Direct and Indirect Effects
(First Draft, Comments Welcome)

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

Abstract

The direct effect of one event on another can be defined and measured by holding constant all intermediate variables between the two. Indirect effects present conceptual and practical difficulties (in nonlinear models), because they cannot be isolated by holding certain variables constant. This paper shows that it is nevertheless possible to define the effect transmitted through a given set of paths without blocking the remaining paths. This permits the assessment of a more representative type of direct and indirect effects, one that is applicable in both linear and nonlinear models. The paper establishes conditions under which such assessments can be estimated consistently from experimental and nonexperimental data, and thus extends path-analytic techniques to nonlinear and nonparametric models. We also provide a policy-related interpretation of indirect effects.

1 INTRODUCTION

The distinction between total, direct, and indirect effects is deeply entrenched in causal conversations, and attains practical importance in many applications, including policy decisions, legal definitions and health care analysis (see Section 2.2 for examples). Structural equation modeling (SEM) provides a principled methodology of defining, identifying, and estimating these three types of causal influence (Goldberger 1972; Duncan 1975). However, the bulk of SEM methodology was developed for linear analysis, and no comparable methodology has been devised to extend these capabilities to structural models involving dichotomous variables or nonlinear dependencies.[1] The prevailing attitude in the SEM literature is that "there is no calculus of path coefficients for loglinear path models" (Hagenaars, 1993, page 17, attributed to Fienberg) and "In loglinear analysis these indirect and total effects cannot be obtained in a simple manner" (ibid, page 49). The purpose of this paper is to show that

[1] A notable exception is the counterfactual analysis of Robins and Greenland (1992) which is applicable to nonlinear models, but does not incorporate path-analytic techniques.

many path analytic techniques concerning effect decomposition can be extended in a natural way to nonlinear models.

A central requirement for any such extension is to detach the notion of "effect" from its algebraic representation as a coefficient in an equation, and redefine "effect" as a general capacity to transmit *changes* among variables. One such extension, based on hypothetical interventions, is presented in Pearl (2000, Chapter 5); it has led to general definitions of total and direct effects, and to new methods of estimating such effects in nonlinear and nonparametric models (that is, models in which the functional form of the equations is unknown).

When applied to linear systems, the generalized definitions coincide with familiar path-analytic expressions of total and direct effects, yet they differ from conventional definitions in several important aspects. First, direct effects are defined in terms of hypothetical experiments in which intermediate variables are held constant by *physical intervention*, not by statistical adjustment (a distinction that is often obscured by the phrase "control for"). Second, total and direct effects are defined independently of each other, as outcomes of two different experiments, thus contrasting conventional definitions (e.g. Bollen 1989, p. 376; Mueller 1996, p. 141; Kline 1998, p. 175) which equate the total effect with sums of products of direct effects.

Finally, the nonlinear generalization of effect decomposition has encountered difficulty extending the notion of "indirect effect" (Pearl, 2000, p. 165). In standard linear analysis, an indirect effect is defined as the difference between the total effect and the direct effects (Bollen 1989). In nonlinear analysis, differences lose their significance, and one must isolate the contribution of mediating paths in some other way. However, since mediating paths cannot be isolated by blocking all other paths in the model (as is done in the isolation of direct effects), the definition of indirect effects has remained incomplete, and, save for asserting inequality between direct and total effects, the very concept of "indirect effect" was deemed void of operational meaning.

This paper shows that it is possible to devise a meaningful generalization of both direct and indirect effects without blocking any paths in the model.[2] This permits the assessment of indirect effects in nonlinear models, and thus extends path-analytic techniques to such models. Additionally, the generalization uncovers a second interpretation of direct effect, here called "descriptive" (see Section 2.2) which concerns the action of causal forces under natural, rather than experimental conditions. This interpretation yields the standard path-coefficients in linear models, but leads to a different definition and different estimation procedures of direct effects in nonlinear models.

Following a conceptual discussion of the descriptive and prescriptive interpretations, (Section 2.2), Sections 3.1 and 3.2 formulate these two interpretations of direct effects, while Section 3.3 establishes conditions under which the descriptive (or "natural") interpretation can be estimated consistently from either experimental or nonexperimental data. Sections 3.4 and 3.5 extend the formulation and identification analysis to indirect effects. In Section 3.6, we generalize the notion of indirect effect to *path-specific effects*, that is, effects transmitted through any specified set of paths in the model. Section 3.7 shows how the various

[2] My interest in seeking such generalization was stimulated by Jacques Hagenaars, who pointed out the importance of quantifying indirect effects in social science.

definitions of direct and indirect effects lead to the standard results when applied to linear systems. Section 4 concludes with an example that illustrates the policy-related significance of effect decomposition and provides a policy related interpretation for indirect effects.

2 TOTAL, DIRECT, AND INDIRECT EFFECTS: CONCEPTUAL ANALYSIS

2.1 Direct versus Total Effects: The Basic Distinction

The causal relationship that is easiest to interpret, define and estimate is the *total effect*. Written as $P(Y_x = y)$, the total effect measures the probability that response variable Y would take on the value y when X is set to x by external intervention.[3] This probability function is what we normally assess in a controlled experiment in which X is randomized and in which the distribution of Y is estimated for each level x of X.

In many cases, however, this quantity does not adequately represent the target of investigation and attention is focused instead on the direct effect of X on Y. The term "direct effect" is meant to quantify an influence that is not mediated by other variables in the model or, more accurately, the sensitivity of Y to changes in X while all other factors in the analysis are held fixed. Naturally, holding those factors fixed would sever all causal paths from X to Y with the exception of the direct link $X \rightarrow Y$, which is not intercepted by any intermediaries.

A classical example of the ubiquity of direct effects (see Hesslow 1976; Cartwright 1989) tells the story of a birth-control pill that is suspect of producing thrombosis in women and, at the same time, has a negative indirect effect on thrombosis by reducing the rate of pregnancies (pregnancy is known to encourage thrombosis). In this example, interest is focused on the direct effect of the pill because it represents a stable biological relationship that, unlike the total effect, is invariant to marital status and other factors that may affect women's chances of getting pregnant or of sustaining pregnancy. This invariance makes the direct effect transportable across cultural and sociological boundaries and, hence, a more useful quantity in scientific explanation and policy analysis.

Another class of examples involves legal disputes over race or sex discrimination in hiring. Here, neither the effect of sex or race on applicants' qualification nor the effect of qualification on hiring are targets of litigation. Rather, defendants must prove that sex and race do not *directly* influence hiring decisions, whatever indirect effects they might have on hiring by way of applicant qualification. This is made quite explicit in the following court ruling:

> "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same." (Carson versus Bethlehem Steel Corp., 70 FEP Cases 921, 7th Cir. (1996), Quoted in Gastwirth 1997.)

[3]The substripted notation Y_x is borrowed from the potential-outcome framework of Neyman (1923) and Rubin (1974). Pearl (2000) used, interchangeably, $P_x(y), P(y|do(x)), P(y|\hat{x})$, and $P(y_x)$, and showed their equivalence to probabilities of subjunctive conditionals: $P((X = x) \Box\!\!\rightarrow (Y = y))$ (Lewis 1973).

Taking this criterion as a guideline, the direct effect of X on Y (in our case X=gender Y=hiring) can roughly be defined as the response of Y to change in X (say from $X = x^*$ to $X = x$) while keeping all other accessible variables at their initial value, namely, the value they would have attained under $X = x^*$.[4] This doubly-hypothetical criterion will be given precise mathematical formulation in Section 3, using the language and semantics of structural counterfactuals (Pearl, 2000; chapter 7).

As a third example, one that illustrates the policy-making ramifications of direct and total effects, consider a drug treatment that has a side effect – headache. Patients who suffer from headache tend to take aspirin which, in turn may have its own effect on the disease or, may strengthen (or weaken) the impact of the drug on the disease. To determine how beneficial the drug is to the population as a whole, under existing patterns of aspirin usage, the total effect of the drug is the target of analysis, and the difference $P(Y_x = y) - P(Y_{x^*} = y)$ may serve to assist the decision, with x and x^* being any two treatment levels. However, to decide whether aspirin should be encouraged or discouraged during the treatment, the *direct* effect of the drug on the disease, both with aspirin and without aspirin, should be the target of investigation. The appropriate expression for analysis would then be the distribution $P(Y_{xz} = y)$, where z stands for any specified level of aspirin intake.[5]

Readers versed in structural equation models (SEMs) will note that, in linear systems, direct effects are fully specified by the corresponding path coefficients, and are independent of the values at which we hold the the intermediate variables (Z in our examples). In nonlinear systems, those values would, in general, modify the effect of X on Y and thus should be chosen carefully to represent the target policy under analysis. For example, the direct effect of a pill on thrombosis would most likely be different for pregnant and nonpregnant women. Epidemiologists call such differences "effect modification" and insist on separately reporting the effect in each subpopulation $Z = z$.

In all these examples, the requirement of holding the mediating variables fixed must be interpreted as (hypothetically) setting these variables to constants by physical intervention, not by analytical means such as selection, conditioning, or adjustment. For example, it will not be sufficient to measure the association between the birth-control pill and thrombosis separately among pregnant and nonpregnant women and then aggregate the results. Instead, we must perform the study among women who became pregnant before the use of the pill and among women who prevented pregnancy by means other than the drug. The reason is that, by conditioning on an intermediate variable (pregnancy in the example), we may create spurious associations between X and Y even when there is no direct effect of X on Y. This can easily be illustrated in the model $X \to Z \leftarrow U \to Y$, where X has no direct effect on Y. Physically holding Z constant would permit no association between X and Y, as can be seen by deleting all arrows entering Z (an operation embodied in the counterfactual expression Y_{xz}.) But if we were to condition on Z, a spurious association would be created through U

[4] Robins and Greenland (1992) have adapted essentially the same criterion (phrased differently) for their interpretation of "direct effect" in epidemiology.

[5] In practice, the difference $E(Y_{x_1} - Y_{x_2})$ is often taken as a measure of the total effect, where x_1 and x_2 are any two treatment levels. This difference can easily be computed from the general distribution $P(Y_x = y)$, in which x and y stand for any arbitrary levels of dosage and outcome, respectively. Similarly, the general distribution $P(Y_{xz} = y)$ can be used for obtaining the (expected) Z-specific effect differences, $E(Y_{x_1 z_1}) - E(Y_{x_2 z_1})$ and $E(Y_{x_1 z_2}) - E(Y_{x_2 z_2})$, where z_1 and z_2 are any two levels of aspirin intake.

