FROM BAYESIAN NETWORKS TO CAUSAL NETWORKS *

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

Abstract

This paper demonstrates the use of graphs as a mathematical tool for expressing independencies, and as a formal language for communicating and processing causal information for decision analysis. We show how complex information about external interventions can be organized and represented graphically and, conversely, how the graphical representation can be used to facilitate quantitative predictions of the effects of interventions.

We first review the theory of Bayesian networks and show that directed acyclic graphs (DAGs) offer an economical scheme for representing conditional independence assumptions and for deducing and displaying all the logical consequences of such assumptions. We then introduce the manipulative account of causation and show that any DAG defines a simple transformation which tells us how the probability distribution will change as a result of external interventions in the system. Using this transformation it is possible to quantify, from non-experimental data, the effects of external interventions and to specify conditions under which randomized experiments are not necessary. As an example, we show how the effect of smoking on lung cancer can be quantified from non-experimental data, using a minimal set of qualitative assumptions.

Finally, the paper offers a graphical interpretation for Rubin's model of causal effects, and demonstrates its equivalence to the manipulative account of causation. We exemplify the tradeoffs between the two approaches by deriving nonparametric bounds on treatment effects under conditions of imperfect compliance.

---

*Portions of this paper were presented at the *49th Session of the International Statistical Institute*, Florence, Italy, August 25 - September 3, 1993.

1 Introduction

Although graphical models are intuitively compelling for conceptualizing statistical associations, the scientific community generally views such models with hesitancy and suspicion. The purpose of this paper is to demonstrate the use of graphs as a precise mathematical tool of great versatility, especially as a formal language for communicating causal information in statistical and decision analysis.

Causal models, no matter how they are represented, discovered, or tested, are generally regarded as more useful than associational models because causal models provide information about the dynamics of the system under study. In other words, a joint distribution tells us how probable events are and how probabilities would change with subsequent observations, but a causal model also tells us how these probabilities would change as a result of external interventions in the system. For this reason, causal models (or "structural models" as they are often called) have been the target of relentless scientific pursuit and, at the same time, the center of much controversy and speculation [Freedman 1987].

Bayesian Belief Networks became an interesting target of investigation precisely for their ability to capture causal information [Pearl 1988]. Unfortunately, the directionality of the arcs in such networks has been treated very cautiously in both the statistical and decision analytic literature [Lauritzen & Spiegelhalter 1988, Cox 1992, Cox & Wermuth 1993, Spiegelhalter etal. 1993, Howard 1990]: the causal interpretation of the directed arcs has been de-emphasized in favor of the safer interpretation in terms of "relevance" and "dependence". This limited interpretation is deficient in several respects. First, causal associations are the primary source of judgments about dependence and relevance; they should therefore guide preformal thinking about the design of statistical studies [Dempster 1990]. Second, rejecting the causal interpretation of directed arcs prevents us from using graphical models for making legitimate predictions about the effects of actions. Such predictions are indispensable in most decision making applications, including policy analysis and treatment management, and often involve actions that were not anticipated in the modeling phase.

The primary aim of this paper is to show how complex information about external interventions can be organized and represented graphically and, conversely, how the graphical representation can be used to facilitate quantitative predictions of the effects of interventions, anticipated and unanticipated alike.

Section 2 will review the use of directed acyclic graphs (DAGs) as a language for communicating conditional independence assumptions. Sections 3 and 4 will define the causal interpretation of DAGs and Section 5 will demonstrate their use in observational studies. Section 6 will demonstrate the equivalence between the language of graphs and Rubin's model of causal effects. Finally, Section 7 applies the two approaches to the analysis of treatment effects in experimental studies with imperfect compliance. Using this example we show how a latent-variable structure can be reduced to an equivalent counterfactual model and how the two approaches can be used to derive nonparametric bounds on the causal effects of treatments, when data is taken under conditions of partial compliance.

2 Directed Graphs and Conditional Independence: A Review

Networks employing *directed acyclic graphs* (DAGs) are used to provide either

1. an economical scheme for representing conditional independence assumptions, or

2. a graphical language for representing causal influences.

This section will focus on the former, since causal influences are discussed in the remaining parts of this paper.

Given a DAG Γ and a joint distribution P over a set $X = \{X_1, ..., X_n\}$ of discrete variables, we say that Γ *represents* P if there is a one-to-one correspondence between the variables in X and the nodes of Γ, such that P admits the recursive product decomposition

$$P(x_1, ..., x_n) = \prod_i P(x_i \mid \mathbf{pa}_i) \tag{1}$$

where $\mathbf{pa}_i$ are the direct predecessors (called *parents*) of X_i in Γ. For example, the DAG in Figure 1 induces the decomposition

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)\ P(x_2|x_1)\ P(x_3|x_1)\ P(x_4|x_2, x_3)\ P(x_5|x_4) \tag{2}$$

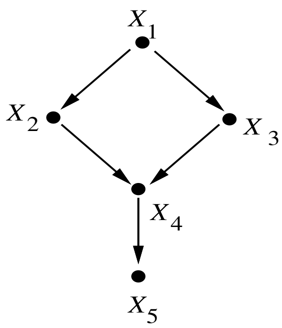

Figure 1: A typical directed acyclic graph (DAG) representing the decomposition of Eq. (2).

The recursive decomposition in Eq. (1) implies that, given its parent set $\mathbf{pa}_i$, each variable X_i is conditionally independent of all its other predecessors $\{X_1, X_2, ..., X_{i-1}\}\backslash\mathbf{pa}_i$. Using Dawid's [1979] notation, we can state this set of independencies as follows:

$$X_i \perp\!\!\!\perp \{X_1, X_2, ..., X_{i-1}\}\backslash\mathbf{pa}_i \mid \mathbf{pa}_i \quad i = 2, ..., n \tag{3}$$

Such a set of independencies will be called *Markovian*, since it reflects the Markovian condition for state transitions: Each state is rendered independent of the past, given its immediately preceding state. For example, the DAG of Figure 1 implies the following Markovian independencies:

$$X_2 \perp\!\!\!\perp \{0\} \mid X_1, \quad X_3 \perp\!\!\!\perp X_2 \mid X_1, \quad X_4 \perp\!\!\!\perp X_1 \mid \{X_2, X_3\}, \quad X_5 \perp\!\!\!\perp \{X_1, X_2, X_3\} \mid X_4 \tag{4}$$

Conversely, any list M of Markovian independencies identifies a DAG Γ (that represents P) because it permits a recursive product decomposition as in Eq. (1). However, such decomposition may imply additional independencies that are not included in M. For example, the decomposition of Eq. (1) implies $X_5 \perp\!\!\!\perp X_3 \mid \{X_1, X_4\}$ (which is not part of M) regardless of the numerical values assigned to the factors of that product. A graphical criterion called d-separation [Pearl 1988] permits us to read off the DAG the sum total of *all* independencies implied by a given decomposition.

Definition 2.1 (*d-separation*) If X, Y, and Z are three disjoint subsets of nodes in a DAG Γ, then Z is said to d-separate X from Y, denoted $d(X, Z, Y)_\Gamma$, if and only if there is no path from a node in X to a node in Y along which the following two conditions hold: (1) every node with converging arrows either is or has a descendant in Z, and (2) every other node is outside Z. A path satisfying the conditions above is said to be *active*; otherwise it is said to be *blocked* (by Z). By *path* we mean a sequence of consecutive edges (of any directionality) in the DAG.

In Figure 1, for example, $X = \{X_2\}$ and $Y = \{X_3\}$ are d-separated by $Z = \{X_1\}$; the path $X_2 \leftarrow X_1 \rightarrow X_3$ is blocked by $X_1 \in Z$, while the path $X_2 \rightarrow X_4 \leftarrow X_3$ is blocked because X_4 and all its descendants are outside Z. Thus $d(X_2, X_1, X_3)$ holds in Γ. However, X and Y are not d-separated by $Z' = \{X_1, X_5\}$ because the path $X_2 \rightarrow X_4 \leftarrow X_3$ is rendered active by virtue of X_5, a descendant of X_4, being in Z. Consequently, $d(X_2, \{X_1, X_5\}, X_3)$ does not hold in Γ; Metaphorically, learning the value of the consequence X_5 renders its causes X_2 and X_3 dependent, as if a pathway were opened along the converging arrows at X_4.

The d-separation criterion has been shown to be both sound and complete relative to the set of distributions that are represented by a DAG Γ [Verma 1986, Geiger & Pearl 1988]. In other words, there is a one-to-one correspondence between the set of independencies implied by the recursive decomposition of Eq. (1) and the set of triples (X, Z, Y) that satisfy the d-separation criterion in Γ. Furthermore, the d-separation criterion can be tested in time linear in the number of edges in Γ. Thus, a DAG can be viewed as an efficient scheme for representing Markovian independence assumptions and for deducing and displaying all the logical consequences of such assumptions. Additional properties of DAGs and their applications to evidential reasoning in expert systems are discussed in [Pearl 1988, Pearl etal. 1990, Geiger 1990, Lauritzen & Spiegelhalter 1988, Spiegelhalter etal. 1993, Pearl 1993a].

