

Causal reversibility in Bayesian networks

MAREK J. DRUZDZEL and HANS VAN LEIJEN*

*Decision Systems Laboratory, School of Information Sciences
and Intelligent Systems Program, University of Pittsburgh,
Pittsburgh, PA 15260, USA*
 e-mail: marek@sis.pitt.edu
 http://www.sis.pitt.edu/~dsl

Abstract. Causal manipulation theorems proposed by Spirtes *et al.* and Pearl in the context of directed probabilistic graphs, such as Bayesian networks, offer a simple and theoretically sound formalism for predicting the effect of manipulation of a system from its causal model. While the theorems are applicable to a wide variety of equilibrium causal models, they do not address the issue of reversible causal mechanisms, i.e. mechanisms that are capable of working in several directions, depending on which of their variables are manipulated exogenously. An example involving reversible causal mechanisms is the power train of a car: normally the engine moves the transmission which, in turn, moves the wheels; when the car goes down the hill, however, the driver may want to use the power train to slow down the car, i.e. let the wheels move the transmission, which then moves the engine. Some probabilistic systems can also be symmetric and reversible. For example, the noise introduced by a noisy communication channel does not usually depend on the direction of data transmission. In this paper, we investigate whether Bayesian networks are capable of representing reversible causal mechanisms. Building on the result of Druzdzel and Simon (1993), which shows that conditional probability tables in Bayesian networks can be viewed as descriptions of causal mechanisms, we study the conditions under which a conditional probability table can represent a reversible causal mechanism. Our analysis shows that conditional probability tables are capable of modelling reversible causal mechanisms but only when they fulfill the condition of *soundness*, which is equivalent to *injectivity* in equations. While this is a rather strong condition, there exist systems where our finding and the resulting framework are directly applicable.

Keywords: causality, manipulation, causal reversibility, Bayesian networks.

1. Introduction

For over a decade, artificial intelligence researchers have tried to incorporate a principled account of causality into existing formalisms for uncertain reasoning, for example in the area of planning (Davidson and Fehling 1994), inferring causal structure from observations (Pearl and Verma 1991, Spirtes *et al.* 1993), explaining causal

* Currently with Universiteit Nyenrode, Straatweg 25, 3621 BG Breukelen, The Netherlands, e-mail: h.vleijen@nyenrode.nl

assumptions in decision analytic models (Druzdzel and Simon 1993, Heckerman and Shachter 1995), experimental design (Pearl 1995), and in modelling counterfactual reasoning (Balke and Pearl 1995).

An explicit representation of causality is important in artificial intelligence because of several reasons.¹ The foremost of these is that causal models allow intelligent agents to predict the effects of their actions. A causal model frees the agent from the need to store a combinatorially large set of pairs *action* \Rightarrow *effect*. Result of external manipulation on the model variables can be predicted directly from the model. In the context of directed graphical probabilistic models, such as Bayesian networks, two foremost formalizations of manipulation are due to Spirtes *et al.* and Pearl. Spirtes *et al.* (1993) proposed a theorem, known as *causal manipulation theorem*, specifying the effect of imposing externally a value on any node in a graphical model. Pearl (1995), who built his recent work on structural equation models, has an equivalent formalization of manipulation. The manipulation theorems offer a simple and theoretically sound formalism for predicting the effect of manipulation of a system from its causal model. They bring much clarity into statistics and such disciplines as experimental design by making a thorough distinction between observation and manipulation. While the theorems are applicable to a variety of equilibrium causal models, they exclude from their framework the issue of reversible causal mechanisms, i.e. mechanisms that are capable of working in several directions, depending on which of their variables are manipulated exogenously.

Both approaches to causal manipulation have been developed in the framework of directed graphical models and imply that if the effect of an external manipulation of a node *n* imposes a deterministic value on it (for the sake of simplicity, we will consider only such interventions in this paper), the effect of this manipulation on other nodes in the network can be predicted by removing the arcs from *n*'s parents to *n*. This operation, popularly known as *arc cutting semantics*, is based on the premise that each mechanism in the model represents an asymmetric relationship among the model's variables. This poses a limitation on modelling reversible causal mechanisms, i.e. mechanisms that are capable of working in several directions, depending on which of their variables are manipulated exogenously. We would like to make a distinction between reversible causal mechanisms and mutual causation. The latter arises in systems that are dynamic in nature and contain feedback loops. For example, research output of a university often depends on the available research funding. Research funding, in turn, depends on the research output. The focus of this paper is on reversible causal mechanisms and our results do not in general apply to mutual causation. Reversible causal mechanisms occur often in equilibrium systems, for example in financial or engineering domains. An example of a reversible causal mechanism is the power train of a typical car: when a car goes up the hill, the engine moves the transmission which, in turn, moves the wheels; when the car goes down the hill, the driver may want to use the power train to slow down the car, i.e. let the wheels move the transmission which then moves the engine. The interactions between the engine and the transmission and between the transmission and the wheels are independent of how the power train is actually used. Knowledge of these interactions can be reused whenever these mechanisms are parts of a larger system. Some probabilistic systems can also be symmetric and reversible. For example, errors introduced by a noisy communication channel do not usually depend on the direction of data transmission. It is the pursuit for a manipulation framework for such mechanisms that provided inspiration for the current paper.