(unobserved) that might be construed as a direct effect of X on Y.

2.2 Direct effects: Descriptive versus prescriptive interpretation

The various usages of direct effects in policy decision, legal disputes and scientific explanation call for a distinction between two types of conceptualizations: *prescriptive* and *descriptive*. We will illustrate this distinction using the treatment-aspirin example described in the last section. In the prescriptive conceptualization, we ask whether a specific untreated patient would improve if treated, while holding the aspirin intake fixed at some predetermined level, say $Z = z$. In the descriptive conceptualization, we ask again whether the untreated patient would improve if treated, but now we hold the aspirin intake fixed at whatever level the patient currently consumes under no-treatment condition. The difference between these two conceptualizations lies in whether we wish to account for the natural relationship between the direct and the mediating cause (that is, between treatment and aspirin) or to modify that relationship to match policy objectives. We call the effect computed from the descriptive perspective the *natural* effect, and the one computed from the prescriptive perspective the *controlled* effect.

To illustrate more sharply the difference between the two conceptions, consider a patient who takes aspirin if and only if treated, and for whom the treatment is effective only when aspirin is present. For such a person, the treatment is deemed to have no natural direct effect (on recovery), because, by keeping the aspirin at the current, pre-treatment level of zero, we ensure that the treatment effect would be nullified. The controlled direct effect, however, is nonzero for this person, because the efficacy of the treatment would surface when we fix the aspirin intake at non-zero level. Note that the descriptive formulation requires knowledge of the individual natural behavior—in our example, whether the untreated patient actually uses aspirin—while the prescriptive formulation requires no such knowledge.

This difference, though trivially bridged at the individual level, becomes a major stumbling block when it comes to estimating *average* direct effects in a population of individuals. At the population level, the prescriptive formulation is pragmatic; we wish to predict the difference in recovery rates between treated and untreated patients when a prescribed dose of aspirin is administered to all patients in the population—the actual consumption of aspirin under uncontrolled conditions need not concern us. In contrast, the descriptive formulation is attributional; we ask whether an observed improvement in recovery rates (again, between treated and untreated patients) is attributable to the treatment itself, as opposed to preferential use of aspirin among treated patients. To properly distinguish between these two contributions, we therefore need to measure the improvement in recovery rates while making each patient take the same level of aspirin that he/she took before treatment. However, as Robins and Greenland (1992) pointed out, such control over individual behavior would require testing the same group of patients twice (i.e., under treatment and no treatment conditions), and cannot be administered in experiments with two different groups, however randomized. (There is no way to determine what level of aspirin an untreated patient would take if treated, unless we actually treat that patient and, then, this patient could no longer be eligible for the untreated group.) Since repeatable tests on the same individuals are rarely feasible, the descriptive measure of the direct effect is not generally estimable from standard experimental studies. In Section 3.3 we will analyze what additional assumptions are required for consistently estimating this measure, the *average natural direct effect*, from either experimental or observational studies.

2.3 Indirect effects: Descriptive versus prescriptive formulation

The descriptive conception of direct effects can easily be transported to the formulation of indirect effects; oddly, the prescriptive formulation is not transportable. Returning to our treatment-aspirin example, if we wish to assess the *natural* indirect effect of treatment on recovery for a specific patient, we withhold treatment and ask, instead, whether that patient would recover if given as much aspirin as he/she would take under treatment. In this way, we insure that whatever changes occur in the patient's condition are due to the aspirin and not to the treatment. Similarly, at the population level, the natural indirect effect of the treatment is interpreted as the improvement in recovery rates if we were to withhold treatment from all patients but, instead, let each patient take the same level of aspirin that he/she would have taken under treatment. As in the descriptive formulation of direct effects, this hypothetical quantity involves nested counterfactuals and will be identifiable only under special circumstances.

The prescriptive formulation of direct effects has no parallel in indirect effects, because there is no way to block the direct effect from operating by holding certain variables constant. In other words, it is impossible to hold a set of variables constant in such a way that the effect of X on Y measured under those conditions would circumvent the direct pathway, if such exists. We could, of course, keep X constant and measure the effect (on Y) of varying the intermediate variable Z between two prescribed levels, z_1 and z_2. But this would represent the (direct) effect of Z, not of X, because we have no guarantee that a given variation in X will indeed cause Z to change from z_1 to z_2.

We will see that, in linear systems, the descriptive and prescriptive formulations of direct effects lead, indeed, to the same expression in terms of structural coefficients. The corresponding linear expression for indirect effects, usually computed as the difference between the total and direct effects, coincides with the descriptive formulation but finds no prescriptive interpretation. This failure to interpret indirect effects as effects that could be realized while holding certain variables fixed led the author to conclude (Pearl, 2000, page 165) that indirect effects lack intrinsic operational meaning. This conclusion assumes that policy interventions are limited to the operation of fixing variables. In Section 4 we will give a broader operational meaning to the descriptive interpretation of indirect effects.

3 FORMAL ANALYSIS

3.0 Notation

Throughout our analysis we will let X be the control variable (whose effect we seek to assess), and let Y be the response variable. We will let Z stand for the set of all intermediate variables between X and Y which, in the simplest case considered, would be a single variable as in Fig. 1. Most of our results will still be valid if we let Z stand for any set of such variables, in particular, the set of Y's parents excluding X.

We will use the counterfactual notation $Y_x(u)$ to denote the value that Y would attain in unit (or situation) $U = u$ under the control regime $do(X = x)$. See appendix, or Pearl (2000, Chapter 7) for formal semantics of these counterfactual utterances. Many concepts associated with direct and indirect effect require comparison to a reference value of X, that is, a value relative to which we measure changes. We will designate this reference value by x^*.

3.1 Controlled Direct Effects (review)

Definition 1 *(Controlled unit-level direct-effect; qualitative)*

A variable X is said to have a controlled direct effect on variable Y in model M and situation $U = u$ if there exists a setting $Z = z$ of the other variables in the model and two values of X, x^* and x, such that

$$Y_{x^*z}(u) \neq Y_{xz}(u) \tag{1}$$

In words, the value of Y under $X = x^*$ differs from its value under $X = x$ when we keep all other variables Z fixed at z. If condition (1) is satisfied for some z, we say that the transition from $X = x^*$ to $X = x$ has a controlled direct-effect on Y.

We recall that X has a direct effect on Y in some situation $U = u$ if and only if X is a parent of Y in the causal graph G associated with M. We likewise recall that the set Z in Definition 1 can be confined to the parents of Y, excluding X. In the sequel, we will use the abbreviation "event $X = x$ has an effect" to mean "the transition from $X = x^*$ to $X = x$ has an effect," keeping the reference point $X = x^*$ implicit.

Definition 2 *(Controlled unit-level direct-effect; quantitative)*
Given a causal model M with causal graph G, the controlled direct effect of $X = x$ on Y in unit $U = u$ and setting $Z = z$ is given by

$$CDE_z(x, x^*; Y, u) = Y_{xz}(u) - Y_{x^*z}(u) \tag{2}$$

where Z stands for all parents of Y (in G) excluding X.

Definition 3 *(Population-level controlled direct effect)*
Given a probabilistic causal model $\langle M, P(u)\rangle$, the controlled direct effect of event $X = x$ on Y is defined as:

$$CDE_z(x, x^*; Y) = E(Y_{xz} - Y_{x^*z}) \tag{3}$$

or, equivalently,

$$\begin{aligned} CDE_z(x, x^*; Y) &= \sum_u DE_z(x, x^*; Y, u)P(u) \qquad (4)\\ &= \sum_y [P(Y_{xz} = y) - P(Y_{x^*z} = y)]y \qquad (5) \end{aligned}$$

It is clear from Eq. 5 that the probability $P(Y_{xz} = y)$, when specified for all x, z, and y, contains all the information needed for characterizing the prescriptive direct effect. We therefore call this distribution the *kernel* of $(DE_z(x, x^*; Y))$.

The kernel distribution $P(Y_{xz} = y)$ can be estimated consistently from experimental studies in which both X and Z are randomized. In nonexperimental studies, the identification of this distribution requires that certain "no-confounding" assumptions hold true in the population tested. Graphical criteria encapsulating these assumptions are described in Pearl (2000, Sections 4.3 and 4.4). We summarize here the main result using both graphical and counterfactual criteria.

Theorem 1 *(Pearl and Robins, 1995)*
The controlled direct effect $DE_z(x, x^*; Y)$ is identified from nonexperimental data if there exist two sets of covariates, W_1 and W_2, such that

1. *W_1 consists of non descendants of X*
2. *W_2 consists of non descendants of Z*
3. *The following separation conditions are satisfied in the causal graph G:*

   (i)
   $$(Y \perp\!\!\!\perp X | W_1)_{G_{\underline{X}\overline{Z}}} \tag{6}$$
   (Read: W_1 d-separates Y from X in the subgraph $G_{\underline{X}\overline{Z}}$ of G, formed by removing all arrows emanating from X and arrows entering Z.)

   (ii)
   $$(Y \perp\!\!\!\perp Z | X, W_1, W_2)_{G_{\underline{Z}}} \tag{7}$$

Moreover, if these conditions are satisfied, $DE_z(x, x^*; Y)$ is given by (5), where

$$P(Y_{xz} = y) = \sum_{w_1, w_2} P(y|w_1, w_2, z, x)P(w_2|w_1, x)P(w_1) \tag{8}$$

Figure 1 illustrates a model that allows for the identification of the controlled direct effects of X on Y, as in Eq. (8). Conditions 3(i) and 3(ii) of Theorem 1 are satisfied using $W_1 = 0$ and $W_2 = \{W\}$, as is shown in the auxiliary graphs of Fig. 1 (b) and (c).

The conditions in Theorem 1 are sufficient, but not necessary; more refined conditions can be devised to exploit missing links among variables in the set Z (Pearl and Robins 1995; Kuroki and Miyakawa 1999).