3 Graphical Models and the Manipulative Account of Causation

The interpretation of DAGs as carriers of independence assumptions does not specifically mention causation and will in fact be valid for any set of Markovian independencies, along any ordering (not necessarily causal or chronological) of the variables.

However, the patterns of independencies portrayed in a DAG are so typical of causal organizations that some of these patterns can only be given meaningful interpretation in terms of causation. For example, we can hardly find a pair of dependent events, E_1 and E_2, that are rendered independent by conditioning on a third event E_3 unless E_3 serves as a cause for either E_1 or E_2 (or both). Indeed, we cannot easily contrive three such events if we constrain E_3 to occur *after* E_1 and E_2, so as to suppress the causal interpretation above. The DAG representation provides a perfect language for such dependencies; it lets E_3 d-separate E_2 from E_1 in the pattern $E_1 \rightarrow E_3 \rightarrow E_2$ or $E_1 \leftarrow E_3 \rightarrow E_2$, but not in the converging pattern $E_1 \rightarrow E_3 \leftarrow E_2$. This distinction is the basis for the Markovian accounts of causation, as exemplified by those of [Granger 1988, Suppes 1970], and by the more elaborate, non-temporal accounts of [Pearl & Verma 1991] and [Spirtes etal. 1993].

However, the Markovian account still leaves open the question of why such intricate patterns of independencies are produced by and become the characteristic signature of causal organizations. A related question is how these patterns are connected with the more basic notions associated with causation, such as influence, manipulation, and control. The connection is made in the mechanism-based account of causation.

The basic idea behind this account goes back to [Simon 1977] and is stated succinctly in his forward to [Glymour etal. 1987]: "The advantage of representing the system by structural equations that describe the direct causal mechanisms is that if we obtain some knowledge that one or more of these mechanisms has been altered, we can use the remaining equations to predict the consequences – the new equilibrium." Here, by "mechanism" Simon means any stable relationship between two or more variables, usually expressed in functional form, that remains invariant to external influences until it falls directly under such influences.

This mechanism-based model was adapted in [Pearl & Verma 1991] for defining probabilistic causal theories; each child-parent family in a DAG Γ represents a deterministic function

$$X_i = f_i(\mathbf{pa}_i, \epsilon_i), \tag{5}$$

where $\mathbf{pa}_i$ are the parents of variable X_i in Γ, and ϵ_i, $0 < i < n$, are mutually independent, arbitrarily distributed random disturbances. Characterizing each child-parent relationship as a deterministic function, instead of the usual conditional probability $P(x_i \mid \mathbf{pa}_i)$, imposes equivalent independence constraints on the resulting distributions and leads to the same recursive decomposition

$$P(x_1, ..., x_n) = \prod_i P(x_i \mid \mathbf{pa}_i) \tag{6}$$

that characterizes DAG models (see Eq. 1). This is so because each ϵ_i is independent on all non-descendants of X_i. However, the functional characterization $X_i = f_i(\mathbf{pa}_i, \epsilon_i)$ also permits us to specify how the resulting distribution would change in response to external interventions, using the convention that each function remains constant unless specifically altered. Moreover, the non-linear character of f_i permits us to treat changes in the function f_i itself as a variable, F_i, by writing

$$X_i = I(\mathbf{pa}_i, F_i, \epsilon_i) \tag{7}$$

where I is a 3-argument function defined by

$$I(a, b, c) = f_i(a, c) \text{ whenever } b = f_i.$$

Thus, any external intervention F_i that alters f_i can be represented graphically as an added parent node of X_i, and the effect of such an intervention can be analyzed by Bayesian conditionalization, that is, by simply setting this added parent variable to the appropriate value f_i.

The simplest type of external intervention is one in which a single variable, say X_i, is forced to take on some fixed value x_i'. Such intervention, which we call *atomic*, amounts to replacing the old functional mechanism $X_i = f_i(\mathbf{pa}_i, \epsilon_i)$ with a new mechanism $X_i = x_i'$ governed by some external force F_i that sets the value x_i'. If we imagine that each variable X_i could potentially be subject to the influence of such an external force F_i, then we can view the causal network Γ as an efficient code for predicting the effects of atomic interventions and of various combinations of such interventions.

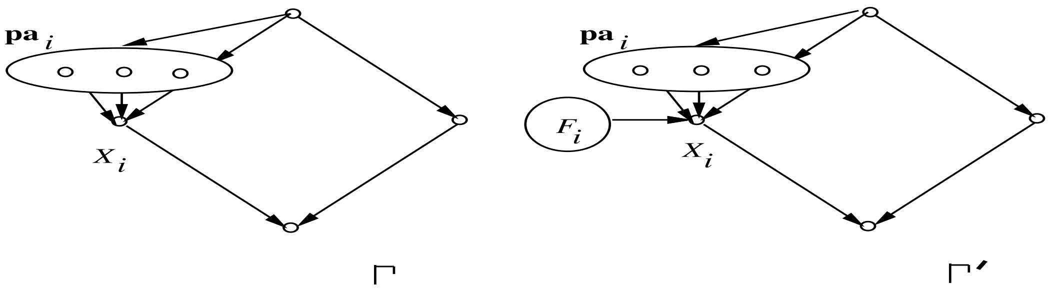

Figure 2: Representing external intervention F_i by an augmented network $\Gamma' = \Gamma \cup \{F_i \rightarrow X_i\}$.

The effect of an atomic intervention $set(X_i = x_i')$ is encoded by adding to Γ a link $F_i \longrightarrow X_i$ (see Figure 2), where F_i is a new variable taking values in $\{set(x_i'),\ idle\}$, x_i' ranges over the domain of X_i, and *idle* represents no intervention. Thus, the new parent set of X_i in the augmented network is $\mathbf{pa}_i' = \mathbf{pa}_i \cup \{F_i\}$, and it is related to X_i by the conditional probability

$$P(x_i \mid \mathbf{pa}_i') = \begin{cases} P(x_i \mid \mathbf{pa}_i) & \text{if } F_i = idle \\ 0 & \text{if } F_i = set(x_i') \text{ and } x_i \neq x_i' \\ 1 & \text{if } F_i = set(x_i') \text{ and } x_i = x_i' \end{cases} \tag{8}$$

The effect of the intervention $set(x_i')$ is to transform the original probability function $P(x_1, ..., x_n)$ into a new function $P_{x_i'}(x_1, ..., x_n)$, given by

$$P_{x_i'}(x_1, ..., x_n) = P'(x_1, ..., x_n \mid F_i = set(x_i')) \tag{9}$$

where P' is the distribution specified by the augmented network $\Gamma' = \Gamma \cup \{F_i \rightarrow X_i\}$ and Eq. (8), with an arbitrary prior distribution on F_i. In general, by adding a hypothetical intervention link $F_i \rightarrow X_i$ to each node in Γ, we can construct an augmented probability function $P'(x_1, ..., x_n; F_1, ..., F_n)$ that contains information about richer types of interventions. Multiple interventions would be represented by conditioning P' on a subset of the F_i's (taking values in their respective $set(x_i')$), while the pre-intervention probability function P would be viewed as the posterior distribution induced by conditioning each F_i in P' on the value *idle*.

4 A Transformation Formula for Interventions

This representation yields a simple and direct transformation between the pre-intervention and the post-intervention distributions:

$$P_{x_i'}(x_1, ..., x_n) = \begin{cases} \frac{P(x_1,...,x_n)}{P(x_i \mid \mathbf{pa}_i)} & \text{if } \ x_i = x_i' \\ 0 & \text{if } \ x_i \neq x_i' \end{cases} \tag{10}$$

This transformation reflects the removal of the term $P(x_i \mid \mathbf{pa}_i)$ from the product decomposition of Eq. (6), since $\mathbf{pa}_i$ no longer influence X_i. Graphically, the removal of this term is equivalent to removing the links between $\mathbf{pa}_i$ and X_i, while keeping the rest of the network intact. Transformations involving conjunctive and disjunctive actions can be obtained by straightforward applications of Eq. (9) [Spirtes etal. 1993, Goldszmidt & Pearl 1992, Goldszmidt 1992]

The transformation (10) exhibits the following properties:

1. An intervention $set(x_i')$ can affect only the descendants of X_i in Γ.

2. For any set $\mathbf{S}$ of variables, we have

$$P_{x_i'}(\mathbf{S} \mid \mathbf{pa}_i) = P(\mathbf{S} \mid x_i', \mathbf{pa}_i). \tag{11}$$

   In other words, given $X_i = x_i'$ and $\mathbf{pa}_i$, it is superfluous to find out whether $X_i = x_i'$ was established by external intervention or not. This can be seen directly from the augmented network Γ' (see Figure 2), since $\{X_i\} \cup \mathbf{pa}_i$ *d*-separates F_i from the rest of the network, thus legitimizing the conditional independence $\mathbf{S} \perp\!\!\!\perp F_i \mid (X_i, \mathbf{pa}_i)$.

