Sufficient Conditions for Causality to be Transitive

Joseph Y. Halpern*
Cornell University
Computer Science Department
Ithaca, NY 14853
halpern@cs.cornell.edu
http://www.cs.cornell.edu/home/halpern

Abstract

Natural conditions are provided that are sufficient to ensure that causality as defined by approaches that use counterfactual dependence and structural equations will be transitive.

*I thank Chris Hitchcock and the anonymous reviewers of the paper for perceptive comments that greatly influenced the structure and story of the paper. Work supported in part by NSF grants IIS-0812045, IIS-0911036, and CCF-1214844, by AFOSR grants FA9550-08-1-0438, FA9550-09-1-0266, and FA9550-12-1-0040, and by ARO grant W911NF-09-1-0281.

1 Introduction

The question of the transitivity of causality has been the subject of much debate. As Paul and Hall [2013] say: "Causality seems to be transitive. If C causes D and D causes E, then C thereby causes E." The appeal to transitivity is quite standard in informal scientific reasoning: we say things like "the billiards expert hit ball A, causing it to hit ball B, causing it to carom into ball C, which then drops into the pocket". It then seems natural to conclude then the pool expert's shot caused ball C to drop into the pocket.

Paul and Hall [2013, p. 215] suggest that "preserving transitivity is a basic desideratum for an adequate analysis of causation". Hall [2000] is even more insistent, saying "That causation is, necessarily, a transitive relation on events seems to many a bedrock datum, one of the few indisputable a priori insights we have into the workings of the concept." Lewis [1986, 2000] imposes transitivity in his influential definition of causality, by taking causality to be the transitive closure ("ancestral", in his terminology) of a one-step causal dependence relation.

But numerous examples have been presented that cast doubt on transitivity. Paul and Hall [2013] give a sequence of such counterexamples; Hall [2000] gives others. I review two such examples in the next section. This leaves us in a somewhat uncomfortable position. It seems so natural to think of causality as transitive. In light of the examples, should we just give up on these intuitions? Paul and Hall [2013] suggest that "What's needed is a more developed story, according to which the inference from "C causes D" and "D causes E" to "C causes E" is safe provided such-and-such conditions obtain—where these conditions can typically be assumed to obtain, except perhaps in odd cases ...". The goal of this paper is to provide sufficient conditions for causality to be transitive. I formalize this using the structural equations framework of Halpern and Pearl [2001, 2005]. The properties that I require suggest that these conditions apply to any definition of causality that depends on counterfactual dependence and uses structural equations (see, for example, [Glymour and Wimberly 2007; Hall 2007; Halpern 2015; Halpern and Pearl 2005; Hitchcock 2001; Hitchcock 2007; Woodward 2003] for examples of such approaches).

These conditions may explain why, although causality is not transitive in general (and is not guaranteed to be transitive according to any of the counterfactual accounts mentioned above), we tend to think of causality as transitive, and are surprised when it is not.

2 Defining causation using counterfactuals

In this section, I review some of the machinery of structural equations needed to define causality. For definiteness, I use the same formalism as that given by Halpern and Pearl [2005].

2.1 Causal structures

Approaches based on structural equations assume that the world is described in terms of random variables and their values. Some random variables may have a causal influence on others. This influence is modeled by a set of *structural equations*. It is conceptually useful to split the random variables into two sets: the *exogenous* variables, whose values are determined by factors outside the model, and the *endogenous* variables, whose values are ultimately determined by the exogenous variables. For example, in a voting scenario, we could have endogenous variables that describe what the voters actually do (i.e., which candidate they vote for), exogenous variables that describe the factors that determine how the voters vote, and a variable describing the outcome (who wins). The structural equations describe how the outcome is determined (majority rules; a candidate wins if A and at least two of B, C, D, and E vote for him; etc.).

Formally, a *causal model* M is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ is a *signature*, which explicitly lists the endogenous and exogenous variables and characterizes their possible values, and $\mathcal{F}$ defines a set of *modifiable structural equations*, relating the values of the variables. A signature $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a set of exogenous variables, $\mathcal{V}$ is a set of endogenous variables, and $\mathcal{R}$ associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y (that is, the set of values over which Y *ranges*). For simplicity, I assume here that $\mathcal{V}$ is finite, as is $\mathcal{R}(Y)$ for every endogenous variable $Y \in \mathcal{V}$. $\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function denoted F_X such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$. This mathematical notation just makes precise the fact that F_X determines the value of X, given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$. If there is one exogenous variable U and three endogenous variables, X, Y, and Z, then F_X defines the values of X in terms of the values of Y, Z, and U. For example, we might have $F_X(u, y, z) = u + y$, which is usually written as $X = U + Y$.[1] Thus, if $Y = 3$ and $U = 2$, then $X = 5$, regardless of how Z is set.

