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 a set of graphoid-like axioms for causal relevance, that is, statements of the form: "Changing X will not affect Y if we hold Z constant". Both a probabilistic and deterministic definition of causal irrelevance are proposed. The probabilistic definition allows for only two axioms, unless stability is assumed. Under the stability assumption, probabilistic causal irrelevance is equivalent to path interception in cyclic graphs. The deterministic definition allows for all of the axioms of path interception in cyclic graphs, with the exception of transitivity.

Introduction

In (Geiger, Verma, & Pearl 1990), a set of axioms was developed for a class of relations called *graphoids*. These axioms characterize informational relevance 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 an agent's epistemic state. Informational relevance 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 relevance is concerned with statements of the form "X is causally irrelevant to Y given Z," which we take to mean: if we physically fix the value of Z, then changing X will not alter the value of Y.

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.

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 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. It supports all the graph interception axioms except transitivity.

Causal Theories

We define causal theories in the following way (see (Pearl 1995a)).

Definition 1 *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*

$$X_i = f_i(PA_i, u) \quad i = 1, \ldots, n \quad PA_i \subseteq V \setminus X_i$$

The members of the set PA_i *(connoting parents) are often 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, and write* $Y = Y_T(u)$*.*

We will consider local concurrent actions of the form $do(X = x)$, which represent external intervention that forces the variables in X to attain the values x.

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.*

We will also assume each variable $Y \in V$ to be a unique function of the disturbances U in any theory T_x, thus, $Y = Y_{T_x}(u)$. For brevity, the subscript T is often omitted, leaving $Y_x(u)$.

For every set of variables $Y \subseteq V$,

$$P(y) = \sum_{\{u \ | \ Y(u)=y\}} P(u).$$

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

$$P(y|\hat{x}) = P(y|do(X = x)) = \sum_{\{u \ | \ Y_x(u)=y\}} P(u).$$

Axioms of Probabilistic Causal Irrelevance

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

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

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

If we remove the "hat" from the definition above, we get the standard definition of conditional independence in probability calculus, denoted $I(X, Z, Y)$, which are governed by the graphoid axioms (Geiger, Verma, & Pearl 1990) below.

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

1.5 (Intersection)
$I(X, ZY, W) \wedge I(X, ZW, Y) \Longrightarrow I(X, Z, YW)$
Intersection requires a strictly positive probability distribution.

One of the most salient difference between informational and causal relevance is the property of symmetry, which is encapsulated in axiom 1.1.

Another basic difference between informational irrelevance and causal irrelevance is that in the former, $\forall x, x' P(y|z, x) = P(y|z, x') \Longrightarrow \forall x P(y|z, x) = P(y|z)$, while in the latter $\forall x, x' P(y|\hat{z}, \hat{x}) = P(y|\hat{z}, \hat{x}') \not\Longrightarrow \forall x P(y|\hat{z}, \hat{x}) = P(y|\hat{z})$.

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.

The only two axioms which we have found to constrain causal irrelevance are the following:

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

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

Many seemingly intuitive properties, however, **do not hold**. For instance, **none** of the following statements 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)
$CI_P(XW, Z, Y) \quad \wedge \quad CI_P(XY, Z, W) \quad \Longrightarrow CI_P(X, Z, Y) \vee CI_P(X, Z, W)$

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

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

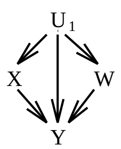
2.5.1 (Right-Intersection) $CI_P(X, ZW, Y) \wedge CI_P(X, ZY, W) \Longrightarrow CI_P(X, Z, WY)$

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

The sentences above were tailored after the graphoid axioms 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 $P(y|\hat{x})$.

The full paper contains counter-examples to properties 2.2.2 – 2.5.1, and 2.6. Here, we give a counter-example for the property 2.2.2, stating (siplified) that if X and W jointly have no effect on Y, then neither X nor W alone will have an effect on Y.

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



$V=\{X,W,Y\}$ binary $\quad f_X(u_1)=u_1$
$U=\{U_1\}$ binary $\quad f_W(u_1)=u_1$
$P(u_1)=0.5 \quad f_Y(x,w,u_1)=Parity(x,w,u_1)$

$$CI_P(XW,\emptyset,Y) \wedge \neg CI_P(X,\emptyset,Y) \wedge \neg CI_P(W,\emptyset,Y)$$

Numeric Constraints

Although Definition 3 imposes only weak constraints (axiom 2.2.1 and 2.5.2) on the notion 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 $P(y|\hat{x},\hat{z}) \geq P(y,z|\hat{x})$ must hold in any causal theory. This can easily be shown by the definition of $P(y,z|\hat{x})$ and $P(y|\hat{x},\hat{z})$. Consider all values u such that $Y_x(u)=y$ and $Z_x(u)=z$. Clearly, for any of these u's, $Y_{xz}(u)=y$, since Z is being held to the same value that it would have obtained without intervention. Additional constraints are explored in (Pearl 1995b).

