In A. Tucker, T. Gonzalez, H. Topi, J. Diaz-Herrera (Eds.) *Computing Handbook* (Renamed), Third Edition, Volume 1, Intelligent Systems section, Chapman and Hall/CRC, 2014.

TECHNICAL REPORT
R-236
May 2014

Graphical Models for Probabilistic and Causal Reasoning

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
(310) 825-3243
(310) 825-2273 Fax
judea@cs.ucla.edu

1 INTRODUCTION

This chapter surveys the development of graphical models known as Bayesian networks, summarizes their semantical basis and assesses their properties and applications to reasoning and planning.

Bayesian networks are directed acyclic graphs (DAGs) in which the nodes represent variables of interest (e.g., the temperature of a device, the gender of a patient, a feature of an object, the occurrence of an event) and the links represent causal influences among the variables. The strength of an influence is represented by conditional probabilities that are attached to each cluster of parents-child nodes in the network.

Figure 1 illustrates a simple yet typical Bayesian network. It describes the causal relationships among the season of the year (X_1), whether rain falls (X_2) during the season, whether the sprinkler is on (X_3) during that season, whether the pavement would get wet (X_4), and whether the pavement would be slippery (X_5). All variables in this figure are binary, taking a value of either true or false, except the root variable X_1 which can take one of four values: Spring, Summer, Fall, or Winter. Here, the absence of a direct link between X_1 and X_5, for example, captures our understanding that the influence of seasonal variations on the slipperiness of the pavement is mediated by other conditions (e.g., the wetness of the pavement).

As this example illustrates, a Bayesian network constitutes a model of the environment rather than, as in many other knowledge representation schemes (e.g., logic, rule-based systems and neural networks), a model of the reasoning process. It simulates, in fact, the causal mechanisms that operate in the environment, and thus allows the investigator to answer a variety of queries, including: associational queries, such as "Having observed A, what can we expect of B?"; abductive queries, such as "What is the most plausible explanation for a given set of observations?"; and control queries; such as "What will happen if we intervene and act on the environment?" Answers to the first type of query depend only on probabilistic

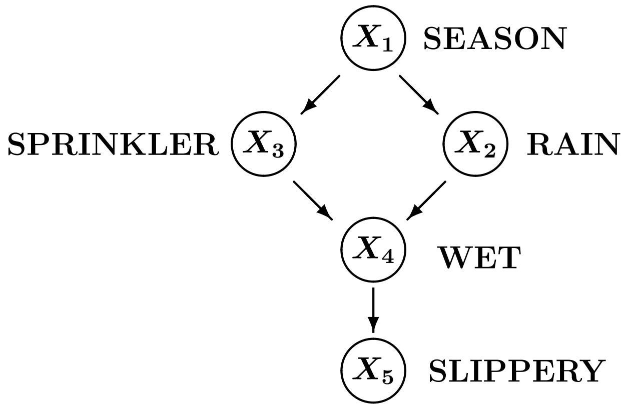

Figure 1: A Bayesian network representing causal influences among five variables.

knowledge of the domain, while answers to the second and third types rely on the causal knowledge embedded in the network. Both types of knowledge, associative and causal, can effectively be represented and processed in Bayesian networks.

The associative facility of Bayesian networks may be used to model cognitive tasks such as object recognition, reading comprehension, and temporal projections. For such tasks, the probabilistic basis of Bayesian networks offers a coherent semantics for coordinating top-down and bottom-up inferences, thus bridging information from high-level concepts and low-level percepts. This capability is important for achieving selective attention, that is, selecting the most informative next observation before actually making the observation. In certain structures, the coordination of these two modes of inference can be accomplished by parallel and distributed processes that communicate through the links in the network.

However, the most distinctive feature of Bayesian networks, stemming largely from their causal organization, is their ability to represent and respond to changing configurations. Any local reconfiguration of the mechanisms in the environment can be translated, with only minor modification, into an isomorphic reconfiguration of the network topology. For example, to represent a disabled sprinkler, we simply delete from the network all links incident to the node "Sprinkler". To represent a pavement covered by a tent, we simply delete the link between "Rain" and "Wet". This flexibility is often cited as the ingredient that marks the division between deliberative and reactive agents, and that enables the former to manage novel situations instantaneously, without requiring retraining or adaptation.

2 HISTORICAL BACKGROUND

Networks employing directed acyclic graphs (DAGs) have a long and rich tradition, starting with the geneticist Sewall Wright (1921). Wright (1921) He developed a method called *Path Analysis* (Wright, 1934), which later became an established representation of causal models in economics (Wold, 1964), sociology (Blalock, Jr., 1971; Kenny, 1979), and psychology (Duncan, 1975). Good (1961) used DAGs to represent causal hierarchies of binary variables with disjunctive causes. *Influence diagrams* represent another application of DAG representation (Howard and Matheson, 1981). Developed for decision analysis, they contain both event nodes and decision nodes. *Recursive models* is the name given to such networks by statisticians seeking meaningful and effective decompositions of contingency tables (Lauritzen, 1982; Wermuth and Lauritzen, 1983; Kiiveri et al., 1984).

The role of the network in the applications above was primarily to provide an efficient description for probability functions; once the network was configured, all subsequent computations were pursued by symbolic manipulation of probability expressions. The potential for the network to work as a computational architecture, and hence as a model of cognitive activities, was noted in Pearl (1982), where a distributed scheme was demonstrated for probabilistic updating on tree-structured networks. The motivation behind this particular development was the modeling of distributed processing in reading comprehension (Rumelhart, 1976), where both top-down and bottom-up inferences are combined to form a coherent interpretation. This dual mode of reasoning is at the heart of Bayesian updating, and in fact motivated Reverend Bayes's original 1763 calculations of posterior probabilities (representing explanations), given prior probabilities (representing causes), and likelihood functions (representing evidence).

Bayesian networks have not attracted much attention in the logic and cognitive modeling circles, but they did in expert systems. The ability to coordinate bi-directional inferences filled a void in expert systems technology of the late 1970s, and it is in this area that Bayesian networks truly flourished. Over the past ten years, Bayesian networks have become a tool of great versatility and power, and they are now the most common representation scheme for probabilistic knowledge (Neapolitan, 1990; Spiegelhalter et al., 1993; Darwiche, 2009; Koller and Friedman, 2009). They have been used to aid in the diagnosis of medical patients (Heckerman, 1991; Andersen et al., 1989; Geiger et al., 1990; Jiang and Cooper, 2010) and malfunctioning systems (Fenton and Neil, 2012); to understand stories (Charniak and Goldman, 1991), to retrieve documents (de Campos et al., 2004), to interpret pictures (Levitt et al., 1990), to perform filtering, smoothing, and prediction (Weiss and Pearl, 2010), to analyze gene expressions (Friedman et al., 2000), genetic counseling (Uebersax, 2004), semantic search (Koumenides and Shadbolt, 2012), error-correcting codes (McEliece et al., 1998), speech recognition (Zweig, 1998), to facilitate planning in uncertain environments (Guestrin, 2003; Beaudry et al., 2010), and to study causation, nonmonotonicity, action, change, and attention. Some of these applications are described in Pearl (1988); Russell and Norvig (2003); Buede (2009).

3 BAYESIAN NETWORKS AS CARRIERS OF PROBABILISTIC INFORMATION

3.1 Formal Semantics

Given a DAG G and a joint distribution P over a set $X = \{X_1, ..., X_n\}$ of discrete variables, we say that G *represents* P if there is a one-to-one correspondence between the variables in X and the nodes of G, 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 G. For example, the DAG in Figure 1 induces the decomposition

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

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 notation (Dawid, 1979), we can state this set of independencies as

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

Such a set of independencies is 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:

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

In addition to these, the decomposition of Eq. (1) implies many more independencies, the sum total of which can be identified from the DAG using the graphical criterion of d-separation (Pearl, 1988):

Definition 1 (d-separation) *Let a* path *in a DAG be a sequence of consecutive edges, of any directionality. A path p is said to be d-separated (or blocked) by a set of nodes Z iff:*

(i) p contains a chain $i \longrightarrow j \longrightarrow k$ or a $fork$ $i \longleftarrow j \longrightarrow k$ such that the middle node j is in Z, or,

(ii) p contains an inverted fork $i \longrightarrow j \longleftarrow k$ such that neither the middle node j nor any of its descendants (in G) are in Z.

If X, Y, and Z are three disjoint subsets of nodes in a DAG G, then Z is said to d-separate X from Y, denoted $(X \perp\!\!\!\perp Y|Z)_G$, iff Z d-separates every path from a node in X to a node in Y.