[1]The fact that X is assigned $U + Y$ (i.e., the value of X is the sum of the values of U and Y) does not imply that Y is assigned $X - U$; that is, $F_Y(U, X, Z) = X - U$ does not necessarily

The structural equations define what happens in the presence of external interventions. Setting the value of some variable X to x in a causal model $M = (\mathcal{S}, \mathcal{F})$ results in a new causal model, denoted $M_{X=x}$, which is identical to M, except that the equation for X in $\mathcal{F}$ is replaced by $X = x$.

Following [Halpern and Pearl 2005], I restrict attention here to what are called *recursive* (or *acyclic*) models. This is the special case where there is some total ordering $\prec$ of the endogenous variables (the ones in $\mathcal{V}$) such that if $X \prec Y$, then X is independent of Y, that is, $F_X(\ldots, y, \ldots) = F_X(\ldots, y', \ldots)$ for all $y, y' \in \mathcal{R}(Y)$. Intuitively, if a theory is recursive, there is no feedback. If $X \prec Y$, then the value of X may affect the value of Y, but the value of Y cannot affect the value of X. It should be clear that if M is an acyclic causal model, then given a *context*, that is, a setting $\vec{u}$ for the exogenous variables in $\mathcal{U}$, there is a unique solution for all the equations. We simply solve for the variables in the order given by $\prec$. The value of the variables that come first in the order, that is, the variables X such that there is no variable Y such that $Y \prec X$, depend only on the exogenous variables, so their value is immediately determined by the values of the exogenous variables. The values of variables later in the order can be determined once we have determined the values of all the variables earlier in the order.

It is sometimes helpful to represent a causal model graphically. Each node in the graph corresponds to one variable in the model. An arrow from one node to another indicates that the former variable figures as a nontrivial argument in the equation for the latter. The graphical representation is useful for visualizing causal models, and will be used in the next section.

2.2 A language for reasoning about causality

To define causality carefully, it is useful to have a language to reason about causality. Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, a *primitive event* is a formula of the form $X = x$, for $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$. A *causal formula (over $\mathcal{S}$)* is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$, where

- φ is a Boolean combination of primitive events,
- $Y_1, \ldots, Y_k$ are distinct variables in $\mathcal{V}$, and
- $y_i \in \mathcal{R}(Y_i)$.

hold.

Such a formula is abbreviated as $[\vec{Y} \leftarrow \vec{y}]\varphi$. The special case where $k = 0$ is abbreviated as φ. Intuitively, $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$ says that φ would hold if Y_i were set to y_i, for $i = 1, \ldots, k$.

A causal formula ψ is true or false in a causal model, given a context. As usual, I write $(M, \vec{u}) \models \psi$ if the causal formula ψ is true in causal model M given context $\vec{u}$. The $\models$ relation is defined inductively. $(M, \vec{u}) \models X = x$ if the variable X has value x in the unique (since we are dealing with acyclic models) solution to the equations in M in context $\vec{u}$ (that is, the unique vector of values for the exogenous variables that simultaneously satisfies all equations in M with the variables in $\mathcal{U}$ set to $\vec{u}$). The truth of conjunctions and negations is defined in the standard way. Finally, $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\varphi$ if $(M_{\vec{Y}=\vec{y}}, \vec{u}) \models \varphi$.

2.3 Defining causality

The basic intuition behind counterfactual definitions of causality is that A is a cause of B if there is counterfactual dependence between A and B: if A hadn't occurred (although it did), then B would not have occurred. It is well known that the counterfactual dependence does not completely capture causality; there are many examples in the literature where people say that A is a cause of B despite the fact that B does not counterfactually depend on A (at least, not in this simple sense). Nevertheless, all the counterfactual definitions of causality (as well as people's causality ascriptions) agree that this simple type of counterfactual dependence gives a *sufficient* condition for causality. For the purposes of this paper, I consider only cases where this counterfactual dependence holds.

More formally, say that $X = x$ is a *but-for cause of* φ *in* $(M, \vec{u})$ (where φ is a Boolean combination of primitive events) if $(M, \vec{u}) \models X = x \land \varphi$ (so both $X = x$ and φ hold in context $\vec{u}$) and there exists some x' such that $(M, \vec{u}) \models [X \leftarrow x']\neg\varphi$. Thus, with a but-for cause, changing the value of X to something other than x changes the truth value of φ; that is, φ counterfactually depends on X.

All the complications in counterfactual approaches to causality arise in how they deal with cases of causality that are not but-for causality. Roughly speaking, the idea is that $X = x$ is a cause of $Y = y$ if the outcome $Y = y$ counterfactually depends on X under the appropriate contingency (i.e., holding some other variables fixed at certain values). While the various approaches to defining causality differ in exactly how this is done, they all agree that a but-for cause should count as a cause. So, for simplicity in this paper, I consider only but-for causality and do not both to give a general definition of causality

3 Sufficient Conditions for Transitivity

In this section I present two different sets of conditions sufficient for transitivity. Before doing that, I give two counterexamples to transitivity, since these motivate the conditions. The first example is taken from (an early version of) Hall [2004], and is also considered by Halpern and Pearl [2005].