Reversible mechanisms are modelled quite naturally in the context of equilibrium structural equation models, as nothing in a structural equation implies asymmetry among its variables. It is the context of a model that introduces asymmetry among variables. Druzdzel and Simon (1993) related the notion of causality in directed probabilistic graphs to structural equations and structural equation models in econometrics. They showed that it is possible to reproduce the joint probability distribution modelled by a Bayesian network (Pearl 1988) by a system of simultaneous equations in such a way that the independence graph of the former coincides with the causal ordering over the latter. They made it quite clear that conditional probability tables in Bayesian networks can be viewed as descriptions of causal mechanisms and, thereby, be equivalent to structural equations. It is natural to ask the question whether Bayesian networks can, similarly to structural equation models, represent reversible causal mechanisms.

In this paper, we study the algebraic conditions under which a conditional probability table can represent a reversible causal mechanism. We demonstrate that representing reversible causal mechanisms in Bayesian networks is possible, although reversibility implies a strong constraint on the conditional probability tables. The conditional probability tables of reversible mechanisms have to be ‘multiply-normalized’, i.e. their numerical probabilities have to add up to one along each dimension. We derive this constraint formally in the paper. We also improve on the notation used by Druzdzel and Simon (1993) and frame their theorems in a more general way. Our approach is unique in that it brings the natural formalization of manipulation in equilibrium structural equation models to Bayesian networks and shows that the latter have the capability to model reversible causal mechanisms, albeit this capability is strongly limited.

We will use lower case letters, such as v , to denote variables, and capital letters, such as V , to denote a set of variables. Furthermore, n_v will denote the number of values of a variable v or, more generally, the number of elements in a finite set. $\pi(v)$ will denote the set of parents of v in the independence graph of a Bayesian network \mathcal{B} . The values of a variable will be called *configurations*, the set of all configurations of a variable v denoted by $C(v)$, while an element of this set will be denoted by c_v . A configuration of a set of variables is an assignment of a value to each of them; the set of all configurations of a set of variables V will be written as $C(V)$, one particular configuration is c_V . The *restriction* $c|_V$ of a configuration c to a set of variables V is the part of c that assigns values to variables in V .

The remainder of this paper is structured as follows. Section 2 reviews the foundations of manipulation within the framework of structural equation models. Section 3 discusses causal reversibility in the framework of structural equation models. Section 4 reviews the fundamentals of Bayesian networks. Sections 5 and 6 study how reversible causal mechanisms can be modelled by Bayesian networks and derive the constraints on the conditional probability tables that have to be fulfilled for a mechanism to be reversible. Finally, section 7 shows an example of a reversible causal mechanism modelled by a Bayesian network.

2. Structural equation models

Systems of simultaneous structural equations, known as *structural equation models* (SEMs) are used throughout science to model so-called *equilibrium systems*, reasonably isolated parts of the real world that are known or assumed to reach a stable

state if outside influences on the system remain invariant. Such outside influences are called *exogenous*. Their values are determined outside the system either because we are not willing or are not able to account for their behaviour. The values of dependent variables, also called *endogenous* variables, are determined inside the model. Structural equation models are usually self-contained in the sense that they have as many equations as they have variables and, by virtue of the fact that they describe an existing system, have at least one solution and at most a denumerably infinite set of solutions.

A system of simultaneous equations can be easily transformed into another system that has the same set of solutions, for example by combining two or more equations together. In general, there are infinitely many systems that are extensionally equivalent (that is, have the same set of solutions). However, a system of equations that is a model of a system in the real world usually has one form that has more intuitive appeal because it portrays the causal structure of the real system. In such a model, each equation represents a conceptually distinct causal mechanism (Simon 1953). The concept of a *structural equation* is a semantic one; it is in general impossible to decide syntactically whether an equation is structural or not, though we can obtain strong clues from experimentation and, to a lesser extent, from observation (Wold 1954).