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, $(X_2 \perp\!\!\!\perp X_3|X_1)_G$ holds in Figure 1. 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, $(X_2 \perp\!\!\!\perp X_3|\{X_1, X_5\})_G$ does not hold; in words, 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 arrows converging at X_4.

The d-separation criterion has been shown to be both necessary and sufficient relative to the set of distributions that are represented by a DAG G (Verma and Pearl, 1990; Geiger et al., 1990). 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 G. Furthermore, the d-separation criterion can be tested in time linear in the number of edges in G. 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.

An important property that follows from the d-separation characterization is a criterion for determining when two dags are observationally equivalent, that is, every probability distribution that is represented by one of the dags is also represented by the other:

Theorem 1 (Verma and Pearl, 1990) *Two dags are observationally equivalent if and only if they have the same sets of edges and the same sets of v-structures, that is, head-to-head arrows with non-adjacent tails.*

The soundness of the d-separation criterion holds not only for probabilistic independencies but for any abstract notion of conditional independence that obeys the semi-graphoid axioms (Verma and Pearl, 1990; Geiger et al., 1990). Additional properties of DAGs and their applications to evidential reasoning in expert systems are discussed in Pearl (1988, 1993a); Pearl et al. (1990); Geiger (1990); Lauritzen and Spiegelhalter (1988); Spiegelhalter et al. (1993); Darwiche (2009); Lauritzen (1996).

3.2 Inference Algorithms

The first algorithms proposed for probability updating in Bayesian networks used message-passing architecture and were limited to trees (Pearl, 1982) and singly connected networks (Kim and Pearl, 1983). The idea was to assign each variable a simple processor, forced to communicate only with its neighbors, and to permit asynchronous back-and-forth message-passing until equilibrium was achieved. Coherent equilibrium can indeed be achieved this way, but only in singly connected networks, where an equilibrium state occurs in time proportional to the diameter of the network.

Many techniques have been developed and refined to extend the tree-propagation method to general, multiply connected networks. Among the most popular are Shachter's (1988) method of node elimination, Lauritzen and Spiegelhalter's (1988) method of clique-tree propagation, loop-cut conditioning (Pearl, 1988, Chapter 4.3), interative conditioning (Darwiche, 2009; Dechter, 1999).

Clique-tree propagation, works as follows. Starting with a directed network representation, the network is transformed into an undirected graph that retains all of its original dependencies. This graph, sometimes called a Markov network (Pearl, 1988, Chapter 3.1), is then triangulated to form local clusters of nodes (cliques) that are tree-structured. Evidence propagates from clique to clique by ensuring that the probability of their intersection set is the same, regardless of which of the two cliques is considered in the computation. Finally, when the propagation process subsides, the posterior probability of an individual variable is computed by projecting (marginalizing) the distribution of the hosting clique onto this variable.

Whereas the task of updating probabilities in general networks is NP-hard (Rosenthal, 1975; Cooper, 1990), the complexity for each of the three methods cited above is exponential in the size of the largest clique found in some triangulation of the network. It is fortunate that these complexities can be estimated prior to actual processing; when the estimates exceed reasonable bounds, an approximation method such as stochastic simulation (Pearl, 1987; Henrion, 1988) and belief propagation (Weiss and Pearl, 2010) can be used instead. Learning techniques have also been developed for systematic updating of the conditional probabilities $P(x_i|\mathbf{pa}_i)$ so as to match empirical data (Spiegelhalter and Lauritzen, 1990; Darwiche, 2009).

3.3 System's properties

By providing graphical means for representing and manipulating probabilistic knowledge, Bayesian networks overcome many of the conceptual and computational difficulties of earlier knowledge-based systems (Pearl, 1988). Their basic properties and capabilities can be summarized as follows:

1. Graphical methods make it easy to maintain consistency and completeness in probabilistic knowledge bases. They also define modular procedures of knowledge acquisition that reduce significantly the number of assessments required (Pearl, 1988; Heckerman, 1991).

2. Independencies can be dealt with explicitly. They can be articulated by an expert, encoded graphically, read off the network, and reasoned about; yet they forever remain robust to numerical imprecision (Geiger, 1990; Geiger et al., 1990; Pearl et al., 1990).

3. Graphical representations uncover opportunities for efficient computation. Distributed updating is feasible in knowledge structures which are rich enough to exhibit intercausal interactions (e.g., "explaining away") (Pearl, 1982; Kim and Pearl, 1983). And, when extended by clustering or conditioning, tree-propagation algorithms are capable of updating networks of arbitrary topology (Lauritzen and Spiegelhalter, 1988; Shachter, 1986; Pearl, 1988).

4. The combination of predictive and abductive inferences resolves many problems encountered by first-generation expert systems and renders belief networks a viable model for cognitive functions requiring both top-down and bottom-up inferences (Pearl, 1988; Shafer and Pearl, 1990).

5. The causal information encoded in Bayesian networks facilitates the analysis of action sequences, their consequences, their interaction with observations, their expected utilities and, hence, the synthesis of plans and strategies under uncertainty (Dean and Wellman, 1991; Pearl, 1993b, 1994b).

6. The isomorphism between the topology of Bayesian networks and the stable mechanisms that operate in the environment facilitates modular reconfiguration of the network in response to changing conditions, and permits deliberative reasoning about novel situations.

3.4 Later Developments

3.4.1 Causal discovery

One of the most exciting prospects in recent years has been the possibility of using the theory of Bayesian networks to discover causal structures in raw statistical data. Several systems have been developed for this purpose (Rebane and Pearl, 1987; Pearl and Verma, 1991; Spirtes et al., 1993, 2000), which systematically search and identify causal structures with hidden variables from empirical data. Technically, because these algorithms rely merely on conditional independence relationships, the structures found are valid only if one is willing to accept weaker forms of guarantees than those obtained through controlled randomized experiments: minimality and stability (Pearl and Verma, 1991; Pearl, 2000, Ch. 2). Minimality guarantees that any other structure compatible with the data is necessarily less specific, and hence less falsifiable and 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 experimental conditions will render that structure no longer compatible with the data. With these forms of guarantees, the theory provides criteria and algorithms for identifying genuine and spurious causes, with or without temporal information.

Alternative methods of identifying structure in data assign prior probabilities to the parameters of the network and use Bayesian updating to score the degree to which a given network fits the data (Cooper and Herskovits, 1991; Heckerman et al., 1994). These methods have the advantage of operating well under small sample conditions, but encounter difficulties coping with hidden variables.

Tian and Pearl (2001a,b) developed yet another method of causal discovery based on the detection of "shocks," or spontaneous local changes in the environment which act like "Nature's interventions," and unveil causal directionality toward the consequences of those shocks.

3.4.2 Plain beliefs

In mundane decision making, beliefs are revised not by adjusting numerical probabilities but by tentatively accepting some sentences as "true for all practical purposes". Such sentences, often named *plain beliefs*, exhibit both logical and probabilistic character. As in classical logic, they are propositional and deductively closed; as in probability, they are subject to retraction and to varying degrees of entrenchment (Spohn, 1988; Goldszmidt and Pearl, 1992).

Bayesian networks can be adopted to model the dynamics of plain beliefs by replacing ordinary probabilities with non-standard probabilities, that is, probabilities that are infinitesimally close to either zero or one. This amounts to taking an "order of magnitude" approximation of empirical frequencies, and adopting new combination rules tailored to reflect this approximation. The result is an integer-addition calculus, very similar to probability calculus, with summation replacing multiplication and minimization replacing addition. A plain belief is then identified as a proposition whose negation obtains an infinitesimal probability (i.e., an integer greater than zero). The connection between infinitesimal probabilities, nonmonotonic logic, and interative belief revision is described in Pearl (1994a); Goldszmidt and Pearl (1996); Darwiche and Pearl (1997); Spohn (2012).

This combination of infinitesimal probabilities with the causal information encoded by the structure of Bayesian networks facilitates linguistic communication of belief commitments, explanations, actions, goals, and preferences, and serves as the basis for qualitative planning under uncertainty (Darwiche and Pearl, 1994; Goldszmidt and Pearl, 1992; Pearl, 1993b; Darwiche and Goldszmidt, 1994). Some of these aspects will be presented in the next section.

4 BAYESIAN NETWORKS AS CARRIERS OF CAUSAL INFORMATION

The interpretation of DAGs as carriers of independence assumptions does not necessarily imply 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 typical of causal organizations and some of these patterns can only be given meaningful interpretation in terms of causation. Consider, for example, two independent events, E_1 and E_2, that have a common effect E_3. This triple represents an intransitive pattern of dependencies: E_1 and E_3 are dependent, E_3 and E_2 are dependent, yet E_1 and E_2 are independent. Such a pattern cannot be represented in undirected graphs because connectivity in undirected graphs is transitive. Likewise, it is not easily represented in neural networks, because E_1 and E_2 should turn dependent once E_3 is known. The DAG representation provides a convenient language for intransitive dependencies via the converging pattern $E_1 \rightarrow E_3 \leftarrow E_2$, which implies the independence of E_1 and E_2 as well as the dependence of E_1 and E_3 and of E_2 and E_3. The distinction between transitive and intransitive dependencies is the basis for the causal discovery systems of Pearl and Verma (1991) and Spirtes et al. (1993) (see Section 3.4.1).