The conditions in Theorem 1 are always satisfied in the special class of models called Markovian (recursive equations with independent error terms), for which Eq. (8) yields (with $W_1 = W_2 = 0$)

$$P(Y_{xz} = y) = P(y|z, x) \tag{9}$$

Thus, in Markovian models, the average controlled direct-effect (Eq. (5)) reduces to a difference between two conditional expectations:

$$CDE_z(x, x^*; Y) = E(Y|x, z) - E(Y|x^*, z) \tag{10}$$

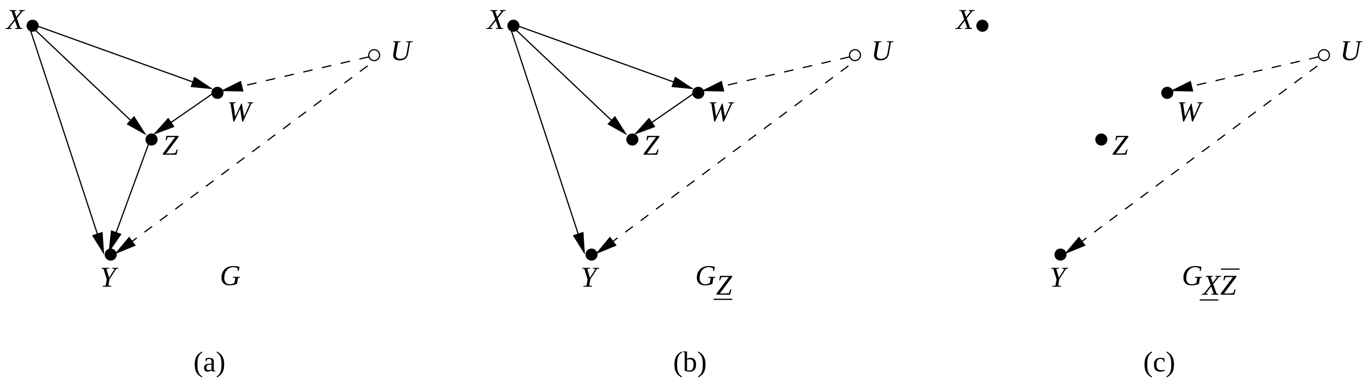

Figure 1: (a) A causal model in which the controlled direct effect of X on Y is identified. (b)–(c) The subgraphs supporting the identification conditions of (6) and (7).

and can be estimated, therefore, using standard multivariate techniques.

We now supplement Theorem 1 with an equivalent condition for identifiability that is expressed in terms of dependencies among counterfactual variables. This condition will be helpful for comparison to other criteria.

Theorem 2 *The controlled direct effect $DE_z(x, x^*; Y)$ is identified from nonexperimental data if, for all x and z, we have:*

1. *$X_z = X$ (i.e., Variables in Z do not affect X), and*
2. *there exist two sets of covariates, W_1 and W_2, such that*
   - *(i)* $(Y_{xz} \perp\!\!\!\perp X | W_1)$
   - *(ii)* $(Y_z \perp\!\!\!\perp Z | X, W_1, W_2)$

These conditions are sufficient for yielding (8).

Proof

$$\begin{array}{rcll} P(Y_{xz} = y) & = & \sum_{w_1} P(Y_{xz} = y | X = x, w_1) P(w_1) & \text{(from (i))} \\ & = & \sum_{w_1} P(Y_{xz} = y | X_z = x, w_1) P(w_1) & \text{(since } X_z = X) \\ & = & \sum_{w_1} P(Y_z = y | X_z = x, w_1) P(w_1) & \text{(from } X_z = x \Rightarrow Y_{xz} = Y_z)) \\ & = & \sum_{w_1} \sum_{w_2} P(Y_z = y | z, x, w_1, w_2) P(w_1) P(w_2 | x, w_1) & \text{(from (ii))} \\ & = & \sum_{w_1} \sum_{w_2} P(y | z, x, w_1, w_2) P(w_2 | x, w_1) P(w) & \text{(from } Z = z \Rightarrow Y_z = Y) \end{array}$$

3.2 Natural Direct Effects: Formulation

Definition 4 *(Unit-level natural direct effect; qualitative)*
An event $X = x$ is said to have a natural direct effect on variable Y in situation $U = u$ if the following inequality holds

$$Y_{x^*}(u) \neq Y_{x, Z_{x^*}(u)}(u) \tag{11}$$

In words, the value of Y under $X = x^*$ differs from its value under $X = x$ even when we keep Z at the same value ($Z_{x^*}(u)$) that Z attains under $X = x^*$.

We can easily extend this definition from events to variables by defining X as having a *natural* direct effect on Y (in model M and situation $U = u$) if there exist two values, x^* and x, that satisfy (11). Note that this definition no longer requires that we specify a value z for Z; that value is determined naturally by the model, once we specify x, x^*, and u. Note also that condition (11) is a direct translation of the court criterion of sex discrimination in hiring (Section 2.1) with $X = x^*$ being a male, $X = x$ a female, and $Y = 1$ a decision to hire.

If one is interested in the magnitude of the natural direct effect, one can take the difference

$$Y_{x, Z_{x^*}(u)}(u) - Y_{x^*}(u) \tag{12}$$

or the ratio

$$Y_{x, Z_{x^*}(u)}(u) / Y_{x^*}(u) \tag{13}$$

or any other relationship between the two components in (1) assuming, of course, that the relationship is well defined for the chosen values of x, x^*, and u.

We will concentrate on the difference measure (12) and designate it by the symbol $NDE(x, x^*; Y, u)$ (acronym for Natural Direct Effect). If we are further interested in assessing the average of this difference in a population of units, we have:

Definition 5 *(Population-level natural direct effect)*
The average natural direct effect of event $X = x$ on a response variable Y, denoted $NDE(x, x^*; Y)$, is defined as

$$\begin{aligned} NDE(x, x^*; Y) &= \sum_u NDE(x, x^*; Y, u) P(u) && (14) \\ &= \sum_y y P(Y_{x, Z_{x^*}} = y) - E(Y_{x^*}) && (15) \\ &= E(Y_{x, Z_{x^*}}) - E(Y_{x^*}) && (16) \end{aligned}$$

where $P(u)$ stands for the probability that unit $U = u$ would be present in the population (or, more generally, the probability that situation $U = u$ would materialize).

3.3 Natural Direct Effects: Identification

Definition 5 requires that, as we vary X from x^* to x, we keep the value of Z for each individual u at its initial value, $Z_{x^*}(u)$, a value that may vary from individual to individual. As noted in Section 2, if we attempt to assess empirically the average natural direct effect we need to ensure that we track the Z values of each individual; an aggregate measures of Z in the population will not suffice. Therefore, we cannot generally evaluate the averaged natural direct-effect from population causal effects, such as those obtained in randomized experiments. Formally, this means that Eq. (14) is not reducible to expressions of the form

$$P(Y_x = y) \quad \text{or} \quad P(Y_{xy} = y);$$

the former stands for the causal effect of X on Y (obtained by randomizing X) and the latter stands for the causal effect of X and Z on Y (obtained by randomizing both X and Z).

We now present conditions under which such reduction is nevertheless feasible.

Theorem 3 *(Experimental identification)*
If there exists a set W of covariates, nondescendants of X or Z, such that

$$Y_{xz} \perp\!\!\!\perp Z_{x^*} | W \qquad \text{for all } z \tag{17}$$

(read: Y_{xz} is conditionally independent of Z_{x^*}, given W), then the average natural direct-effect is experimentally identifiable, and it is given by

$$NDE(x, x^*; Y) = \sum_{w,y,z} yP(Y_{x,z} = y|W = w)P(Z_{x^*} = z|W = w]P(W = w) - E(Y_{x^*}) \tag{18}$$

$$= \sum_{w,z} [E(Y_{xz}|w) - E(Y_{x^*z}|w)]P(Z_{x^*} = z|w)P(w) \tag{19}$$

Proof
The probability term in (15) yields (using (17)):

$$P(Y_{x,Z_{x^*}} = y) = \sum_w \sum_z P(Y_{x,z} = y|Z_{x^*} = z, W = w)P(Z_{x^*} = z|W = w]P(W = w) \tag{20}$$

$$= \sum_w \sum_z P(Y_{x,z} = y|W = w)P(Z_{x^*} = z|W = w)P(W = w) \tag{21}$$

Each factor in (21) is identifiable; $P(Y_{x,z} = y|W = w)$, by randomizing X and Z in each stratum of W, and $P(Z_{x^*} = z|W = w)$ by randomizing X in each stratum of W. This proves the assertion in the theorem. Substituting (21) into (15) yields (18) and, using the law of composition $E(Y_{x^*}) = E(Y_{x^* Z_{x^*}})$ (Pearl 2000, p. 229) gives (19), and completes the proof of Theorem 3. □

The conditional independence relation in Eq. (17) might seem intimidating at first, however, this condition can easily be verified from the causal graph associated with the model, using a graphical interpretation of counterfactuals (Pearl, 2000, p. 214-5). Cast in terms of the treatment-aspirin example of Section 2, condition (17) requires that factors controlling the intake of aspirin under no-treatment be independent of factors controlling recovery under treatment and aspirin.

Figure 2(a) illustrates a typical graph associated with estimating the direct effect of X on Y. Z_x represents all direct causes of Z excluding X, which in Fig. 2(a) amount to $\{W, U_4\}$. $Y_{x',z}$ represents all direct causes of Y, excluding X and Z, which in Fig. 2(a) amounts to $\{U_1, U_2\}$. We see indeed that condition (17) is satisfied in this graph—$\{W, U_4\}$ and $\{U_1, U_2\}$ are conditionally independent given W. If W were unobservable, we would not be able to condition on W to gain the desired independence between factors determining Z_x and those determining Y_{xz}; the natural direct effect would not be identified then.

The identification condition of Eq. (17) can be given simple graphical interpretation which reads:

$$(Y \perp\!\!\!\perp Z|W)_{G_{\underline{XZ}}} \tag{22}$$

In words, W d-separates Y from Z in the graph formed by deleting all (solid) arrows emanating from X and Z. This subgraph is shown in Fig. 2(b), and illustrates how W separates Y from Z. The separation condition in (22) is stronger than (17), since the former implies the latter for every pair of values, x and x^*, of X (see (Pearl 2000, p. 214)).

