Axioms of Causal Relevance

David Galles and Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
galles@cs.ucla.edu judea@cs.ucla.edu

Abstract

This paper develops axioms and formal semantics for statements of the form "X is causally irrelevant to Y in context Z," which we interpret to mean "Changing X will not affect Y if we hold Z constant." The axiomization of causal irrelevance is contrasted with the axiomization of informational irrelevance, as in "Learning X will not alter our belief in Y, once we know Z." Two versions of causal irrelevance are analyzed, probabilistic and deterministic. We show that, unless stability is assumed, the probabilistic definition yields a very loose structure, that is governed by just two trivial axioms. Under the stability assumption, probabilistic causal irrelevance is isomorphic to path interception in cyclic graphs. Under the deterministic definition, causal irrelevance complies with all of the axioms of path interception in cyclic graphs, with the exception of transitivity. We compare our formalism to that of [Lewis, 1973], and offer a graphical method of proving theorems about causal relevance.

1 Introduction

In [Geiger *et al.*, 1990], a set of axioms was developed for a class of relations called *graphoids*. These axioms characterize informational relevance[1] among observed events based on the semantics of conditional independence in probability calculus. This paper develops a parallel set of axioms for *causal relevance*, that is, the tendency of certain events to affect the occurrence of other events in the physical world, independent of the observer-reasoner. Informational irrelevance is concerned with statements of the form "X is conditionally independent of Y given Z," which means that, given the value of Z, gaining information about X gives us no new information about Y. Causal irrelevance is concerned with statements of the form "X is causally irrelevant to Y given Z," which we take to mean "Changing X will not alter the value of Y, if Z is fixed."

The notion of causal relevance has its roots in the philosophical works of [Good, 1961], [Suppes, 1970] and [Salmon, 1984], who attempted to give probabilistic interpretations to

[1] The term "relevance" will be used primarily as a generic name for the relationship of being relevant or irrelevant. It will be clear from the context when "relevance" is intended to negate "irrelevance."

cause effect relationships, and recognized the need to distinguish causal from statistical relevance. Although these attempts have not produced an algorithmic definition of causal relevance, they led to methods of testing the consistency of relevance statements against a given probability distribution and a given temporal ordering among the variables [Cartwright, 1989, Eells, 1991, Pearl, 1996b]. The current paper aims at axiomatizing relevance statements in themselves, with no reference to underlying probabilities or temporal orderings.

Axiomatic characterization of causal relevance may serve as a normative standard for theories of action as well as a guide for developing representation schemes (e.g., graphical models) for planning and decision-making applications. For example, instead of explicitly storing all possible effects of an action, as in STRIPS [Fikes and Nilsson, 1972], such representation schemes should enable an agent to examine only direct effects of actions, and infer which actions are relevant for a given goal, and which actions cease to be relevant once others are implemented.

An axiomization of causal relevance could also be useful to experimental researchers in domains where exact causal models do not exist. If we know, through experimentation, that some variables have no causal influence on others in a system, we may wish to determine whether other variables will gain such influence, perhaps under different experimental conditions, or may ask what additional experiments could provide such information. For example, suppose we find that a rat's diet has no effect on tumor growth while the amount of exercise is kept constant and, conversely, that exercise has no effect on tumor growth while diet is kept constant. We would like to be able to infer that controlling only diet (while paying no attention to exercise) would still have no influence on tumor growth. A more subtle inference problem is whether changing cage temperature could have an effect on the rat's physical activity, having established that temperature has no effect on activity when diet is kept constant and that temperature has no effect on (the rat's choice of) diet when activity is kept constant.

We provide two formal definitions of causal irrelevance, a probabilistic definition and a deterministic definition. The probabilistic definition, which equates causal irrelevance with inability to change the probability of the effect variable, has intuitive appeal but is inferentially very weak; it does not support a very expressive set of axioms unless further assumptions are made about the underlying causal theory. If we add the stability assumption (i.e., that no irrelevance can be destroyed by changing the nature of the individual processes in the system), then we obtain the same set of axioms for probabilistic causal irrelevance as the one governing path interception in directed graphs. The deterministic definition, which equates causal irrelevance with inability to change the effect variable (in any state of the world), allows for a richer set of axioms without making any assumptions about the causal theory. All of the path interception axioms for directed graphs, with the exception of transitivity, hold for deterministic causal irrelevance.

In Section 2, we define causal theories, a formal model for interpreting causal statements. In Section 3 we provide a definition of probabilistic causal irrelevance, and determine which of the graphoid axioms hold under this definition. Finally, in Section 4, we give a non-probabilistic definition of causal irrelevance, and offer a graphical method of proving statements about causal irrelevance.

2 Causal Theories

A causal theory is a fully specified model of the causal relationships that govern a given domain, namely, a mathematical object that provides an interpretation (and computation) of every causal query about the domain. Following [Pearl, 1995a] we will adopt here a definition that generalizes most causal models used in engineering and economics.

Definition 1 *(Causal Theory) A causal theory is a 4-tuple*

$$T = < V, U, P(u), \{f_i\} >$$

where

(i) $V = \{X_1, \ldots, X_n\}$ *is a set of endogenous variables determined within the system,*

(ii) $U = \{U_1, \ldots, U_m\}$ *is a set of exogenous variables that represent disturbances, abnormalities, assumptions, or boundary conditions,*

(iii) $P(u)$ *is a distribution function over* $U_1, \ldots, U_m$*, and*

(iv) $\{f_i\}$ *is a set of n deterministic, non-trivial functions, each of the form*

$$x_i = f_i(\mathbf{pa}_i, u) \quad i = 1, \ldots, n \tag{1}$$

where $\mathbf{pa}_i$ *are the values of a set of variables* $PA_i \subseteq V \setminus X_i$ *(connoting parents), called the* direct causes *of* X_i*. We will assume that the set of equations in (iv) has a unique solution for* $X_1, \ldots, X_n$*, given any value of the disturbances* $U_1, \ldots, U_m$*. Thus we can consider each variable* $Y \in V$ *to be a function of the disturbances* U *in the causal theory* T*:* $Y = Y_T(u)$*.*

The uniqueness assumption is equivalent to the requirement that $\{f_i\}$ represent a deterministic physical system in equilibrium. Assuming that all relevant boundary conditions U were accounted for, such a system can only be in one state. Systems with feedback, however, can have several equilibrium states. For example, consider the equations $x = y \lor u$ and $y = x \lor u$. The state $U = 0$ permits two possible solutions for X and Y — $(X = 1, Y = 1)$ and $(X = 0, Y = 0)$ — so such functions would be disallowed in a causal theory. Nonuniqueness, however indicates dependency on other factors, not modeled in U. Such factors often can be summarized by the notion of "previous state", and incorporated into our analysis as a third kind of variables supplementing V and U [Galles, 1996a].

Drawing arrows between the variables PA_i and X_i defines a directed graph $G(T)$, which we call the *causal graph* of T. In general, $G(T)$ can be cyclic. Figure 1 illustrates a simple yet typical causal graph. It describes the causal relationships among the season of the year (X_1), whether rain falls (X_2) during the season, whether the sprinkler is on (X_3) during the season, whether the pavement is wet (X_4), and whether the pavement is slippery (X_5). All variables in this figure are binary, taking a value of either "True" or "False," except the root variable X_1 which can take one of four values: "Spring," "Summer," "Fall," or "Winter." Here, the absence of a direct link between X_1 and X_5, for example, captures our understanding that the influence of seasonal variations on the slipperiness of the pavement is mediated by other

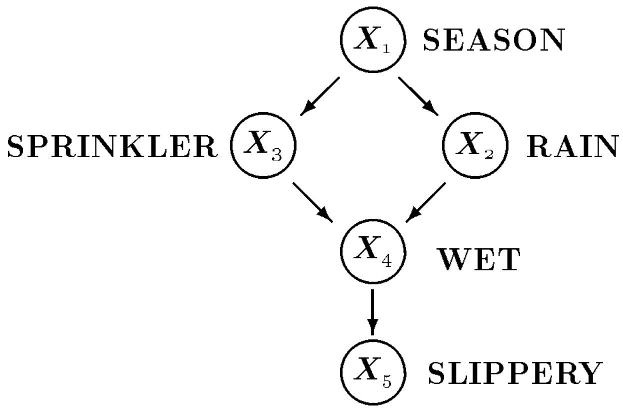

Figure 1: A diagram representing a causal theory on five variables.

conditions (e.g., the wetness of the pavement). The corresponding theory consists of five functions, each representing an autonomous mechanism:

$$\begin{aligned} x_1 &= U_1 \\ x_2 &= f_2(X_1, U_2) \\ x_3 &= f_3(X_1, U_3) \\ x_4 &= f_4(X_3, X_2, U_4) \\ x_5 &= f_5(X_4, U_5) \end{aligned} \tag{2}$$

The disturbances $U_1, \ldots, U_5$ are not shown explicitly in Figure 1, but are understood to govern the uncertainties associated with the causal relationships. A typical specification of the functions $\{f_1, \ldots, f_5\}$ and the disturbance terms is given by the Boolean theory below:

$$\begin{aligned} x_2 &= [(X_1 = \text{Winter}) \vee (X_1 = \text{Fall}) \vee ab_2] \wedge \neg ab'_2 \\ x_3 &= [(X_1 = \text{Summer}) \vee (X_1 = \text{Spring}) \vee ab_3] \wedge \neg ab'_3 \\ x_4 &= (x_2 \vee x_3 \vee ab_4) \wedge \neg ab'_4 \\ x_5 &= (x_4 \vee ab_5) \wedge \neg ab'_5 \end{aligned} \tag{3}$$

where x_i stands for $X_i = true$, and ab_i and ab'_i stand, respectively, for triggering and inhibiting abnormalities. For example, ab_4 stands for (unspecified) events that might cause the ground to get wet (x_4) when the sprinkler is off ($\neg x_2$) and it does not rain ($\neg x_3$), while $\neg ab'_4$ stands for (unspecified) events that will keep the ground dry in spite of the rain, the sprinkler and ab_4, say covering the ground with plastic sheet.

Definition 1 merely provides a description of the mathematical objects that enter into a causal theory. To meet our requirement that a causal theory be capable of computing answers for all causal queries, we need to supplement Definition 1 with an interpretation of the sentence "$X = x$ causes $Y = y$." In ordinary discourse, such a sentence is normally interpreted to mean that we can bring about the condition $Y = y$ by externally enforcing the condition $X = x$. Thus, Definition 1 needs to be supplemented with a formal interpretation of the notion "enforcing $X = x$" that is compatible with its usage in the language.