However, the Markovian account still leaves open the question of how such intricate patterns of independencies relate to the more basic notions associated with causation, such as influence, manipulation, and control, which reside outside the province of probability theory. The connection is made in the mechanism-based account of causation.

The basic idea behind this account goes back to structural equations models (Wright, 1921; Haavelmo, 1943; Simon, 1953) and it was adapted in Pearl and Verma (1991) for defining probabilistic causal theories, as follows. Each child-parents family in a DAG G 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 G, and $\epsilon_i,\ 0 < i < n$, 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 that characterizes DAG models (see Eq. (1)). However, the functional characterization $X_i = f_i(\mathbf{pa}_i, \epsilon_i)$ also specifies how the resulting distributions would change in response to external interventions, since each function is presumed to represent a stable mechanism in the domain and therefore remains constant unless specifically altered. Thus, once we know the identity of the mechanisms altered by the intervention and the nature of the alteration, the overall effect of an intervention can be predicted by modifying the appropriate equations in the model of Eq. (5) and using the modified model to compute a new probability function of the observables.

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 *atomic* intervention 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 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, then we can view each Bayesian network as an efficient code for predicting the effects of atomic interventions and of various combinations of such interventions, without representing these interventions explicitly.

4.1 Causal theories, actions, causal effect, and identifiability

Definition 2 *A causal theory is a 4-tuple*

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

where

(i) $V = \{X_1, \ldots, X_n\}$ *is a set of observed variables*

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

(iii) $P(\mathbf{u})$ *is a distribution function over* $U_1, \ldots, U_n$*, and*

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

$$X_i = f_i(PA_i, u) \quad i = 1, \ldots, n \tag{6}$$

where PA_i *is a subset of* V *not containing* X_i*.*

The variables PA_i (connoting parents) are considered the direct causes of X_i and they define a directed graph G which may, in general, be cyclic. Unlike the probabilistic definition of "parents" in Bayesian networks (Eq. (1)), PA_i is selected from V by considering functional mechanisms in the domain, not by conditional independence considerations. We will assume that the set of equations in (6) has a unique solution for $X_i, \ldots, X_n$, given any value of the disturbances $U_1, \ldots, U_n$. Therefore the distribution $P(\mathbf{u})$ induces a unique distribution on the observables, which we denote by $P_T(\mathbf{v})$.

We will consider concurrent actions of the form $do(X = x)$, where $X \subseteq V$ is a set of variables and x is a set of values from the domain of X. In other words, $do(X = x)$ represents a combination of actions that forces the variables in X to attain the values x.

Definition 3 *(effect of actions) The effect of the action* $do(X = x)$ *on a causal theory* T *is given by a subtheory* T_x *of* T*, where* T_x *obtains by deleting from* T *all equations corresponding to variables in* X *and substituting the equations* $X = x$ *instead.*

The framework provided by Definitions 2 and 3 permits the coherent formalization of many subtle concepts in causal discourse, such as causal influence, causal effect, causal relevance, average causal effect, identifiability, counterfactuals, exogeneity, and so on. Examples are:

$***$ X **influences** Y in context u if there are two values of X, x and x', such that the solution for Y under $U = u$ and $do(X = x)$ is different from the solution under $U = u$ and $do(X = x')$.

$***$ X **can potentially influence** Y if there exist both a subtheory T_z of T and a context $U = u$ in which X influences Y.

$***$ Event $X = x$ **is the (singular) cause of event** $Y = y$ if (i) $X = x$ and $Y = y$ are true, and (ii) in every context u compatible with $X = x$ and $Y = y$, and for all $x' \neq x$, the solution of Y under $do(X = x')$ is not equal to y.

The definitions above are deterministic. Probabilistic causality emerges when we define a probability distribution $P(u)$ for the U variables, which, under the assumption that the equations have a unique solution, induces a unique distribution on the endogenous variables for each combination of atomic interventions.

Definition 4 *(causal effect) Given two disjoint subsets of variables, $X \subseteq V$ and $Y \subseteq V$, the causal effect of X on Y, denoted $P_T(y|\hat{x})$, is a function from the domain of X to the space of probability distributions on Y, such that*

$$P_T(y|\hat{x}) = P_{T_x}(y) \tag{7}$$

for each realization x of X. In other words, for each $x \in dom(X)$, the causal effect $P_T(y|\hat{x})$ gives the distribution of Y induced by the action $do(X = x)$.

Note that causal effects are defined relative to a given causal theory T, though the subscript T is often suppressed for brevity.

Definition 5 *(identifiability) Let $Q(T)$ be any computable quantity of a theory T; Q is* identifiable *in a class M of theories if for any pair of theories T_1 and T_2 from M, $Q(T_1) = Q(T_2)$ whenever $P_{T_1}(v) = P_{T_2}(v)$.*

Identifiability is essential for estimating quantities Q from P alone, without specifying the details of T, so that the general characteristics of the class M suffice. The question of interest in planning applications is the identifiability of the causal effect $Q = P_T(y|\hat{x})$ in the class M_G of theories that share the same causal graph G. Relative to such classes we now define:

Definition 6 *(causal-effect identifiability) The* causal effect *of X on Y is said to be* identifiable *in M_G if the quantity $P(y|\hat{x})$ can be computed uniquely from the probabilities of the observed variables, that is, if for every pair of theories T_1 and T_2 in M_G such that $P_{T_1}(v) = P_{T_2}(v)$, we have $P_{T_1}(y|\hat{x}) = P_{T_2}(y|\hat{x})$.*

The identifiability of $P(y|\hat{x})$ ensures that it is possible to infer the effect of action $do(X = x)$ on Y from two sources of information:

(i) passive observations, as summarized by the probability function $P(v)$,

(ii) the causal graph, G, which specifies, qualitatively, which variables make up the stable mechanisms in the domain or, alternatively, which variables participate in the determination of each variable in the domain.

Simple examples of identifiable causal effects will be discussed in the next subsection.

4.2 Acting vs. Observing

Consider the example depicted in Figure 1. The corresponding theory consists of five functions, each representing an autonomous mechanism:

$$\begin{aligned} X_1 &= U_1 \\ X_2 &= f_2(X_1, U_2) \\ X_3 &= f_3(X_1, U_3) \\ X_4 &= f_4(X_3, X_2, U_4) \\ X_5 &= f_5(X_4, U_5) \end{aligned} \tag{8}$$

To represent the action "turning the sprinkler ON", $do(X_3 = \mathrm{ON})$, we delete the equation $X_3 = f_3(x_1, u_3)$ from the theory of Eq. (8), and replace it with $X_3 = \mathrm{ON}$. The resulting subtheory, $T_{X_3=\mathrm{ON}}$, contains all the information needed for computing the effect of the actions on other variables. It is easy to see from this subtheory that the only variables affected by the action are X_4 and X_5, that is, the descendant, of the manipulated variable X_3.

The probabilistic analysis of causal theories becomes particularly simple when two conditions are satisfied:

1. The theory is recursive, i.e., there exists an ordering of the variables $V = \{X_1, \ldots, X_n\}$ such that each X_i is a function of a subset PA_i of its predecessors

$$X_i = f_i(PA_i, U_i), \qquad PA_i \subseteq \{X_1, \ldots, X_{i-1}\} \tag{9}$$

2. The disturbances $U_1, \ldots, U_n$ are mutually independent, that is,

$$P(u) = \prod_i P(u_i) \tag{10}$$

These two conditions, also called Markovian, are the basis of the independencies embodied in Bayesian networks (Eq. 1) and they enable us to compute causal effects directly from the conditional probabilities $P(x_i|\mathbf{pa}_i)$, without specifying the functional form of the functions f_i, or the distributions $P(u_i)$ of the disturbances. This is seen immediately from the following observations: The distribution induced by any Markovian theory T is given by the product in Eq. (1)

$$P_T(x_1, \ldots, x_n) = \prod_i P(x_i|\mathbf{pa}_i) \tag{11}$$

where $\mathbf{pa}_i$ are (values of) the parents of X_i in the diagram representing T. At the same time, the subtheory $T_{x'_j}$, representing the action $do(X_j = x'_j)$ is also Markovian, hence it also induces a product-like distribution

$$P_{T_{x'_j}}(x_1, \ldots, x_n) \;\; = \begin{cases} \prod_{i \neq j} P(x_i|\mathbf{pa}_i) = \frac{P(x_1,\ldots,x_n)}{P(x_j|\mathbf{pa}_j)} & \text{if} \;\; x_j = x'_j \\ 0 & \text{if} \;\; x_j \neq x'_j \end{cases} \tag{12}$$

where the partial product reflects the surgical removal of the

$$X_j = f_j(\mathbf{pa}_j, U_j)$$

from the theory of equation (9) (see (Pearl, 1993a)).