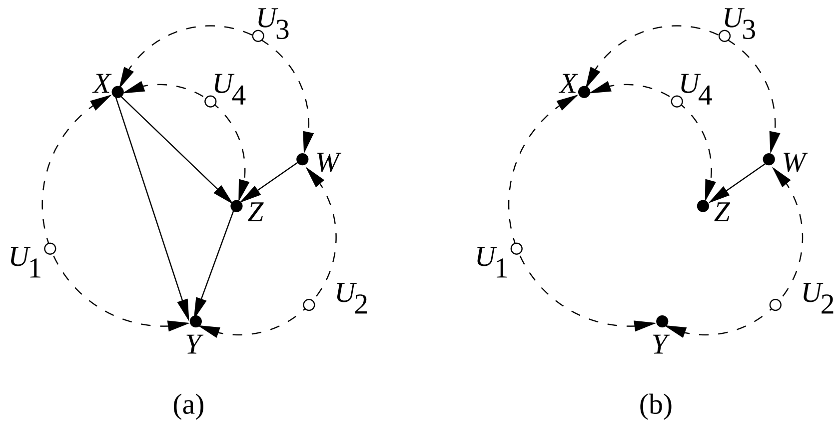

Figure 2: (a) A causal model with latent variables (U's) where the natural direct effect can be identified in experimental studies. (b) The subgraph $G_{\underline{xz}}$ illustrating the criterion of experimental identifiability (Eq. 22): W separates Y from Z.

We are ready now to tackle the harder problem of identifying the natural direct effect from *nonexperimental* data. From Eq. (18) we see that it is sufficient to identify the two (conditional) total effects: $P(Y_{x,z} = y|W = w)$ and $P(Z_{x^*} = z|W = w]$, where W is any set of covariates that satisfies Eq. (17) (or (22)). This yields the following criterion for identification:

Theorem 4 *(Nonexperimental identification)*
The average natural direct-effect $NDE(x, x^*; Y)$ is identifiable in nonexperimental studies if there exists a set W of covariates, nondescendants of X or Z, such that, for all values z and w we have:

(i) $Y_{xz} \perp\!\!\!\perp Z_{x^*}|W$

(ii) $P(Y_{x,z} = y|W = w)$ *and* $P(Y_{x^*z} = y|W = w)$ *are identifiable*

(iii) $P(Z_{x^*} = z|W = w)$ *is identifiable*

Moreover, if conditions (i)-(iii) are satisfied, the natural direct effect is given by (19).

Explicating the identification conditions for (ii) and (iii) per Theorem 2 yields the following two corollaries:

Corollary 1 *(Counterfactual identification criterion)*
The average natural direct-effect $NDE(x, x^*; Y)$ is identifiable in nonexperimental studies if there exist four sets of covariates, W_0, W_1, W_2, and W_3, such that, for all z,

(i) $Y_{xz} \perp\!\!\!\perp Z_{x^*} | W_0$

(ii) $Y_{xz} \perp\!\!\!\perp X | W_0, W_1$

(iii) $Y_z \perp\!\!\!\perp Z | X, W_0, W_1, W_2$

(iv) $Z_{x^*} \perp\!\!\!\perp X | W_0, W_3$

Moreover, if conditions (i)-(iii) are satisfied, the natural direct effect is given by (20), with the following substitutions:

$$P(Y_{xz} = y) = \sum_{w_1, w_2} P(y|w_0, w_1, w_2, z, x) P(w_2|w_0, w_1, x) P(w_1|w_0)$$

and

$$P(Z_{x^*} = z | W_0 = w_0) = \sum_{w_3} P(z|w_0, w_3, x^*) P(w_3|w_0)$$

Corollary 2 *(Graphical identification criterion)*
The average natural direct-effect $NDE(x, x^*; Y)$ is identifiable in nonexperimental studies if there exist four sets of covariates, W_0, W_1, W_2, and W_3, such that

(i) $(Y \perp\!\!\!\perp Z | W_0)_{G_{\underline{XZ}}}$

(ii) $(Y \perp\!\!\!\perp X | W_0, W_1)_{G_{\underline{X}\overline{Z}}}$

(iii) $(Y \perp\!\!\!\perp Z | X, W_0, W_1, W_2)_{G_{\underline{Z}}}$

(iv) $(Z \perp\!\!\!\perp X | W_0, W_3)_{G_{\underline{X}}}$

(v) W_0, W_1, and W_3 contain no descendant of X and W_2 contains no descendant of Z.

As an example for applying these criteria, consider Figure 2(a), and assume that all variables (including the U's) are observable. Conditions (i)-(iv) of Corollary 2 are satisfied if we choose:

$$W_0 = \{W\},\ W_1 = \{U_1, U_2\},\ W_2 = \emptyset \text{ and } W_4 = \{U_4\}$$

or, alternatively,

$$W_0 = \{U_2\},\ W_1 = \{U_1\},\ W_2 = \emptyset \text{ and } W_3 = \{U_3, U_4\}$$

In contrast, the model depicted in Fig. 1 does not permit the identification of the natural direct effect (assuming U is unobserved). The reason is that the only set W_0 satisfying (i) is $W_0 = W$, which violates condition (v) of Corollary 2, because W is a descendant of X.

Comparing Theorems 1 and 2, we conclude that that the identification of natural direct effects requires more stringent assumptions than that of the controlled direct effect. In addition to the added conditions (i) and (iv), conditions (ii)-(iii) must be satisfied at every level w_0 of W_0, where W_0 is selected so as to satisfy condition (i).

For comparison, it is instructive to examine the form that expression (19) takes in Markovian models, where condition (17) is always satisfied with $W = \emptyset$ (since Y_{xz} is independent of all variables in the model). Moreover, in Markovian models we also have the following three relationships:

$$P(Y_{xz} = y) = P(y|x, z) \tag{23}$$

since $X \cup Z$ is the set of Y's parents,

$$P(Z_{x^*} = z) = \sum_s P(z|x^*, s)P(s), \tag{24}$$

$$P(Y_{x,Z_{x^*}} = y) = \sum_s \sum_z P(y|x, z)P(z|x^*, s)P(s) \tag{25}$$

where S stands for the parents of Z, excluding X (or the parents of X, or any other set satisfying the back-door criterion (Pearl 2000, p. 79)). This yields the following corollary of Theorem 3:

Corollary 3 *The average natural direct effect in Markovian models is identifiable from non-experimental data, and it is given by*

$$\begin{aligned} NDE(x, x^*; Y) &= \sum_s \sum_z \sum_y y[P(y|x, z) - P(y|x^*, z)]P(z|x^*, s)P(s) & (26) \\ &= \sum_s \sum_z [E(Y|x, z) - E(Y|x^*, z)]P(z|x^*, s)P(s) & (27) \end{aligned}$$

where S stands for the parents of Z excluding X.

Eq. (27) follows by substituting (25) into (16) and using the identify $E(Y_{x^*}) = E(Y_{x^*Z_{x^*}})$.

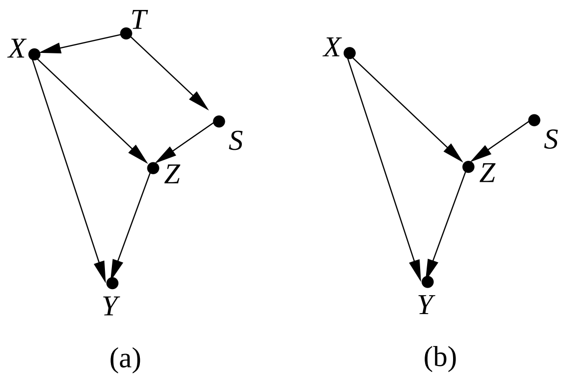

Figure 3: Simple Markovian models for which the natural direct effect is given by Eq. (27) (for (a)) and Eq. (29) (for (b)).

Further insight can be gained by examining simple Markovian models in which the effect of X on Z is not confounded, that is,

$$P(Z_{x^*} = z) = P(z|x^*) \tag{28}$$

In such models, a simple version of which is illustrated in Fig. 3, Eq. (24) can be replace by (28) and (27) simplifies to

$$NDE(x, x^*; Y) = \sum_z [E(Y|x, z) - E(Y|x^*, z)]P(z|x^*) \tag{29}$$

This expression has a simple interpretation as a weighted average of the controlled direct effect $E(Y|x, z) - E(Y|x^*, z)$, where the intermediate value z is chosen according to its distribution under $X = x^*$.

3.4 Natural Indirect Effects: Formulation

As we discussed in Section 2.3, the prescriptive formulation of "controlled direct effect" has no parallel in indirect effects; we therefore use the descriptive formulation, and define *natural* indirect effects at both the unit and population levels. Lacking the controlled alternative, we will drop the title "natural" from discussions, of indirect effects, unless it serves to convey a contrast.

Definition 6 *(Unit-level indirect effect; qualitative)*
An event $X = x$ is said to have an indirect effect on variable Y in situation $U = u$ if the following inequality holds

$$Y_{x^*}(u) \neq Y_{x^*, Z_x(u)}(u) \tag{30}$$

In words, the value of Y changes when we keep X fixed at its reference level $X = x^*$ and change Z to a new value, $Z_x(u)$, the same value that Z would attain under $X = x$.

Taking the difference between the two sides of Eq. (30), we can define the unit level indirect effect as

$$NIE(x, x^*; Y, u) = Y_{x^*, Z_x(u)}(u) - Y_{x^*}(u) \tag{31}$$

and proceed to define its average in the population:

Definition 7 *(Population-level indirect effect)*
The average indirect effect of event $X = x$ on variable Y, denoted $NIE(x, x^*; Y)$, is defined as

$$\begin{aligned} NIE(x, x^*; Y) &= \sum_u NIE(x, x^*; Y, u)P(u) && (32) \\ &= \sum_y P(Y_{x^*, Z_x} = y) y P(y) - E(Y_{x^*}) && (33) \\ &= E(Y_{x^*, Z_x}) - E(Y_{x^*}) && (34) \end{aligned}$$

Comparing Eqs. (16) and (34), we see that the indirect effect associated with the transition from x^* to x is closely related to the direct effect associated with the reverse transition, from x to x^*. In fact, recalling that the difference $E(Y_x) - E(Y_{x^*})$ equals the total effect of $X = x$ on Y,

$$TE(x, x^*; Y) = E(Y_x) - E(Y_{x^*}) \tag{35}$$

we obtain the following theorem:

Theorem 5 *The total, direct and indirect effects obey the following relationships*

$$TE(x, x^*; Y) = NIE(x, x^*; Y) - NDE(x^*, x; Y) \tag{36}$$
$$TE(x, x^*; Y) = NDE(x, x^*; Y) - NIE(x^*, x; Y) \tag{37}$$

In words, the total effect (on Y) associated with the transition from x^* to x is equal to the difference between the indirect effect associated with this transition and the (natural) direct effect associated with the reverse transition, from x to x^*.