External intervention normally implies changing some mechanisms in the domain. In a logical circuit, for example, the act of enforcing the condition $X_i = 0$ by connecting some intermediate variable X_i to ground amounts to changing the mechanism that normally determines X_i. If X_i is the output of an OR gate, then after the intervention X_i would no longer be determined by the OR gate but by a new mechanism (involving the ground) which clamps X_i to 0 regardless of the input to the OR gate. In the equational representation, this amounts to replacing the equation $X_i = f_i(\mathbf{pa}_i, u)$ with a new equation, $X_i = 0$, that represents the grounding of X_i.

The replacement of just one equation, not several, reflects the principle of locality in the common understanding of imperative sentences such as: "Raise taxes" or "Make him laugh." When told to clean his face, a child does not ask for a razor, nor does he jump into the swimming pool. The proper interpretation of the modal sentence "*do p*" corresponds to a a minimal perturbation of the existing state of affairs, and this, in the context of Definition 1, corresponds to the replacement of a minimal set of equations necessary to make p compatible with U.

In general, we will consider concurrent action of the form $do(X = x)$, where X involves several variables in V.[2] This leads to the following definition:

Definition 2 (Effect of Actions) *The effect of the action $do(X = x)$ on a causal theory T is given by a subtheory T_x of T, where T_x is obtained by deleting from T all equations corresponding to variables in X and substituting the equations $X = x$ instead.*

The syntactical transformation described in Definition 2 corresponds to replacing the old functional mechanisms $x_i = f_i(PA_i, u)$ with new mechanisms $X_i = x_i$ that represent the external forces that set the values x_i for each $X_i \in X$. As before, we will assume each variable $Y \in V$ to be a unique function of the disturbances U in any theory T_x: $Y = Y_{T_x}(u)$. For brevity, the subscript T is often omitted, leaving $Y_x(u)$.

The assumption that there is a unique solution for $X_1, \ldots, X_n$ imposes some restrictions on the functions f_i. However, the equations do not need to be recursive to ensure uniqueness. For example, the causal theory given by Figure 2 dictates unique values for X and Y for $U_1 = 0$ and $U_1 = 1$. The subtheories of T also dictate unique solutions; there is a unique value for Y (for both values of U_1) in $T_{X=0}$ and $T_{X=1}$, and a unique value for X (for both values of U_1) in $T_{Y=0}$ and $T_{Y=1}$.

$V = \{X, Y\}$ binary
$U = \{U_1\}$ binary

$x = u_1 \ \vee \ \neg y$
$y = u_1 \ \& \ x$
$P(u_1) = 0.5$

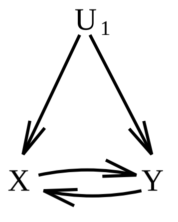

Figure 2: A valid nonrecursive causal theory, with unique values for X and Y for all values of U.

[2] The formalization of conditional actions of the form "$do(X = x)$ if $Z = z$" is straightforward [Pearl, 1994].

Returning to the example of Figure 1, represent the action "turning the sprinkler ON," or $do(X_3 = \text{ON})$, we delete the equation $X_3 = f_3(X_1, U_3)$ from the theory of Eq. (2) and replace it with $X_3 = \text{ON}$. The resulting subtheory, $T_{X_3=\text{ON}}$, contains all the information needed for computing the effect of the action on other variables. It is easy to see from this subtheory that the only variables affected by the action are X_4 and X_5, that is, the descendants of the manipulated variable X_3. Note, however, that the operation $do(X_3 = \text{ON})$ stands in marked contrast to that of *finding* the sprinkler ON; the latter involves making the substitution *without* removing the equation for X_3, and therefore may potentially influence (the belief in) every variable in the network. This mirrors indeed the difference between seeing and doing: after observing that the sprinkler is ON, we wish to infer that the season is dry, that it probably did not rain, and so on; no such inferences should be drawn in evaluating the effects of the contemplated action "turning the sprinkler ON."

The notation $Y_x(u)$ is sometimes used in the statistical literature [Rubin, 1974] to stand for the counterfactual sentence "The value that Y would take in person u, had X been x," where X stands for a type of treatment that a person can receive. There is a strong connection between the the sentence above and our interpretation of $Y_x(u)$ [Pearl, 1995a]. Definition 2 interprets the abstract, counterfactual sentence above in terms of the processes responsible for Y taking on the value $Y_x(u)$ as X changes to x. It treats u not merely as an index of an individual but, rather, as the set of attributes u that characterize the individual, the experimental conditions under study, and so on. In Section 4, we will show that the process-based semantics given in Definition 2 will uncover new properties of $Y_x(u)$ that were not formalized in the statistical literature.

An explicit translation of intervention into "wiping out" equations from the causal model was first proposed in [Strotz and Wold, 1960], and used in [Fisher, 1970] and [Sobel, 1990]. Graphical ramifications were explicated in [Spirtes *et al.*, 1993] and [Pearl, 1993]. Interpretations of causal and counterfactual utterances in terms of $Y_x(u)$ are given in [Pearl, 1996a].Other formulations of causality, in terms of event trees are given in [Robins, 1987] and [Shafer, 1996].

Note that $Y_x(u)$ is well defined even when $U = u$ and $X = x$ are incompatible in T, thus allowing for actions to enforce propositions that are not realized under normal conditions. For example, if T describes a logic circuit we might wish to intervene and set some voltage X to x, even though the input dictates $X \neq x$. It is for this reason that one must invoke some notion of mechanism breakdown or "surgery" in the definition of interventions.

The unique feature of our formulation of actions, which sets it apart from the formulations in control theory or decision analysis [Savage, 1954, Heckerman and Shachter, 1995], is that an action is treated as a *modality*, namely, it is not given an explicit name but acquires the names of the propositions that it enforces as true. This enables the model to predict the effect of a huge number of action combinations without the modeler having to attend to such combinations. Instead, the causal theory is constructed by specifying the characteristics of each individual mechanism under normal conditions, free of intervention. Likewise, the distribution $P(u)$ need only characterize normal fluctuations in boundary conditions, excluding abnormal eventualities such as interventions.

3 Probabilistic Causal Irrelevance

The fact that each endogenous variable is a function of U and that T specifies a probability distribution over U defines a probability distribution over the endogenous variables. That is, for every set of variables $Y \subseteq V$, we have

$$P(y) = \sum_{\{u \mid Y(u)=y\}} P(u) \tag{4}$$

The probability induced by the action $do(X = x)$ is defined in the same manner, through the function $Y_x(u)$ induced by the subtheory T_x. Using $\hat{x}$ to abbreviate $do(X = x)$, we obtain

$$P(y|\hat{x}) \doteq P(y|do(X = x)) = \sum_{\{u \mid Y_x(u)=y\}} P(u) \tag{5}$$

The existence of a probability distribution over all variables leads to a natural definition of the probabilistic version of causal irrelevance.

Definition 3 (Probabilistic Causal Irrelevance) *X is probabilistically causally irrelevant to Y, given Z, written $CI_P(X, Z, Y)$, iff*

$$\forall x, x', y, z \quad P(y|\hat{z}, \hat{x}) = P(y|\hat{z}, \hat{x}') \tag{6}$$

Read: Once we hold Z fixed (at z), changing X will not affect the probability of Y.

3.1 Comparison to Informational Relevance

If we remove the "hats" from Definition 3 above, we get the standard definition of conditional independence in probability calculus, denoted $I(X, Z, Y)$, which is governed by the graphoid axioms [Geiger *et al.*, 1990] given in Figure 3

1.1 (Symmetry) $I(X, Z, Y) \Longrightarrow I(Y, Z, X)$

1.2 (Decomposition) $I(X, Z, YW) \Longrightarrow I(X, Z, Y)$

1.3 (Weak union) $I(X, Z, YW) \Longrightarrow I(X, ZW, Y)$

1.4 (Contraction) $I(X, Z, Y)$ & $I(X, ZY, W) \Longrightarrow I(X, Z, YW)$

1.5 (Intersection) $I(X, ZY, W)$ & $I(X, ZW, Y) \Longrightarrow I(X, Z, YW)$

Intersection requires a strictly positive probability distribution.

Figure 3: The graphoid axioms.

These axioms, a special form of which was introduced in [Dawid, 1979] and [Spohn, 1980], were rediscovered by [Pearl and Paz, 1987] who conjectured them to be complete. The conjecture has been refuted by [Studeny, 1990], who also proved that conditional independence in probability theory has no finite axiomatization. Nevertheless, the graphoid axioms capture the most important features of informational relevance, "Learning irrelevant information should not alter the relevance status of other propositions in the system; what was relevant remains relevant, and what was irrelevant remains irrelevant." [Pearl, 1988]

One of the most salient difference between informational and causal relevance is the property of symmetry, axiom 1.1. Informational relevance is symmetric, stating that if X is relevant to Y, then Y is relevant to X as well. For example, learning whether the sprinkler is on provides information on whether the grass is wet and, vice versa, learning whether the grass is wet provides information on whether the sprinkler is on. This property is clearly violated in causal theories: turning a sprinkler on tends to make the grass wet, so turning on the sprinkler gives us information about the state of the grass. Conversely, wetting the grass has no physical effect on the state of the sprinkler, and gives us no information about whether the sprinkler was on or off.

Another basic difference between informational and causal relevance is that in the former the rule of hypothetical middle [Pearl, 1988, p. 17] always holds:

$$\mathrm{MIN}_x P(y|x) \leq P(y) \leq \mathrm{MAX}_x P(y|x) \tag{7}$$

In causal relevance, $P(y)$ might greater than $\mathrm{MAX}_x P(y|\hat{x})$, or less than $\mathrm{MIN}_x P(y|\hat{x})$. Figure 4 illustrates such a possibility.

$V = \{X, Y\}$ binary
$U = \{U_1\}$ binary

$$x = u_1$$
$$y = \begin{cases} 1 & \text{if } x = u_1 \\ 0 & \text{otherwise} \end{cases}$$
$$P(u_1) = 0.5$$

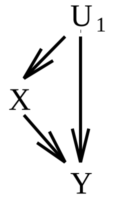

Figure 4: An example of $P(y) > \mathrm{MAX}_x P(y|\hat{x})$.

