mathbf{c}_*)$  that is either inconsistent or not identifiable from  $\mathbb{Z}$ .*

*Proof.* Consider the counterfactual graph  $\mathcal{G}'$  produced by  $\text{make-cg}(\mathcal{G}, (\mathbf{Y}_* = \mathbf{y}_*))$ , where  $\text{make-cg}$  is as described in [22]. We will argue that  $\mathbf{V}(\mathcal{G}') = An(\mathbf{Y}_*)$ . First, notice that for every  $W_{\mathbf{z}} \in An(Y_{\mathbf{x}} \in \mathbf{Y}_*)$  there is an equivalent variable  $\mathbf{W}_{\mathbf{x}}$  in the parallel worlds graph, where  $\mathbf{z} = \mathbf{x} \cap An(W)_{\mathcal{G}_{\overline{\mathbf{x}}}}$ . This equivalence follows by rule 3 of do calculus as  $(\mathbf{X} \setminus An(\mathbf{W}) \perp\!\!\!\perp W)$  in  $\mathcal{G}_{\overline{\mathbf{x}}}$ . Next, for the sake of contradiction suppose there exists some  $W_{\mathbf{x}} \in \mathbf{V}(\mathcal{G}')$  not equivalent to any  $An(\mathbf{Y}_*)$ . Since  $\text{make-cg}$  returns an ancestral graph, there must be a path from  $W_{\mathbf{x}}$  to some  $Y_{\mathbf{t}}$  that corresponds to some element of  $Y_{\mathbf{r}} \in \mathbf{Y}_*$  or a node where it was merged. By the same reasoning as before, we can find an equivalent  $An(\mathbf{Y}_*)$  for every variable in such path, which contradicts  $W_{\mathbf{x}}$  not being equivalent to any  $An(\mathbf{Y}_*)$ . We conclude that  $\mathbf{V}(\mathcal{G}') = An(\mathbf{Y}_*)$ .

The expression in line 7 is equal to  $\sum_{\mathbf{w} \setminus \mathbf{y}_*} P(\mathbf{W}_* = \mathbf{w}_*)$  where  $\mathbf{W}_* = \mathbf{V}(\mathcal{G}') = An(\mathbf{Y}_*)$  is a ctf-factor. This is by Theorem 1.

Then, to connect  $P(\mathbf{W}_* = \mathbf{w}_*)$  with the product in this line, notice first that by Theorem 2  $P(\mathbf{W}_* = \mathbf{w}_*) = \prod_i P(\mathbf{C}_{i*} = \mathbf{c}_{i*})$ , where  $(\mathbf{C}_{i*} = \mathbf{c}_{i*})$  are ctf-factors associated with each c-component of  $\mathcal{G}[\mathbf{V}(\mathbf{W}_*)]$  which are the same as the c-components defined based on  $\mathcal{G}'$  as the unobservable variables are shared the same way. For each variable  $S_{\mathbf{t}} \in \mathbf{S}_i$ , line 7 is fixing the values of  $\mathbf{V}(\mathcal{G}') \setminus \mathbf{S}_i$  consistently the values of those variables. If all parents of  $S$  are fixed by  $\mathbf{t}$  and the values  $\mathbf{v}(\mathcal{G}') \setminus \mathbf{s}_i$ , we have that

$$S_{\mathbf{t} \cup \mathbf{v}(\mathcal{G}') \setminus \mathbf{s}_i} = S_{\mathbf{pa}_s} \quad (121)$$

by exclusion restriction, where  $\mathbf{pa}_s$  is consistent with the subscript in the left-hand side.

If there is some  $Z \in \mathbf{Pa}_S$  that has not been fixed, then  $Z_r \in \mathbf{S}_i$  with  $\mathbf{R} = \mathbf{Z} \cap \text{An}(Z)$  and has been fixed as  $Z_{\mathbf{r} \cup \mathbf{v}(\mathcal{G}) \setminus \mathbf{s}_i}$ . Then, we can use composition to fix  $Z$  in the subscript of  $S$  to obtain once again  $S_{\mathbf{pa}_s}$ . Then we can define  $(\mathbf{C}_* = \mathbf{c}_*) = (\bigwedge_{S_t \in \mathbf{S}_i} \mathbf{S}_{\mathbf{pa}_s} = s)$ .

Finally, if  $\text{GID}^*((\mathbf{S}_i)_{\mathbf{v}(\mathcal{G}) \setminus \mathbf{s}_i}, \mathbf{y}'_*(\mathbf{S}_i), \mathbb{Z}, \mathcal{G}[\mathbf{S}_i])$  fails at line 10 in the recursion, it follows that  $(\mathbf{C}_* = \mathbf{c}_*)$  is also inconsistent.  $\square$

**Corollary 2.** *A counterfactual probability  $P(\mathbf{Y}_* = \mathbf{y}_*)$  is identifiable from  $\mathbb{Z}$  and  $\mathcal{G}$  if and only if  $\text{GID}^*$  returns an expression for it.*

*Proof.* This follows from the exact same reasoning as the necessity part of the proof for Theorem 5 as it starts from the fact that there exists a ctf-factor  $P(\mathbf{C}_* = \mathbf{c}_*)$  corresponding to a c-component of  $\mathcal{G}[\mathbf{V}(\text{An}(\mathbf{Y}_*))]$  that is inconsistent or not identifiable from  $\mathbb{Z}$ . By Lemma 9, we have that if  $\text{GID}^*$  fails on  $P(\mathbf{Y}_* = \mathbf{y}_*)$  given  $\mathcal{G}$  and  $\mathbb{Z}$  in line 10 then the ctf-factor is inconsistent. Moreover, if the failure comes from the call to  $\text{GID}$  in line 12, then the same ctf-factor is not identifiable from  $\mathbb{Z}$ .  $\square$