As strange as these relationships appear, they produce the standard, additive relation

$$TE(x, x^*; Y) = NIE(x, x^*; Y) + NDE(x, x^*; Y) \tag{38}$$

when applied to linear models (see Section 4). The reason is clear; in linear system the effect of the transition from x^* to x is proportional to $x - x^*$, hence it is always equal and of opposite sign to the effect of the reverse transition. Thus, adding (36) and (37) yields (38).

The main significance of Theorem 5 is that all the information about the kernel $P(Y_{x^*, Z_x} = y)$ of the indirect effect is contained in the kernel of the natural direct effect, $P(Y_{x, Z_{x^*}} = y)$, when viewed as a function of two parameters, x and x^*.

3.5 Natural Indirect Effects: Identification

As with the natural direct effect, we cannot generally evaluate the average indirect-effect from randomized experiments. However, Eqs. (36) and (37) show that the indirect effect is identified whenever both the total and the (natural) direct effect are identified for all x and x^*. Moreover, the identification conditions and the resulting expressions for indirect effects are identical to the corresponding ones for direct effects (Theorems 2 and 3), save for a simple exchange of the indices x and x^*.

The following theorem explicates the main aspects of this relationship.

Theorem 6 *If there exists a set W of covariates, nondescendants of X or Z, such that*

$$Y_{x^*z} \perp\!\!\!\perp Z_x | W \tag{39}$$

for all values z, then the average indirect-effect is experimentally identifiable, and it is given by

$$NIE(x, x^*; Y) = \sum_{w,y,z} yP(Y_{x^*z} = y|W = w)P(Z_x = z|W = w)P(W = w) - E(Y_{x^*}) \tag{40}$$
$$= \sum_{w,z} E(Y_{x^*z}|w)[P(Z_x = z|w) - P(Z_{x^*} = z|w)]P(w) \tag{41}$$

Moreover, the average indirect effect is identified in nonexperimental studies whenever the following causal effects are identified for all z and w:

$$E(Y_{x^*z}|w), \ P(Z_x = z|w) \text{ and } P(Z_{x^*} = z|w),$$

with W satisfying Eq. (39)

The proof follows the same steps as in the proof of Theorem 2. We should note, though, that the expression in Eq. (41) can no longer be interpreted as a weighted average of the controlled direct effect $E(Y_{xz}) - E(Y_{x^*z})$, as obtained for the natural direct effect (see 19) and (29). Rather, (41) should be interpreted as a difference between two weighted averages of the controlled expectation $E(Y_{x^*z})$, one weighing Z according to the distribution of Z_x, the second according to the distribution of Z_{x^*}. For example, in the simple Markovian model depicted in Fig. 3, Eq. (41) reduces to

$$NIE(x, x^*; Y) = \sum_z E(Y|x^*, z)[P(z|x) - P(z|x^*)] \tag{42}$$

$$= \sum_z E(Y|x^*, z)P(z|x) - E(Y|x^*) \tag{43}$$

Contrasted with Eq. (29), we see that the kernel of both equations are identical, save for exchanging x and x^* (see also (16) and (34)). The kernel of the indirect effect (43) reads

$$E(Y_{x^*Z_x}) = \sum_z E(Y|x^*, z)P(z|x)$$

while that of the direct effect (29) reads

$$E(Y_{xZ_{x^*}}) = \sum_z E(Y|x, z)P(z|x^*)$$

Both expressions average the expectation of Y, conditioned on X and Z. However, the expression for the indirect effect fixes X at the reference value x^*, and lets z vary according to its distribution under the post-transition value of $X = x$. The expression for the direct effect fixes X at x, and lets z vary according to its distribution under the reference conditions $X = x^*$.

3.6 General Path-specific Effects

The analysis of the last section suggests that effect decomposition can best be understood in terms of a *path-deactivation process*, where a selected set of paths, rather than nodes, are forced to remain inactive during the transition from $X = x^*$ to $X = x$. In Fig. 4(b) for example, the indirect effect of X on Y is computed by deactivating the direct link $X \to Y$, keeping the links $X \to Z$ and $Z \to Y$ active, and then examining the effect (on Y) of the transition from x^* to x. The deactivation process cannot be accomplished by fixing the value of X (for that would deactivate the path $X \to Z$ as well) but, rather, by "freezing" the X-input in the function $y = f(x, z, u)$ at a constant $X = x^*$ while, at the same time, allowing the X-input in the function $z = F_Z(x, u)$ to vary (see Fig. 4(b)). Such selective freezing cannot be expressed in term of a straightforward application of the $do(X = x)$ operator (as in (2)), and necessitates the use of nested counterfactuals (as in (31)).

Selective deactivation can also be read into the definition of natural direct effects. In the model of Fig. 4, for example, the direct effect $X \to Y$ is measured by freezing the inputs to the links $X \to Z$ and $Z \to Y$ at their pre-transition levels of x^* and $z^*(u) = Z_{x^*}(u)$, respectively, as shown in Fig. 4(c). Whereas the $do(X = x)$ operator (as represented by the subscript notation $Y_x(u)$) deactivates uniformly all inputs to a given node, selective

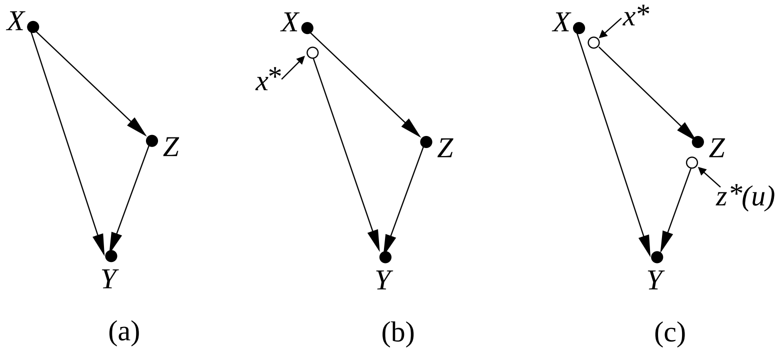

Figure 4: The indirect and direct effects (of X on Y) in model (a) are interpreted as path-deactivation schema in models (b) and (c), respectively.

deactivation permits some of the inputs to vary freely under the influence of their parents in the model. Moreover, whereas the $do(X = x)$ operator fixes variables to predetermined constants, selective deactivation ties inputs to values (e.g., $z^*(u)$) that are u-dependent.

In general, a convenient way of formulating this process is to consider the deactivation operation as creating a new model, in which each structural function f_i in M is replaced with a new function of a smaller set of arguments, since some of the arguments are replaced by constants. This formulation will allow us to define the effect transmitted through any selected set of paths as the total effect (of x on Y) in a new model, formed by deactivating the remaining paths in the diagram. The following definition expresses this idea formally.

Definition 8 *(path-specific effect)*
*Let G be the the causal graph associated with model M, and let g be an edge-subgraph of G containing the paths selected for effect analysis. The g-*specific effect *of x on Y (relative to reference x^*) is defined as the total effect of x on Y in a modified model M^*_g formed as follows. Let each parent set PA_i in G be partitioned into two parts*

$$PA_i = \{PA_i(g), PA_i(\overline{g})\} \tag{44}$$

where $PA_i(g)$ represents those members of PA_i that are linked to X_i in g, and $PA_i(\overline{g})$ represents the complementary set, from which there is no link to X_i in g. We replace each function $f_i(pa_i, u)$ with a new function $f^*_i(pa_i, u; g)$, defined as

$$f^*_i(pa_i, u; g) = f_i(pa_i(g), pa^*_i(\overline{g}), u) \tag{45}$$

where $pa^*_i(\overline{g})$ stands for the values that the variables in $PA_i(\overline{g})$ would attain (in M and u) under $X = x^*$ (that is, $pa^*_i(\overline{g}) = PA_i(\overline{g})_{x^*}$). The g-specific effect of x on Y, denoted $SE_g(x, x^*; Y, u)_M$ is defined as

$$SE_g(x, x^*; Y, u)_M = TE(x, x^*; Y, u)_{M^*_g}. \tag{46}$$

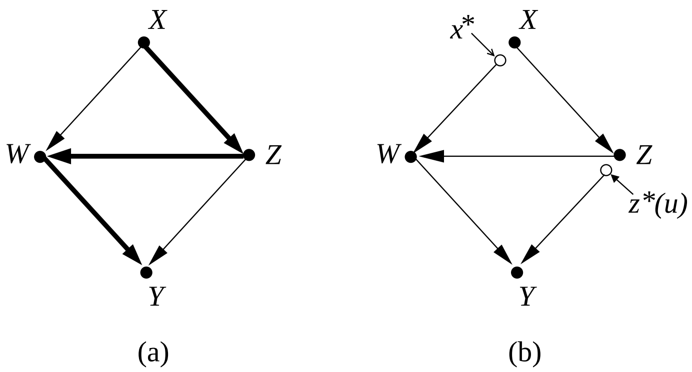

Figure 5: The path-specific effect transmitted through $X \to Z \to W \to Y$ (heavy lines) in (a) is equal to the total effect transmitted through the model in (b), where x^* and $z^*(u)$ are constants. (By convention, u is not shown in the diagram.)

We demonstrate this construction in the model of Fig. 5 which stands for the equations:

$$\begin{aligned} z &= f_Z(x, u_Z) \\ w &= f_W(z, x, u_W) \\ y &= f_Y(z, w, u_Y) \end{aligned}$$

where u_Z, u_W, and u_Y are the components of u that enter the corresponding equations. Assuming that u_Z, u_W, and u_Y are statistically independent (the Markov assumption), we wish to evaluate the path-specific effect (of X on Y) defined by the subgraph $g : X \to Z \to W \to Y$. Clearly, we cannot isolate this subgraph by holding Z or W constant, for both must vary in the process. Rather, we isolate the desired effect by fixing the appropriate subset of arguments in each equation. In other words, we replace $f_W(z, x, u_W)$ with $f_W(z, x^*, u_W)$ and $f_Y(z, w, u_Y)$ with $f_Y(z^*(u), w, u_Y)$, where $z^*(u) = Z_{x^*}(u)$.