In the example of Figure 1, the pre-action distribution is given by the product

$$P_T(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_4) \tag{13}$$

while the surgery corresponding to the action $do(X_3 = \text{ON})$ amounts to deleting the link $X_1 \rightarrow X_3$ from the graph and fixing the value of X_3 to ON, yielding the post-action distribution:

$$P_T(x_1, x_2, x_4, x_5|do(X_3 = \text{ON})) = P(x_1)\; P(x_2|x_1)\; P(x_4|x_2, X_3 = \text{ON})\; P(x_5|x_4) \tag{14}$$

Note the difference between the action $do(X_3 = \text{ON})$ and the observation $X_3 = \text{ON}$. The latter is encoded by ordinary Bayesian conditioning, while the former by conditioning a mutilated graph, with the link $X_1 \rightarrow X_3$ removed. This mirrors indeed the difference between seeing and doing: after observing that the sprinkler is ON, we wish to infer that the season is dry, that it probably did not rain, and so on; no such inferences should be drawn in evaluating the effects of the deliberate action "turning the sprinkler ON". The amputation of $X_3 = f_3(X_1, U_3)$ from (8) ensures the suppression of any abductive inferences from X_3, the action's recipient.

Note also that Equations (11) through (14) are independent of T, in other words, the pre-actions and post-action distributions depend only on observed conditional probabilities but are independent of the particular functional form of $\{f_i\}$ or the distribution $P(\mathbf{u})$ which generate those probabilities. This is the essence of identifiability as given in Definition 6, which stems from the Markovian assumptions (9) and (10). Section 4.3 will demonstrate that certain causal effects, though not all, are identifiable even when the Markovian property is destroyed by introducing dependencies among the disturbance terms.

Generalization to multiple actions and conditional actions are reported in Pearl and Robins (1995). Multiple actions $do(X = x)$, where X is a compound variable result in a distribution similar to (12), except that all factors corresponding to the variables in X are removed from the product in (11). Stochastic conditional strategies (Pearl, 1994b) of the form

$$do(X_j = x_j) \text{ with probability } P^*(x_j|\mathbf{pa}^*_j) \tag{15}$$

where $\mathbf{pa}^*_j$ is the support of the decision strategy, also result in a product decomposition similar to (11), except that each factor $P(x_j|\mathbf{pa}_j)$ is *replaced* with $P^*(x_j|\mathbf{pa}^*_j)$.

The surgical procedure described above is not limited to probabilistic analysis. The causal knowledge represented in Figure 1 can be captured by logical theories as well, for example,

$$\begin{aligned}
x_2 &\iff [(X_1 = \text{Winter}) \vee (X_1 = \text{Fall}) \vee ab_2] \wedge \neg ab'_2 \\
x_3 &\iff [(X_1 = \text{Summer}) \vee (X_1 = \text{Spring}) \vee ab_3] \wedge \neg ab'_3 \\
x_4 &\iff (x_2 \vee x_3 \vee ab_4) \wedge \neg ab'_4 \\
x_5 &\iff (x_4 \vee ab_5) \wedge \neg ab'_5
\end{aligned} \tag{16}$$

where x_i stands for $X_i = true$, and ab_i and ab'_i stand, respectively, for trigerring and inhibiting abnormalities. The double arrows represent the assumption that the events on the r.h.s. of each equation are the *only* direct causes for the l.h.s, thus identifying the surgery implied by any action.

It should be emphasized though that the models of a causal theory are not made up merely of truth value assignments which satisfy the equations in the theory. Since each equation represents an autonomous process, the content of each individual equation must be specified in any model of the theory, and this can be encoded using either the graph (as in Figure 1) or the generic description of the theory, as in (8). Alternatively, we can view a model of a causal theory to consist of a mutually consistent set of submodels, with each submodel being a standard model of a single equation in the theory.

4.3 Action Calculus

The identifiability of causal effects demonstrated in Section 4.1 relies critically on the Markovian assumptions (9) and (10). If a variable that has two descendants in the graph is unobserved, the disturbances in the two equations are no longer independent, the Markovian property (9) is violated and identifiability may be destroyed. This can be seen easily from Eq. (12); if any parent of the manipulated variable X_j is unobserved, one cannot estimate the conditional probability $P(x_j|\mathbf{pa}_j)$, and the effect of the action $do(X_j = x_j)$ may not be predictable from the observed distribution $P(x_1, \ldots, x_n)$. Fortunately, certain causal effects are identifiable even in situations where members of $\mathbf{pa}_j$ are unobservable (Pearl, 1993a) and, moreover, polynomial tests are now available for deciding when $P(x_i|\hat{x}_j)$ is identifiable, and for deriving closed-form expressions for $P(x_i|\hat{x}_j)$ in terms of observed quantities (Galles and Pearl, 1995).

These tests and derivations are based on a symbolic calculus (Pearl, 1994a, 1995) to be described in the sequel, in which interventions, side by side with observations, are given explicit notation, and are permitted to transform probability expressions. The transformation rules of this calculus reflect the understanding that interventions perform "local surgeries" as described in Definition 3, i.e., they overrule equations that tie the manipulated variables to their pre-intervention causes.

Let X, Y, and Z be arbitrary disjoint sets of nodes in a DAG G. We denote by $G_{\overline{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X. Likewise, we denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X. To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{X}\underline{Z}}$.

Finally, the expression $P(y|\hat{x}, z) \stackrel{\Delta}{=} P(y, z|\hat{x})/P(z|\hat{x})$ stands for the probability of $Y = y$ given that $Z = z$ is observed and X is held constant at x.

Theorem 2 *Let G be the directed acyclic graph associated with a Markovian causal theory, and let $P(\cdot)$ stand for the probability distribution induced by that theory. For any disjoint subsets of variables $X, Y, Z,$ and W we have:*

Rule 1 *(Insertion/deletion of observations):*

$$P(y|\hat{x}, z, w) = P(y|\hat{x}, w) \quad \textit{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}} \tag{17}$$

Rule 2 *(Action/observation exchange):*

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, z, w) \quad \textit{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}} \tag{18}$$

Rule 3 *(Insertion/deletion of actions):*

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, w) \quad \textit{if} \quad (Y \perp\!\!\!\perp Z|X,\ W)_{G_{\overline{X},\ \overline{Z(W)}}} \tag{19}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

Each of the inference rules above follows from the basic interpretation of the "$\hat{x}$" operator as a replacement of the causal mechanism that connects X to its pre-action parents by a new mechanism $X = x$ introduced by the intervening force. The result is a submodel characterized by the subgraph $G_{\overline{X}}$ (named "manipulated graph" in Spirtes et al. (1993)) which supports all three rules.

Corollary 1 A causal effect q: $P(y_1, ..., y_k|\hat{x}_1, ..., \hat{x}_m)$ is identifiable in a model characterized by a graph G if there exists a finite sequence of transformations, each conforming to one of the inference rules in Theorem 2, which reduces q into a standard (i.e., hat-free) probability expression involving observed quantities. □

Although Theorem 2 and Corollary 1 require the Markovian property, they can also be applied to non Markovian, recursive theories because such theories become Markovian if we consider the unobserved variables as part of the analysis, and represent them as nodes in the graph. To illustrates, assume that variable X_1 in Figure 1 is unobserved, rendering the disturbances U_3 and U_2 dependent since these terms now include the common influence of X_1. Theorem 2 tells us that the causal effect $P(x_4|\hat{x}_3)$ is identifiable, because:

$$P(x_4|\hat{x}_3) = \sum_{x_2} P(x_4|\hat{x}_3, x_2) P(x_2|\hat{x}_3)$$

Rule 3 permits the deletion

$$P(x_2|\hat{x}_3) = P(x_2), \text{ because } (X_2 \perp\!\!\!\perp X_3)_{G_{\overline{X}_3}},$$

while Rule 2 permits the exchange

$$P(x_4|\hat{x}_3, x_2) = P(x_4|x_3, x_2), \text{ because } (X_4 \perp\!\!\!\perp X_3|X_2)_{G_{\underline{X_3}}}.$$

This gives

$$P(x_4|\hat{x}_3) = \sum_{x_2} P(x_4|x_3, x_2)P(x_2)$$

which is a "hat-free" expression, involving only observed quantities.