In Figure 4, there are two endogenous variables X and Y, as well as an exogenous variable U_1. Without any intervention, X will always have the same value as U_1, hence, Y will have the value 1. If X and U_1 have different values, then Y will have the value 0. If we intervene and set X to 1, then Y will have the value 1 when $U_1 = 1$, which has a probability 0.5, and Y will have the value 0 when $U_1 = 0$, which has a probability 0.5: $P(Y = 0|set(X = 1)) = P(Y = 1|set(X = 1)) = 0.5$. Similarly, we can see that $P(Y = 0|set(X = 0)) = P(Y = 0|set(X = 1)) = 0.5$. Thus, $MAX_x P(y|\hat{x}) = 0.5$, and $P(Y = 1) = 1 > 0.5 = \mathrm{MAX}_x P(y|\hat{x})$.

Note that, in view of the violation of the rule of the hypothetical middle (Eq. (7)), Definition 3 is not equivalent to

$$\forall x, y, z \quad P(y|\hat{z}, \hat{x}) = P(y|\hat{z}) \tag{8}$$

Read: Once we hold Z fixed (at z), controlling X will not affect the probability of Y. In fact, Definition 3 is stronger than Eq. (8), since statement 2.5.2 (left-intersection of Theorem 1 below) follows from the former and not from the latter.

The notion of probabilistic causal irrelevance may bring to mind a related concept of *ignorability* [Rosenbaum and Rubin, 1983] which is extremely important in analyzing the effectiveness of treatments (e.g., drugs, diet, educational programs) from uncontrolled studies. The two concepts are however different. Ignorability allows us to ignore HOW X obtained its value x, while irrelevance allows us to ignore which value X actually obtained. Ignorability is defined as the condition

$$P(Y_x = y|z) = P(Y = y|z, x) \tag{9}$$

which in our notation reads:

$$P(y|z, \hat{x}) = P(y|z, x) \tag{10}$$

It allows an investigator to relate the response Y_x to observable conditional probabilities. A central question in experimental design is to select a set of observables Z that would make Eq. 9 true, given causal knowledge of the domain. Ignorability in itself does not provide such a criterion; though it states the problem in formal counterfactual language: "Z can be selected if, for every x, the value that Y would obtain had X been x is conditionally independent of X, given Z." A criterion for selecting Z can be obtained from the graph $G(T)$ underlying a causal theory, as given by the "back-door criterion" in [Pearl, 1995a].

The question we attempt to answer in this section is whether the relation of causal irrelevance, $CI_P(\cdot)$, is governed by a set of axioms similar to those governing informational irrelevance $I(\cdot)$. An extreme way of motivating this question would be to ask whether there are any constraints that prohibit the assignment of arbitrary functions $P(y|\hat{x})$ to any pair (X, Y) of variable sets in V, in total disregard of the fact that $P(y|\hat{x})$ represents the probability of $(Y = y)$ induced by physically setting X to x in some causal theory T. Our finding indicate that, although the assignment $P(y|\hat{x})$ is not totally arbitrary, it is only weakly constrained by axioms of causal irrelevance.

3.2 Axioms of Probabilistic Causal Irrelevance

We have found only two axioms that constrain causal irrelevance.

Theorem 1 *For any causal theory, the following two properties must hold :*

2.2.1 (Right-Decomposition) $CI_P(X, Z, YW) \Longrightarrow CI_P(X, Z, Y) \,\&\, CI_P(X, Z, W)$

2.5.2 (Left-Intersection) $CI_P(X, ZW, Y) \,\&\, CI_P(W, ZX, Y) \Longrightarrow CI_P(XW, Z, Y)$

Property 2.2.1 reads: If changing X has no effect on Y and W considered jointly, then it has no effect on either Y or W considered separately. This follows trivially from the fact that $P(\cdot)$ is a probability function, but it does not reflect any quality of causation.

Property 2.5.2 reads: If changing X cannot affect $P(y)$ when W is fixed, and changing W cannot affect $P(y)$ when X is fixed, then changing X and W together cannot affect $P(y)$.

Many seemingly intuitive properties, however, do not hold. For instance, none of the following sentences hold for all causal theories.

2.2.2 (Left-Decomposition-1) $CI_P(XW, Z, Y) \Longrightarrow CI_P(X, Z, Y) \vee CI_P(W, Z, Y)$

2.2.3 (Left-Decomposition-2) $CI_P(XW, Z, Y) \Longrightarrow CI_P(X, Z, Y) \vee CI_P(X, Z, W)$

2.2.4 (Left-Decomposition-3) $CI_P(XW, Z, Y)$ & $CI_P(XY, Z, W) \Longrightarrow CI_P(X, Z, Y) \vee CI_P(X, Z, W)$

2.3 (Weak Union) $CI_P(X, Z, WY) \Longrightarrow CI_P(X, ZW, Y)$

2.4 (Contraction) $CI_P(X, Z, Y)$ & $CI_P(X, ZY, W) \Longrightarrow CI_P(X, Z, WY)$

2.5.1 (Right-Intersection) $CI_P(X, ZW, Y)$ & $CI_P(X, ZY, W) \Longrightarrow CI_P(X, Z, WY)$

2.6 (Transitivity) $CI_P(X, Z, Y) \Longrightarrow CI_P(a, Z, Y) \vee CI_P(X, Z, a) \quad \forall a \notin X \cup Z \cup Y$

The sentences above were tailored after the graphoid axioms (Figure 3) with the provision that symmetry does not hold, thus requiring left and right versions. Many of these sentences have intuitive appeal and yet are not sound relative to the semantics of $P(y|\hat{x})$. For example; property 2.2.2 states that if changing X has an effect on Y, and changing W has an effect on Y, then changing X and W simultaneously should also affect Y. It is hard to come up with a simple real-life example that refutes this assertion. Still, as will be shown in the Section 3.4 and in Appendix A, each of these sentences is refuted by some specific causal theory.

3.3 Proofs of Axioms of Probabilistic Causal Irrelevance

We now prove the two sentences of Theorem 1.

2.2.1 $CI_P(X, Z, YW) \Longrightarrow CI_P(X, Z, Y)$&$CI_P(X, Z, W)$ holds trivially. $CI_P(X, Z, YW) \Longrightarrow P(yw|\hat{z}, \hat{x}) = P(yw|\hat{z}, \hat{x}')$. We can sum over W to get $P(y|\hat{z}, \hat{x}) = P(y|\hat{z}, \hat{x}')$, which implies $CI_P(X, Z, Y)$. □

2.5.2 (By contradiction) Assume $CI_P(X, ZW, Y)$ & $CI_P(W, ZX, Y)$ & $\neg CI_P(XW, Z, Y)$. Since $\neg CI_P(XW, Z, Y)$, by definition of $CI_P(\cdot)$, $\exists y, x, x', w, w', z \quad P(y|\hat{z}, \hat{w}, \hat{x}) \neq P(y|\hat{z}, \hat{w}', \hat{x}')$. However, $CI_P(X, ZW, Y)$ implies $\forall y, x, x', z, w \quad P(y|\hat{z}, \hat{x}, \hat{w}) = P(y|\hat{z}, \hat{x}', \hat{w})$. Furthermore, $CI_P(W, ZX, Y)$ implies $\forall y, x', w, w', z \quad P(y|\hat{z}, \hat{x}', \hat{w}) = P(y|\hat{z}, \hat{x}', \hat{w}')$, so $\forall x, x', w, w', z \quad P(y|\hat{z}, \hat{x}, \hat{w}) = P(y|\hat{z}, \hat{x}', \hat{w}) = P(y|\hat{z}, \hat{x}', \hat{w}')$. Thus $\forall x, x', w, w', z \quad P(y|\hat{z}, \hat{x}, \hat{w}) = P(y|\hat{z}, \hat{x}', \hat{w}')$, which contradicts $\exists x, x', w, w', z \quad P(y|\hat{z}, \hat{x}, \hat{w}) \neq P(y|\hat{z}, \hat{x}', \hat{w}')$. □

3.4 Counterexample to Property 2.2.2

We now disprove property 2.2.2 by counterexample. This counterexample is not necessarily meant to model a common, real-life situation. Rather, it disproves the claim that *all possible* causal theories must conform to the property.

2.2.2 $CI_P(XW, Z, Y) \Longrightarrow CI_P(X, Z, Y) \vee CI_P(W, Z, Y)$.

Figure 5 shows a counterexample to this sentence. In this theory, $CI_P(XW, \emptyset, Y)$ & $\neg CI_P(X, \emptyset, Y)$ & $\neg CI_P(W, \emptyset, Y)$

$V = \{X, W, Y\}$ binary
$U = \{U_1\}$ binary

$x = u_1$
$w = u_1$
$y = Parity(x, w, u_1)$
$P(u_1) = 0.5$

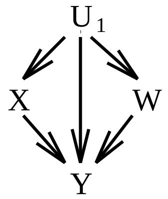

Figure 5: Counterexample to property 2.2.2.

This counterexamples is more clear when we consider the contrapositive form of the claim. In this example, changing W can affect the probability of Y, and changing X can affect the probability of Y, but changing W and X simultaneously has no effect on the probability of Y. This is extremely counterintuitive, if tweaking X has an effect on Y, and tweaking W has an effect on Y, we would expect the more flexible option of changing X and W simultaneously to also affect Y.

The key to this counterexample is the fact that setting W removes the connection between W and U_1. When we intervene on only X, W takes on the same value as U_1, and Y will always have the value of X. When we intervene on both X and W, there is no longer any connection between U_1 and W. Thus, the probability that W and U_1 will have the same value is 0.5, and $P(y) = 0.5$

Counterexamples to the other properties (2.2.3, 2.2.4, 2.3, 2.4, 2.5.1, 2.6) are in Appendix A.

3.5 Numeric Constraints

Although Definition 3 imposes only weak constraints (axiom 2.2.1 and 2.5.2) on the structure of probabilistic causal irrelevance, the probability assignments $P(y|\hat{x})$, which describe the effects of actions in the domain, are constrained nevertheless by non-trivial numerical bounds. For instance, the inequality

$$(y|\hat{x}, \hat{z}) \geq P(y, z|\hat{x}) \tag{11}$$

must hold in any causal theory. This can easily be shown by the definition of $P(y, z|\hat{x})$ and $P(y|\hat{x}, \hat{z})$. Recall from Eq. (5) that

$$P(y, z|\hat{x}) = \sum_{\{u \ | \ Y_x(u)=y \ \& \ Z_x(u)=z\}} P(u)$$

and

$$P(y|\hat{x}, \hat{z}) = \sum_{\{u \ | \ Y_{xz}(u)=y\}} P(u)$$

Consider U^{yz}, the set of all values u of U such that $Y_x(u) = Y$ and $Z_x(u) = z$, and U_z^y, the set of all values u' of U such that $Y_{xz}(u) = y$. Since all values u of U^{yz} already constrain Z

to have the value z, fixing Z at z will not affect the value of Y. Thus, for all values u of U^{yz}, $Y_{xz}(u) = y$. Hence, $U_z^y \supseteq U^{yz}$ and $P(y|\hat{x}, \hat{z}) \geq P(yz|\hat{x})$. This can be shown more formally using Theorem 6, which is proven below, in Section 4.2. Additional constraints are explored in [Pearl, 1995b].