3. A sufficient condition for an external intervention $set(X_i = x_i')$ to have the same effect on X_j as the passive observation $X_i = x_i'$ is that X_i *d*-separates $\mathbf{pa}_i$ from X_j, that is,

$$P_{x_i'}(x_j) = P(x_j \mid x_i') \ \text{ iff } \ X_j \perp\!\!\!\perp \mathbf{pa}_i \mid X_i. \tag{12}$$

The immediate implication of Eq. (10) is that, given the structure of the causal network Γ, one can infer post-intervention distributions from pre-intervention distributions; hence, we can reliably estimate the effects of interventions from passive (i.e., non-experimental) observations. Of course, Eq. (10) does not imply that we can always substitute observational studies for experimental studies, as this would require an estimation of $P(x_i \mid \mathbf{pa}_i)$. The mere identification of $\mathbf{pa}_i$ (i.e., the direct causal factors of X_i) requires substantive causal knowledge of the domain which is often unavailable. Moreover, even when we have sufficient substantive knowledge to structure Γ, some members of $\mathbf{pa}_i$ may be unobservable, or *latent*. Fortunately, there are conditions for which an unbiased estimate of $P_{x_i'}(x_j)$ can be obtained even when the $\mathbf{pa}_i$ variables are latent and, moreover, a simple graphical criterion can tell us when these conditions are satisfied.

5 Eliminating Confounding Bias

5.1 Strong Ignorability and the Back-door Criterion

Assume we are given a causal network Γ together with non-experimental data on a subset $\mathbf{X}_o$ of observed variables in Γ and we wish to estimate what effect the intervention $set(X_i = x_i')$ would have on some response variable X_j. In other words, we seek to estimate $P_{x_i'}(x_j)$ from a sample estimate of $P(\mathbf{X}_o)$. Applying Eq. (9), we can write

$$\begin{aligned} P_{x_i'}(x_j) &= P'(x_j \mid F_i = set(x_i')) \\ &= \sum_{\mathbf{S}} P'(x_j \mid \mathbf{S}, X_i = x_i', F_i = set(x_i'))P'(\mathbf{S} \mid F_i = set(x_i')) \end{aligned} \tag{13}$$

where $\mathbf{S}$ is any set of variables. Clearly, if $\mathbf{S}$ satisfies

$$\mathbf{S} \perp\!\!\!\perp F_i \quad \text{and} \quad X_j \perp\!\!\!\perp F_i \mid (X_i, \mathbf{S}) \tag{14}$$

then Eq. (13) can be reduced to

$$\begin{aligned} P_{x_i'}(x_j) &= \sum_{\mathbf{S}} P(x_j \mid \mathbf{S}, x_i')P(\mathbf{S}) \\ &= E_{\mathbf{S}}[P(x_j \mid \mathbf{S}, x_i')] \end{aligned} \tag{15}$$

Thus, if we find a set $\mathbf{S} \subseteq \mathbf{X}_o$ of observables satisfying Eq. (14), we can estimate $P_{x_i'}(x_j)$ by taking the expectation (over $\mathbf{S}$) of $P(x_j \mid \mathbf{S}, x_i')$, and the latter can easily be estimated from non-experimental data. It is also easy to verify that Eq. (14) is satisfied by any set $\mathbf{S}$ that meets the following *back-door*[1] *criterion*:

1. No node in $\mathbf{S}$ is a descendant of X_i, and

2. $\mathbf{S}$ d-separates X_i from X_j along every path containing an arrow into X_i.

In Figure 3, for example, the sets $\mathbf{S}_1 = \{X_3, X_4\}$ and $\mathbf{S}_2 = \{X_4, X_5\}$ meet the back-door criterion, but $\mathbf{S}_3 = \{X_4\}$ does not because X_4 does not d-separate X_i from X_j along the path $(X_i, X_3, X_1, X_4, X_2, X_5, X_j)$. Thus, we have obtained a simple graphical criterion for finding a set of observables for estimating (by conditioning) the effect of interventions from purely non-experimental data.

It is interesting that the conditions formulated in Eq. (14) are equivalent to those known as *strongly ignorable treatment assignment* (SITA) conditions in Rubin's model for causal effect [Rosenbaum & Rubin 1983] (see Section 6 for detailed comparison). Reducing the SITA conditions to a graphical back-door criterion facilitates computer-aided search for an optimal conditioning set $\mathbf{S}$ and significantly simplifies the judgments required for ratifying the validity of such conditions in practical situations.

[1] The name "back-door" echoes condition 2, which requires that only indirect paths from X_i to X_j be d-separated; these paths can be viewed as entering X_i through the back door.

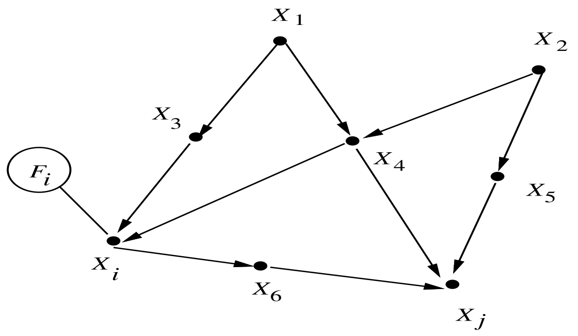

Figure 3: A DAG representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ (or $\{X_4, X_5\}$) yields an unbiased estimate of $P(x_j \mid set(x_i'))$.

5.2 Other Graphical Criteria

The control of confounding bias does not end with the back-door estimand of Eq. (15); an orthogonal estimand, worthy of the name "the front-door criterion", may complement the latter in cases where we cannot find observed covariates **S** satisfying the back-door conditions. Consider variable X_6 in Figure 3, and assume that it is the only observed variable in the graph, beside X_i and X_j. Clearly, X_6 does not satisfy any of the back-door conditions because (1) it is a descendant of X_i, and (2) it does not block any of the back-door paths between X_i and X_j. We shall now show that measurements of X_6 can nevertheless facilitate a consistent estimation of $P(x_j|set(x_i))$. This can be shown either using the algebra of the "set" operator or, more directly, by reducing the expression for $P(x_j|set(x_i))$ to formulae computable from the observed distribution function $P(x_i, x_6, x_j)$. To that end, let us denote by U the compound variable consisting of all confounding variables between X_i and X_j (i.e., $U = \{X_1, ..., X_5\}$ in Figure 3), and further denote X_i by X and X_j by Y. All together, we now have a structure depicted in Figure 4 below, containing one unobserved variable, U, three observed variables X, Z, Y, with Z mediating the interaction between X and Y. We will also assume that $P(x, z) > 0$ for all values of x and z.

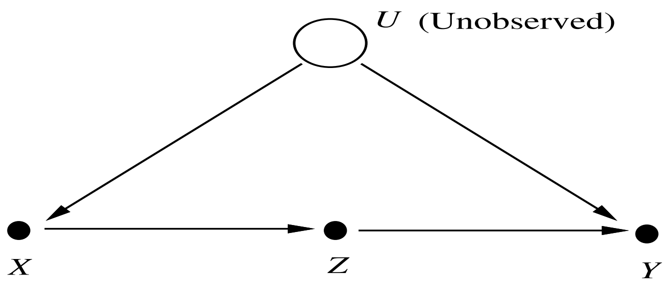

Figure 4

The joint distribution function of all four variables is given by the product

$$P(x, y, z, u) = P(y|z, u)P(z|x)P(x|u)P(u)$$

and our task is to compute the causal effect of X on Y, given by

$$P(y|set(x)) = \sum_u P(y|x,u)P(u) \tag{16}$$

Using the two conditional independence claims embodied in the graph of Figure 4, it is possible to eliminate u from the rhs of (16) and obtain:

$$P(y|set(x)) = \sum_z P(z|x) \sum_{x'} P(y|x',z)P(x') \tag{17}$$

Since all factors on the r.h.s of (17) are consistently estimable from non-experimental data, it follows that $P(y|set(x))$ is estimable as well. Thus, we are in the possession of a consistent non-parametric estimator for the causal effect of a potential cause X on a potential response Y, assuming of course that we find a mediating instrument Z that meets the conditions of Figure 4.

We summarize this result by a theorem, following a formal definition *identifiability.*

Definition 5.1 *The causal effect of X on Y is said to be identifiable, if the quantity $P(y|set(x))$, can be computed from the joint distribution of the observed variables. Identifiability means that the quantity $P(y|set(x))$ can be estimated consistently from an arbitrarily large sample, randomly drawn from the joint distribution.*

Theorem 5.2 *If a variable Z satisfies the following conditions relative to an ordered pair of variables (X, Y),*

1. *Z intercepts all direct paths from X to Y*
2. *There is no back-door path between X and Z, nor between Z and Y.*
3. *The relation between X and Z is non-deterministic, i.e., $P(x, z) > 0$*

then the causal effect of X on Y is identifiable and is given by the formula in Eq. (17).

The graphical criterion of Theorem 5.2 uncovers many new structures that permit the identification of causal effects from nonexperimental observations. In contrast, most of the literature on statistical experimentation considers the measurement of intermediate variables, affected by the action, to be useless, if not harmful, for causal inference [Cox 1958, Pratt & Schlaifer 1988]. The relevance of such structures in practical situations can be seen, for instance, if we identify X with smoking, Y with lung cancer, Z with the amount of tar deposited in a subject's lungs, and U with an unobserved carcinogenic genotype that, according to the tobacco industry, also induces an inborn craving for nicotine. In this case, Eq. (17) would provide us with the means to quantify, from nonexperimental data, the causal effect of smoking on cancer. (Assuming, of course, that the data $P(x, y, z)$ is made available and that we believe that smoking does not have any direct causal effect on lung cancer except that mediated by tar deposits).