Example 1: Consider the following scenario:

> Billy contracts a serious but nonfatal disease so is hospitalized. Suppose that Monday's doctor is reliable, and administers the medicine first thing in the morning, so that Billy is fully recovered by Tuesday afternoon. Tuesday's doctor is also reliable, and would have treated Billy if Monday's doctor had failed to. Given that Monday's doctor treated Billy, it's a good thing that Tuesday's doctor did not treat him: one dose of medication is harmless, but two doses are lethal.

Suppose that we are interested in Billy's medical condition on Wednesday. We can represent this using a causal model M_B with three variables:

- MT for Monday's treatment (1 if Billy was treated Monday; 0 otherwise);
- TT for Tuesday's treatment (1 if Billy was treated Tuesday; 0 otherwise); and
- BMC for Billy's medical condition (0 if Billy feels fine on Wednesday; 1 if Billy feels sick on Wednesday; 2 if Billy is dead on Wednesday).

We can then describe Billy's condition as a function of the four possible combinations of treatment/nontreatment on Monday and Tuesday. I omit the obvious structural equations corresponding to this discussion; the causal graph is shown in Figure 1:

In the context where Billy is sick and Monday's doctor treats him, $MT = 1$ is a but-for cause of $TT = 0$—because Billy is treated Monday, he is not treated on Tuesday morning. And $TT = 0$ is a but-for cause of Billy's being alive ($BMC = 0 \vee BMC = 1$). However, $MT = 1$ is *not* a cause of Billy's being alive. It is clearly not a but-for cause; Billy will still be alive if MT is set to 0. Indeed, it is not even a cause under the more general definitions of causality, according to all the approaches mentioned above; no setting of the other variables will lead to a counterfactual dependence between MT and $BMC \neq 2$. This shows that causality is not transitive according to these approaches. Although $MT = 1$ is a

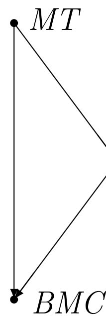

Figure 1: Billy's medical condition.

cause of $TT = 0$ and $TT = 0$ is a cause of $BMC = 0 \lor BMC = 1$, $MT = 1$ is not a cause of $BMC = 0 \lor BMC = 1$. (Of course, according to Lewis [1986, 2000], who takes the transitive closure of the one-step dependence relation, $MT = 1$ is a cause of $BMC = 0 \lor BMC = 1$.) ∎

Although this example may seem somewhat forced, there are many quite realistic examples of lack of transitivity with exactly the same structure. Consider the body's homeostatic system. An increase in external temperature causes a short-term increase in core body temperature, which in turn causes the homeostatic system to kick in and return the body to normal core body temperature shortly thereafter. But if we say that the increase in external temperature happened at time 0 and the return to normal core body temperature happened at time 1, we certainly would not want to say that the increase in external temperature at time 0 caused the body temperature to be normal at time 1![2]

There is another reason that causality is intransitive, which is illustrated by the following example, due to McDermott [1995].

Example 2: Suppose that a dog bites Jim's right hand. Jim was planning to detonate a bomb, which he normally would do by pressing the button with his right forefinger. Because of the dog bite, he presses the button with his left forefinger. The bomb still goes off.

Consider the causal model M_D with variables DB (the dog bites, with values 0 and 1), P (the press of the button, with values 0, 1, and 2, depending on whether the button is not pressed at all, pressed with the right hand, or pressed with the

[2] I thank Richard Scheines [personal communication, 2013] for this example.

left hand), and B (the bomb goes off). We have the obvious equations: DB is determined by the context, $P = DB + 1$, and $B = 1$ if P is either 1 or 2. In the context where $DB = 1$, it is clear that $DB = 1$ is a but-for cause of $P = 2$ (if the dog had not bitten, P would have been 1), and $P = 2$ is a but-for cause of $B = 1$ (if P were 0, then B would be 0), but $DB = 1$ is not a but-for cause of $B = 1$. And again, $DB = 1$ is not a cause of $B = 1$ even under a more general notion of causation. Whether or not the dog had bitten Jim, the button would have been pressed and the bomb would have detonated. ∎

As I said, I believe that we feel that causality is transitive because, in typical settings, it is. My belief is based mainly on introspection here and informal polling of colleagues. Even when told that causality is not transitive, people seem to find it hard to construct counterexamples. This suggests that when they think about their everyday experience of causality, they come up with examples where causality is transitive. If there were many counterexamples available in everyday life, it would be easier to generate them.