In general, it can be shown (Pearl, 1995) that:

1. The effect of interventions can often be identified (from nonexperimental data) without resorting to parametric models,

2. The conditions under which such nonparametric identification is possible can be determined by simple graphical tests[1], and,

3. When the effect of interventions is not identifiable, the causal graph may suggest nontrivial experiments which, if performed, would render the effect identifiable.

The ability to assess the effect of interventions from nonexperimental data has immediate applications in the medical and social sciences, since subjects who undergo certain treatments often are not representative of the population as a whole. Such assessments are also important in AI applications where an agent needs to predict the effect of the next action on the basis of past performance records, and where that action has never been enacted under controlled experimental conditions, but in response to environmental needs or to other agent's requests.

4.4 Recent Developments

4.4.1 Complete identification results

A key identification condition, which operationalizes Theorem 2 has been derived by Jin Tian. It reads:

Theorem 3 *(Tian and Pearl, 2002) A sufficient condition for identifying the causal effect $P(y|do(x))$ is that there exists no bi-directed path (i.e., a path composed entirely of bi-directed arcs) between X and any of its children.*[2]

[1]These graphical tests offer, in fact, a complete formal solution to the "covariate-selection" problem in statistics: finding an appropriate set of variables that need be adjusted for in any study which aims to determine the effect of one factor upon another. This problem has been lingering in the statistical literature since Karl Pearson, the founder of modern statistics, discovered (1899) what in modern terms is called the "Simpson's paradox"; any statistical association between two variables may be reversed or negated by including additional factors in the analysis (Aldrich, 1995).

[2]Before applying this criterion, one may delete from the causal graphs all nodes that are not ancestors of Y.

Remarkably, the theorem asserts that, as long as every child of X (on the pathways to Y) is not reachable from X via a bi-directed path, then, regardless of how complicated the graph, the causal effect $P(y|do(x))$ is identifiable. In Figure 1, for example, adding unobserved confounders for the pairs $(X_1, X_3), (X_2, X_4), (X_3, X_5)$, and (X_2, X_5) will still permit the identification of $P(X_5|\hat{x}_3)$, but adding (X_2, X_4) to this list will prohibit identification.

Tian and Pearl (2002) further showed that the condition is both sufficient and necessary for the identification of $P(v|do(x))$, where V includes all variables except X. A necessary and sufficient condition for identifying $P(w|do(z))$, with W and Z two arbitrary sets, was established by Shpitser and Pearl (2006b). Subsequently, a complete graphical criterion was established for determining the identifiability of *conditional* interventional distributions, namely, expressions of the type $P(y|do(x), z)$ where X, Y, and Z are arbitrary sets of variables (Shpitser and Pearl, 2006a).

These results constitute a complete characterization of causal effects in graphical models. They provide us with polynomial time algorithms for determining whether an arbitrary quantity invoking the $do(x)$ operator is identified in a given semi-Markovian model and, if so, what the estimand of that quantity is. Remarkably, one corollary of these results also states that the *do*-calculus is complete, namely, a quantity $Q = P(y|do(x), z)$ is identified if and only if it can be reduced to a *do*-free expression using the three rules of Theorem 2.[3] This turns corollary 1 into an "if-and-only-if" assertion. Tian and Shpitser (2010) provide a comprehensive summary of these results.

4.5 Transportability and Transfer Learning

In applications involving identification, the role of the *do*-calculus is to remove the *do*-operator from the query expression. We now discuss a totally different application, to decide if experimental findings can be transported to a new, potentially different environment, where only passive observations can be performed. This problem, labeled "transportability" in (Pearl and Bareinboim, 2011) is at the heart of every scientific investigation. For example, experiments performed in the laboratory are invariably intended to be used elsewhere, where conditions differ substantially from the laboratory. A robot trained by a simulator should be able to use the knowledge acquired through training to a new environment, in which experiments are costly or infeasible.

To formalize problems of this sort, Pearl and Bareinboim devised a graphical representation called "selection diagrams" which encodes knowledge about differences and commonalities between populations. A selection diagram is a causal diagram annotated with new variables, called S-nodes, which point to the mechanisms where discrepancies between the two populations are suspected to take place. The task of deciding if transportability is feasible now reduces to a syntactic problem (using the *do*-calculus) aiming to separate the *do*-operator from a the S-variables.

Theorem 4 *(Pearl and Bareinboim, 2011) Let D be the selection diagram characterizing two populations, π and π^*, and S a set of selection variables in D. The relation $R = P^*(y|do(x), z)$ is transportable from π and π^* if and only if the expression $P(y|do(x), z, s)$*

[3] This was independently established by Huang and Valtorta (2006).

is reducible, using the rules of do-calculus, to an expression in which S appears only as a conditioning variable in do-free terms.

Theorem 4 does not specify the sequence of rules leading to the needed reduction when such a sequence exists. Bareinboim and Pearl (2012) established a complete and effective graphical procedure of confirming transportability which also synthesizes the transport formula whenever possible. Each transport formula determines for the investigator what information need to be taken from the experimental and observational studies and how they ought to be combined to yield an unbiased estimate of R.

A generalization of transportability theory to multi-environment has let to a principled solution to "Meta Analysis." "Meta Analysis" is a data fusion problem aimed at combining results from many experimental and observational studies, each conducted on a different population and under a different set of conditions, so as to synthesize an aggregate measure of effect size that is "better," in some sense, than any one study in isolation. This fusion problem has received enormous attention in the health and social sciences, where data are scarce and experiments are costly (Pearl, 2012a,c).

Using multiple "selection diagrams" to encode commonalities among studies, Bareinboim and Pearl (2013) were able to "synthesize" an estimator that is guaranteed to provide unbiased estimate of the desired quantity based on information that each study share with the target environment.

4.6 Inference with Missing Data

It is commonly assumed that causal knowledge is necessary only when interventions are contemplated and that in purely predictive tasks, probabilistic knowledge suffices. Not so. One predictive task in which causal information is essential and which have thus far been treated in purely statistical terms is the problem of drawing valid inferences from statistical data in which some items are "missing" or failed to be recorded.

The "missing data" problem is pervasive in machine learning and every experimental science; users fail to answer certain items on a questionnaire, sensors may miss certain signal due to bad weather, and so on. The problem is causal in nature, because the mechanism that determines the *reasons* for missingness makes a difference in whether/how we can recover from the data, and that mechanism requires causal language to be properly described, statistics is insufficient.

Mohan and Pearl (2013) have shown that current practices of handling missing data, by relying exclusively on statistical considerations (Rubin, 1976; Little and Rubin, 2002) are deficient in several key areas, and could be remedied using causal diagrams. In particularly, they showed that

1. It is possible to define precisely what relations can be *recovered* from missing data and what causal and statistical knowledge need to be assumed to ensure bias-free recovery.

2. Different causal assumptions lead to different routines for recovering a relation from the data.

3. Adding auxiliary variables to the data, a popular technique in current practice (Schafer and Graham, 2002) may or may not help the recovery process, depending on conditions that can be read from the causal diagram.

4. It is possible to determine when the conditions necessary for bias-free recovery have testable implications.

4.7 Historical Remarks

An explicit translation of interventions to "wiping out" equations from linear econometric models was first proposed by Strotz and Wold (1960) and later used in Fisher (1970) and Sobel (1990). Extensions to action representation in nonmonotonic systems were reported in Goldszmidt and Pearl (1992); Pearl (1993a). Graphical ramifications of this translation were explicated first in Spirtes et al. (1993) and later in Pearl (1993b). A related formulation of causal effects, based on event trees and counterfactual analysis was developed by Robins (1986, pp. 1422–25). Calculi for actions and counterfactuals based on this interpretation are developed in Pearl (1994b) and Balke and Pearl (1994), respectively.

5 Counterfactuals

A counterfactual sentence has the form

If A were true, then C would have been true?

where A, the counterfactual antecedent, specifies an event that is contrary to one's real-world observations, and C, the counterfactual consequent, specifies a result that is expected to hold in the alternative world where the antecedent is true. A typical example is "If Oswald were not to have shot Kennedy, then Kennedy would still be alive" which presumes the factual knowledge of Oswald's shooting Kennedy, contrary to the antecedent of the sentence.