Generalizations to more elaborate structures, including multiple Z variables, nested combinations of back-door and front-door patterns, and concurrent "set" operations, are natural extensions which we discuss elsewhere [Pearl 1994].

5.3 Related Topics

Eq. (10) was derived under the assumption that the pre-intervention probability P is given by the product of Eq. (6), which represents a joint distribution prior to making any observations. To predict the effect of action F_i *after* observing C, we must also invoke assumptions about persistence, so as to distinguish properties that will terminate as a result of F_i from those that will persist despite F_i. Such a model of persistence was invoked in [Pearl 1993b]; there, it was assumed that only those properties that are not under any causal influence to terminate should persist. This assumption yields formulas for the effect of *conditional interventions* (conditioned on a pre-action observation C). Again, given Γ, these effects can be estimated from non-experimental data.

[Spirtes etal. 1993] have explored a more ambitious task – estimation of the effect of intervention when the structure of Γ is not available and must also be inferred from the data. Recent developments in graphical models [Pearl & Verma 1991, Spirtes etal. 1993] have produced methods that, under certain conditions, permit us to infer plausible causal structures from non-experimental data, albeit such structures have a weaker set of guarantees than those obtained through controlled randomized experiments. These guarantees fall into two categories: minimality and stability [Pearl & Verma 1991]. Minimality guarantees that any other causal structure compatible with the data is necessarily more redundant, and hence less trustworthy, than the one(s) inferred. Stability ensures that any alternative structure compatible with the data must be less stable than the one(s) inferred; namely, slight fluctuations in the parameters of the functions f_i (Eq. (7)) will render an alternative structure no longer compatible with the data.

When the structure of Γ is to be inferred under these guarantees, the formulas governing the effects of interventions and the conditions required for estimating these effects become rather complex [Spirtes etal. 1993]. Alternatively, one can produce bounds on the effects of interventions by taking representative samples of inferred structures and estimating $P_{x'_i}(x_j)$ according to Eq. (10) (or Eq. (15)) for each such sample.

6 Relation to Rubin's Model of Causal Effects

So far, our discussion of causal graphs has focused on the manipulative account of causation which, as was shown in Section 3, coincides with Simon's mechanism-based account. Another view of these basic accounts is provided by the counterfactual model developed by [Rubin 1974] and [Holland 1986, Rosenbaum & Rubin 1983, Pratt & Schlaifer 1988], the roots of which date back to [Neyman 1935] and [Fisher 1935].

In Rubin's model, we imagine that an intervention Z (or "treatment" as it is often called) can be applied at various levels $1, 2, ..., T$ to any experimental subject (called a "unit") and that it is possible to record the values of the response observed in conjunction with the different levels of the treatment. The correspondence between the applied levels of the treatment and the recorded levels of the response would then constitute the "causal effect" associated with the particular subject, as it char-

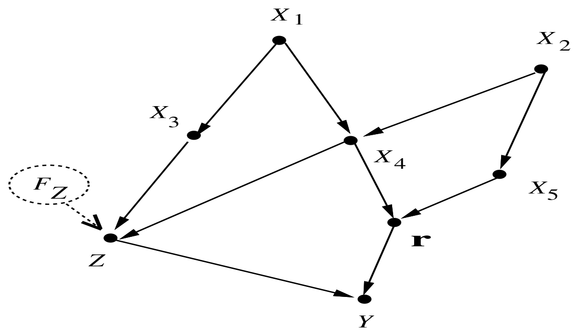

Figure 5: Graphical representation of Rubin's model, showing the observed response (Y) *as a function of treatment* (Z) *and causal effect* $(\mathbf{r})$ *variables.*

acterizes the potential impact of the treatment *if* applied (counterfactually) to that subject. The target of causal-inference analysis is then the estimation, from statistical data, of the properties of the potential-response vector $\mathbf{r} = (r_1, r_2, ..., r_t, ..., r_T)$, where r_t stands for the response that the subject would exhibit if the t-th level of the treatment were applied. For any given subject, r_t is considered a deterministic (albeit unobservable) entity, as it determines precisely the response of the subject, had he/she been given the treatment $Z = t$. However, for a subject randomly drawn from a population, we can view r_t as a random variable, and therefore, we can attempt to estimate its distribution, its expectation, or the expectation of the difference $r_t - r_{t'}$.

The distribution of r_t, using the language of the manipulative account, is equal to the distribution of the observed response Y, conditioned on the intervention $F_Z = set(Z = t)$, namely,

$$P(r_t = y) = P'(Y = y|set(Z = t)) = P_{Z=t}(y) \tag{18}$$

and

$$E(r_t - r_{t'}) = E(Y|set(Z = t)) - E(Y|set(Z = t')) \tag{19}$$

The reason we must condition on the action $F_Z = set(Z = t)$ and not on the observation $Z = t$ is that, to comply with the interpretation of r_t as the subject's hypothetical response to treatment $Z = t$, we must suppress any information that the assignment $Z = t$ may provide on the nature of subject.

The translation provided by Eqs. (18)-(19) implies that the causal effect defined as $E(r_t - r_{t'})$ can be computed from the manipulative account defined in Section 3 and its associated transformations, as given in Eq. (10). This translation also permits us to devise a graphical representation to Rubin's model, thus displaying the functional role of r_t. For example, if in Figure 3 we take X_i to be the treatment variable Z and X_j to be the observed response Y, then the graph associated with Rubin's model would correspond to the one in Figure 5. The arc from X_3 to Z represents a non-randomized treatment assignment policy, where the assignment of subjects to treatment Z may depends on the factor X_3.

The main difference between the two figures is that in the counterfactual model of Figure 5 $\mathbf{r}$ is treated explicitly as a variable, whereas in the manipulative model of

Figure 3 $\mathbf{r}$ is represented implicitly as a function that connects Y to its direct causal factors: X_4, X_5, and X_6. The two alternative representations of $\mathbf{r}$ are in line with the transformation defined in Eq. (7). Note also that the counterfactual reading of $\mathbf{r}$ is an integral part of the mechanism-based reading of causation; the semantics of the function $f_i(\mathbf{pa}_i, \epsilon_i)$ is intrinsically counterfactual because it defines the value of X_i for any hypothetical value combination of $\mathbf{pa}_i$ and ϵ_i.

It is not hard to verify that Figures 3 and 5 are empirically equivalent, in the sense that they imply the same statistical and manipulative behavior for all observed variables (with the exception of X_6, which is maginalized out). For example, our back-door criterion between X_i and X_j (see Eq. (14)) translates to an equivalent back-door criterion between Z and $\mathbf{r}$,

$$Z \perp\!\!\!\perp \mathbf{r} \mid \mathbf{S} \tag{20}$$

This is precisely the SITA condition defined in [Rosenbaum & Rubin 1983]. Moreover, since Dawid's [1979] axioms for conditional independence are faithfully encoded in the d-separation criterion, we can immediately translate the condition in Eq. (20) into equivalent graphical criteria, all of which are vividly displayed in the graph. For example, stated in terms of the unobserved set of variables $\mathbf{U}$ ($\mathbf{U} = \{X_1, X_2\}$ in Figure 5), our back-door criterion (Eq. (20)) reads

$$Z \perp\!\!\!\perp \mathbf{U} \mid \mathbf{S} \quad \text{or} \quad \mathbf{U} \perp\!\!\!\perp \mathbf{r} \mid \mathbf{S} \tag{21}$$

These are precisely the alternative conditions for (X, U)-*adjustable treatment assignment* given in [Rosenbaum 1989].

The main attraction of Rubin's model has been the facility to precisely define the causal quantities we wish to estimate without specifying the inference methods used in obtaining these estimates. As a result, the model exposes the fundamental assumptions needed to make the desired estimates feasible, and we are often able to reduce these assumptions to statements about independencies which, at least in principle, can be submitted to judgmental verification.

Since quantities defined in Rubin's model can be translated to equivalent quantities in the manipulative account of causation (see Eqs. (18)-(19)), it is clear that the latter should enjoy similar advantages. Moreover, considering that graphical models provide a calculus for processing manipulative statements (through the introduction of hypothetical action variables, as shown in Section 3), it is not surprising that graphical techniques are applicable for processing statements articulated in Rubin's model.

The current popularity of Rubin's model is in part a reaction to basic inadequacies of the structural equations framework, which forces the analyst to commit to a particular regression model, governed by a particular set of random variables, *iid* disturbances, and hypothetical parameters. While the graphical framework indeed commits the analyst to treating quantities as random variables, often latent, it does not require any assumption of *iid* or parametric structure. The analyst is committed *only* to the qualitative structure behind causal thinking which, we conjecture, is the very structure an analyst must consult when judging assumptions about Rubin's potential-response vector, such as Eq. (20).

It is not surprising, then, that the two approaches yield identical conclusions in all cases where such conclusions can be stated formally or tested empirically. In cases where the conclusions involve human judgment (e.g., confirming the SITA conditions), the two approaches provide complementary languages for phrasing the judgments required. However, in problems requiring substantial use of domain knowledge, the causal-network approach offers definite advantages.