3.6 Axioms of Causal Irrelevance for Stable Theories

The set of axioms we obtained for causal irrelevance was much smaller than we would expect from our intuition of causal effect relations. We have two explanations for this discrepancy. One possibility is that probabilistic causal irrelevance does not capture our intuition of causal mechanisms. This possibility will be explored in Section 4, which gives a deterministic definition of causal irrelevance and yields a more complete set of axioms. The other possibility is that the type of examples exploited in Section 3.4 and Appendix A are not commonly observed in everyday life. This section explores what assumptions need to be made for probabilistic causal irrelevance to have a more expressive set of axioms.

A more expressive set of causal irrelevance axioms is obtained if we confine the analysis to *stable* causal theories, that is, causal theories whose irrelevances are implied by the structure of the causal theory, and, hence, remain invariant to changes in the forms of each individual functions f_i. We will define stability through the concept of a *replacement class*. A replacement class τ is a set of all theories that have the same variables V and U, and the same functional arguments. In other words, the functions are allowed to change between members of τ, but the arguments of these functions are not allowed to vary. Formally, for any two theories $T_1, T_2 \in \tau$ and any two functions $f_i(PA_i) \in T_1$ and $f'_i(PA'_i) \in T_2$, $PA_i = PA'_i$. The class $\tau(T)$ represents the replacement class that contains the theory T.

We now define stability using replacement classes, similar to [Pearl and Verma, 1991a][3].

Definition 4 (Stability) *Let T be a causal theory. An irrelevance $CI_P(X, Z, Y)$ in T is stable if it is shared by all theories in $\tau(T)$. The theory T is stable if all of the irrelevances in T are stable.*

Stability requires irrelevance to be determined by the structure of the equations, not merely by the parameters of the functions. Thus, a causal theory is not stable if we can remove an irrelevance relationship by replacing an equation or set of equations to obtain a new theory with fewer irrelevance statements. In each of the examples in Section 3.4 and Appendix A, for instance, a minor change in the form of one of the equations would destroy an irrelevance. Note that none of the theories presented in Figure 5 or the appendix is stable.

There are, however, many stable causal theories. All monotonic linear systems, for example, are stable. One might think that any causal theory that contained only additive, monotonic functions f_i would be stable. The causal theory of Figure 16, however, refutes that conjecture.

Definition 5 (Path-interception) *Let $int(X, Z, Y)_G$ stand for the statement "every directed path from X to Y in graph G contains at least one element in Z"*

[3]The probabilistic notion of stability (also called "DAG-isomorphism," "nondegeneracy" [Pearl, 1988, p. 391], and "faithfulness" [Spirtes *et al.*, 1993]) was used by Pearl and Verma [1991] to emphasize the invariance of certain independencies to functional form.

Theorem 2 *If a causal theory T is stable, then X is probabilistically causally irrelevant to Y, given Z, in T iff Z intercepts all directed paths from X to Y in the graph $G(T)$ defined by T. That is,*

$$CI_P(X,Z,Y) \Longleftrightarrow int(X,Z,Y)_{G(T)}$$

Proof:

(i) $\quad CI_P(X,Y,Z) \Longrightarrow int(X,Y,Z)_{G(T)}$

Assume that there exists a stable causal theory T that induces a probabilistic causal irrelevance relation $CI_P(\cdot)$, and assume that, for some sets of variables X, Y, Z, $CI_P(X,Z,Y)$ and $\neg int(X,Z,Y)_{G(T)}$. Since there is a directed path from X to Y that is not intercepted by Z in $G(T)$, we can easily construct a theory T' such that $G(T') = G(T)$ and $\neg CI_P(X,Z,Y)$ in T'. We can do this by changing all of the functions that lie on the path from X to Y to disjunctions and then modifying the other functions to ensure that $P(y|\hat{z}) < 1$. Thus, if we force X to have the value 1, Y will also have the value 1, and $P(y|\hat{z},\hat{x}) \neq P(y|\hat{z})$. By assumption, $CI_P(X,Z,Y)$, so $CI(T) \not\subseteq CT(T')$. Thus, T is not a stable causal theory, a contradiction.

(ii) $\quad int(X,Z,Y)_{G(T)} \Longrightarrow CI_P(X,Z,Y)$

We will use the following lemma:

Lemma 1 *For any structural equation f_Y in a causal theory T, if a series of functional substitutions results in a new function g_Y such that X is an argument of g_Y, then there must be a directed path from X to Y in $G(T)$.*

We will prove this lemma by induction on the number of functional substitutions.

Base Case: If we make no substitutions into f_Y, then every argument X of f_Y must be a parent of Y in $G(T)$, by our definition of $G(T)$. Thus, there is a directed path from each argument of f_Y to Y in $G(T)$

Inductive Case: Assume that $n-1$ functional substitutions into f_Y always result in the new function g_Y such that for each argument X of g_Y, there is a directed path from X to Y in $G(T)$. We use this assumption to prove that after n substitutions resulting in g'_Y, there is a directed path from every argument of g'_Y to Y in $G(T)$, as follows: When we do a single substitution, we replace a variable with a function of its parents in $G(T)$. So, for any new argument X' that is introduced to g'_Y by substituting in for X, X' must be a parent of X in $G(T)$. By the inductive hypothesis, there must be a directed path from X to Y in $G(T)$. Thus, there must be a directed path from X' to Y in $G(T)$.

We can now prove the implication $int(X,Z,Y)_{G(T)} \Longrightarrow CI_P(X,Z,Y)$. We will consider f_Y, the functional equation for Y in T_z. After we do a functional substitution for all variables in f_Y except for X and Z, we are left with a new function g_Y. By Lemma 1, since there is no directed path from X to Y in $G(T_z)$, X is not an argument of g_Y, so g_Y is a function of only Z and U. Since g_Y is a function of only Z and U, and not of X, $Y_{xz}(u) = Y_z(u)$, so $P(y|\hat{x},\hat{z}) = P(y|\hat{z})$, and $CI_P(X,Z,Y)$. $\square$

Since $CI_P(X,Y,Z) \Longleftrightarrow int(X,Y,Z)_{G(T)}$ in stable causal theories, probabilistic causal irrelevance is completely characterized by the axioms of path interception in directed graphs. A complete set of such axioms was developed in [Paz and Pearl, 1994, Paz *et al.*, 1996] and is given in Figure 6.

3.2.1 (Right-Decomposition) $int(X, Z, YW)_G \Longrightarrow int(X, Z, Y)_G \;\&\; int(X, Z, W)_G$

3.2.2 (Left-Decomposition) $int(XW, Z, Y)_G \Longrightarrow int(X, Z, Y)_G \;\&\; int(W, Z, Y)_G$

3.4 (Strong Union) $int(X, Z, Y)_G \Longrightarrow int(X, ZW, Y)_G \quad \forall\, W$

3.5.1 (Right-Intersection) $int(X, ZW, Y)_G \;\&\; int(X, ZY, W)_G \Longrightarrow int(X, Z, YW)_G$

3.5.2 (Left-Intersection) $int(X, ZW, Y)_G \;\&\; int(W, ZX, Y)_G \Longrightarrow int(XW, Z, Y)_G$

3.6 (Transitivity) $int(X, Z, Y)_G \Longrightarrow int(a, Z, Y)_G \vee int(X, Z, a)_G \quad \forall a \notin X \cup Z \cup Y$

Figure 6: Sound and complete axioms for path interception in directed graphs.

4 Causal Irrelevance

The notion of causal irrelevance obtains a deterministic definition when we consider the effects of an action conditioned on a specific state of the world u.

Definition 6 (Causal Irrelevance) *X is causally irrelevant to Y, given Z, in a causal theory T, $CI_T(X, Z, Y)$, if*

$$\forall u, z, x, x' \;\; Y_{xz}(u) = Y_{x'z}(u) \tag{12}$$

in every subtheory of T_z.

Note that unlike the probabilistic definition of causal irrelevance (see Eq. (8)), this definition is equivalent to

$$\forall u, z, x \;\; Y_{xz}(u) = Y_z(u) \tag{13}$$

This definition captures the intuition "If X is causally irrelevant to Y, then X cannot affect Y under any circumstance." It is stronger than the probabilistic definition, in that $CI_T(X, Z, Y) \Longrightarrow CI_P(X, Z, Y)$.

This definition of irrelevance bears some similarity to the idea of limited unresponsiveness presented in [Heckerman and Shachter, 1995]. However, whereas Heckerman and Shacter define causality in terms of limited unresponsiveness to a specific set of actions, we view irrelevance as a property of a causal theory. In fact, a version of their definition of causality, translated into our language, will be shown to be a theorem of causal irrelevance in Section 4.6.2 (see Eq. (18)).

To see why we require the equality $Y_{xz}(u) = Y_{x'z}(u)$ to hold in every subtheory of T_z, consider the causal theory of Figure 7. In this example, Z follows X and, hence, Y follows X, that is, $Y_{X=0}(u) = Y_{X=1}(u) = u_2$. However, since f_y is a nontrivial function of X, X is perceived to be causally relevant to Y. Only holding Z constant would reveal the causal influence of X on Y. To capture this intuition, we must therefore consider all subtheories in Definition 6.

$V = \{X, Z, Y\}$ binary
$U = \{U_1, U_2\}$ binary

$$x = u_1$$
$$y = \begin{cases} u_2 & \text{if } x = z \\ x & \text{otherwise} \end{cases}$$
$$z = x$$
$$P(u_1) = P(u_2) = 0.5$$

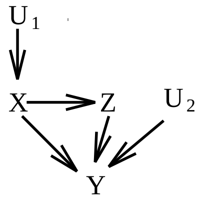

Figure 7: Example of a causal theory that requires the examination of subtheories to determine causal relevance.