The majority of the philosophers who have examined the semantics of counterfactual sentences have resorted to some version of Lewis' "closest world" approach; "C if it were A" is true, if C is true in worlds that are "closest" to the real world yet consistent with the counterfactuals antecedent A Lewis (1973). Ginsberg (1986), followed a similar strategy. While the "closest world" approach leaves the precise specification of the closeness measure almost unconstrained, causal knowledge imposes very specific preferences as to which worlds should be considered closest to any given world. For example, considering an array of domino tiles standing close to each other. The manifestly closest world consistent with the antecedent "tile i is tipped to the right" would be a world in which just tile i is tipped, while all the others remain erect. Yet, we all accept the counterfactual sentence "Had tile i been tipped over to the right, tile $i+1$ would be tipped as well" as plausible and valid. Thus, distances among worlds are not determined merely by surface similarities but require a distinction between disturbed mechanisms and naturally occurring transitions. The local surgery paradigm expounded in Section 4.1 offers a concrete explication of the closest world approach which respects causal considerations. A world w_1 is "closer" to w than a world w_2 is, if the the set of atomic surgeries needed for transforming w into w_1 is a subset of those needed for transforming w into w_2. In the domino example, finding tile i tipped and $i+1$ erect requires the breakdown of two mechanism (e.g., by two external actions) compared with one mechanism for the world in which all j-tiles, $j > i$ are tipped. This paradigm conforms to our perception of causal influences and lends itself to economical machine representation.

5.1 Formal underpinning

The structural equation framework offers an ideal setting for counterfactual analysis.

Definition 7 *(context-based potential response): Given a causal theory T and two disjoint sets of variables, X and Y, the potential response of Y to X in a context u, denoted $Y(x,u)$ or $Y_x(u)$, is the solution for Y under $U = u$ in the subutheory T_x. $Y(x,u)$ can be taken as the formal definition of the counterfactual English phrase: "the value that Y would take in context u, had X been x,"*[4]

Note that this definition allows for the context $U = u$ and the proposition $X = x$ to be incompatible in T. For example, if T describes a logic circuit with input U it may well be reasonable to assert the counterfactual: "Given $U = u$, Y would be high if X were low", even though the input $U = u$ may preclude X from being low. It is for this reason that one must invoke some notion of intervention (alternatively, a theory change or a "miracle" (Lewis, 1973)) in the definition of counterfactuals.

If U is treated as a random variable, then the value of the counterfactual $Y(x,u)$ becomes a random variable as well, denoted as $Y(x)$ of Y_x. Moreover, the distribution of this random variable is easily seen to coincide with the causal effect $P(y|\hat{x})$, as defined in Eq. (7), i.e.,

$$P((Y(x) = y) = P(y|\hat{x})$$

The probability of a counterfactual conditional $x \rightarrow y \mid o$ may then be evaluated by the following procedure:

- Use the observations o to update $P(u)$ thus forming a causal theory $T^o =< V, U, \{f_i\}, P(u|o) >$
- Form the mutilated theory T_x^o (by deleting the equation corresponding to variables in X) and compute the probability $P_{T^o}(y|\hat{x})$ which T_x^o induces on Y.

Unlike causal effect queries, counterfactual queries are not identifiable even in Markovian theories, but require that the functional-form of $\{f_i\}$ be specified. In Balke and Pearl (1994) a method is devised for computing sharp bounds on counterfactual probabilities which, under certain circumstances may collapse to point estimates. This method has been applied to the evaluation of causal effects in studies involving noncompliance, and to the determination of legal liability.

[4]The term *unit* instead of *context* is often used in the statistical literature Rubin (1974), where it normally stands for the identity of a specific individual in a population, namely, the set of attributes u that characterize that individual. In general, u may include the time of day, the experimental conditions under study, and so on. Practitioners of the counterfactual notation do not explicitly mention the notions of "solution" or "intervention" in the definition of $Y(x,u)$. Instead, the phrase "the value that Y would take in unit u, had X been x," viewed as basic, is posited as the definition of $Y(x,u)$.

5.2 Applications to Policy Analysis

Counterfactual reasoning is at the heart of every planning activity, especially real-time planning. When a planner discovers that the current state of affairs deviates from the one expected, a "plan repair" activity need be invoked to determine what went wrong and how it could be rectified. This activity amounts to an exercise of counterfactual thinking, as it calls for rolling back the natural course of events and determining, based on the factual observations at hand, whether the culprit lies in previous decisions or in some unexpected, external eventualities. Moreover, in reasoning forward to determine if things would have been different a new model of the world must be consulted, one that embodies hypothetical changes in decisions or eventualities, hence, a breakdown of the old model or theory.

The logic-based planning tools used in AI, such as STRIPS and its variants or those based on the situation calculus, do not readily lend themselves to counterfactual analysis; as they are not geared for coherent integration of abduction with prediction, and they do not readily handle theory changes. Remarkably, the formal system developed in economics and social sciences under the rubric "structural equations models" does offer such capabilities but, as will be discussed below, these capabilities are not well recognized by current practitioners of structural models. The analysis presented in this chapter could serve both to illustrate to AI researchers the basic formal features needed for counterfactual and policy analysis, and to call the attention of economists and social scientists to capabilities that are dormant within structural equations models.

Counterfactual thinking dominates reasoning in political science and economics. We say, for example, "If Germany were not punished so severely at the end of World War I, Hitler would not have come to power," or "If Reagan did not lower taxes, our deficit would be lower today." Such thought experiments emphasize an understanding of generic laws in the domain and are aimed toward shaping future policy making, for example, "defeated countries should not be humiliated," or "lowering taxes (contrary to Reaganomics) tends to increase national debt."

Strangely, there is very little formal work on counterfactual reasoning or policy analysis in the behavioral science literature. An examination of a number of econometric journals and textbooks, for example, reveals a glaring imbalance: while an enormous mathematical machinery is brought to bear on problems of estimation and prediction, policy analysis (which is the ultimate goal of economic theories) receives almost no formal treatment. Currently, the most popular methods driving economic policy making are based on so-called *reduced-form* analysis: to find the impact of a policy involving decision variables X on outcome variables Y, one examines past data and estimates the conditional expectation $E(Y|X = x)$, where x is the particular instantiation of X under the policy studied.

The assumption underlying this method is that the data were generated under circumstances in which the decision variables X act as exogenous variables, that is, variables whose values are determined outside the system under analysis. However, while new decisions should indeed be considered exogenous for the purpose of evaluation, past decisions are rarely enacted in an exogenous manner. Almost every realistic policy (e.g., taxation) imposes control over some endogenous variables, that is, variables whose values are determined by other variables in the analysis. Let us take taxation policies as an example. Economic data are generated in a world in which the government is reacting to various indicators and various pressures; hence, taxation is endogenous in the data-analysis phase of the study. Taxation becomes exogenous when we wish to predict the impact of a specific decision to raise or lower taxes. The reduced-form method is valid only when past decisions are non-responsive to other variables in the system, and this, unfortunately, eliminates most of the interesting control variables (e.g., tax rates, interest rates, quotas) from the analysis.

This difficulty is not unique to economic or social policy making; it appears whenever one wishes to evaluate the merit of a plan on the basis of the past performance of other agents. Even when the signals triggering the past actions of those agents are known with certainty, a systematic method must be devised for selectively ignoring the influence of those signals from the evaluation process. In fact, the very essence of *evaluation* is having the freedom to imagine and compare trajectories in various counterfactual worlds, where each world or trajectory is created by a hypothetical implementation of a policy that is free of the very pressures that compelled the implementation of such policies in the past.

Balke and Pearl (1995) demonstrate how linear, nonrecursive structural models with Gaussian noise can be used to compute counterfactual queries of the type: "Given an observation set O, find the probability that Y would have attained a value greater than y, had X been set to x". The task of inferring "causes of effects", that is, of finding the probability that $X = x$ is the cause for effect E, amounts to answering the counterfactual query: "Given effect E and observations O, find the probability that E would not have been realized, had X not been x". The technique developed in Balke and Pearl (1995) is based on probability propagation in dual networks, one representing the actual world, the other the counterfactual world. The method is not limited to linear functions but applies whenever we are willing to assume the functional form of the structural equations. The noisy OR-gate model (Pearl, 1988) is a canonical example where such functional form is normally specified. Likewise, causal theories based on Boolean functions (with exceptions), such as the one described in Eq. (16) lend themselves to counterfactual analysis in the framework of Definition 7.