The modified model $M^*(g)$ now reads:

$$\begin{aligned} z &= f_Z(x, u_Z) \\ w &= f_W(z, x^*, u_W) \text{ and} \\ y &= f_Y(z^*(u), w, u_Y) \end{aligned} \tag{47}$$

and our task amounts to computing the total effect of x on Y in $M^*(g)$, or

$$\begin{aligned} &TE(x, x^*; Y, u)_{M_g^*} = \\ &\quad = f_Y(z^*(u), f_W(f_Z(x, u_Z), x^*, u_W), u_Y) - Y_{x^*}(u) \\ &\quad = f_Y(z^*(u), f_W(f_Z(x, u_Z), x^*, u_W), u_Y) - f_Y(z^*(u), f_W(f_Z(x^*, u_Z), x^*, u_W), u_Y) \end{aligned} \tag{48}$$

where $z^*(u) = f_Z(x^*, u_Z)$.

The average g-specific effect is obtained by taking the expectation (over u) of (48), yielding

$$
\begin{aligned}
TE(x,x^*;Y)_{M_g^*} &= \\
&= E_u[f_Y(z^*(u), f_W(f_Z(x,u),x^*,u),u)] - E_u(Y_{x^*}(u)) \quad (49)\\
&= E_u[f_Y(z^*(u_Z), f_W(f_Z(x,u_Z),x^*,u_W),u_Y)] - E_u[f_Y(z^*(u_Z), f_W(f_Z(x^*,u_Z),x^*,u_W),u_Y)] \quad (50)
\end{aligned}
$$

where $z^*(u) = f_Z(x^*, u_Z)$.

The nested functions in (48) and (50) can be given counterfactual interpretation, using nested subscripts. For example, Eq. (50) can be written:

$$TE(x,x^*;Y)_{M_g^*} = E(Y_{Z_{x^*}W_{Z_x,x^*}}) - E(Y_{x^*}) \tag{51}$$

We can now ask under what conditions this expression is identifiable. Using the same algebra as we did in Section 3.5, we obtain

$$P(Y_{Z_{x^*},W_{Z_x,x^*}}) \;=\; \sum_{z,w} P(Y_{zw}=y|Z_{x^*}=z, W_{Z_x,x^*}=w)P(Z_{x^*}=z, W_{Z_x,x^*}=w)$$

and observe that, unfortunately, this expression is not identifiable even in Markovian models. Whereas the first term is reducible (in Markovian models) to $P(Y_{zw}=y)$ by virtue of the independence

$$Y_{z,w} \perp\!\!\!\perp Z_{x^*}, W_{Z_x,x^*},$$

the second term involves an irreducible, joint probability of counterfactuals. Each of the two variables, Z_{x^*} and W_{Z_x,x^*}, depends directly on the exogenous variable U_Z (i.e., the disturbance term associated with the equation of Z) and, hence, there is no set of covariates that can render these variables conditionally independent of one another.

This example suggests a general recursive scheme of expressing and analyzing the g-specific effect associated with a subgraph g of G. Denoting by $P_g^*(Y_x=y)$ the probability of $Y=y$ in model M_g^*, and partitioning the parents of Y into $PA_Y = \{S,T\}$, where $S = PA_Y(\overline{g})$ and $T = PA_Y(g)$ we can write:

$$P_g^*(Y_x=y) = P(Y_{S_{x^*},T_x^*(g)}=y)$$

where $T_x^*(g)$ stands for the value of T_x in M_g^*. Equivalently,

$$
\begin{aligned}
P_g^*(Y_x=y) \;&=\; \sum_{s,t} P(Y_{st}=y|S_{x^*}=s, T_x^*(g)=t)\\
&\qquad P(S_{x^*}=s, T_x^*(g)=t)
\end{aligned}
$$

Further assuming that U_y is independent of other U's in the model, we obtain.

$$P_g^*(Y_x=y) = \sum_{s,t} P(Y_{st}=y)P(S_{x^*}=s, T_x^*(g)=t) \tag{52}$$

Thus, Eq. (52) can be computed recursively, starting with Y, proceeding to the variables in T, then to T's parents, and so on. The result would be an expression for the kernel of the g-specific effects which, combined with $E(Y_{x^*}$, gives the average g-specific effect of x on Y, as in (51). However, as we observed in the example of Fig. 5, the second term in Eq. (52) is not identifiable even in Markovian models, although the first term can immediately be reduced to

$$P(Y_{st} = y) = P(y|st)$$

(because the union of S and T make up all the parents of Y).

To appreciate the generality of Eq. (52), we note that the expressions we obtained in the preceding sections, for the direct and indirect effects can be viewed as special cases of (52). For the indirect effect, we consider a subgraph g formed by deleting from G the arrow $X \rightarrow Y$, that is, $S = PA_Y(\overline{g}) = X$, and $T = PA_Y(g) = PA_Y \backslash X = Z$. Substituting $S = X$ and $T = Z$ in (52), we have

$$P(S_{x^*} = s) = \begin{cases} 1 & \text{if } s = x^* \\ 0 & \text{otherwise} \end{cases}$$

and

$$P^*_g(T_x = t) = P(Z_x = z)$$

Therefore,

$$\begin{aligned} P^*_g(Y_x = y) &= \sum_z P(Y_{x^*z} = y)P(Z_x = z) \\ &= \sum_z P(y|x^*z)P(Z_x = z) \end{aligned}$$

from which Eq. (43) follows.

For the direct effect, we consider a subgraph g in which X is the only parent of Y, that is,

$$T = PA_Y(g) = X \text{ and } S = PA_Y(\overline{g}) = Z.$$

Substituting $T = X$ and $S = Z$ in (52) we have

$$P^*_g(T_x = t) = \begin{cases} 1 & \text{if } t = x \\ 0 & otherwise \end{cases}$$

Therefore,

$$\begin{aligned} P^*_g(Y_x = y) &= \sum_z P(Y_{xz} = y)P(Z_{x^*} = z) \\ & \sum_z P(y|xz)P(Z_{x^*} = z) \end{aligned}$$

from which (27) follows.

3.7 Relations to Linear Models

Example 1 *Let M stand for the structural model*

$$\begin{aligned} x &= h(w) \\ z &= g(x, u) \\ y &= f(x, z, v) \end{aligned}$$

where w, v, and u are (possibly correlated) error terms, and h, f, and g are arbitrary functions.

The total effect of X on Y is given (see (35)) by

$$TE(x, x^*; Y) = E[f(x, g(x, u), v)] - E[f(x^*, g(x^*, u), v],$$

the controlled direct effect is given (see (3)) by

$$CDE_z(x, x^*; Y) = E[f(x, z, v)] - E[f(x^*, z, v)],$$

while the average natural direct and indirect effects are given (see (16) and (34)) by

$$\begin{aligned} NDE(x, x^*; Y) &= E[f(x, g(x *, u), v)] - E[f(x^*, g(x^*, u), v)] \\ NIE(x, x *; Y) &= E[f(x^*, g(x, u), v)] - E[f(x^*, g(x^*, u), v)] \end{aligned}$$

where the expectations are taken over u and v (distributed as in M).

If M is linear, we have

$$\begin{aligned} x &= w \\ z &= ax + u \\ y &= bx + cz + v \end{aligned}$$

and the effects simplify to:

$$\begin{aligned} TE(x, x^*; Y) &= (b + ca)(x - x^*) \\ CDE_z(x, x^*; Y) &= b(x - x^*) \\ NDE(x, x^*; Y) &= b(x - x^*) \\ NIE(x, x^*; Y) &= ac(x - x^*) \end{aligned}$$

As expected, we see that the controlled and natural direct effects coincide, and are both governed by the path coefficient b. Likewise, we see that the natural indirect effect is governed by the product ac, and that the total effect is the sum of the direct and indirect effect.

4 Policy Implications: An Example

In this section we explicate the policy implications of direct and indirect effect through an example by Jacques A. Hagenaars.[6]

Example 2 *The influence of Education (X) on Political Preferences (Y) is mediated through Economic Status (Z), since higher educated people get the better jobs and earn more money, and also through a Cultural Mechanism (S) which has to do with the contents of the education and the accompanying socialization processes at school. It is important to know what the causal directions (signs) of these two processes are and which one is the dominant one. (At least in The Netherlands they did tend to go into different directions, the first leading to a right wing preference, the second to a left wing).*

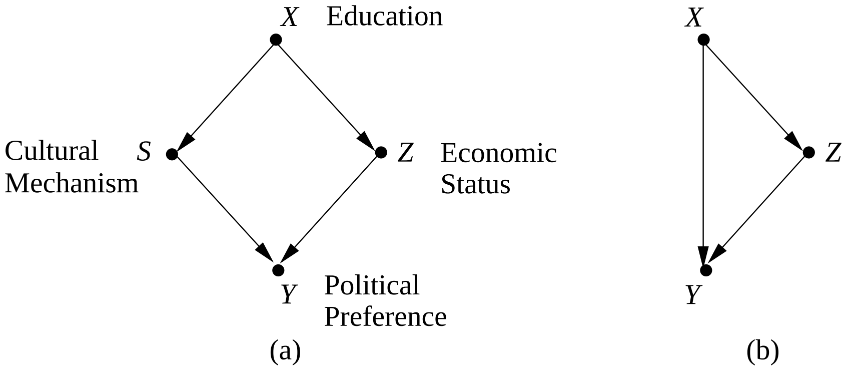

Figure 6: (a) The effect of Education on Political Preference mediated by two mechanisms. (b) Same when S is not observed.

A simple Markovian model of this example is shown in Fig. 6, where it is assumed that all exogenous variables (not shown explicitly) are jointly independent. If all four variables are observed, we can easily compute the relative contributions of the two indirect paths:

$$X \rightarrow Z \rightarrow Y \text{ and } X \rightarrow S \rightarrow Y.$$

The first is measured through $E(Y_{xz}) = E(Y|x,z)$ and the second through $E(Y_{xz}) = E(Y|x,s)$—the contribution of each path is measured by blocking the other. These expectations would depend of course on the level at which we hold the blocking variable, and that would depend on the policy options that we consider.

The difficulty surfaces in cases where a significant direct effect is present. In such cases, since the direct effect cannot be blocked by controlling variables, the indirect effect cannot be assessed by blocking, and the expectations $E(Y_{xs})$ and $E(Y_{xz})$ no longer represent indirect effects. In our example, if variable S was unmeasurable (or not modeled), its mediation would be conceived as a direct effect between X and Y (Fig. 6(b)), the quantity $E(Y_{xz})$ would give us the controlled direct effect (as in (3)), but the quantity $E(Y_{xs})$ would not be expressible in our model. This means that we would not be able to compute the contribution of the indirect path $X \rightarrow Z \rightarrow Y$, except by comparing the total and direct effects. Such comparison, however, would not give us meaningful reading on the indirect effect through Z because, in nonlinear systems, the total and direct effect can both be equal to zero while the indirect effect may be significant.