7 Example: Causal Effects Under Partial Compliance

To demonstrate the interplay between the counterfactual and the latent-variable models, we will present an analysis of a well-known practical problem using the two approaches.

7.1 The problem

Consider an experimental study in which random assignment has taken place but compliance is not perfect, that is, the treatment received is different from that assigned. It is well known that under such conditions a bias may be introduced, in the sense that the true causal effect of the treatment may deviate substantially from that computed by simply comparing subjects receiving the treatment with those not receiving the treatment. For example, subjects who did not comply with the assigned treatment may be precisely those who would have responded adversely to the treatment, so the treatment, when applied uniformly to the population, might actually be substantially less effective than the study reveals.

In an attempt to compensate for such bias, economists have devised correctional formulas, called "instrumental variables" [Bowden & Turkington 1984], which, in general, do not hold outside the linear regression model. A recent analysis by Efron and Feldman [1991] represents a healthy departure from the linear regression model, yet it still makes restrictive commitments to a particular mode of interaction between compliance and response. Angrist et al. [1993], invoking Rubin's model, have identified a set of assumptions under which the "instrumental variable" formula is valid, but have not provided an alternative, assumption-free formula. We now derive correctional formulas that rely solely on observed quantities and are universal, that is, they are valid no matter what model actually governs the interactions between compliance and response.

7.2 The latent-structure approach

The canonical partial-compliance setup can be represented by the following network:

We assume that Z, D, and Y are observed binary variables where, using conventional terminology (e.g., [Angrist et al. 1993]), Z represents the (randomized) "treatment assignment", D is the treatment actually received, and Y is the observed

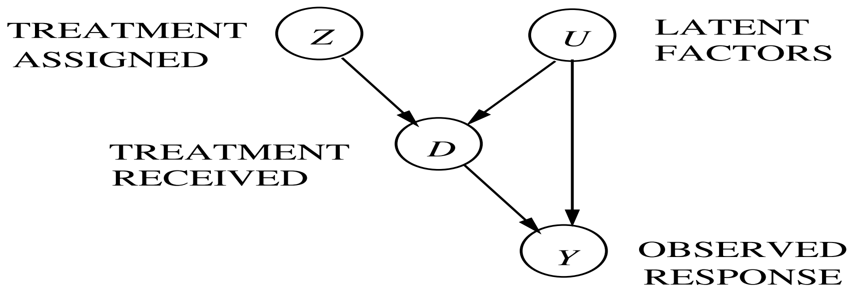

Figure 6: Graphical representation of causal dependencies in a randomized clinical trial with partial compliance.

response. U represents all unobserved and unknown factors which, as Figure 6 shows, may influence the outcome Y and the treatment D. To simplify the notation, we let z, d, and y represent, respectively, the values taken by the variables Z, D, and Y, with the following interpretation:
$z \in \{z_0, z_1\}$, z_1 asserts that treatment has been assigned (z_0, its negation)
$d \in \{d_0, d_1\}$, d_1 asserts that treatment has been administered (d_0, its negation)
$y \in \{y_0, y_1\}$, y_1 asserts a positive observed response (y_0, its negation)
The domain of U remains unspecified and may, in general, combine the spaces of several random variables, both discrete and continuous.

The graphical model reflects two assumptions:

1. The treatment assignment does not influence Y directly, only through the actual treatment D, that is,
$$Z \perp\!\!\!\perp Y \mid \{D, U\} \tag{22}$$
In practice, any direct effect Z might have on Y would be adjusted for through the use of a placebo.

2. Z and U are marginally independent, that is, $Z \perp\!\!\!\perp U$. This independence is partly ensured through the randomization of Z, which rules out any common cause for both Z and U. The absence of a direct path from Z to U represents the assumption that latent factors, U, which represent a person's initial characteristics, are not in themselves affected by the assignment.

These assumptions impose the following decomposition on the joint distribution
$$P(y, d, z, u) \;=\; P(y|d, u)\; P(d|z, u)\; P(z)\; P(u) \tag{23}$$
which, of course, cannot be observed directly. However, the marginal distribution $P(y, d, z)$ and, in particular, the conditional distribution $P(y, d|z), z = z_0, z_1$ are observed, and the challenge is to estimate the causal effect of D on Y from these distributions.

For any two binary variables X and Y, define the causal effect $R(X \rightarrow Y)$ of X on Y as
$$R(X \rightarrow Y) = P(y_1|set(x_1)) - P(y_1|set(x_0)) \tag{24}$$

Thus, for the experimental design depicted in Figure 6, we seek an estimate of

$$\begin{aligned} R(D \rightarrow Y) &= P(y_1|set(d_1) - P(y_1|set(d_0) \\ &= \sum_u [P(y_1|d_1, u) - P(y_1|d_0, u)]P(u) \end{aligned} \tag{25}$$

given the observed probabilities $P(y, d|z_0)$ and $P(y, d|z_1)$.

A few algebraic manipulations of (25) (see Appendix) yields an alternative expression for $R(D \rightarrow Y)$

$$R(D \rightarrow Y) = E\left[\frac{P(y_1|z_1, u) - P(y_1|z_0, u)}{P(d_1|z_1, u) - P(d_1|z_0, u)}\right] \tag{26}$$

where E stands for the expectation taken over u.

If we think of u as an index characterizing the experimental units (i.e., the subjects) the result is simple and intuitive. It says that for each individual unit u, the indirect causal effect along the chain $Z \rightarrow D \rightarrow Y$ is equal to the product of the individual causal effects along the two links of the chain. If all units were to exhibit the same difference in compliance probabilities, $P(d_1|z_1, u) - P(d_1|z_0, u)$, we would have the celebrated instrumental variable formula

$$R(D \rightarrow Y) = \frac{R(Z \rightarrow Y)}{R(Z \rightarrow D)} \tag{27}$$

which says that the causal effect $R(Z \rightarrow Y)$ associated with the intent-to-treat needs to be adjusted upward, through division by the partial compliance $R(Z \rightarrow D)$. This ratio formula is indeed valid in linear regression models and was derived by econometricians as far back as 1940 [Angrist et al. 1993]. In general, however, since the quantities on the r.h.s. of Eq. (26) cannot be observed directly (only in expectation), the expression for R can become as low as zero and even negative. Still, when an almost-perfect compliance is observed, the unknown quantities $P(y|d, u), P(d|z, u)$, and $P(u)$ do not have the freedom to render $R(D \rightarrow Y)$ substantially different from $R(Z \rightarrow Y)$, and meaningful bounds can then be obtained on the actual causal effect of the treatment.

The analysis presented in the Appendix yields the following bounds for the two terms on the r.h.s. of (25)[2]:

$$\max[P(y_1, d_1|z_1);\ P(y_1, d_1|z_0)] \leq P(y_1|set(d_1)) \leq 1 - \max[P(y_0, d_1|z_0);\ P(y_0, d_1|z_1)] \tag{28}$$

$$\max[P(y_1, d_0|z_0);\ P(y_1, d_0|z_1)] \leq P(y_1|set(d_0)) \leq 1 - \max[P(y_0, d_0|z_0);\ P(y_0, d_0|z_1)] \tag{29}$$

Choosing appropriate terms to bound the difference $P(y_1|set(d_1)) - P(y_1|set(d_0))$, we obtain a useful lower bound on the causal effect of D on Y:

$$R(D \rightarrow Y) \geq R(Z \rightarrow Y) - P(y_1, d_0|z_1) - P(y_0, d_1|z_0) \tag{30}$$

[2] These bounds were independently obtained by Robins (1989) and Manski (1990), and were significantly improved by Balke and Pearl (1993).

This bound guarantees that the difference between the causal effect of the intent-to-treat and the causal effect of the actual treatment could never exceed the sum of two measurable quantities, $P(y_1, d_0|z_1) + P(y_0, d_1|z_0)$.

Before continuing to Rubin's approach, we should mention that the structural model of Figure 6 imposes definite constraints, obtained directly from Eq. (28)-(29), on the observed distributions $P(y, d|z_0)$ and $P(y, d|z_1)$:

$$\begin{aligned} P(y_1, d_1|z_1) &\leq 1 - P(y_0, d_1|z_0) \\ P(y_1, d_1|z_0) &\leq 1 - P(y_0, d_1|z_1) \\ P(y_1, d_0|z_1) &\leq 1 - P(y_0, d_0|z_0) \\ P(y_1, d_0|z_0) &\leq 1 - P(y_0, d_0|z_1) \end{aligned} \tag{31}$$

These constraints constitute necessary and sufficient conditions for a marginal probability $P(y, d, z)$ to be generated by the structure of the model given in Figure 6 and therefore may serve as an operational test for the consistency of that structure with the observed data.

7.3 The counterfactual approach

A peculiar feature of the graphical model discussed so far is its capacity for producing meaningful results while keeping the latent variable U totally unspecified. U may be finite or unbounded, discrete or continuous, ordered or unstructured. Although this generality has the advantage of freeing the analyst from commitment to a particular parametric model, it may turn into an inconvenience when finer mathematical details, such as tighter bounds or maximum likelihood estimates, are needed.