I now give two sets of simple conditions that are sufficient to guarantee transitivity. Specifically, I give conditions to guarantee that if $X_1 = x_1$ is a but-for cause of $X_2 = x_2$ in $(M, \vec{u})$ and $X_2 = x_2$ is a but-for cause of $X_3 = x_3$ in $(M, \vec{u})$, then $X_1 = x_1$ is a but-for cause of $X_3 = x_3$ in $(M, \vec{u})$.

The first set of conditions assumes that X_1, X_2, and X_3 each has a default setting. We can think of the default setting as the result of doing nothing. This makes sense, for example, in the billiards example at the beginning of the paper, where we can take the default setting for the shot to be the expert doing nothing, and the default setting for the balls to be that they are not in motion. Let the default setting be denoted by the value 0.

Proposition 3.1: *Suppose that (a)* $X_1 = x_1$ *is a but-for cause of* $X_2 = x_2$ *in* $(M, \vec{u})$*, (b)* $X_2 = x_2$ *is a but-for cause of* $X_3 = x_3$ *in* $(M, \vec{u})$*, (c)* $x_3 \neq 0$*, (d)* $(M, \vec{u}) \models [X_1 \leftarrow 0](X_2 = 0)$*, and (e)* $(M, \vec{u}) \models [X_1 \leftarrow 0, X_2 \leftarrow 0](X_3 = 0)$*. Then* $X_1 = x_1$ *is a but-for cause of* $X_3 = x_3$ *in* $(M, \vec{u})$*.*

Proof: If $X_2 = 0$ is the unique solution to the equations in the causal model $M_{X_1 \leftarrow 0}$ in context $\vec{u}$ and $X_3 = 0$ in the unique solution to the equations in $M_{X_1 \leftarrow 0, X_2 \leftarrow 0}$ in context $\vec{u}$, then it is immediate that $X_3 = 0$ in the unique solution to the equations in $M_{X_1 \leftarrow 0}$ in context $\vec{u}$. That is, $(M, \vec{u}) \models [X_1 \leftarrow 0](X_3 = 0)$. It follows from assumption (a) that $(M, \vec{u}) \models X_1 = x_1$. We must thus have $x_1 \neq 0$, since otherwise $(M, \vec{u}) \models X_1 = 0 \wedge [X_1 \leftarrow 0](X_3 = 0)$, so $(M, \vec{u}) \models X_3 = 0$, which contradicts assumptions (b) and (c). Thus, $X_1 = x_1$ is a but-for cause of $X_3 = x_3$, since the value of X_3 depends counterfactually on that of X_1. ∎

Although the conditions of Proposition 3.1 are clearly rather specialized, they arise often in practice. Conditions (d) and (e) say that if X_1 remains in its default state, then so will X_2, and if both X_1 and X_2 remain in their default states, then so will X_3. (These assumptions are very much in the spirit of the assumptions that make a causal network *self-contained*, in the sense defined by Hitchcock [2007].) Put another way, this says that the reason for X_2 not being in its default state is X_1 not being in its default state, and the reason for X_3 not being in its default state is X_1 and X_2 both not being in their default states. The billiard example can be viewed as a paradigmatic example of when these conditions apply. It seems reasonable to assume that if the expert does not shoot, then ball A does not move; and if the expert does not shoot and ball A does not move (in the context of interest), then ball B does not move, and so on.

Of course, the conditions on Proposition 3.1 do not apply in either Example 1 or Example 2. The obvious default values in Example 1 are $MT = TT = 0$, but the equations say that in all contexts $\vec{u}$ of the causal model M_B for this example, we have $(M_B, \vec{u}) \models [MT \leftarrow 0](TT = 1)$. In the second example, if we take $DB = 0$ and $P = 0$ to be the default values of DB and P, then in all contexts $\vec{u}$ of the causal model M_D, we have $(M_D, \vec{u}) \models [DB \leftarrow 0](P = 1)$.

While Proposition 3.1 is useful, there are many examples where there is no obvious default value. When considering the body's homeostatic system, even if there is arguably a default value for core body temperature, what is the default value for the external temperature? But it turns out that the key ideas of the proof of Proposition 3.1 apply even if there is no default value. Suppose that $X_1 = x_1$ is a but-for cause of $X_2 = x_2$ in $(M, \vec{u})$ and $X_2 = x_2$ is a but-for cause of $X_3 = x_3$ in $(M, \vec{u})$. Then to get transitivity, it suffices to find values x_1', x_2', and x_3' such that $x_3 \neq x_3'$, $(M, \vec{u}) \models [X_1 \leftarrow x_1'](X_2 = x_2')$, and $(M, \vec{u}) \models [X_1 \leftarrow x_1, X_2 \leftarrow x_2'](X_3 = x_3')$. The argument in the proof of Proposition 3.1 then shows that $(M, \vec{u}) \models [X_1 \leftarrow x_1'](X_3 = x_3')$.[3] It then follows that $X_1 = x_1$ is a but-for cause of $X_3 = x_3$ in $(M, \vec{u})$. In Proposition 3.1, x_1', x_2', and x_3' were all 0, but there is nothing special about the fact that 0 is a default value here. As long as