4.1 Causal Irrelevance Axioms

With this definition of Causal Irrelevance, we have the following theorem:

Theorem 3 *For any causal theory, the following sentences must hold:*

4.2.1 (Right-Decomposition) $CI_T(X, Z, YW) \Longrightarrow CI_T(X, Z, Y)$ & $CI_T(X, Z, W)$

4.2.2 (Left-Decomposition) $CI_T(XW, Z, Y) \Longrightarrow CI_T(X, Z, Y)$ & $CI_T(W, Z, Y)$

4.4 (Strong Union) $CI_T(X, Z, Y) \Longrightarrow CI_T(X, ZW, Y) \quad \forall\, W$

4.5.1 (Right-Intersection) $CI_T(X, ZW, Y)$ & $CI_T(X, ZY, W) \Longrightarrow CI_T(X, Z, YW)$

4.5.2 (Left-Intersection) $CI_T(X, ZW, Y)$ & $CI_T(W, ZX, Y) \Longrightarrow CI_T(XW, Z, Y)$

The following sentence, however, **does not** hold in every causal theory:

4.6 (Transitivity) $CI_T(X, Z, Y) \Longrightarrow CI_T(a, Z, Y) \vee CI_T(X, Z, a) \quad \forall a \notin X \cup Z \cup Y$

4.2 Theorems of Causal Statements

To prove the causal irrelevance axioms, we will use some of the following theorems and definitions.

Definition 7 (Null Action) *For any variable* X, $X_\emptyset(u) = X(u)$.

Definition 7 provides an interpretation for a null subscript, which will be needed in the proofs below.

Theorem 4 (Degeneracy) *For all variables* X *and* W, $X_{xw}(u) = x$.

Proof:

This theorem follows from Definition 1, where $Y_x(u)$ is interpreted as the unique solution for Y of a set of equations under $X = x$. It is included for completeness. □

Theorem 5 (Uniqueness) *For any variable X, set of variables Z, and value z, there exists a unique value x of X such that $X_z(u) = x$.* $\Box$

Proof:

This theorem follows directly from the definition of causal theories, which required a unique value for any variable X, given any value of the disturbances U, in any subtheory. $\Box$

Theorem 6 (Composition) *For any variables Y and W, and set of variables XW in a causal theory,*

$$W_x(u) = w \Longrightarrow Y_x(u) = Y_{xw}(u) \tag{14}$$

Proof:

Since $Y_x(u)$ has a unique solution, forming T_x and substituting out all other variables would yield a unique solution for Y, regardless of the order of substitution. So, we will form T_x and examine the structural equation for Y in T_x : $Y_x = f_Y(x, z, w, u)$, where Z stands for the rest of the parent set of Y. We now solve for Z by substituting out all variables except X,Y, and W. That is, we substitute out all variables in T_x, avoiding substitutions into X, W and Y, and express Z as a function of x, w, and u. We then plug this solution into f_Y to get $Y_x = f_Y(x, w, Z(x, w, u), u)$, which we can write as $Y_x = f(x, w, u)$. At this point, we can solve for W, substituting out all variables in T_X other than X, which leaves $Y_x = f(x, W(u, x), u)$. We can now see that if $w = W_x(u)$, then $Y_x(u) = Y_{xw}(u)$. $\Box$

This proof is still valid in cases where $X = \emptyset$.

Corollary 1 (Consistency) *For any variables Y and X in a causal theory,*

$$X(u) = x \Longrightarrow Y(u) = Y_x(u) \tag{15}$$

Proof:

Corollary 15 follows directly from Composition. Substituting X for W and $\emptyset$ for X in Eq. (14), we obtain $X_\emptyset(u) = x \Longrightarrow Y_\emptyset(u) = Y_x(u)$. Null Action allows us to drop the $\emptyset$, leaving $X(u) = x \Longrightarrow Y(u) = Y_x(u)$. $\Box$

The implication in Eq. (15) was called *Consistency* by [Robins, 1987]. [4]

Theorem 7 (Reversibility) *For any variables X and W, and set of variables X,*

$$(Y_{xw}(u) = y) \ \&\ (W_{xy}(u) = w) \Longrightarrow Y_x(u) = y \tag{16}$$

Proof:

Reversibility follows from the assumption that the solution for V in every subtheory is unique. Since $Y_x(u)$ has a unique solution, forming T_x and substituting out all other variables would yield a unique solution for Y, regardless of the order of substitution. So, we will form T_x and examine the structural equation for Y in T_x, which might in general be a function of

[4] This property and Composition were tacitly used in economics [Manski, 1990] and statistics within the so-called Rubin's model [Rubin, 1974]. To the best of our knowledge, Robins was the first to state Consistency formally and to use it to derive other properties of counterfactuals.

X,W,U, and additional variables : $Y_x = f_Y(x, w, z, u)$, where Z stands for parents of Y not contained in $X \cup W \cup U$. We now solve for Z by substituting out all variables except X,Y, and W. That is, we substitute out all variables in T_x, avoiding substitutions into X, W and Y, and express Z as a function of x, w, and u. We then plug this solution into f_Y to get $Y_x = f_Y(x, w, Z(x, w, u), u)$, which we can write as $Y_x = f(x, w, u)$. We now consider what would happen if we solved for Y in T_{xw}. Since we avoided substituting anything into W when we solved for Y in T_x, we will get the same result as before, namely, $Y_{xw} = f(x, w, u)$. In the same way, we can show that $W_x = g(x, y, u)$ and $W_{xy} = g(x, y, u)$. So, solving for $y = Y_x(u)$, $w = W_x(u)$ is the same as solving for $y = f(x, w, u)$ and $w = g(x, y, u)$, which is the same as solving for $y = Y_{xw}(u)$, $w = W_{xy}(u)$. Thus, any solution y to $y = Y_{xw}(u), w = W_{xy}(u)$ would also be a solution to $y = Y_x(u)$. □

Reversibility reflects memoryless behavior – the state of the system, V, tracks the state of U, regardless of its history. A typical example of irreversibility is a system of two agents (as in the prisoners' dilemma) who adhere to a "tit-for-tat" strategy. Such a system has two stable solutions, cooperation and defection, under the same external conditions U and, therefore, does not satisfy the Reversibility condition; forcing either one of the agents to cooperate results in the other agent's cooperation ($Y_w(u) = y, W_y(u) = w$), yet this does not guarantee cooperation from the start ($Y(u) = y, W(u) = w$). Irreversibility, in such examples, is a product of using too coarse a state description, where not all of the factors which determine the ultimate state of the system are included in U. In the tit-for-tat example, such factors should include the previous actions of the players. Reversibility is restored once these factors are included.

The properties of Null Action, Degeneracy and Composition are complete for recursive systems. In non-recursive systems, Null Action, Degeneracy, Composition, and Reversibility are not complete. If, however, we replace Reversibility with a slightly stronger property, we obtain a complete (but not sound) set of properties [Galles, 1996b].

4.3 Proofs of Causal Irrelevance Axioms

Using the theorems from the previous section, we can prove the axioms of causal irrelevance.

4.2.1 Holds trivially. □

4.2.2 (By contradiction) Assume that there exists a causal theory such that $CI_T(XW, Z, Y) \& \neg(CI_T(X, Z, Y) \;\&\; CI_T(W, Z, Y))$. So, either $CI_T(XW, Z, Y) \;\&\; \neg CI_T(X, Z, Y)$, or $CI_T(XW, Z, Y) \& \neg CI_T(W, Z, Y)$. First, we consider $CI_T(XW, Z, Y) \& \neg CI_T(X, Z, Y)$. By our definition of $CI_T(\cdot)$, $\neg CI_T(X, Z, Y)$ implies that there exists two values x, x' of X and some value u of U such that $Y_{xz}(u) \neq Y_{x'z}(u)$. Now, let us consider the x, x', z, u such that $Y_{xz}(u) \neq Y_{x'z}(u)$. Using these values, we can determine a w and w' as follows : Let $w = W_{xz}(u)$, and $w' = W_{x'z}(u)$. It does not matter whether $w = w'$ or $w \neq w'$. By Composition, $Y_{xzw}(u) \neq Y_{x'zw}(u)$. Thus, $\exists x, w, z, u \;\; Y_{xwz}(u) \neq Y_{x'w'z}(u)$, which contradicts $CI_T(XW, Z, Y)$. Thus, $CI_T(XW, Z, Y) \& \neg CI_T(X, Z, Y)$ leads to a contradiction. We can use a symmetric argument to show that $CI_T(XW, Z, Y) \& \neg CI_T(W, Z, Y)$ also leads to a contradiction. □

4.4 By our definition of $CI_T(\cdot)$, $CI_T(X,Z,Y) \Longrightarrow Y_{xz}(u) = Y_{x'z}(u)$ for all subtheories of T_{xz}. For an arbitrary W, we consider the subtheory T_w where W is forced to have the value w. By our definition of causal irrelevance, $Y_{xzw}(u) = Y_{x'zw}$ for all values w. In addition, since $CI_T(X,Z,Y) \Longrightarrow Y_{xz}(u) = Y_{x'z}(u)$ for all subtheories of T, $Y_{xzw}(u) = Y_{x'zw}$ for all subtheories of T_w. Since W was arbitrary, $CI_T(X,Z,Y) \Longrightarrow CI_T(X,ZW,Y)$ for all W. $\square$

4.5.1 (By contradiction) Assume $CI_T(X,ZW,Y)$ & $CI_T(X,ZY,W)$ & $\neg CI_T(X,Z,YW)$. $\neg CI_T(X,Z,YW)$ implies $\exists x,x',z \quad (Y_{xz}(u) \neq Y_{x'z}(u)) \vee (W_{xz}(u) \neq W_{x'z}(u))$. Since W and Y are symmetric, we will only consider Y. Consider the values of x,x',z,u such that $Y_{xz}(u) \neq Y_{x'z}(u)$. Let $y = Y_{xz}(u)$ and $y' = Y_{x'z}(u)$.