The question also arises whether it is at all meaningful to speak about indirect effects in situations where we cannot block the direct effect by some policy. In our example, if we cannot observe or control the variable S, it is hard to think of a policy problem that would benefit from knowing the indirect effects.

Example 2 helps us envision a sense in which the indirect effect is relevant to policy making, albeit not to straightforward interventional policies. Instead of blocking the direct effect, we may envision a change in the educational program that would influence *only* economic

[6] This example was part of Hagenaars comments on my paper (Pearl 1998), posted on ⟨http://bayes.cs.ucla.edu/BOOK-2K/hagenaars.html⟩.

status but not the associated cultural mechanism. Such a change could be implemented, for example, if we enhance the earning-specific aspects of education (say, by emphasizing job-related skills, or giving instructions on job-interview techniques), while keeping other aspects of education unaltered. Such fine-grained policy options are not represented in our coarse-grained model, since education is designated a single variable with levels representing *years* of schooling rather than *type* of schooling; every change in educational level will be transmitted simultaneously through both paths.

This consideration exposes the weakness of variable fixing as a sole representative of policy making. Our fine-grain policy question should be posed in terms of the natural indirect effect, not in terms of controlled indirect effect (with S controlled). In other words, without thinking about S, we should ask (and answer) the following question:

Q: What will the distribution of political preferences be among low-education people, had they received high earning-related education (similar to the one provided in the high-education section), but otherwise, all other aspects of education will be kept unaltered?

An equivalent way of phrasing the question is:

Q': What will the distribution of political preferences be among low-education people, had they received high education but, otherwise, each person would go through the same cultural process (S) as he/she actually did?

Both question are formalized through the average natural indirect effect, $NIE(x, x^*; Y)$, as defined in (34). Since our model is Markovian, Theorem 6 ensures that NIE is identifiable from observational studies, and it is given in (42) and (43). Letting x_0 = low education, and x_1 = high education, we substitute x_0 for x^*, x_1 for x, and we obtain:

$$NIE(x_1, x_0; Y) = \sum_z E(Y|x_0, z)[P(z|x_1) - P(z|x_0)] \tag{53}$$

To illustrate the computation of $NIE(x, x^*; Y)$ in concrete numbers, we will make some simplifying assumptions. First, we assume that both Z and Y are binary, with $Z = \{z_0, z_1\} = \{$*low income, high income*$\}$, $Y = \{y_0, y_1\} = \{$*left preference, right preference*$\}$. Next we will assume the (fictitious) data shown in Table 1, which displays separately the probabilities $P(y, z|x_0)$ and $P(y, z|x_1)$ for each of the four (y, z) cells.

The table embodies strong correlation between education and income, as well as strong dependency of political preferences on both income and education; right-preference seems to be strong only among people with both high-income and high-education.

To make our calculations meaningful, it is convenient to extract from the table the following parameters:

- $p = P(z_1|x_0) = 0.10$ is the probability that a person with low education (x_0) would still be rich (z_1).
- $p' = P(z_0|x_1) = 0.05$ is the probability that a person with high education (x_1) would still be poor (z_0).
- $q = P(y_1|x_0, z_1) = 0.15$ is the probability that a rich person (z_1) with low education (x_0) would lean to the right (y_1).

	Low Education (x_0)		High Education (x_1)	
	Low Income (z_0)	High Income (z_1)	Low Income (z_0)	High Income (z_1)
Left-preference (y_0)	0.90	0.085	0.05	0.19
Right-preference (y_1)	0	0.015	0	0.76

Table 1: Fictitious data for Example 2, showing the percentage of people in each income/preference category, for High-Education and Low-Education groups.

- $q' = P(y_0|x_1, z_1) = 0.20$ is the probability that a rich person (z_1) with high education (x_1) would lean to the left (y_0).
- No poor person would lean to the right, i.e., $P(y_1|x, z_0) = 0$ for all x.

Substituting in (53), we obtain

$$\begin{aligned}&\sum_z P(y_1|x_0, z)P(z|x_1)\\ &\quad = P(y_1|x_0, z_0)P(z_0|x_1) + P(y_1|x_0, z_1)P(z_1|x_1)\\ &\quad = q(1 - p') = 0.15(1 - 0.05) = 0.1425\end{aligned}$$

and

$$\begin{aligned}\sum_z P(y_1|x_0, z)P(z|x_0) &= P(y_1|x_0)\\ &= P(y_1|x_0, z_0)P(z_0|x_0) + P(y_1|x_0, z_1)P(z_1|x_0)\\ &= qp = 0.15 \times 0.10 = 0.015\end{aligned}$$

Thus, the probability of right-preference among people with low education would increase ten-fold, from $qp = 0.015$ to $q(1 - p') = 0.1425$, due to the indirect effect of higher earnings associated with education.

In comparison, the direct effect of higher education is given by (see (29));

$$\begin{aligned}NDE(x_1, x_0; Y) &= \sum_z P(y|x_1, z)P(z|x_0) - P(x_1|x_0)\\ &\quad = P(y_1|x_1, z_0)P(z_0|x_0) + P(y_1|x_1, z_1)P(z_1|x_0) - P(y_1|x_0)\\ &\quad = q'p - qp = 0.20 \times 0.10 - 0.015 = 0.005\end{aligned}$$

The left term, $q'p$, represents the exceptional cases of those who got rich despite low education, and who would have become right-leaning had they been given higher education.

Both the direct and indirect effects are small, being on the order of the fractions of exceptional cases. For comparison, consider the total effect of education on Y.

$$
\begin{aligned}
T[x_0, x_1; Y] &= P(y_1|x_1) - P(y_1|x_0) \\
&= \sum_z P(y_1|x_1, z)P(z|x_1) - P(y_1|x_0) \\
&= P(y_1|x_1, z_0)P(z_0|x_1) + P(y_1|x_1, z_1)P(z_1|x_1) - qp \\
&= 0 + (1-q')(1-p') - qp = (1-0.20)(1-0.05) - 0.015 = 0.745
\end{aligned}
$$

This is a substantial effect; the fraction of right-leaning persons (among the uneducated) would increase from 0.015 to 0.76, if given higher education. This illustrates the nature of nonlinear interactions: both direct and indirect effects may be small and, still, the total effect can be large.

5 Appendix - Causal Models and Counterfactuals

This appendix presents a brief summary of the structural-equation semantics of counterfactuals as defined in Balke and Pearl (1995), Galles and Pearl (1997, 1998), and Halpern (1998). Related approaches have been proposed in Simon and Rescher (1966) (see footnote 11) and Robins (1986). For detailed exposition of the structural account and its applications see [Pearl, 2000].

Causal models are generalizations of the structural equations used in engineering, biology, economics and social science.[7] World knowledge is represented as a collection of stable and autonomous relationships called "mechanisms," each represented as a function, and changes due to interventions or hypothetical changes are treated as local modifications of these functions.

A causal model is a mathematical object that assigns truth values to sentences involving causal relationships, actions, and counterfactuals. We will first define causal models, then discuss how causal sentences are evaluated in such models. We will restrict our discussion to recursive (or feedback-free) models; extensions to non-recursive models can be found in Galles and Pearl (1997, 1998) and Halpern (1998).

Definition 9 (*Causal model*)
A causal model *is a triple*

$$M = \langle U, V, F \rangle$$

where

(i) *U is a set of variables, called* exogenous. *(These variables will represent background conditions, that is, variables whose values are determined outside the model.)*

(ii) *V is an ordered set $\{V_1, V_2, \ldots, V_n\}$ of variables, called* endogenous. *(These represent variables that are determined in the model, namely, by variables in $U \cup V$.)*

[7]Similar models, called "neuron diagrams" [Lewis, 1986, p. 200; Hall, 1998] are used informally by philosophers to illustrate chains of causal processes.

(iii) *F is a set of functions $\{f_1, f_2, \ldots, f_n\}$ where each f_i is a mapping from $U \times (V_1 \times \ldots \times V_{i-1})$ to V_i. In other words, each f_i tells us the value of V_i given the values of U and all predecessors of V_i. Symbolically, the set of equations F can be represented by writing* [8]

$$v_i = f_i(pa_i, u_i) \quad i = 1, \ldots, n$$

where pa_i is any realization of the unique minimal set of variables PA_i in V (connoting parents) *sufficient for representing f_i*[9]*. Likewise, $U_i \subseteq U$ stands for the unique minimal set of variables in U that is sufficient for representing f_i.*

Every causal model M can be associated with a directed graph, $G(M)$, in which each node corresponds to a variable in V and the directed edges point from members of PA_i toward V_i (by convention, the exogenous variables are usually not shown explicitly in the graph). We call such a graph the *causal graph* associated with M. This graph merely identifies the endogenous variables PA_i that have direct influence on each V_i but it does not specify the functional form of f_i.

For any causal model, we can define an *action* operator, $do(x)$, which, from a conceptual viewpoint, simulates the effect of external action that sets the value of X to x and, from a formal viewpoint, transforms the model into a *submodel*, that is, a causal model containing fewer functions.

Definition 10 (*Submodel*)
Let M be a causal model, X be a set of variables in V, and x be a particular assignment of values to the variables in X. A submodel *M_x of M is the causal model*

$$M_x = \langle U, V, F_x \rangle$$

where

$$F_x = \{f_i : V_i \notin X\} \cup \{X = x\} \tag{54}$$

In words, F_x is formed by deleting from F all functions f_i corresponding to members of set X and replacing them with the set of constant functions $X = x$.

If we interpret each function f_i in F as an independent physical mechanism and define the action $do(X = x)$ as the minimal change in M required to make $X = x$ hold true under any u, then M_x represents the model that results from such a minimal change, since it differs from M by only those mechanisms that directly determine the variables in X. The transformation from M to M_x modifies the algebraic content of F, which is the reason for the name *modifiable structural equations* used in [Galles and Pearl, 1998].[10]

[8] We use capital letters (e.g., X, Y) as names of variables and sets of variables, and lower-case letters (e.g., x, y) for specific values (called realizations) of the corresponding variables.

[9] A set of variables X is *sufficient* for representing a given function $y = f(x, z)$ if f is trivial in Z—that is, if for every x, z, z' we have $f(x, z) = f(x, z')$.