5.3 The Algorithmization of Counterfactuals

Prospective counterfactual expressions of the type $P(Y_x = y)$ are concerned with predicting the average effect of hypothetical actions and policies and can, in principle, be assessed from experimental studies in which X is randomized. Retrospective counterfactuals, on the other hand, consist of variables at different hypothetical worlds (different subscripts) and these may or may not be testable experimentally. In epidemiology, for example, the expression $P(Y_{x'} = y'|x, y)$ may stand for the fraction of patients who recovered (y) under treatment (x) that would not have recovered (y') had they not been treated (x'). This fraction cannot be assessed in experimental study, for the simple reason that we cannot re-test patients twice, with and without treatment. A different question is therefore posed: which counterfactuals can be tested, be it in experimental or observational studies. This question has been given a mathematical solution in (Shpitser and Pearl, 2007). It has been shown, for example, that in linear systems, $E(Y_x|e)$ is estimable from experimental studies whenever the prospective effect $E(Y_x)$ is estimable in such studies. Likewise, the counterfactual probability $P(Y_{x'}|x)$, also known as the effect of treatment on the treated (ETT) is estimable from observational studies whenever an admissible S exists for $P(Y_x = y)$ (Shpitser and Pearl, 2009).

Retrospective counterfactuals have also been indispensable in conceptualizing direct and indirect effects (Robins and Greenland, 1992; Pearl, 2001), which require nested counterfactuals in their definitions. For example, to evaluate the direct effect of treatment $X = x'$ on individual u, un-mediated by a set Z of intermediate variables, we need to construct the nested counterfactual $Y_{x',Z_x(u)}$ where Y is the effect of interest, and $Z_x(u)$ stands for whatever values the intermediate variables Z would take had treatment not been given.[5] Likewise, the average *indirect effect*, of a transition from x to x' is defined as the expected change in Y affected by holding X constant, at $X = x$, and changing Z, hypothetically, to whatever value it would have attained had X been set to $X = x'$.

This counterfactual formulation has enabled researchers to derive conditions under which direct and indirect effects are estimable from empirical data (Pearl, 2001; Petersen et al., 2006) and to answer such questions as: "Can data prove an employer guilty of hiring discrimination?" or, phrased counterfactually, "what fraction of employees owes its hiring to sex discrimination?"

These tasks are performed using a general estimator, called the Mediation Formula (Pearl, 2001, 2009, 2012b), which is applicable to nonlinear models with discrete or continuous variables, and permits the evaluation of path-specific effects with minimal assumptions regarding the data-generating process.

Acknowledgments

This research was supported in parts by grants from NSF #IIS-1249822 and ONR #N00014-13-1-0153.

References

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. Forthcoming *Statistical Science*.

Andersen, S., Olesen, K., Jensen, F. and Jensen, F. (1989). HUGIN - a shell for building Bayesian belief universes for expert systems. In *Eleventh International Joint Conference on Artificial Intelligence*.

Balke, A. and Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds, and applications. In *Uncertainty in Artificial Intelligence 10* (R. L. de Mantaras and D. Poole, eds.). Morgan Kaufmann, San Mateo, CA, 46–54.

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

Bareinboim, E. and Pearl, J. (2012). Transportability of causal effects: Completeness results. In *Proceedings of the 26th AAAI Conference*. Toronto, Ontario, Canada.

[5]Note that conditioning on the intermediate variables in Z would generally yield the wrong answer, due to unobserved "confounders" affecting both Z and Y. Moreover, in non linear systems, the value at which we hold Z constant will affect the result (Pearl, 2000, pp. 126-132).

Bareinboim, E. and Pearl, J. (2013). Meta-transportability of causal effects: A formal approach. Tech. Rep. R-407, <http://ftp.cs.ucla.edu/pub/stat_ser/r407.pdf>, University of California Los Angeles, Computer Science Department, CA. Submitted.

Beaudry, E., Kabanza, F. and Michaud, F. (2010). Planning for concurrent action executions under action duration uncertainty using dynamically generated Bayesian networks. In *Proceedings of the Twentieth International Conference on Automated Planning and Scheduling (ICAPS 2010*. Toronto, Canada.

Blalock, Jr., H. (1971). *Causal Models in the Social Sciences*. Macmillan, London.

Buede, D. M. (2009). *The Engineering Design of Systems: Models and Methods*. 2nd ed. Wiley, Hoboken,, NJ.

Charniak, E. and Goldman, R. (1991). A probabilistic model of plan recognition. In *Proceedings, AAAI-91*. AAAI Press/The MIT Press, Anaheim, CA, 160–165.

Cooper, G. (1990). Computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence* **42** 393–405.

Cooper, G. and Herskovits, E. (1991). A Bayesian method for constructing Bayesian belief networks from databases. In *Proceedings of Uncertainty in Artificial Intelligence Conference, 1991* (B. D'Ambrosio, P. Smets and P. Bonissone, eds.). Morgan Kaufmann, San Mateo, 86–94.

Darwiche, A. (2009). *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, New York.

Darwiche, A. and Goldszmidt, M. (1994). On the relation between kappa calculus and probabilistic reasoning. In *Uncertainty in Artificial Intelligence* (R. L. de Mantaras and D. Poole, eds.), vol. 10. Morgan Kaufmann, San Francisco, CA, 145–153.

Darwiche, A. and Pearl, J. (1994). Symbolic causal networks for planning under uncertainty. In *Symposium Notes of the 1994 AAAI Spring Symposium on Decision-Theoretic Planning*. AAAI Press, Stanford, CA, 41–47.

Darwiche, A. and Pearl, J. (1997). On the logic of iterated belief revision. *Artificial Intelligence* **89** 1–29.

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

de Campos, L., Fernandez-Luna, J. and Huete, J. (2004). Bayesian networks and information retrieval: An introduction to the special issue. *Information Processing and Management* **40** 727–733.

Dean, T. and Wellman, M. (1991). *Planning and Control*. Morgan Kaufmann, San Mateo, CA.

Dechter, R. (1999). Bucket elimination: A unifying framework for reasoning. *Artificial Intelligence* **113** 41–85.

Duncan, O. (1975). *Introduction to Structural Equation Models*. Academic Press, New York.

Fenton, N. and Neil, M. (2012). *Risk Assessment and Decision Analysis with Bayesian Networks*. CRC Press, Boca Raton, FL.

Fisher, F. (1970). A correspondence principle for simultaneous equations models. *Econometrica* **38** 73–92.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology* **7** 601–620.

Galles, D. and Pearl, J. (1995). Testing identifiability of causal effects. In *Uncertainty in Artificial Intelligence 11* (P. Besnard and S. Hanks, eds.). Morgan Kaufmann, San Francisco, 185–195.

Geiger, D. (1990). Graphoids: A qualitative framework for probabilistic inference. Ph.D. thesis, University of California, Los Angeles, Department of Computer Science.

Geiger, D., Verma, T. and Pearl, J. (1990). Identifying independence in Bayesian networks. In *Networks*, vol. 20. John Wiley, Sussex, England, 507–534.

Ginsberg, M. (1986). Counterfactuals. *Artificial Intelligence* **30**.

Goldszmidt, M. and Pearl, J. (1992). Rank-based systems: A simple approach to belief revision, belief update, and reasoning about evidence and actions. In *Proceedings of the Third International Conference on Knowledge Representation and Reasoning* (B. Nebel, C. Rich and W. Swartout, eds.). Morgan Kaufmann, San Mateo, CA, 661–672.

Goldszmidt, M. and Pearl, J. (1996). Qualitative probabilities for default reasoning, belief revision, and causal modeling. *Artificial Intelligence* **84** 57–112.

Good, I. (1961). A causal calculus (I). *British Journal for the Philosophy of Science* **11** 305–318.

Guestrin, C. E. (2003). *Planning under uncertainty in complex structured environments*. Ph.D. thesis, Stanford University, Department of Computer Science.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11** 1–12. Reprinted in D.F. Hendry and M.S. Morgan (Eds.), *The Foundations of Econometric Analysis*, Cambridge University Press, 477–490, 1995.

Heckerman, D. (1991). Probabilistic similarity networks. *Networks* **20** 607–636.

Heckerman, D., Geiger, D. and Chickering, D. (1994). Learning Bayesian networks: The combination of knowledge and statistical data. In *Uncertainty in Artificial Intelligence 10* (R. L. de Mantaras and D. Poole, eds.). Morgan Kaufmann, San Mateo, CA, 293–301.

Henrion, M. (1988). Propagation of uncertainty by probabilistic logic sampling in Bayes' networks. In *Uncertainty in Artificial Intelligence 2* (J. Lemmer and L. Kanal, eds.). Elsevier Science Publishers, North-Holland, Amsterdam, Netherlands, 149–164.

Howard, R. and Matheson, J. (1981). Influence diagrams. *Principles and Applications of Decision Analysis* Strategic Decisions Group, Menlo Park, CA.

Huang, Y. and Valtorta, M. (2006). Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (R. Dechter and T. Richardson, eds.). AUAI Press, Corvallis, OR, 217–224.

Jiang, X. and Cooper, G. (2010). A bayesian spatio-temporal method for disease outbreak detection. *Journal of the American Medical Informatics Association* **17** 462–471.