By Composition, $Y_{xz}(u) = Y_{xzw}(u)$ for $w = W_{xz}(u)$. By assumption, $Y_{xzw}(u) = Y_{x'zw}(u)$. Also by Composition, $W_{xz}(u) = W_{xzy}(u)$ for $y = Y_{xz}(u)$. By assumption, $W_{xzy}(u) = W_{x'zy}(u)$. By Reversibility, since y is a solution to the simultaneous equations $y = Y_{x'zw}$ and $w = W_{x'zy}$, then y must also be a solution to $Y_{x'z}(u)$. Thus $y = y'$, a contradiction. We can use a symmetric argument to show that $W_{xz}(u) \neq W_{x'z}(u)$ also leads to a contradiction. $\square$

4.5.2 (By contradiction) Assume $CI_T(X,ZW,Y)$ & $CI_T(W,ZX,Y)$ & $\neg CI_T(XW,Z,Y)$. Since $\neg CI_T(XW,Z,Y)$, by definition of $CI_T(\cdot)$, $\exists x,x',w,w',z \quad Y_{xwz}(u) \neq Y_{x'w'z}(u)$. However, $CI_T(X,ZW,Y)$ implies $\forall x,x',z,w \quad Y_{xzw}(u) = Y_{x'zw}(u)$. Furthermore, $CI_T(W,ZX,Y)$ implies $\forall x',w,w',z \quad Y_{x'wz}(u) = Y_{x'w'z}(u)$ So, $\forall x,x',w,w',z \quad Y_{xwz}(u) = Y_{x'wz}(u) = Y_{x'w'z}(u)$, thus $\forall x,x',w,w',z \quad Y_{xwz}(u) = Y_{x'w'z}(u)$. This contradicts $\exists x,x',w,w',z \quad Y_{xwz}(u) \neq Y_{x'w'z}(u)$. $\square$

4.4 Causal Relevance and Lewis's Counterfactuals

It is instructive to compare our framework to that of [Lewis, 1973]. We give here a version of Lewis's logic for counterfactual sentences (from [Lewis, 1981]).

Rules

(1) If A and $A \Longrightarrow B$ are theorems, so is B.

(2) If $(B_1 \,\&\, \ldots) \Longrightarrow C)$ is a theorem, so is $((A \,\Box\!\!\rightarrow B_1)\ldots) \Longrightarrow (A \,\Box\!\!\rightarrow C)$

Axioms

(1) All truth-functional tautologies

(2) $A \,\Box\!\!\rightarrow A$

(3) $(A \,\Box\!\!\rightarrow B) \,\&\, (B \,\Box\!\!\rightarrow A) \Longrightarrow (A \,\Box\!\!\rightarrow C) \equiv (B \,\Box\!\!\rightarrow C)$

(4) $((A \vee B) \,\Box\!\!\rightarrow A) \vee ((A \vee B) \,\Box\!\!\rightarrow B) \vee (((A \vee B) \,\Box\!\!\rightarrow C) \equiv (A \,\Box\!\!\rightarrow C) \,\&\, (B \,\Box\!\!\rightarrow C))$

(5) $A \,\Box\!\!\rightarrow B \Longrightarrow A \Longrightarrow B$

(6) $A \,\&\, B \Longrightarrow A \,\Box\!\!\rightarrow B$

Where the statement $A \,\Box\!\!\rightarrow B$ stands for "in all closest worlds where A holds, B holds as well". Lewis is careful not to put any restrictions on definitions of closest worlds, except for the obvious requirement that world w be no further from itself than any other $w' \neq w$. In essence, causal theories with local interventions define an ordering among worlds that gives a metric by which to define what worlds are closest. As such, all of the axioms of Lewis are true for causal theories, and follow from Degeneracy, Composition, and Reversibility.[5]

In order to relate Lewis's axioms to our own, we need a translation from his syntax to ours:

$$A \;\Box\!\!\rightarrow B \equiv Y^B_A(u) = B \tag{17}$$

where Y^B is a variable with values $\{B, \overline{B}\}$. We can now examine each of Lewis's axioms in turn.

(1) Trivially True.

(2) This axiom is the same as Degeneracy, and it is stated in our formalism as $X_x(u) = x$.

(3) This follows directly from Reversibility.

(4) Since actions in causal theories are restricted to conjunctions of literals, this axiom does not apply. However, under the interpretation $do(A \vee B) \equiv do(A) \vee do(B)$, this property does hold.

(5) This axiom follows directly from Composition.

(6) This axiom follows directly from Composition.

We see that Lewis's axioms are more general, hence less powerful. Composition is a consequence of Lewis's axiom (5) and rule (1). Reversibility, however, is not enforced by the Lewis framework. Lewis's axiom (3), while similar, is not as strong as Reversibility. $Y = y$ may hold in all closest w-worlds, $W = w$ may hold in all closest y-worlds and, still, $Y = y$ may not hold in our world.

4.5 Why Transitivity Fails in Causal Relevance

Causal transitivity is a property that makes intuitive sense, which we would like to explain with an axiomatic definition. If a variable A has a causal influence on B, and B has a causal influence on C, one would think that A would have causal influence on C. However, this is not always the case, even in deterministic causality. Consider the causal theory described in Figure 16 of Appendix A, reprinted here as Figure 8.

In this example, X is not causally irrelevant to W, and W is not causally irrelevant to Y, but X is causally irrelevant to Y. The intuition behind this example is that changing X can only cause a minor change in W, while Y only responds to large changes in W. However, the failure of transitivity is deeper then that. Even when X has more complete control over the intermediate variable W, we may still not be able to achieve transitivity. Consider the causal theory of Figure 9.

This theory is the same as the theory of Figure 8, except that W has now been split into $W_1 \ldots W_4$, corresponding to W's four possible values. That is, W_1 is true if $x + u_2 = 0$, W_2

[5]Steps toward applying Lewis's system to actions and decisions were taken by Gibbard and Harper [Gibbard and Harper, 1981] but this work has not been extended to cover the additional properties, such as Reversibility, that causal theories dictate.

$$V = \{X, W, Y\},\ x, y \in \{0,1\},\ w \in \{0,1,2,3\}$$
$$U = \{U_1, U_2\}\ u_1, u_2 \in \{0,1\}$$

$$w = x + 2 * u_2$$
$$x = u_1$$
$$y = (w > 1)$$
$$P(u_1) = P(u_2) = 0.5$$

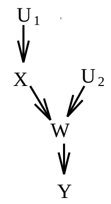

Figure 8: Counterexample to transitivity in causal irrelevance.

$$V = \{X, W_1, W_2, W_3, W_4, Y\} \text{ binary}$$
$$U = \{U_1, U_2\} \text{ binary}$$

$$x = u_1$$
$$w_1 = \neg x \,\&\, \neg u_2$$
$$w_2 = x \,\&\, \neg u_2$$
$$w_3 = \neg x \,\&\, u_2$$
$$w_4 = x \,\&\, u_2$$
$$y = (w_3 \,\&\, \neg w_1 \,\&\, \neg w_2) \vee (w_4 \,\&\, \neg w_1 \,\&\, \neg w_2)$$
$$P(u_1) = P(u_2) = 0.5$$

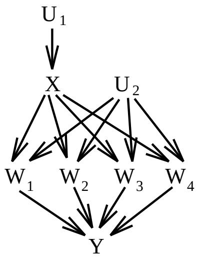

Figure 9: Transitivity fails, even when a variable is more completely controlled by its parents than in the prior example.

is true if $x + u_2 = 1$, W_3 is true if $x + u_2 = 2$, and W_4 is true of $x + u_2 = 3$. Now, by fixing X, we can cause any of the intermediate variables $W_1 \ldots W_4$ to be false in any given state of the world u. Likewise, each of the intermediate variables $W_1 \ldots W_4$ can affect Y in any given state of the world u. However, X has no effect on Y in any state of the world u.

4.6 Causal Relevance and Directed Graphs

4.6.1 Causal Graphs as Irrelevance-Maps

Comparing axioms 3.2–3.5 to 4.2–4.5, we see that causal irrelevance is quite similar to path interception in directed graphs. Since people (and machines) can easily reason about graphs it would be useful to create a graph that represents all of the causal relevances and irrelevances of a given causal theory. That is, we would like to create a graph $G'(T)$ such that

(i) Each variable X in T corresponds to exactly one node X' in $G'(T)$

(ii) For all subsets of nodes X', Y', Z' in $G'(T)$, $int(X', Z', Y')_{G'(T)} \Longrightarrow CI_T(X, Z, Y)$.

(iii) For all subsets of variables X, Y, Z in T, $CI_T(X, Z, Y) \Longrightarrow int(X', Z', Y')_{G'(T)}$.

In such a graph $G'(T)$, if all directed paths from X' to Y' were intercepted by some variables in Z, then X would be causally irrelevant to Y in the theory T. Likewise, if a set of variables X was causally irrelevant to a set Y given fixed Z, then all paths from nodes in X' to nodes in Y' would be intercepted by some variables in Z.

The obvious choice for $G'(T)$ is $G(T)$, the graph associated with the causal theory itself, as defined by Eq. (1). In fact, if we use $G'(T) = G(T)$, then the implication (ii) holds, since in Section 3.6 we showed that $int(X, Z, Y)_{G(T)} \Longrightarrow Y_{xz}(u) = Y_z(u)$, and thus $CI_T(X, Z, Y)$. However, since transitivity holds in $int(\cdot)_G$ and not always in $CI_T(\cdot)$, there might not be a graph $G'(T)$ that implications (ii) and (iii) hold simultaneously. Nonetheless, we can, use directed graphs to validate candidate theorems of causal irrelevance, as shown below.

4.6.2 Graphs as Theorem Provers

Consider an oracle that takes in statements about path interception and returns YES if the statement holds in all directed graphs and NO otherwise. We will show that such an oracle can be used to validate or refute sentences about causal irrelevance.

First, let the *canonical form* for sentences in the language of causal irrelevance be an implication, whose antecedent consists of a conjunction of non-negated literals, and consequent consists of non-negated literals. For instance, consider the sentence[6]

$$CI_T(X, Z, Y) \,\&\, \neg CI_T(X, \emptyset, Y) \Longrightarrow \neg CI_T(Z, \emptyset, Y) \tag{18}$$

This sentence is not in canonical form because the second conjunct in the antecedent is negated and the statement in the consequent is negated. The canonical form of this sentence is

$$CI_T(X, Z, Y) \,\&\, CI_T(Z, \emptyset, Y) \Longrightarrow CI_T(X, \emptyset, Y) \tag{19}$$

Any causal irrelevance sentence can be written in a unique canonical form using standard logical procedures.

Definition 8 *(Horn Component) A* Horn component *H of a causal irrelevance sentence S is a sentence H such that :*

- *H is in canonical form,*
- *The consequent of H contains no disjunctions, and*
- *$H \Longrightarrow S$.*

If a sentence S is in the canonical form $a_1 \,\&\, a_2 \,\&\, \ldots a_i \Longrightarrow b_1 \vee b_2 \vee \ldots b_k$, then a Horn component of S is any sentence of the form $a_1 \,\&\, a_2 \,\&\, \ldots a_i \Longrightarrow b_j$. For example, Eq. (19) has no disjunctions in its consequent, hence is itself a Horn component.