The structure of Figure 7 is similar to that of Figure 6, with the difference that the latent variables R R' have only four states each. We will now show that every model that fits into the general latent structure of Figure 6 can also fit into the finite-variable structure of Figure 7 and, moreover, that the states of the variables R and R' correspond precisely to the components of the causal-effect vector in Rubin's model.

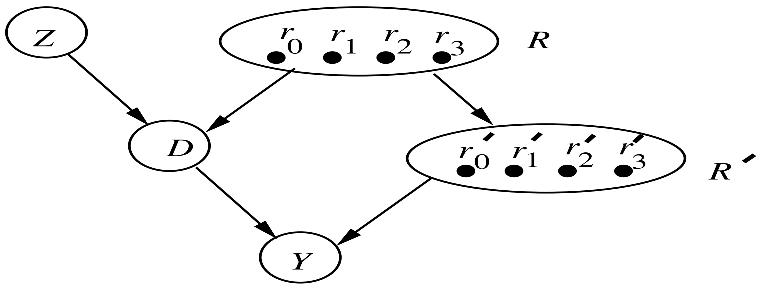

Figure 7: A structure equivalent to that of Figure 6 employing two latent variables, R and R', with 4-states each.

Our first step is to convert each conditional probability term $P(x_i|\mathbf{pa}_i)$ in Eq. (23) into an equivalent functional form, $x_i = f_i(\mathbf{pa}_i, \epsilon_i)$, as in Eq. (7). This can be accomplished by the standard method of simulating probability distributions, letting ϵ_i be uniformly distributed over $[0,1]$, and defining

$$X_i = f_i(\mathbf{pa}_i, \epsilon_i) \quad = \begin{cases} 1 & \text{if } \epsilon_i \leq P(x_i = 1|\mathbf{pa}_i) \\ 0 & \text{otherwise} \end{cases} \tag{32}$$

It is clear that f_i induces the specified conditional probability relation between $\mathbf{pa}_i$ and X_i.

The next step is to convert the functional equations involving the hypothetical variable U to ones involving variables with a finite number of states. Consider the conditional probability $P(d|z,u)$ in its functional form $d = f_D(z,u,\epsilon_D)$. No matter how complex U and ϵ_D might be, their impact on D cannot amount to more than a modification of the functional relationship between D and Z and, since there are exactly four functions relating two binary variables, each (u, ϵ_D) pair selects one of the four functions. Thus, the impact of the random pair (u, ϵ_D) can be simulated by a four-state variable $r \in \{r_0, r_1, r_2, r_3\}$, together with the appropriate distribution over r's states, with each state selecting one of the four binary functions.

Formally, if $dom(U)$ and $dom(\epsilon_D)$ are the domains of U and ϵ_D, respectively, define the mapping $R : dom(U) \times dom(\epsilon_D) \rightarrow \{r_0, r_1, r_2, r_3\}$ as follows:

$$R = \begin{cases} r_0 & \text{if} \quad f_D(z_0,u,\epsilon_D) = 0 \quad \text{and } f_D(z_1,u,\epsilon_D) = 0 \\ r_1 & \text{if} \quad f_D(z_0,u,\epsilon_D) = 0 \quad \text{and } f_D(z_1,u,\epsilon_D) = 1 \\ r_2 & \text{if} \quad f_D(z_0,u,\epsilon_D) = 1 \quad \text{and } f_D(z_1,u,\epsilon_D) = 0 \\ r_3 & \text{if} \quad f_D(z_0,u,\epsilon_D) = 1 \quad \text{and } f_D(z_1,u,\epsilon_D) = 1 \end{cases} \tag{33}$$

We can now write D as a function of the variables Z and R:

$$d = F_D(z,r) = \begin{cases} d_0 & \text{if} \quad r = r_0 \\ \\ d_0 & \text{if} \quad r = r_1 \quad z = z_0 \\ d_1 & \text{if} \quad r = r_1 \quad z = z_1 \\ \\ d_1 & \text{if} \quad r = r_2 \quad z = z_0 \\ d_0 & \text{if} \quad r = r_2 \quad z = z_1 \\ \\ d_1 & \text{if} \quad r = r_3 \end{cases} \tag{34}$$

Repeating the same transformation on the factor $P(y|d,u)$ or its functional form $y = f_Y(d,u,\epsilon_Y)$ permits us to express Y as a function of D and a second four-state variable R':

$$y = F_Y(d,r') = \begin{cases} y_0 & \text{if} \quad r' = r'_0 \\ \\ y_0 & \text{if} \quad r' = r'_1 \quad d = d_0 \\ y_1 & \text{if} \quad r' = r'_1 \quad d = d_1 \\ \\ y_1 & \text{if} \quad r' = r'_2 \quad d = d_0 \\ y_0 & \text{if} \quad r' = r'_2 \quad d = d_1 \\ \\ y_1 & \text{if} \quad r' = r'_3 \end{cases} \tag{35}$$

where $R' : dom(U) \times dom(\epsilon_Y) \rightarrow \{r'_0, r'_1, r'_2, r'_3\}$. Since U influences both R and R', the two variables are not independent, hence the arrow $R \rightarrow R'$ in Figure 7. The join distribution over $R \times R'$ requires 15 independent parameters, and these parameters are sufficient for specifying the model of Figure 7, since Y and D stand in a functional relation to their parents.

The correspondence between the states of variables R and R' and the potential response vectors in the Rubin's model is rather transparent: each state corresponds to a counterfactual statement specifying how a unit in the population (e.g., a person) would have reacted to any given input. For example, r_1 represents units with perfect compliance, while r_2 represents units with perfect defiance. Similarly, r'_1 represents units with perfect response to treatment, while r'_0 represents units with no response ($y = y_0$) regardless of treatment. The counterfactual variables Y_1 and Y_0 usually invoked in Rubin's model can be obtained from R' as follows:

$$
\begin{aligned}
Y_1 = \quad \{Y \text{ if } D = d_1\} \quad &= \begin{cases} 1 & \text{if } R' = r'_1 \text{ or } R' = r'_3 \\ 0 & \text{otherwise} \end{cases} \\
Y_0 = \quad \{Y \text{ if } D = d_0\} \quad &= \begin{cases} 1 & \text{if } R' = r'_2 \text{ or } R' = r'_3 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

The transformations shown in Eqs. (32) - (35) demonstrate that the so-called counterfactual events (or "potential-response" variables) emerge in a natural way from a purely mathematical exercise aimed at reducing the domain of the latent variables to the bare minimum (see Balke & Pearl 1994).

The causal effect of the treatment can now be obtained directly from Eq. (35), giving

$$
\begin{aligned}
P(y_1|set(d_1)) &= P(r' = r'_1) + P(r' = r'_3) \\
P(y_1|set(d_0)) &= P(r' = r'_2) + P(r' = r'_3)
\end{aligned}
\tag{36}
$$

and

$$
R(D \rightarrow Y) = P(r' = r'_1) - P(r' = r'_2) \tag{37}
$$

The computational advantage of this scheme is two-fold. First, upper and lower bounds on $R(D \rightarrow Y)$ can now be produced by minimizing a linear function over a 15-dimensional vector space, rather than by dealing with the unspecified domain of U. Second, the constraints that the data $P(y, d|z_0)$ and $P(y, d|z_1)$ induce on the parameters of $P(r, r')$ are linear, while the constraints induced on the parameters $P(d|z, u)$ and $P(y|d, u)$ in the previous model are non-convex (see Eq. (50) in the Appendix). These advantages enables the use of linear programming techniques to obtain tighter bounds on the causal effect $R(D \rightarrow Y)$ [Balke & Pearl 1993]; such bounds are much harder to obtain in a model where U remains unspecified.

7.4 The causal effects of treatments on the treated

Much of the statistical literature assumes that

$$
R(D \rightarrow Z) \;=\; E[P(y_1|d_1, u) - P(y_1|d_0, u)] \tag{38}
$$

is the parameter of interest in experimental studies, since it predicts the impact of applying the treatment uniformly (or randomly) over the entire population. Heckman and Robb [1985] have argued that $R(D \rightarrow Z)$ is not a useful evaluation parameter if future treatment policies will involve selection decisions by the agents. In such cases, the parameter of interest should measure the impact of the treatment *on the treated*,

$$\alpha^* = E[P(y_1|d_1, u) - P(y_1|d_0, u)|D = d_1] \tag{39}$$

namely, the change of the mean response of the treated subjects compared to the mean response of these same subjects had they not been treated.

The analysis presented in Appendix II shows that, under the conditions specified in Figure 6, α^* can be assessed with much greater accuracy than $R(D \rightarrow Z)$. More remarkably, under the additional condition of "no intrusion" (namely, $P(d_1|z_0) = 0$ as in most clinical trials), α^* can be identified precisely.