[10] Structural modifications date back to Marschak (1950) and Simon (1953). An explicit translation of interventions into "wiping out" equations from the model was first proposed by Strotz and Wold (1960) and later used in Fisher (1970), Sobel (1990), Spirtes et al. (1993), and Pearl (1995). A similar notion of sub-model is introduced in Fine (1985), though not specifically for representing actions and counterfactuals.

Definition 11 (*Effect of action*)
Let M be a causal model, X be a set of variables in V, and x be a particular realization of X. The effect of action $do(X = x)$ *on M is given by the submodel M_x.*

Definition 12 (*Potential response*)
Let Y be a variable in V, let X be a subset of V, and let u be a particular value of U. The potential response *of Y to action $do(X = x)$ in situation u, denoted $Y_x(u)$, is the (unique) solution for Y of the set of equations F_x.*

We will confine our attention to actions in the form of $do(X = x)$. Conditional actions, of the form "$do(X = x)$ if $Z = z$" can be formalized using the replacement of equations by functions of Z, rather than by constants [Pearl, 1994]. We will not consider disjunctive actions, of the form "$do(X = x$ or $X = x')$", since these complicate the probabilistic treatment of counterfactuals.

Definition 13 (*Counterfactual*)
Let Y be a variable in V, and let X be a subset of V. The counterfactual expression "The value that Y would have obtained, had X been x" is interpreted as denoting the potential response $Y_x(u)$.

Definition 5 thus interprets the counterfactual phrase "had X been x" in terms of a hypothetical external action that modifies the actual course of history and enforces the condition "$X = x$" with minimal change of mechanisms. This is a crucial step in the semantics of counterfactuals [Balke and Pearl, 1994], as it permits x to differ from the actual value $X(u)$ of X without creating logical contradiction; it also suppresses abductive inferences (or backtracking) from the counterfactual antecedent $X = x$.[11]

It can easily be shown [Galles and Pearl, 1997] that the counterfactual relationship just defined, $Y_x(u)$, satisfies the following two properties:
Effectiveness:
For any two disjoint sets of variables, Y and W, we have

$$Y_{yw}(u) = y. \tag{55}$$

In words, setting the variables in W to w has no effect on Y, once we set the value of Y to y.
Composition:
For any two disjoint sets of variables X and W, and any set of variables Y,

$$W_x(u) = w \Longrightarrow Y_{xw}(u) = Y_x(u). \tag{56}$$

In words, once we set X to x, setting the variables in W to the same values, w, that they would attain (under x) should have no effect on Y. Furthermore, effectiveness and composition are *complete* whenever M is recursive (i.e., $G(M)$ is acyclic)

[11] Simon and Rescher (1966, p. 339) did not include this step in their account of counterfactuals and noted that backward inferences triggered by the antecedents can lead to ambiguous interpretations.

[Galles and Pearl, 1998, Halpern, 1998], that is, every property of counterfactuals that follows from the structural model semantics can be derived by repeated application of effectiveness and composition.

A corollary of composition is a property called *consistency* by [Robins, 1987]:

$$(X(u) = x) \Longrightarrow (Y_x(u) = Y(u)) \tag{57}$$

Consistency states that, if in a certain context u we find variable X at value x, and we intervene and set X to that same value, x, we should not expect any change in the response variable Y. Composition and consistency are used in several derivations of Section 3.

The structural formulation generalizes naturally to probabilistic systems, as is seen below.

Definition 14 (*Probabilistic causal model*)
A probabilistic causal model *is a pair*

$$\langle M, P(u) \rangle$$

where M is a causal model and $P(u)$ is a probability function defined over the domain of U.

$P(u)$, together with the fact that each endogenous variable is a function of U, defines a probability distribution over the endogenous variables. That is, for every set of variables $Y \subseteq V$, we have

$$P(y) \triangleq P(Y = y) = \sum_{\{u \ | \ Y(u)=y\}} P(u) \tag{58}$$

The probability of counterfactual statements is defined in the same manner, through the function $Y_x(u)$ induced by the submodel M_x. For example, the *causal effect* of x on y is defined as:

$$P(Y_x = y) = \sum_{\{u \ | \ Y_x(u)=y\}} P(u) \tag{59}$$

Likewise, a probabilistic causal model defines a joint distribution on counterfactual statements, i.e., $P(Y_x = y, Z_w = z)$ is defined for any sets of variables Y, X, Z, W, not necessarily disjoint. In particular, $P(Y_x = y, X = x')$ and $P(Y_x = y, Y_{x'} = y')$ are well defined for $x \neq x'$, and are given by

$$P(Y_x = y, X = x') = \sum_{\{u | Y_x(u)=y \ \& \ X(u)=x'\}} P(u) \tag{60}$$

and

$$P(Y_x = y, Y_{x'} = y') = \sum_{\{u \ | \ Y_x(u)=y \ \& \ Y_{x'}(u)=y'\}} P(u). \tag{61}$$

When x and x' are incompatible, Y_x and $Y_{x'}$ cannot be measured simultaneously, and it may seem meaningless to attribute probability to the joint statement "Y would be y if $X = x$ *and* Y would be y' if $X = x'$." Such concerns have been a source of recent objections to treating counterfactuals as jointly distributed random variables [Dawid, 2000]. The definition of Y_x and $Y_{x'}$ in terms of two distinct submodels, driven by a standard probability space over U, demonstrates that joint probabilities of counterfactuals have solid mathematical and conceptual underpinning and, moreover, these probabilities can be encoded rather parsimoniously using $P(u)$ and F.

References

[Balke and Pearl, 1994] A. Balke and J. Pearl. Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume I, pages 230–237. MIT Press, Menlo Park, CA, 1994.

[Balke and Pearl, 1995] A. Balke and J. Pearl. Counterfactuals and policy analysis in structural models. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 11–18. Morgan Kaufmann, San Francisco, 1995.

[Bollen, 1989] K.A. Bollen. *Structural Equations with Latent Variables.* John Wiley, New York, 1989.

[Cartwright, 1989] N. Cartwright. *Nature's Capacities and Their Measurement.* Clarendon Press, Oxford, 1989.

[Dawid, 2000] A.P. Dawid. Causal inference without counterfactuals (with comments and rejoinder). *Journal of the American Statistical Association*, 95(450):407–448, June 2000.

[Duncan, 1975] O.D. Duncan. *Introduction to Structural Equation Models.* Academic Press, New York, 1975.

[Fine, 1985] K. Fine. *Reasoning with Arbitrary Objects.* B. Blackwell, New York, 1985.

[Fisher, 1970] F.M. Fisher. A correspondence principle for simultaneous equations models. *Econometrica*, 38(1):73–92, January 1970.

[Galles and Pearl, 1997] D. Galles and J. Pearl. Axioms of causal relevance. *Artificial Intelligence*, 97(1-2):9–43, 1997.

[Galles and Pearl, 1998] D. Galles and J. Pearl. An axiomatic characterization of causal counterfactuals. *Foundation of Science*, 3(1):151–182, 1998.

[Gastwirth, 1997] J.L. Gastwirth. Statistical evidence in discrimination cases. *Journal of the Royal Statical Society, Series A*, 160(Part 2):289–303, 1997.

[Goldberger, 1972] A.S. Goldberger. Structural equation models in the social sciences. *Econometrica: Journal of the Econometric Society*, 40:979–1001, 1972.

[Hagenaars, 1993] J. Hagenaars. *Loglinear Models with Latent Variables.* Sage Publications, Newbury Park, CA, 1993.

[Hall, 1998] N. Hall. Two concepts of causation, 1998. In press.

[Halpern, 1998] J.Y. Halpern. Axiomatizing causal reasoning. In G.F. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence*, pages 202–210. Morgan Kaufmann, San Francisco, CA, 1998.

[Hesslow, 1976] G. Hesslow. Discussion: Tow notes on the probabilistic approach to causality. *Philosophy of Science*, 43:290–292, 1976.

[Kline, 1998] R.B. Kline. *Principles and Practice of Structural Equation Modeling.* The Guilford Press, New York, 1998.

[Kuroki and Miyakawa, 1999] M. Kuroki and M. Miyakawa. Identifiability criteria for causal effects of joiont interventions. *Journal of the Japan Statistical Society*, 29(2):105–117, 1999.

[Lewis, 1973] D. Lewis. Counterfactuals and comparative probability. *Journal of Philosophical Logic*, 2, 1973.

[Lewis, 1986] D. Lewis. *Philosophical Papers.* Oxford University Press, New York, 1986.

[Marschak, 1950] J. Marschak. Statistical inference in economics. In T. Koopmans, editor, *Statistical Inference in Dynamic Economic Models*, pages 1–50. Wiley, New York, 1950. Cowles Commission for Research in Economics, Monograph 10.

[Mueller, 1996] R.O. Mueller. *Basic Principles of Structural Equation Modeling.* Springer, New York, 1996.

[Neyman, 1923] J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–480, 1923.

[Pearl and Robins, 1995] J. Pearl and J.M. Robins. Probabilistic evaluation of sequential plans from causal models with hidden variables. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 444–453. Morgan Kaufmann, San Francisco, 1995.

[Pearl, 1994] J. Pearl. A probabilistic calculus of actions. In R. Lopez de Mantaras and D. Poole, editors, *Uncertainty in Artificial Intelligence 10*, pages 454–462. Morgan Kaufmann, San Mateo, CA, 1994.

[Pearl, 1995] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, December 1995.

[Pearl, 1998] J. Pearl. Graphs, causality, and structural equation models. *Sociological Methods and Research*, 27(2):226–284, 1998.

[Pearl, 2000] J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, New York, 2000.

[Robins and Greenland, 1992] J.M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

[Robins, 1986] J.M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

[Robins, 1987] J.M. Robins. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases*, 40(Suppl 2):139S–161S, 1987.

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

[Simon and Rescher, 1966] H.A. Simon and N. Rescher. Cause and counterfactual. *Philosophy and Science*, 33:323–340, 1966.

[Simon, 1953] H.A. Simon. Causal ordering and identifiability. In Wm. C. Hood and T.C. Koopmans, editors, *Studies in Econometric Method*, pages 49–74. Wiley and Sons, Inc., 1953.

[Sobel, 1990] M.E. Sobel. Effect analysis and causation in linear structural equation models. *Psychometrika*, 55(3):495–515, 1990.

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

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