[3]The analogous statement is also valid in standard conditional logic. That is, taking $A > B$ to represent "if A were the case then B would be the case", using standard closest-world semantics [Lewis 1973], $(A > B) \wedge ((A \wedge B) > C) \Rightarrow (A > C)$ is valid. I thank two of the anonymous reviewers of this paper for encouraging me both to note that this idea is the key argument of the paper and to relate it to the Lewis approach.

we can find some values x_1', x_2', and x_3', these conditions apply. I formalize this as Proposition 3.2, which is a straightforward generalization of Proposition 3.1.

Proposition 3.2: *Suppose that there exist values x_1', x_2', and x_3' such that (a) $X_1 = x_1$ is a but-for cause of $X_2 = x_2$ in $(M, \vec{u})$, (b) $X_2 = x_2$ is a but-for cause of $X_3 = x_3$ in $(M, \vec{u})$, (c) $x_3 \neq x_3'$, (d) $(M, \vec{u}) \models [X_1 \leftarrow x_1'](X_2 = x_2')$, and (e) $(M, \vec{u}) \models [X_1 \leftarrow x_1', X_2 \leftarrow x_2'](X_3 = x_3')$. Then $X_1 = x_1$ is a but-for cause of $X_3 = x_3$ in $(M, \vec{u})$.*

To see how these ideas apply, suppose that a student receive an A+ in a course, which causes her to be accepted at Cornell University (her top choice, of course!), which in turn causes her to move to Ithaca. Further suppose that if she had received an A in the course she would have gone to university U_1 and as a result moved to city C_1, and if she gotten anything else, she would have gone to university at U_2 and moved to city C_2. This story can be captured by a causal model with three variables, G for her grade, U for the university she goes to, and C for the city she moves to. There are no obvious default values for any of these three variables. Nevertheless, we have transitivity here: The student's A+ was a cause of her being accepted at Cornell and being accepted at Cornell was a cause of her move to Ithaca; it seems like a reasonable conclusion that the student's A+ was a cause of her move to Ithaca. And, indeed, transitivity follows from Proposition 3.2. We can take the student getting an A to be x_1', take the student being accepted at university U_1 to be x_2', and the student moving to C_1 to be x_3' (assuing that U_1 is not Cornell and that C_1 is not Ithaca, of course).

The conditions provided in Proposition 3.2 are not only sufficient for causality to be transitive, they are necessary as well, as the following result shows.

Proposition 3.3: *If $X_1 = x_1$ is a but-for cause of $X_3 = x_3$ in $(M, \vec{u})$, then there exist values x_1', x_2', and x_3' such that $x_3 \neq x_3'$, $(M, \vec{u}) \models [X_1 \leftarrow x_1'](X_2 = x_2')$, and $(M, \vec{u}) \models [X_1 \leftarrow x_1', X_2 \leftarrow x_2'](X_3 = x_3')$.*

Proof: Since $X_1 = x_1$ is a but-for cause of $X_3 = x_3$ in $(M, \vec{u})$, there must exist values $x_1' \neq x_1$ and $x_3 \neq x_3'$ such that $(M, \vec{u}) \models [X_1 \leftarrow x_1'](X_3 = x_3')$. Let x_2' be such that $(M, \vec{u}) \models [X_1 \leftarrow x_1'](X_2 = x_2')$. Since $(M, \vec{u}) \models [X_1 \leftarrow x_1'](X_2 = x_2' \land X_3 = x_3')$, it easily follows that $(M, \vec{u}) \models [X_1 \leftarrow x_1', X_2 = x_2'](X_3 = x_3')$. ∎

In light of Propositions 3.2 and 3.3, understanding why causality is so often taken to be transitive comes down to finding sufficient conditions to guarantee the assumptions of Proposition 3.2. I now present another set of conditions sufficient to guarantee the assumptions of Proposition 3.2 (and thus, sufficient to make causality transitive), motivated by the two examples showing that causality is not transitive. To deal with the problem in Example 2, I require that for every value x_2' in the range of X_2, there is a value x_1' in the range of X_1 such that $(M, \vec{u}) \models [X_1 \leftarrow x_1'](X_2 = x_2')$. This requirement holds in many cases of interest; it is guaranteed to hold if $X_1 = x_1$ is a but-for cause of $X_2 = x_2zi$ and X_2 is a binary variable (i.e., takes on only two values), since but-for causality requires that two different values of X_1 result in different values of X_2. But this requirement does not hold in Example 2; no setting of DB can force P to be 0.