The bounds governing α^* are:

$$\alpha^* \geq \frac{P(y_1|z_1) - P(y_1|z_0)}{P(d_1)/P(z_1)} - \frac{P(y_0, d_1|z_0)}{P(d_1)} \tag{40}$$

$$\alpha^* \leq \frac{P(y_1|z_1) - P(y_1|z_0)}{P(d_1)/P(z_1)} + \frac{P(y_1, d_1|z_0)}{P(d_1)} \tag{41}$$

The width the α^* bounds is equal to $\frac{P(d_1|z_0)}{P(d_1)} = \frac{P(z_0|d_1)}{1-P(z_1)}$ which, for small $P(z_1)$ approximates to $P(z_0|d_1)$: the fraction of treated persons that were not assigned treatment. Clearly, in situations where treatment may only be obtained by those designated by the assignment, α^* is perfectly identifiable and is given by:

$$\alpha^* = \frac{P(y_1|z_1) - P(y_1|z_0)}{P(d_1|z_1)} \quad \text{if} \quad P(d_1|z_0) = 0 \tag{42}$$

Unlike $R(D \rightarrow Z)$, α^* is not an intrinsic property of the treatment, as it varies with the strictness with which compliance is enforced. The significance of the α^* measure emerges primarily in studies where it is desired to evaluate the efficacy of an existing program on its current participants under a constant recruitment policy. Eq. (42) states that accurate evaluation does not require that participation be mandated at random; it is sufficient that notification about the program (or other types of encouragement) were sent to a randomly selected sample. If no participation has taken place without notification then α^* can be identified precisely (using Eq. (42)) by taking the mean response difference between the notified and unnotified populations, divided by the participation rate $P(d_1|z_1)$.

In cases where some unnotified participation may take place, the bounds of Eqs. (40)-(41) need be invoked and some additional parameters need be measured: $P(y_0, d_1|z_0)$, $P(y_1, d_1|z_0)$, and $P(d_1)$, all of which can be estimated from the participating population.

7.5 Remarks

Balke and Pearl (1993) have obtained significantly tighter bounds on the average treatment effect $R(D \rightarrow Y)$ relative to those shown in Eqs. (28)-(28). In particular, they have shown that the difference between the upper and lower bound cannot exceed the rate of noncompliance, $P(d_1|z_0) + P(d_0|z_1)$, and that it may in some cases collapse to a point estimate, even when as many as 50% of subjects switch over to unassigned treatments. Remarkably, the statistical establishment responsible for treatment evaluations has remained totally oblivious to such possibilities; in practice, the effectiveness of drugs and other treatments are still being assessed by an intent-to-treat analysis, namely by the $R(Z \rightarrow Y)$ measure.

The identifiability of α^* has been shown and discussed in [Angrist & Imbens 1991] but its significance has apparently been overlooked in the rest of the literature. Heckman [1992] has advocated the use of randomized eligibility in the evaluation of social programs and has shown that under such eligibility policy the mean-difference measure, $P(y_1|d_1) - P(y_1|d_0)$, is always less biased (relative to α^*) than in studies in which eligibility is not randomized. Remarkably, Eq. (42) shows that randomized eligibility, as a general experimental methodology, is more efficacious than expected; it does not just result in a lower bias – it eliminates bias altogether and permits the precise determination of α^*. The bounds presented in Eqs. (40) - (41) have apparently not been reported before.

8 Conclusions

I hope this paper convinces the reader that DAGs can be used not only for specifying assumptions of conditional independence but also as a formal language for organizing claims about external interventions and their interactions. I hope to have demonstrated as well that DAGs can serve as an analytical tool for quantifying, from non-experimental data, the effect of actions (given qualitative causal structure), for specifying and testing conditions under which randomized experiments are not necessary, and for aiding experimental design and model selection.

Most statisticians and decision analysts are reluctant to deal with problems involving causal considerations because we lack the mathematical notation for distinguishing causal influence from statistical association. The $set(x)$ notation introduced in this paper makes this distinction explicit, and permits us to evaluate probability expressions involving both observations and interventions. In [Pearl 1994] it is further shown that such evaluations can be performed by symbolic manipulations guided by the causal network. The resulting calculus yields amazingly simple solutions to a number of interesting problems in causal inference and should allow rank-and-file researchers to tackle practical problems that are generally considered too hard, or impossible. It is hoped therefore that the semantics and notation introduced in this paper will reinstate causality as a legitimate form of human knowledge, worthy of explicit formal representation and mathematical analysis.

Acknowledgement

This work benefitted from discussions with Alex Balke, David Chickering, David Cox, Arthur Dempster, Thomas Ferguson, David Galles, Moisés Goldszmidt, Stefan Lauritzen, Charles Manski, John Pratt, Paul Rosenbaum, James Robins, Don Rubin, Glenn Shafer, Michael Sobel, Pat Suppes, and Nanny Wermuth. The research was partially supported by Air Force grant #AFOSR 90 0136, NSF grant #IRI-9200918, Northrop Micro grant #92-123, and Rockwell Micro grant #92-122. Sections 1-6 will appear in the *Proceedings of 49th Session, International Statistical Institute: Invited papers*, Florence, Italy, August 1993.

BIBLIOGRAPHY

Angrist, J.D. and Imbens, G.W. (1991) Source of Identifying Information in Evaluation Models. *Discussion Paper 1568*, Department of Economics, Harvard University, Cambridge, MA.

Angrist, J.D., Imbens, G.W., and Rubins D.B. (1993) Identification of Causal effects Using instrumental Variables. *Technical Report*, Department of Economics, Harvard University, Cambridge , MA.

Balke, A. and Pearl, J. (1993) Nonparametric bounds on treatment effects in partial compliance studies. *Technical Report R-197*, UCLA Computer Science Department. (In preparation.)

Balke, A. and Pearl, J. (1994) Counterfactual probabilities: Computational methods, bounds and applications. *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, (eds R. Lopez de Mantaras and D. Poole), Morgan Kaufmann, San Mateo, CA, 46-54.

Bowden, R.J. and Turkington, D.A. (1984) *Instrumental Variables*, Cambridge University Press, Cambridge, MA.

Cox, D.R. and N. Wermuth (1993) Linear dependencies represented by chain graphs. *Statistical Science*, **8** (3), 204-218.

Cox, D.R. (1958) *The Planning of Experiments*, New York: John Wiley and Sons.

Cox, D.R. (1992) Causality: Some statistical aspects. *Journal of the Royal Statistical Society*, Series A, **155**, 291-301.

Dawid, A.P., (1979) Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series A*, 1-31.

Dempster, A.P. (1990) Causality and statistics. *Journal of Statistics Planning and Inference*, **25**, 261-278.

Efron, B. and Feldman D. (1991) Compliance as an Explanatory Variable in Clinical Trials. *Journal of the American Statistical Association*, **86**, 9-26.

Fisher, R.A. (1935) *The Design of Experiments*, Hafner, New York.

Freedman, D. (1987) As others see us: A case study in path analysis (with discussion). *Journal of Educational Statistics*, **12**, 101-223.

Geiger, D. (1990) Graphoids: A qualitative framework for probabilistic inference. Ph.D. Dissertation, University of California, Los Angeles, CA.

Geiger, D. and J. Pearl (1988) On the logic of causal models. *Proceedings of the 4th Workshop on Uncertainty in Artificial Intelligence*, St Paul, MN, pp. 136-147. Also in L. Kanal, etal. (eds.) (1990) *Uncertainty in Artificial Intelligence*, 4, North-Holland Publishing Co., Amsterdam, 3-14.

Glymour, C., R. Scheines, P. Spirtes and K. Kelly (1987) *Discovering Causal Structure*, Academic Press, Orlando, FL.

Goldszmidt, M. (1992) Qualitative Probabilities: A Normative Framework for Commonsense Reasoning. *Technical Report R-190*, UCLA Cognitive Systems Laboratory, Ph.D. Thesis.

Goldszmidt, M. and J. Pearl (1992) Default ranking: A practical framework for evidential reasoning, belief revision and update. in *Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning*, Morgan Kaufmann, San Mateo, CA, pp. 661-672.

Granger, C.W.J. (1988) Causality testing in a decision science, in *Causation in decision, Belief Change and Statistics I* (eds W. Harper and B. Skyrms), Kluwer Academic Publishers, pp. 1-20.

Heckman, J.J. (1992) Randomization and Social Policy Evaluation, in *Evaluations Welfare and Training Programs*, (eds C. Manski and I. Garfinkle), Harvard University Press, pp. 201-230.

Heckman, J. and R. Robb (1985) Alternative Methods for Evaluating the Impact of Interventions, in *Longitudinal Analysis of Labor Market Data*, (eds J. Heckman and B. Singer), Cambridge University Press, New York, NY.

Holland, P.W. (1986) Statistics and causal inference. *Journal of the American Statistics Association*, 81, 945-968.

Howard, R. (1990) From influence diagrams to relevance to knowledge. *Influence Diagrams, Belief Nets and Decision Analysis*, (eds R.M. Oliver and J.Q. Smith), Johy Wiley and Sons, Inc., New York, NY, 3-23.

Lauritzen, S.L. and D.J. Spiegelhalter (1988) Local computations with probabilities on graphical structures and their applications to expert systems. *Proceedings of the Royal Statistical Society*, Series B, **50**, 154-227.

Manski, C.F. (1990) Nonparametric bounds on treatment effects, *American Economic Review, Papers and Proceedings,* 80, 319-323, May 1990.

Neyman, J. (1935) Statistical problems in agricultural experimentation (with discussion). *Journal of the Royal Statistical Society*, **2**, 107-180.

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

Pearl, J. (1993) Belief networks revisited. *Artificial Intelligence*, **59**, 49-56.