A generic form of an equation that we will use throughout this paper is

$$f(x_1, x_2, \dots, x_n, \mathcal{E}) = 0, \tag{1}$$

where f is some algebraic function, its arguments x_1, x_2, \dots, x_n are various system variables and \mathcal{E} is an error variable. This form is usually called an *implicit function*. In order to obtain a variable x_i ($1 \leq x_i \leq n$) as a function of the remaining variables, we must solve equation (1) for x_i . We say that the function

$$x_i = g(x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n, \mathcal{E}) \tag{2}$$

found in this way is defined *implicitly* by (1) and that the solution of this equation gives us the function explicitly. Often, the solution can be stated explicitly in terms of elementary functions. In other cases, the solution can be obtained in terms of an infinite series or other limiting process; that is, one can approximate (2) as closely as desired.

Example: The following system of equations is our perception of the relationship between heart disease (h), blood cholesterol level (c), wine consumption (w), average age (a), and amount of fat in the diet (d) of a population.

$$\begin{cases} w = c_0 \\ d = c_1 \\ a = c_2 \\ c = \alpha d + \beta w + c_3 \\ h = \gamma w + \delta c + \epsilon a + c_4 \end{cases}$$

□

The relations among variables in a self-contained model consisting of structural equations are asymmetric. For example, if we increase w , h will be affected; but modifying the last equation will leave w untouched. This is because *heart disease* (h) depends (indirectly) on *wine consumption* (w) but not vice versa. Examining

the dependencies captured by equations, and asymmetries that they imply for the model variables, leads to an ordering of the variables which is represented by the graph on the right hand side. Simon (1953) developed an algorithm for explicating this ordering, and argued that, if all equations in the model are structural and all exogenous variables in the model are truly exogenous in the system, the resulting graph can be interpreted causally.

For establishing dependency relationships between variables, only occurrence of variables in equations is important; the exact form of those equations is less important. That is, knowing that $\beta < 0$ in the example above (knowing that wine has an adverse effect on heart disease) is not needed for answering the question whether drinking wine has any influence on heart disease at all. Because knowing which variables occur in which equations is sufficient, the causal ordering algorithm treats equations as mere sets of variables. Therefore, in the sequel of this paper we will refer to equations as if they are sets.

Causal ordering takes as input a self-contained set $S = \{E_1, \dots, E_n\}$, where each element is an equation $E_i = \{v_{i,1}, \dots, v_{i,n_{v_i}}\}$ of n_{v_i} variables. Its output is a directed acyclic graph $G = (V(G), A(G))$, where the nodes are sets of equations. The following description is taken from Simon (1953) and adapted to incrementally build the causal graph instead of defining it after applying the algorithm. Proofs of assertions can be found in that paper.

A set of equations is *self-contained* if it has as many equations as variables and if every subset of equations has at least as many variables as equations. So a set of n equations is self-contained if it contains n unknowns and if every subset of m equations has at least m variables. Mechanisms in real-world systems often involve a small number of elements and a set of structural equations will usually contain subsets that are self-contained (i.e. they also consist of as many equations as variables). The first step is to identify minimal self-contained subsets, subsets of equations that are self-contained but do not have self-contained subsets themselves. Because, as Simon proves, intersections of self-contained subsets are self-contained, minimal self-contained subsets are disjunct as are the unions of the variables they contain. These are called *complete subsets of 0th order* and a node is created for each. Next, the complete subsets of 0th order are removed from S . Furthermore, all occurrences of variables in complete subsets of 0th order are removed from remaining equations. This step corresponds respectively to solving the subsets of 0th order and substitution of the thus obtained values into remaining equations. The new system is called the *derived system of first order*. Note that these variables are exogenous to the subsets identified at a later stage. The procedure stops when no self-contained subsets can be identified. A self-contained subset is defined to be of *order k* if it was identified and solved at step k of this procedure. The order in which the subsets were identified defines the causal ordering of the equations and variables. Variables exogenous to a subset are the direct causal predecessors of the variables in that subset.

The recursive step consists of three parts.

- (1) Identify self-contained subsets in the derived system of k th order and create a node for each.
- (2) Add an arc to $A(G)$ from a complete subset of the previous order to a complete subset of the present order if variables appearing in the former were removed from equations that make up the latter. This corresponds to acknowledging the fact that we were able to solve a present complete subset by substituting values of variables obtained by solving a previous complete subset.

- (3) Remove the identified complete subsets from the system and remove variables appearing in them from remaining equations.

The algorithm halts when the system is empty. Now, each node in G corresponds to the union of variables appearing in a complete subset of some order. The variables within each node are inherently simultaneous, they cannot be solved through direct substitutions. In the worst case there is only one node. In such a strongly interconnected system, causal ordering cannot find any asymmetric relationships. In most cases however, nodes will be small because mechanisms in real-world systems are usually fairly isolated. Because of our focus on the relationship between structural equation models and Bayesian networks, in the remainder of this paper we will focus on causal graphs in which each node corresponds to a single variable.