Kenny, D. (1979). *Correlation and Causality*. Wiley, New York.

Kiiveri, H., Speed, T. and Carlin, J. (1984). Recursive causal models. *Journal of Australian Math Society* **36** 30–52.

Kim, J. and Pearl, J. (1983). A computational model for combined causal and diagnostic reasoning in inference systems. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI-83)*. Karlsruhe, Germany.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA.

Koumenides, C. L. and Shadbolt, N. R. (2012). Combining link and content-based information in a bayesian inference model for entity search. In *Proceedings of the 1st Joint International Workshop on Entity-Oriented and Semantic Search*. JIWES '12, ACM, New York, NY, USA.

Lauritzen, S. (1982). *Lectures on Contingency Tables*. 2nd ed. University of Aalborg Press, Aalborg, Denmark.

Lauritzen, S. (1996). *Graphical Models*. Clarendon Press, Oxford.

Lauritzen, S. and Spiegelhalter, D. (1988). Local computations with probabilities on graphical structures and their application to expert systems(with discussion). *Journal of the Royal Statistical Society* ***Series B*****, 50** 157–224.

Levitt, T., Agosta, J. and Binford, T. (1990). Model-based influence diagrams for machine vision. In *Uncertainty in Artificial Intelligence 5* (M. Henrion, R. Shachter, L. Kanal and J. Lemmer, eds.). North Holland, Amsterdam, 371–388.

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

Little, R. J. and Rubin, D. B. (2002). *Statistical analysis with missing data*. 2nd ed. Wiley, New York.

McEliece, R., MacKay, D. and Cheng, J. (1998). Trubo decoding as an instance of Pearl's belief propagation algorithm. *IEEE Journal of Selected Areas in Communication* **16** 140–152.

Mohan, K. and Pearl, J. (2013). Recoverability from missing data: A formal approach. Tech. Rep. R-410, <http://ftp.cs.ucla.edu/pub/stat_ser/r410.pdf>, University of California Los Angeles, Computer Science Department, CA. Submitted.

Neapolitan, R. (1990). *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. Wiley, New York.

Pearl, J. (1982). Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings, AAAI National Conference on AI*. Pittsburgh, PA.

Pearl, J. (1987). Bayes decision methods. In *Encyclopedia of AI*. Wiley Interscience, New York, 48–56.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Pearl, J. (1993a). From Bayesian networks to causal networks. In *Proceedings of the Adaptive Computing and Information Processing Seminar*. Brunel Conference Centre, London. See also *Statistical Science*, 8(3):266–269, 1993.

Pearl, J. (1993b). From conditional oughts to qualitative decision theory. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence* (D. Heckerman and A. Mamdani, eds.). Morgan Kaufmann, 12–20.

Pearl, J. (1994a). From Adams' conditionals to default expressions, causal conditionals, and counterfactuals. In *Probability and Conditionals* (E. Eells and B. Skyrms, eds.). Cambridge University Press, Cambridge, MA, 47–74.

Pearl, J. (1994b). A probabilistic calculus of actions. In *Uncertainty in Artificial Intelligence 10* (R. L. de Mantaras and D. Poole, eds.). Morgan Kaufmann, San Mateo, CA, 454–462.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, 411–420.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys* **3** 96–146. <http://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf>.

Pearl, J. (2012a). *do*-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence* (N. de Freitas and K. Murphy, eds.). AUAI Press, Corvallis, OR.

Pearl, J. (2012b). The mediation formula: A guide to the assessment of causal pathways in nonlinear models. In *Causality: Statistical Perspectives and Applications* (C. Berzuini, P. Dawid and L. Bernardinelli, eds.). John Wiley and Sons, Ltd, Chichester, UK, 151–179.

Pearl, J. (2012c). Some thoughts concerning transfer learning, with applications to meta-analysis and data-sharing estimation. Tech. Rep. R-387, <http://ftp.cs.ucla.edu/pub/stat_ser/r387.pdf>, Department of Computer Science, University of California, Los Angeles, CA.

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)*. Menlo Park, CA. Available at: <http://ftp.cs.ucla.edu/pub/stat_ser/r372a.pdf>.

Pearl, J., Geiger, D. and Verma, T. (1990). The logic and influence diagrams. In *Influence Diagrams, Belief Nets and Decision Analysis* (R. Oliver and J. Smith, eds.). Wiley, 67–87.

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

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

Petersen, M., Sinisi, S. and van der Laan, M. (2006). Estimation of direct causal effects. *Epidemiology* **17** 276–284.

Rebane, G. and Pearl, J. (1987). The recovery of causal poly-trees from statistical data. In *Proceedings of the Third Workshop on Uncertainty in AI*. Seattle, WA.

Robins, J. (1986). 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.

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

Rosenthal, A. (1975). A computer scientist looks at reliability computations. In *Reliability and Fault Tree Analysis* (R. E. Barlow, J. B. Fussell and N. D. Singpurwalla, eds.). SIAM, Philadelphia, 133–152.

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

Rubin, D. (1976). Inference and missing data. *Biometrika* **63** 581592.

Rumelhart, D. (1976). Toward an interactive model of reading. Tech. Rep. CHIP-56, University of California, La Jolla.

Russell, S. J. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, NJ.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods* **7** 147–177.

Shachter, R. (1986). Evaluating influence diagrams. *Operations Research* **34** 871–882.

Shachter, R. (1988). Probabilistic inference and influence diagrams. *Operations Research* **36** 589–604.

Shafer, G. and Pearl, J. E. (1990). *Readings in Uncertain Reasoning*. Morgan Kaufmann, San Mateo, CA.

Shpitser, I. and Pearl, J. (2006a). Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (R. Dechter and T. Richardson, eds.). AUAI Press, Corvallis, OR, 437–444.

Shpitser, I. and Pearl, J. (2006b). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*. AAAI Press, Menlo Park, CA, 1219–1226.

Shpitser, I. and Pearl, J. (2007). What counterfactuals can be tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Vancouver, BC, Canada, 352–359. Also, *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Shpitser, I. and Pearl, J. (2009). Effects of treatment on the treated: Identification and generalization. In *Proceedings of the Twenty-Fifth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-09)*. AUAI Press, Corvallis, Oregon.

Simon, H. (1953). Causal ordering and identifiability. In *Studies in Econometric Method* (W. C. Hood and T. Koopmans, eds.). Wiley and Sons, Inc., New York, NY, 49–74.

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

Spiegelhalter, D. and Lauritzen, S. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks* **20** 579–605.

Spiegelhalter, D., Lauritzen, S., Dawid, P. and Cowell, R. (1993). Bayesian analysis in expert systems (with discussion). *Statistical Science* **8** 219–283.

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

Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, Prediction, and Search.* 2nd ed. MIT Press, Cambridge, MA.

Spohn, W. (1988). A general non-probabilistic theory of inductive reasoning. In *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence*. Minneapolis, MN.

Spohn, W. (2012). *The Laws of Belief: Ranking Theory and its Philosophical Applications*. Oxford University Press, UK.

Strotz, R. and Wold, H. (1960). Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica* **28** 417–427.

Tian, J. and Pearl, J. (2001a). Causal discovery from changes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, 512–521.

Tian, J. and Pearl, J. (2001b). Causal discovery from changes: A Bayesian approach. Tech. Rep. R-285, Computer Science Department, UCLA.

Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*. AAAI Press/The MIT Press, Menlo Park, CA, 567–573.

Tian, J. and Shpitser, I. (2010). On identifying causal effects. In *Heuristics, Probability and Causality: A Tribute to Judea Pearl* (R. Dechter, H. Geffner and J. Halpern, eds.). College Publications, UK, 415–444.

Uebersax, J. (2004). *Genetic Counseling and Cancer Risk Modeling: An Application of Bayes Nets*. Ravenpack International, Marbella, Spain.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*. Cambridge, MA. Also in P. Bonissone, M. Henrion, L.N. Kanal and J.F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 6*, Elsevier Science Publishers, B.V., 255–268, 1991.

Weiss, Y. and Pearl, J. (2010). Belief propagation – perspectives. *Communications of the ACM* **53** 1.

Wermuth, N. and Lauritzen, S. (1983). Graphical and recursive models for contingency tables. *Biometrika* **70** 537–552.

Wold, H. (1964). *Econometric Model Building*. North-Holland, Amsterdam.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* **20** 557–585.

Wright, S. (1934). The method of path coefficients. *Ann. Math. Statist.* **5** 161–215.

Zweig, G. (1998). *Speech Recognition with Dynamic Bayesian Networks*. Ph.D. thesis, University of California at Berkeley, Computer Science Division.