Imposing this requirement still does not deal with the problem in Example 1. To do that, we need one more condition. Say that a variable Y *depends on* X if there is some setting of all the variables in $\mathcal{U} \cup \mathcal{V}$ other than X and Y such that varying the value of X in that setting results in Y's value varying; that is, there is a setting $\vec{z}$ of the variables other than X and Y and values x and x' of X such that $F_Y(x, \vec{z}) \neq F_Y(x', \vec{z})$.

Up to now I have used the phrase "causal path" informally; I now make it more precise. A *causal path* in a causal model M is a sequence $(Y_1, \ldots, Y_k)$ of variables such that Y_{j+1} depends on Y_j for $j = 1, \ldots, k-1$. Since there is an edge between Y_j and Y_{j+1} in the causal graph for M exactly if Y_{j+1} depends on Y_j, a causal path is just a path in the causal graph. A *causal path from* X_1 *to* X_2 is just a causal path whose first node is X_1 and whose last node is X_2. Finally, Y *lies on a causal path from* X_1 *to* X_2 if Y is a node (possibly X_1 or X_2) on a directed path from X_1 to X_2.

The additional condition that I require for transitivity is that X_2 must lie on every causal path from X_1 to X_3. Roughly speaking, this says that all the influence of X_1 on X_3 goes through X_2. This condition does not hold in Example 1; as Figure 1 shows, there is a direct causal path from MT to BMC that does not include TT. On the other hand, this condition does hold in many examples of interest. Going back to the example of the student's grade, the only way that the student's grade can influence which city the student moves to is via the university that accepts the student.

The following result summarizes the second set of conditions sufficient for transitivity.

Proposition 3.4: *Suppose that $X_1 = x_1$ is a but-for cause of $X_2 = x_2$ in the causal setting $(M, \vec{u})$, $X_2 = x_2$ is a but-for cause of $X_3 = x_3$ in $(M, \vec{u})$, and the following two conditions hold:*

(a) for every value $x'_2 \in \mathcal{R}(X_2)$, there exists a value $x'_1 \in \mathcal{R}(X_1)$ such that $(M, \vec{u}) \models [X_1 \leftarrow x'_1](X_2 = x'_2)$;

(b) X_2 is on every causal path from X_1 to X_3.

Then $X_1 = x_1$ is a but-for cause of $X_3 = x_3$.

The proof of Proposition 3.4 is not hard, although we must be careful to get all the details right. The high-level idea of the proof is easy to explain though. Suppose that $X_2 = x_2$ is a but-for cause of $X_3 = x_3$ in $(M, \vec{u})$. Then there must be some values $x_2 \neq x'_2$ and $x_3 \neq x'_3$ such that $(M, \vec{u}) \models [X_2 \leftarrow x'_2](X_3 = x'_3)$. By assumption, there exists a value $x'_1 \in \mathcal{R}(X_1)$ such that $(M, \vec{u}) \models [X_1 \leftarrow x'_1](X_2 = x'_2)$. The requirement that X_2 is on every causal path from X_1 to X_3 guarantees that $[X_2 \leftarrow x'_2](X_3 = X_3)$ implies $[X_1 \leftarrow x'_1, X_2 \leftarrow x'_2](X_3 = X_3)$. Roughly speaking, X_2 "screens off" the effect of X_1 on X_3, since it is on every causal path from X_1 to X_3. Now we can apply Proposition 3.2. I defer the formal argument to the appendix.

It is easy to construct examples showing that the conditions of Proposition 3.4 are not necessary for causality to be transitive. Suppose that $X_1 = x_1$ causes $X_2 = x_2$, $X_2 = x_2$ causes $X_3 = x_3$, and there are several causal paths from X_1 to X_3. Roughly speaking, the reason that $X_1 = x_1$ may not be a but-for cause of $X_3 = x_3$ is that the effects of X_1 on X_3 may "cancel out" along the various causal paths. This is what happens in the homeostasis example. If X_2 is is on all the causal paths from X_1 to X_3, then, as we have seen, all the effect of X_1 on X_3 is mediated by X_2, so the effect of X_1 on X_3 on different causal paths cannot "cancel out". But even if X_2 is not on all the causal paths from X_1 to X_3, the effects of X_1 on X_3 may not cancel out along the causal paths; and $X_1 = x_1$ may still be a cause of $X_3 = x_3$. That said, it seems difficult to find a weakening of the condition in Proposition 3.4 that is simple to state and suffices for causality to be transitive.

A Proof of Proposition 3.4

To prove Proposition 3.4, I need a preliminary result, which states a key (and obvious!) property of causal paths: if there is no causal path from X to Y, then changing the value of X cannot change the value of Y. Although it is intuitively obvious, proving it carefully requires a little bit of work.