For any causal irrelevance statement A of the form $CI_T(X, Z, Y)$, we will consider A_g to be the corresponding path-interception statement $int(X, Z, Y)_{G(T)}$. Using this convention, we can define

[6] A version of this theorem was chosen in [Heckerman and Shachter, 1995] as the definition of causality.

Theorem 8 (Graphical Theorem Verification) *A causal irrelevance sentence S is true for all causal theories if and only if there exists a Horn component H of S such that H_g is true for all graphs.*

For example, consider the sentence in Eq. (18). The canonical form of this sentence is given in Eq. (19). The sentence is itself a Horn component. The corresponding sentence for path interception in directed graphs, $int(X, Z, Y)_G \ \& \ int(Z, \emptyset, Y)_G \Longrightarrow int(X, \emptyset, Y)_G$, states that if all paths from X to Y are intercepted by Z, and there are no paths from Z to Y, then there is no path from X to Y. This sentence is true for all directed graphs, hence Eq. (18) is a valid theorem.

Next, consider transitivity, stated as $CI_T(X, Z, Y) \Longrightarrow CI_T(a, Z, Y) \vee CI_T(X, Z, a)$. The Horn components of this sentence are

$$H^1 \ : CI_T(X, Z, Y) \Longrightarrow CI_T(a, Z, Y) \tag{20}$$

$$H^2 \ : CI_T(X, Z, Y) \Longrightarrow CI_T(X, Z, a). \tag{21}$$

Looking at each of the corresponding path interception sentences in turn, we find that H_g^1 : $int(X, Z, Y)_G \Longrightarrow int(a, Z, Y)_G$ is not true for all directed graphs, and H_g^2 : $int(X, Z, Y)_G \Longrightarrow int(X, Z, a)_G$ is also not true for all directed graphs, that is, if Z intercepts all paths from X to Y, it is not the case that either Z intercepts all paths from any other variable to Y or Z intercepts all paths from X to any other variable. Thus, transitivity is not a theorem in all causal theories.

Proof (of Theorem 8)

First, we prove that if there are no disjunctions in the consequent of a canonical form sentence, then the sentence is true if and only if the corresponding sentence is true for path interception in directed graphs.

We will prove this by contradiction. Assume that there exists some theorem $A \Longrightarrow B$, where A and B are conjunctions of literals such that :

- $A \Longrightarrow B$ is not a theorem in causal irrelevance
- $A_g \Longrightarrow B_g$ is a theorem in path interception in directed graphs

Since $A_g \Longrightarrow B_g$ is a theorem in path interception, then we must be able to generate B_g from A_g using the axioms of path interception in directed graphs.

Also, since $A \Longrightarrow B$ is not a theorem in causal irrelevance, every such generation of B_g from A_g must include the application of the axiom of transitivity. When the axiom of transitivity is used, a disjunction is created. This disjunction must be used in the generation of B_g. By assumption, B_g does not contain a disjunction. Also, none of the antecedents of any of the axioms of path interception contain disjunctions. Thus the only way to use this disjunction in the generation of B_g is to to resolve the disjunction with a negated clause. Since A_g started with no negated statements, and none of the axioms of path interception can be used to create negated statements, we cannot resolve the disjunction with anything. Thus the generation of B_g from A_g did not require an application of transitivity, a contradiction.

Next, we prove that if a theorem $A \Longrightarrow B \vee C$ is a theorem in causal irrelevance, then either $A \Longrightarrow B$ is a theorem in causal irrelevance or $A \Longrightarrow C$ is a theorem in causal irrelevance. If $A \Longrightarrow B \vee C$ is a theory in causal irrelevance, then we must be able to generate $B \vee C$ from A using the axioms of causal irrelevance. Since no axiom creates a disjunction, the only way to generate $B \vee C$ from A is to either generate B from A and add C, or generate C from A and add B.

Thus, a causal irrelevance sentence is a theorem if and only if there is a path interception theorem that corresponds to one of the Horn components of the original sentence. $\square$

5 Conclusions

How do scientists predict the outcome of one experiment from the results of other experiments run under totally different conditions? Such transfer of experimental knowledge, though it is essential to scientific progress, involves inferences that cannot easily be formalized in the standard languages of logic, physics, or probability.

The formalization of such inferences requires a language within which the experimental conditions prevailing in one experiment can be represented, and the outcome of that experiment can be posed as constraint in the design and analysis of the next experiment. The description of experimental conditions, in turn, involves both observational and manipulative sentences, and requires that manipulative phrases (e.g., "having no effect on," "holding Z fixed"), as distinct from observational phrases (e.g., "being independent of," "conditioning on Z"),[7] be given formal notation, semantical interpretation, and axiomatic characterization. It turns out that standard algebras, including the algebra of equations, Boolean algebra, and probability calculus, are all geared to serve observational sentences, but not manipulative sentences.

This paper bases the semantics of manipulative sentences on a set of structural equations that we call a *causal theory*. Unlike ordinary algebraic equations, a causal theory treats every equation as an independent mathematical object attached to one and only one variable. Actions are treated as modalities and are interpreted as the nonalgebraic operator of replacing equations.

This semantics permits us to develop an axiomatic characterization of manipulative statements of the form "Changing X will not affect Y if we hold Z constant," that we propose as the meaning of causal irrelevance: "X is causally irrelevant to Y in context Z." This axiomatization highlights the differences between causal and informational irrelevance, as in "Finding X will not affect our belief in Y, once we know Z." The former shows a closer affinity to graphical representation than the latter. Under the deterministic definition, causal irrelevance complies with all of the axioms of path interception in cyclic graphs, with the exception of transitivity. This affinity leads to graphical methods of proving theorems about causal relevance and explains, in part, why graphs are so prevalent in causal talk and causal modeling.

[7] Philosophers, statisticians, and economists have been notoriously sloppy about confusing "holding Z constant" with "conditioning on a given Z" [Pearl, 1995a].

Acknowledgments

This research was partially supported by Air Force grant #AFOSR/F496209410173, NSF grant #IRI-9420306, and Rockwell/Northrop Micro grant #94-100. We thank Joe Halpern for commenting on the first draft of this paper and for noting that property 4.5.1 does not hold in Lewis' closest-world framework.

A Appendix : Counterexamples

2.2.3 $CI_P(XW, Z, Y) \Longrightarrow CI_P(X, Z, Y) \vee CI_P(X, Z, W)$.

$V = \{X, W, Y\}$ binary	$x = u_1$
$U = \{U_1\}$ binary	$w = (x \ \& \ u_1)$
	$y = Parity(x, w, u_1)$
	$P(u_1) = 0.5$

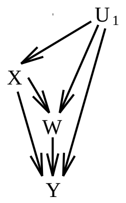

Figure 10: Counterexample to sentence 2.2.3.

In Figure 10, we can see that $CI_P(XW, \emptyset, Y) \,\&\, \neg CI_P(X, \emptyset, W) \,\&\, \neg CI_P(X, \emptyset, Y)$.

In this example, changing X can affect the probability of Y, and changing X can affect the probability of W, but changing X and W together cannot affect the probability of Y. Since changing X affects the value of W, it makes sense to think that intervening on W while intervening on X would not interfere with the effect that X has on Y. However, X does not completely control W. That is, when we only intervene on X, U_1 still has some effect on W. Controlling both X and Y removes the influence of U_1 on W. As in the property 2.2.2, removing the connection between U_1 and W prevents X from having an effect on Y.

2.2.4 $CI_P(XW, Z, Y) \,\&\, CI_P(XY, Z, W) \Longrightarrow CI_P(X, Z, Y) \vee CI_P(X, Z, W)$.

In Figure 11, we can see that

$P(w) = P(y) = 0.5$;

$P(w|set(X = 1)) = P(y|set(X = 1)) = 0.75$;

$P(w|\hat{x}, \hat{y}) = 0.5$ for all values of $\hat{x}, \hat{y}$; and

$P(y|\hat{x}, \hat{w}) = 0.5$ for all values of $\hat{x}, \hat{w}$

Thus, $CI_P(XW, \emptyset, Y) \,\&\, CI_P(XY, \emptyset, W) \,\&\, \neg(CI_P(X, \emptyset, Y) \vee CI_P(X, \emptyset, W))$.

This example actually contains two causal theories, each similar to the theory of example 2.2.2. In one, W is a function of X, Y, and U_1, and Y is a function of U_1. As in example 2.2.2, X can affect W when Y has the same value as U_2, but it has no

$V = \{X, W, Y\}$ binary
$U = \{U_1, U_2\}$ binary

$$x = u_1$$
$$y = \begin{cases} u_1 & \text{if } u_2 = 0 \\ Parity(x, w, u_1) & \text{if } u_2 = 1 \end{cases}$$
$$w = \begin{cases} Parity(x, y, u_1) & \text{if } u_2 = 0 \\ u_1 & \text{if } u_2 = 1 \end{cases}$$
$$P(u_1) = P(u_2) = 0.5$$

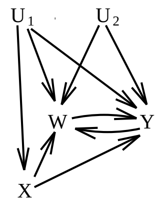

Figure 11: Counterexample to property 2.2.4.

effect on $P(w)$ when Y is held constant. In the other, W is a function of U_1, and Y is a function of X, W, and U_1. Also as in 2.2.2, X can affect Y when W has the same value as U_1, but it has no effect on $P(w)$ when W is fixed. U_2 determines which model is in effect at any given time. While intervening on only X can affect $P(w)$ and $P(y)$, simultaneously changing X and Y together have no effect on $P(w)$, and simultaneously changing X and W together have no effect on $P(y)$.

2.3 $CI_P(X, Z, WY) \Longrightarrow CI_P(X, ZW, Y)$.

$V = \{X, W, Y\}$ binary
$U = \{U_1, U_2\}$ binary

$$x = u_1$$
$$w = u_2$$
$$y = (x \;\&\; (w \text{ XOR } u_2))$$
$$P(u_1) = P(u_2) = 0.5$$

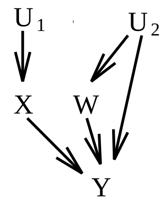

Figure 12: Counterexample to property 2.3.

In the causal theory of Figure 12, $CI_P(X, \emptyset, YW) \;\&\; \neg CI_P(X, W, Y)$.