Pearl, J. (1993) From conditional oughts to qualitative decision theory, in *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence*, (eds D. Heckerman and A. Mamdani), Morgan Kaufmann, San Mateo, CA, pp. 12-20.

Pearl, J. (1993) Causal diagrams for empirical research, to appear in *Biometrika.*

Pearl, J. and T. Verma (1991) A theory of inferred causation, in *Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference*, (eds J.A. Allen, R. Fikes and E. Sandewall), Morgan Kaufmann, San Mateo, CA, pp. 441-452.

Pearl, J., D. Geiger and T. Verma (1990) The logic of influence diagrams, in *Influence Diagrams, Belief Nets and Decision Analysis*, (eds R.M. Oliver and J.Q. Smith), John Wiley and Sons, Inc., New York, NY, pp. 67-87.

Pratt, J.W. and R. Schlaifer (1988) On the interpretation and observation of laws. *Journal of Econometrics*, **39**, 23-52.

Robins, J.M. (1989) The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies, in *Health Service Research Methodology: A Focus on AIDS*, (eds L. Sechrest, H. Freeman, and A. Mulley), NCHSR, U.S. Public Health Service, 113-159, 1989.

Rosenbaum, P.R. (1989) The role of known effects in observational studies. *Biometrics*, **45**, 557-569.

Rosenbaum, P. and D. Rubin (1983) The central role of propensity score in observational studies for causal effects. *Biometrica*, **70**, 41-55.

Rubin, D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688-701.

Simon, H.A. (1977) *Models of Discovery: and Other Topics in the Methods of Science*, D. Reidel, Dordrecht, Holland.

Spiegelhalter, D.J., S.L. Lauritzen, P.A. Dawid and R.G. Cowell (1993), Bayesian analysis in expert systems. *Statistical Science*, **8** (3), 219-247.

Spirtes, P., C. Glymour and R. Schienes (1993) *Causation, Prediction, and Search*, Springer-Verlag, New York.

Suppes, P. (1970) *A Probabilistic Theory of Causation*, North Holland, Amsterdam.

Verma, T.S. (1986) Causal networks: Semantics and expressiveness. *Technical Report R-65*, UCLA Cognitive Systems Laboratory. (Also in *Uncertainty in Artificial Intelligence*, (eds R. Shachter *et al.*), Elsevier Science Publishers, **4**, 69-76, 1990.

Appendix I (Derivation of Eqs. (26) and (30))

To prove (26), we use the conditional independence assumption of (22), and write

$$
\begin{aligned}
P(y|z,u) &= \sum_d P(y|z,d,u)\ P(d|z,u) && (43)\\
&= \sum_d P(y|d,u)\ P(d|z,u) && (44)
\end{aligned}
$$

which amounts to two equations,

$$
\begin{aligned}
P(y_1|z_1,u) &= P(y_1|d_1,u)\ P(d_1|z_1,u) + P(y_1|d_0,u)[1-P(d_1|z_1,u)]\\
P(y_1|z_0,u) &= P(y_1|d_1,u)\ P(d_1|z_0,u) + P(y_1|d_0,u)[1-P(d_1|z_0,u)] \qquad (45)
\end{aligned}
$$

Solving for $P(y_1|d_1,u)$ and $P(y_1|d_0,u)$, and taking their difference, gives

$$
P(y_1|d_1,u) - P(y_1|d_0,u) = \frac{P(y_1|z_1,u) - P(y_1|z_0,u)}{P(d_1|z_1,u) - P(d_1|z_0,u)} \qquad (46)
$$

Finally, taking the expectation (over u) on both sides, gives Eq. (26).

To prove (30), we write

$$
P(y,d|z) = \sum_u P(y|d,u)\ P(d|z,u)\ P(u) \qquad (47)
$$

and define the following four functions:

$$
\begin{aligned}
f_0(u) &= P(y_1|d_0,u) \qquad g_0(u) = P(d_1|u,z_0) && (48)\\
f_1(u) &= P(y_1|d_1,u) \qquad g_1(u) = P(d_1|u,z_1) && (49)
\end{aligned}
$$

This permits us to express six independent components of $P(y,d|z)$ as expectations of these functions:

$$
\begin{aligned}
P(y_1,d_0|z_0) &= E[f_0(1-g_0)] = a\\
P(y_1,d_0|z_1) &= E[f_0(1-g_1)] = b\\
P(d_1|z_0) &= E(g_0) = c\\
P(d_1|z_1) &= E(g_1) = d\\
P(y_1,d_1|z_0) &= E[f_1 \cdot g_0] = e\\
P(y_1,d_1|z_1) &= E[f_1 \cdot g_1] = h
\end{aligned}
\qquad (50)
$$

For any two random variables X and Y such that $0 \leq X \leq 1, 0 \leq Y \leq 1$ we have

$$
1 + E(XY) - E(Y) \geq E(X) \geq E(XY) \qquad (51)
$$

since $E[(1-X)(1-Y)] \geq 0$. This inequality holds for any pair of f, g functions (since they lie between 0 and 1) and we can write:

$$
\begin{aligned}
1 + E(f_1 g_0) - E(g_0) &\geq E(f_1) \geq E(f_1 g_0)\\
1 + E(f_1 g_1) - E(g_1) &\geq E(f_1) \geq E(f_1 g_1)\\
1 + E[f_0(1-g_0)] - E(1-g_0) &\geq E(f_0) \geq E[f_0(1-g_0)]\\
1 + E[f_0(1-g_1)] - E(1-g_1) &\geq E(f_0) \geq E[f_0(1-g_1)]
\end{aligned}
\qquad (52)
$$

or,

$$\begin{aligned} \max[h;e] \ &\leq E(f_1) \ \leq \min[(1+e-c);(1+h-d)] \\ \max[a;b] \ &\leq E(f_0) \ \leq \min[(a+c);(b+d)] \end{aligned} \tag{53}$$

Substituting back the $P(y,d|z)$ expressions from (48) and (49), gives Eqs. (28) and (29). Finally, lower bounding $E(f_1)$ and upper bounding $E(f_0)$ provides a lower bound for their difference

$$\begin{aligned} E(f_1) - E(f_0) \ &\geq \ \max[e;h] - \min[(a+c);(b+d)] \\ &\geq \ h - (a+c) \end{aligned} \tag{54}$$

from which Eq. (30) follows.

Appendix II (Derivation of Eqs. (40)-(41))

To evaluate

$$\alpha^* \;=\; E\{[P(y_1|d_1,u) - P(y_1|d_0,u)]|D \;=\; d_1\} \tag{55}$$

we define

$$\Delta(u) \;\equiv\; P(y_1|d_1,u) - P(y_1|d_0,u) \;=\; f_1(u) - f_0(u) \tag{56}$$

$$q \;\equiv\; P(z_1) \tag{57}$$

and write (using Eqs. (40)-(42)):

$$\begin{aligned}
\alpha^* &= E[\Delta(u)|D = d_1] \\
&= \textstyle\sum_u \Delta(u)P(u|d_1) \\
&= \tfrac{1}{P(d_1)}\textstyle\sum_u \Delta(u)P(d_1|u)P(u) \\
&= \tfrac{1}{P(d_1)}\textstyle\sum_u \sum_z \Delta(u)P(d_1|u,z)P(z)P(u) \\
&= \tfrac{1}{P(d_1)}\textstyle\sum_u \Delta(u)P(u)[P(z_1)g_1(u) + P(z_0)g_0(u)] \\
&= \tfrac{1}{P(d_1)}E\{[f_1(u) - f_0(u)][qg_1(u) + (1-q)g_0(u)]\} \\
&= \tfrac{1}{P(d_1)}E[qf_1g_1 + (1-q)f_1g_0 - qf_0g_1 - (1-q)f_0g_0] \\
&= \tfrac{1}{P(d_1)}[qh + (1-q)e - qE(f_0g_1) - (1-q)E(f_0g_0)] \\
&= \tfrac{1}{P(d_1)}[qh + (1-q)e - q(E(f_0) - b) - (1-q)(E(f_0) - a)] \\
&= \tfrac{1}{P(d_1)}[q(h+b) + (1-q)(e+a) - E(f_0)]
\end{aligned} \tag{58}$$

Substituting the expressions for $(h+b)$ and $(e+a)$ from (42), and using

$$a \leq E(f_0) < a + c \tag{59}$$

from (45), we obtain upper and lower bounds on α^*:

$$\frac{1}{P(d_1)}[P(y_1) - P(d_1|z_0) - P(y_1, d_0|z_0)] \leq \alpha^* \leq \frac{1}{P(d_1)}[P(y_1) - P(y_1, d_0|z_0)] \tag{60}$$

Alternatively, collecting common terms on both sides of (60), we get

$$-\frac{P(y_0, d_1|z_0)}{P(d_1)} \;\leq\; \alpha^* - \frac{P(y_1|z_1) - P(y_1|z_0)}{P(d_1)/P(z_1)} \;\leq\; \frac{P(y_1, d_1|z_0)}{P(d_1)} \tag{61}$$

which proves Eqs. (40)-(41)). Therefore,

$$\alpha^* = \frac{P(y_1|z_1) - P(y_1|z_0)}{P(d_1|z_1)} \quad \text{if} \quad P(d_1|z_0) = 0 \tag{62}$$