Lemma A.1: *If Y and all the variables in $\vec{X}$ are endogenous, $Y \notin \vec{X}$, and there is no causal path from a variable in $\vec{X}$ to Y, then for all sets $\vec{W}$ of variables disjoint from $\vec{X}$ and Y, and all settings $\vec{x}$ and $\vec{x}'$ for $\vec{X}$, y for Y, and $\vec{w}$ for $\vec{W}$, we have*

$$(M,u) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}](Y = y) \text{ iff } (M,u) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}](Y = y)$$

and

$$(M,u) \models [\vec{X} \leftarrow \vec{x}](Y = y) \text{ iff } (M,u) \models Y = y.$$

Proof: Define the *maximum distance* of a variable Y in a causal model M, denoted $maxdist(Y)$, to be the length of the longest causal path from an exogenous variable to Y. We prove the result by induction on $maxdist(Y)$. If $maxdist(Y) = 1$, then the value of Y depends only on the values of the exogenous variables, so the result trivially holds. If $maxdist(Y) > 1$, let $Z_1, \ldots, Z_k$ be the endogenous variables on which Y depends. These are the endogenous parents of Y in the causal graph (i.e., these are exactly the endogenous variables Z such that there is an edge from Z to Y in the causal graph). For each $Z \in \{Z_1, \ldots, Z_k\}$, $maxdist(Z) < maxdist(Y)$: for each path from an exogenous variable to Z, there is a longer path to Y, namely, the one formed by adding the edge from Z to Y. Moreover, there is no path from a variable in $\vec{X}$ to any of $Z_1, \ldots, Z_k$ nor is any of $Z_1, \ldots, Z_k$ in $\vec{X}$ (for otherwise there would be a path from a variable in $\vec{X}$ to Y, contradicting the assumption of the lemma). Thus, the inductive hypothesis holds for each of $Z_1, \ldots, Z_k$. Since the value of each of $Z_1, \ldots, Z_k$ does not change when we change the setting of $\vec{X}$ from $\vec{x}$ to $\vec{x}'$, and the value of Y depends only on the values of $Z_1, \ldots, Z_k$ and $\vec{u}$ (i.e., the values of the exogenous variables), the value of Y cannot change either. ∎

I can now prove Proposition 3.4. I restate it here for the convenience of the reader.

Proposition 3.4: *Suppose that $X_1 = x_1$ is a but-for cause of $X_2 = x_2$ in the causal setting $(M, \vec{u})$, $X_2 = x_2$ is a but-for cause of $X_3 = x_3$ in $(M, \vec{u})$, and the following two conditions hold:*

(a) for every value $x_2' \in \mathcal{R}(X_2)$, there exists a value $x_1' \in \mathcal{R}(X_1)$ such that $(M, \vec{u}) \models [X_1 \leftarrow x_1'](X_2 = x_2')$;

(b) X_2 is on every causal path from X_1 to X_3.

Then $X_1 = x_1$ is a but-for cause of $X_3 = x_3$.

Proof: Since $X_2 = x_2$ is a but-for cause of $X_3 = x_3$ in $(M, \vec{u})$, there must exist $x_2' \neq x_2$ and $x_3' \neq x_3$ such that $(M, \vec{u}) \models [X_2 \leftarrow x_2'](X_3 = x_3')$. By assumption, there exists a value x_1' such that $(M, \vec{u}) \models [X_1 \leftarrow x_1'](X_2 = x_2')$. I claim that $(M, \vec{u}) \models [X_1 \leftarrow x_1'](X_3 = x_3')$. This follows from a more general claim. I show that if Y is on a causal path from X_2 to X_3, then

$$(M, \vec{u}) \models [X_1 \leftarrow x_1'](Y = y) \text{ iff } (M, \vec{u}) \models [X_2 \leftarrow x_2'](Y = y). \quad (1)$$

Although it is not obvious, this is essentially the argument sketched in the main part of the text. Literally the same argument as that given below for the proof of (1) also shows that

$$(M, \vec{u}) \models [X_1 \leftarrow x_1'](Y = y) \text{ iff } (M, \vec{u}) \models [X_1 \leftarrow x_1' \wedge X_2 \leftarrow x_2'](Y = y).$$

Define a partial order $\prec$ on endogenous variables that lie on a causal path from X_2 to X_3 by taking $Y_1 \prec Y_2$ if Y_1 precedes Y_2 on some causal path from X_2 to X_3. Since M is a recursive model, if $Y_1 \prec Y_2$, we cannot have $Y_2 \prec Y_1$ (otherwise there would be a cycle). I prove (1) by induction on the $\prec$ ordering. The least element in this ordering is clearly X_2; X_2 must come before every other variable on a causal path from X_2 to X_3. By assumption, $(M, \vec{u}) \models [X_1 \leftarrow x_1'](X_2 = x_2')$, and clearly $(M, u) \models [X_2 \leftarrow x_2'](X_2 = x_2')$. Thus, (1) holds for X_2. Thus completes the base case of the induction.