In this example, X does not have any effect on Y since $P(y) = 0$, and X can only act as an inhibitor if Y. When we intervene on W, then it is possible for Y to have the value 1, and X can affect the probability of Y. Thus, X can only affect Y when we intervene on W, and X has no effect on W.

2.4 $CI_P(X, Z, Y) \;\&\; CI_P(X, ZY, W) \Longrightarrow CI_P(X, Z, WY)$.

In the causal theory in Figure 13, $CI_P(X, \emptyset, Y) \;\&\; CI_P(X, Y, W) \;\&\; \neg CI_P(X, \emptyset, WY)$.

While changing X can affect $P(w)$ (and hence $P(y, w)$) when Y is not held fixed, and changing X has no effect on $P(y)$, fixing Y blocks the effect that X has on W.

2.5.1 $CI_P(X, ZW, Y) \;\&\; CI_P(X, ZY, W) \Longrightarrow CI_P(X, Z, WY)$.

$V = \{X, W, Y\}$ binary
$U = \{U_1, U_2\}$ binary

$$x = u_1$$
$$y = u_2$$
$$w = Parity(x, y, u_2)$$
$$P(u_1) = P(u_2) = 0.5$$

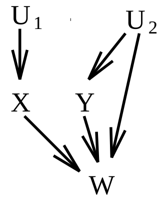

Figure 13: Counterexample to property 2.4.

$V = \{X, W, Y\}$ binary
$U = \{U_1, U_2\}$ binary

$$x = u_1$$
$$y = u_2$$
$$w = Parity(x, y, u_2)$$
$$P(u_1) = P(u_2) = 0.5$$

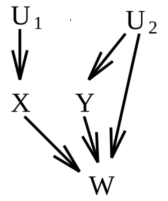

Figure 14: Counterexample to property 2.5.1.

In Figure 14, $CI_P(X, W, Y)$ & $CI_P(X, Y, W)$ & $\neg CI_P(X, \emptyset, WY)$.

Fixing W prevents X from altering the probability of Y, and fixing Y prevents X from altering the probability of W, but X can change the probability of W (and hence the probability of W & Y) if there is no intervention on Y.

Up to this point, all of the counterexamples have relied on some exogenous variable from U having two different children in V. Obviously, this is not essential, since we could always create similar examples in which each exogenous variable has exactly one child. For example, in the theory of Figure 14, we could replace U_2 with Z to get the theory of Figure 15.

In this theory, all of the exogenous variables U have exactly one child, yet property 2.5.1 still does not hold. There is still an undirected cycle in the underlying causal graph, which is required for property 2.5.1 to be false. Properties 2.2.1 – 2.6 are all true for all causal theories whose causal graphs are trees. In addition, properties 2.2.1–2.5.2 are true for all causal theories whose causal graphs are polytrees. Property 2.6, as we will see now, is not always true, even when we restrict its causal graph to be a polytree.

2.6 $CI_P(X, Z, Y) \Longrightarrow CI_P(a, Z, Y) \vee CI_P(X, Z, a) \quad \forall a \notin X \cup Z \cup Y$.

In the causal theory of Figure 16, $CI_P(X, \emptyset, Y) \& \neg CI_P(W, \emptyset, Y) \& \neg CI_P(X, \emptyset, W) \& W \notin X \cup Z \cup Y$.

X can only cause a minor change in W, while a large change in W is required to affect Y. Thus, X can affect W, and W can affect Y, but X has no effect on W. Even if we

$V = \{X, W, Y, Z\}$ binary
$U = \{U_1, U_2\}$ binary

$x = u_1$
$y = u_2$
$w = Parity(x, y, z)$
$z = u_2$
$P(u_1) = P(u_2) = 0.5$

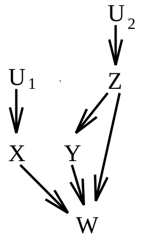

Figure 15: Counterexample to 2.5.1, such that each variable in U has a single child.

$V = \{X, W, Y\}$, $x, y \in \{0, 1\}$, $w \in \{0, 1, 2, 3\}$
$U = \{U_1, U_2\}$ $u_1, u_2 \in \{0, 1\}$

$x = u_1$
$w = x + 2 * u_2$
$y = (w > 1)$
$P(u_1 = 1) = P(u_2 = 1) = 0.5$

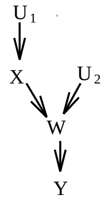

Figure 16: Counterexample to property 2.6.

restrict all variables to be binary, transitivity will not hold. For this counterexample, W could be split into 4 binary variables $W_1 \ldots W_4$, with $f_{W_1} = \neg(x \vee u_2)$, $f_{W_2} = x \,\&\, \neg u_2$, $f_{W_3} = \neg x \,\&\, u_2$, $f_{W_4} = x \,\&\, u_2$, $f_y = w_3 \vee w_4$. Section 4.5 further elaborates this example.

References

[Cartwright, 1989] N. Cartwright. *Nature Capacities and Their Measurements.* Clarendon Press, Oxford, 1989.

[Dawid, 1979] A.P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series A*, 41:1–31, 1979.

[Eells, 1991] E. Eells. *Probabilistic Causality.* Cambridge University Press, Cambridge, MA, 1991.

[Fikes and Nilsson, 1972] R.E. Fikes and Nils Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 3:251–284, 1972.

[Fisher, 1970] F.M. Fisher. A correspondence proncipple for simultaneous equation models. *Econometrica*, 38:73–92, 1970.

[Galles, 1996a] D. Galles. Causal diagrams : A formalism for modeling action and intervention. Technical report, Computer Science Department, UCLA, 1996.

[Galles, 1996b] D. Galles. On the completeness of axioms for counterfactual statements. Technical report, Computer Science Department, UCLA, 1996. In Progress.

[Geiger *et al.*, 1990] D. Geiger, T.S. Verma, and J. Pearl. Identifying independence in Bayesian networks. In *Networks*, volume 20, pages 507–534. John Wiley and Sons, Sussex, England, 1990.

[Gibbard and Harper, 1981] A. Gibbard and L. Harper. Counterfactuals and two kinds of expected utility. In W.L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs*. D. Reidel, Dordrecht: Holland, 1981.

[Good, 1961] I.J. Good. A causal calculus. *Philosophy of Science*, 11:305–318, 1961.

[Heckerman and Shachter, 1995] David Heckerman and Ross Shachter. A definition and graphical representation of causality. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 262–273, San Mateo, CA, 1995. Morgan Kaufmann.

[Lewis, 1973] D. Lewis. *Counterfactuals*. Harvard University Press, Cambridge, MA, 1973.

[Lewis, 1981] David Lewis. Counterfactuals and comparative possibility. In W.L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs*. D. Reidel, Dordrecht: Holland, 1981.

[Manski, 1990] C.F. Manski. Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, 80:319–323, 1990.

[Paz and Pearl, 1994] A. Paz and J. Pearl. Axiomatic characterization of directed graphs. Technical Report R-234, Computer Science Department, UCLA, 1994.

[Paz *et al.*, 1996] A. Paz, J. Pearl, and S. Ur. A new characterization of graphs based on interception relations. *Journal of Graph Theory*, 22(2):125–136, 1996.

[Pearl and Paz, 1987] J. Pearl and A. Paz. Graphoids: A graph-based logic for reasoning about relevance relations. In B. Du Boulay et. al., editor, *Advances in Artificial Intelligence-II*, pages 357–363. North-Holland Publishing Co., 1987.

[Pearl and Verma, 1991a] J. Pearl and T. Verma. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452, San Mateo, CA, 1991. Morgan Kaufmann.

[Pearl and Verma, 1991b] J. Pearl and T. Verma. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference*, pages 441–452, San Mateo, CA, 1991. Morgan Kaufmann.

[Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligence Systems*. Morgan Kaufmann, San Mateo, CA, 1988. (Revised 2nd printing, 1992).

[Pearl, 1993] J. Pearl. Graphical models, causality, and intervention. *Statistical Science*, 8(3):266–273, 1993.

[Pearl, 1994] J. Pearl. A probabilistic calculus of actions. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 454–462, San Mateo, CA, 1994. Morgan Kaufmann Publishers.

[Pearl, 1995a] J. Pearl. Causal diagrams for empirical research (with discussion). *Biometrika*, 82(4):669–709, 1995.

[Pearl, 1995b] J. Pearl. On the testability of causal models with latent and instrumental variables. In D. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 435–443, San Francisco, CA, 1995. Morgan Kaufmann.

[Pearl, 1996a] J. Pearl. Causation, action, and counterfactuals. In *Proceedings of the Sixth Conference on Theoretical Aspects of Rationality and Knowledge (TARK IV 96)*, pages 51–73, The Netherlands, March 1996.

[Pearl, 1996b] J. Pearl. Structural and probabilistic causality. *The Psychology of Learning and Motivation*, 34, 1996.

[Robins, 1987] J. Robins. Addendum to 'a new approachto causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect'. *Comp. Math. Applic.*, 14:923–45, 1987.

[Rosenbaum and Rubin, 1983] P. Rosenbaum and D. Rubin. The central role of propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[Rubin, 1974] D.B. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[Salmon, 1984] W. Salmon. *Scientific explanation and the causal structure of the world.* Princeton University Press, Princeton, 1984.

[Savage, 1954] L.J. Savage. *The Foundations of Statistics*, volume 1. John Wiley and Sons, Inc., New York, 1954.

[Shafer, 1996] G. Shafer. *The Art of Causal Conjecture.* MIT Press, Cambridge, MA, 1996.

[Sobel, 1990] M.E. Sobel. Effect analysis and causation in linear structural equation models. *Psychometrika*, 55:495–515, 1990.

[Spirtes *et al.*, 1993] P. Spirtes, C. Glymour, and R. Schienes. *Causation, Prediction, and Search.* Springer-Verlag, New York, 1993.

[Spohn, 1980] W. Spohn. Stochastic independence, causal independence, and shieldability. *Journal of Philosophical logic*, 9:73–99, 1980.

[Strotz and Wold, 1960] R.H. Strotz and O.A. Wold. Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica*, 28:417–427, 1960.

[Studeny, 1990] M. Studeny. Conditional independence relations have no complete characterization. In *Proceedings of 11-th Prague Conference on Information Theory, Statistical Decision Foundation and Random Processes*, 1990.

[Suppes, 1970] P. Suppes. *A Probabilistic Theory of Causation.* Amsterdam, North Holland, 1970.