Axioms of Causal Irrelevance for Stable Distributions

A more expressive set of causal irrelevance axioms is obtained if we confine the analysis to causal theories that represent *stable* distributions, that is, distributions 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 and probability $P(u)$. The functions f_i of a causal theory T define a directed (possibly cyclic) graph $G(T)$ in the following way: Each node in $G(T)$ represents a variable in V, and there is a directed arc from X to Y in $G(T)$ iff $X \in PA_Y$. We can define a stability condition, similar to (Pearl & Verma 1991), as follows:

Definition 4 (Stability) *Let T be a causal theory, $G(T)$ be the directed graph described by T, and $CI(T)$ represent all probabilistic causal irrelevances in T. A causal theory T is* **stable** *iff $\forall T'$ such that $G(T)=G(T')$, we have $CI(T) \subseteq CI(T')$.*

Stability requires irrelevance to be determined by the structure of the equations, and 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 the counter-example to 2.2.2, for example, a minor change in the form of one of the equations would destroy an irrelevance, thus the theory is not stable.

Theorem 1 *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:*

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

Since $CI_P(X,Z,Y) \Longleftrightarrow R_{G(T)}(X,Z,Y)$ 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 & Pearl 1994) and is given below:

3.2.1 (Right-Decomposition) $CI_P(X,Z,YW) \Longrightarrow CI_T(X,Z,Y) \wedge CI_T(X,Z,W)$

3.2.2 (Left-Decomposition) $CI_P(XW,Z,Y) \Longrightarrow CI_P(X,Z,Y) \wedge CI_P(W,Z,Y)$

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

3.5.1 (Right-Intersection) $CI_P(X,ZW,Y) \wedge CI_P(X,ZY,W) \Longrightarrow CI_P(X,Z,YW)$

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

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

Axioms of Deterministic Causal Irrelevance

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

Definition 5 (Causal Irrelevance) *X is causally irrelevant to Y given Z in a causal theory T, $CI_T(X,Z,Y)$, if*

$$\forall u,z,x,x' \quad Y_{xz}(u)=Y_{x'z}(u)$$

in every subtheory of T_z.

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

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

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

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

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

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

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

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

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

The proofs of these axioms will use the following theorems :

Theorem 2 (composition) *For any variables W, X, Y in a causal theory, $W_x(u) = w \Longrightarrow Y_x(u) = Y_{xw}(u)$*

Theorem 3 (reversibility) *For any variables X, Y and W, $(Y_{xw} = y, W_{xy} = w \Longrightarrow Y_x = y)$.*

Note that reversibility does not hold in Lewis' closest-world framework (Lewis 1973). $Y = y$ 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.

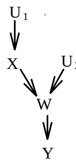
We give a proof for only one of the axioms, the others are similar.

4.5.1 (By Contradiction) Assume $CI_T(X, ZW, Y) \wedge CI_T(X, ZY, W) \wedge \neg CI_T(X, Z, YW)$. $\neg CI_T(X, Z, YW)$ 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 arguement to show that $W_{xz}(u) \neq W_{x'z}(u)$ also leads to a contradiction.

Why Transitivity Fails in Deterministic Causality

Causal transitivity is a property that makes intuitive sense. 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 following causal theory :



$V = \{X, W, Y\}$, $x, y \in \{0, 1\}$, $w \in \{0, 1, 2, 3\}$
$U = \{U_1, U_2\}$ $u_1, u_2 \in \{0, 1\}$
$P(u_1) = P(u_2) = 0.5 \quad f_W(x, u_2) = x + 2 * u_2$
$f_X(u_1) = u_1 \quad f_Y(w) = (w > 1)$

$CI_T(X, \emptyset, Y) \wedge \neg CI_T(W, \emptyset, Y) \wedge \neg CI_T(X, \emptyset, W) \wedge W \notin X \cup Z \cup Y$

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, and Y only responds to large changes in W.

Deterministic Causal Irrelevance and Directed Graphs

Comparing axioms 3.2-3.5 to 4.2-4.5, we see that deterministic causal irrelevance is quite similar to path interception in a directed graph. In fact, the two notions are related in the following way :

Theorem 4 *If Z blocks all directed paths between X and Y in the graph $G(T)$ defined by T, then X is causally irrelevant to Y given Z.*

One would like, of course, to have a graph for which the implication goes both ways. However, transitivity holds in graph interception and not in $CI_T(\cdot)$.

Thus the graph $G(T)$ defined by a causal theory T is an Irrelevance-Map of T (every path interception in the graph corresponds to a valid irrelevance in the theory), but it is not a perfect map, (not every irrelevance in the theory corresponds to path interception in the graph.)

Acknowledgements 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 Axiom 4.5.1 does not hold in Lewis' closest-world framework.

References

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

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

Geiger, D.; Verma, T.; and Pearl, J. 1990. Identifying independence in Bayesian networks. In *Networks*, volume 20. Sussex, England: John Wiley and Sons. 507–534.

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

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

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

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

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

Pearl, J. 1995a. Causal diagrams for empirical research. Technical Report R-218-B, Computer Science Department, UCLA. To appear in *Biometrika*, December 1995.

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

Simon, H. 1953. Causal ordering and identifiability. In Hood, W., and Koopmans, T., eds., *Studies in Econometric Method.* New York: John Wiley and Sons.

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

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