For the inductive step, let Y be a variable that lies on a causal path from X_2 and X_3, and suppose that (1) holds for all variables Y' such that $Y' \prec Y$. Let $Z_1, \ldots, Z_k$ be the endogenous variables that Y depends on in M. For each of these variables Z_i, either there is a causal path from X_1 to Z_i or there is not. If there is, then the path from X_1 to Z_i can be extended to a directed path P from X_1 to X_3, by going from X_1 to Z_i, from Z_i to Y, and from Y to X_3 (since Y lies on a causal path from X_2 to X_3). Since, by assumption, X_2 lies on every causal path from X_1 to X_3, X_2 must lie on P. Moreover, X_2 must precede Y on P. (Proof: Since Y lies on a path P' from X_2 to X_3, X_2 must precede Y on P'. If Y precedes X_2 on P, then there is a cycle, which is a contradiction.) Since Z_i precedes Y on P, it follows that $Z_i \prec Y$, so by the inductive hypothesis, $(M, \vec{u}) \models [X_1 \leftarrow x_1'](Z_i = z_i)$ iff $(M, \vec{u}) \models [X_2 \leftarrow x_2'](Z_i = z_i)$.

Now if there is no causal path from X_1 to Z_i, then there also cannot be a causal path P from X_2 to Z_i (otherwise there would be a causal path from X_1 to Z_i formed by appending P to a causal path from X_1 to X_2, which must exist since, if not, it easily follows from Lemma A.1 that $X_1 = x_1$ would not be a cause of $X_2 = x_2$). Since there is no causal path from X_1 to Z_i, by Lemma A.1, we must have that $(M,\vec{u}) \models [X_1 \leftarrow x_1'](Z_i = z_i)$ iff $(M,\vec{u}) \models Z_i = z_i$ iff $(M,\vec{u}) \models [X_2 \leftarrow x_2'](Z_i = z_i)$.

Since the value of Y depends only on the values of $Z_1, \ldots, Z_k$ and $\vec{u}$, and I have just shown that $(M,\vec{u}) \models [X_1 \leftarrow x_1'](Z_1 = z_1 \wedge \ldots \wedge Z_k = z_k)$ iff $(M,\vec{u}) \models [X_2 \leftarrow x_2'](Z_1 = z_1 \wedge \ldots \wedge Z_k = z_k)$, it follows that $(M,\vec{u}) \models [X_1 \leftarrow x_1'](Y = y)$ iff $(M,\vec{u}) \models [X_2 \leftarrow x_2'](Y = y)$. This completes the proof of the induction step. Since X_3 is on a causal path from X_2 to X_3, it follows that $(M,\vec{u}) \models [X_1 \leftarrow x_1'](X_3 = x_3')$ iff $(M,\vec{u}) \models [X_2 \leftarrow x_2'](X_3 = x_3')$. Since $(M,\vec{u}) \models [X_2 \leftarrow x_2'](X_3 = x_3')$ by construction, we have that $(M,\vec{u}) \models [X_1 \leftarrow x_1'](X_3 = x_3')$, as desired. Thus, $X_1 = x_1$ is a but-for cause for $X_3 = x_3$. ∎

References

Glymour, C. and F. Wimberly (2007). Actual causes and thought experiments. In J. Campbell, M. O'Rourke, and H. Silverstein (Eds.), *Causation and Explanation*, pp. 43–67. Cambridge, MA: MIT Press.

Hall, N. (2000). Causation and the price of transitivity. *Journal of Philosophy XCVII*(4), 198–222.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, and L. A. Paul (Eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT Press.

Hall, N. (2007). Structural equations and causation. *Philosophical Studies 132*, 109–136.

Halpern, J. Y. (2015). A modification of the Halpern-Pearl definition of causality. In *Proc. 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pp. 3022–3033.

Halpern, J. Y. and J. Pearl (2001). Causes and explanations: A structural-model approach. Part I: Causes. In *Proc. Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI 2001)*, pp. 194–202.

Halpern, J. Y. and J. Pearl (2005). Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for Philosophy of Science 56*(4), 843–887.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy XCVIII*(6), 273–299.

Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *Philosophical Review 116*, 495–532.

Lewis, D. (1986). Causation. In *Philosophical Papers*, Volume II, pp. 159–213. New York: Oxford University Press. The original version of this paper, without numerous postscripts, appeared in the *Journal of Philosophy* **70**, 1973, pp. 113–126.

Lewis, D. (2000). Causation as influence. *Journal of Philosophy XCVII*(4), 182–197.

Lewis, D. K. (1973). *Counterfactuals*. Cambridge, Mass.: Harvard University Press.

McDermott, M. (1995). Redundant causation. *British Journal for the Philosophy of Science 40*, 523–544.

Paul, L. A. and N. Hall (2013). *Causation: A User's Guide*. Oxford University Press.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford, U.K.: Oxford University Press.