The significance of causal ordering is that it determines the direction of causality on the basis of otherwise undirected equations. That is, there is nothing inherent in an equation that suggests a direction of causality. This direction depends not on the equation itself but on the way the equation is embedded in a context. Simon also showed that there is no such thing as ‘internal causal ordering’, of the endogenous equations without regard to a particular set of exogenous equations. It is the choice of exogenous variables that determines the direction of causality among variables.

The advantage of a structural form is that it supports prediction of the effects of changes to the system. Such changes are modelled by replacing or modifying those mechanisms that are affected by the intervention. For example, a government campaign to increase wine consumption would be modelled by increasing c_0 while invention of a new medicine to increase resistance to heart disease would be modelled by modifying the last equation. The ability of SEMs to predict the effect of manipulation, or, as it is often called, *change in structure*, has been known in econometrics since the inception of the concept of structural equation models. The seminal work of Simon (1953) explicated this property and gave it a causal interpretation. We will refer to this view of causality as *mechanism based*.

3. Causal reversibility

The formalizations proposed by Spirtes *et al.* (1993) and Pearl (1995) assume that imposing a value on a variable renders that variable independent of its direct causes. In other words, taking action on a variable can never be relevant to that variable’s causes. This may seem evident in situations where there is a strong asymmetric relationship between the variable and its causes; for example, wearing a raincoat protects one from getting wet but it does not make the rain go away. But in many areas where a cause and effect variable are more on a par with each other, one cannot be sure the cause variable will not be affected by tampering with the (former) effect variable. For example, acting on the wheels of a car does not normally break their link to transmission; a modern electric train will use its engine as a generator when braking, thereby feeding electricity back to the power lines.

Structural equation models have no difficulty with modelling this type of dependencies. Replacement of one equation by another can in general reverse the direction of causal ordering between variables. To model a decision, or an action, we add an equation that describes the intervention. When the intervention is so strong that it sets the manipulated variable to a specific value, we add an exogenous equation, i.e. an equation that sets a variable to a constant value. After adding this equation, the

system is overconstrained and needs to be relaxed by removing another equation, in general not necessarily the one that determined the value of the manipulated variable in the original causal graph. Depending on the choice of this relaxing equation, the causal ordering of the modified system can be quite different; in particular, the direction of causation between any pair of variables can be reversed.

Example: Consider a cylinder filled with ideal gas. The condition of the gas can be described by its temperature (t), pressure (p) and volume (v), and the relationship between these three is given by the perfect gas law $pv/t = C$, where C is a constant. Suppose that we set the volume and the temperature externally. This situation is captured by the following system of equations and its corresponding causal graph:

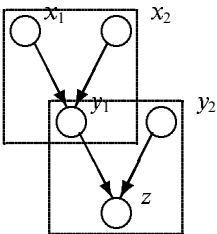
$$\left\{ \begin{array}{l} t = c_0 \\ v = c_1 \\ pv/t = c_3 \end{array} \right.$$


Causal ordering will reveal that both temperature and volume affect pressure. Changing temperature, for example, will affect pressure but not volume. However, if we change the context in which the mechanism that ties temperature, volume and pressure is embedded, for example by allowing the cylinder to expand so that the pressure will be equal to the outside pressure, the resulting system will be:

$$\left\{ \begin{array}{l} t = c_0 \\ p = c_2 \\ pv/t = c_3 \end{array} \right.$$


In this system, we have replaced the equation that fixed the volume with an equation that fixes the pressure, leaving the mechanism that ties them intact. Now the new volume can be computed by substituting the first two equations in the third and solving for v . The resulting new causal ordering will show that volume has become endogenous, being affected by temperature and pressure. In effect, manipulation has led to reversal of the direction of causation. It is worth noting that applying the arc-cutting semantics without regard for the reversibility of the mechanisms would lead to a graph in which all three nodes were disconnected. The prediction obtained from such a graph would be that when pressure is manipulated, the temperature and volume remain unchanged. \square

Example: A change in structure can in general propagate through the entire system of equations, yielding changes in all reversible mechanisms. Consider the following self-contained system of five equations and a causal graph obtained from this system by the procedure of causal ordering

$$\left\{ \begin{array}{l} e_1 : f_1(x_1) = 0 \\ e_2 : f_2(x_2) = 0 \\ e_3 : f_3(x_1, x_2, y_1) = 0 \\ e_4 : f_4(y_2) = 0 \\ e_5 : f_5(y_1, y_2, z) = 0 \end{array} \right.$$


The rectangles contain two individual mechanisms that belong to this system,

modelled by equations e_3 and e_5 . Suppose that we decide to manipulate on variable z and release the variable x_1 . This amounts to introducing a new equation e_6 which fixes the value of z and at the same time to removing the equation e_1 that determined the value of x_1 previously. These operations are always guided by semantics constraints on the effects of the real world manipulation. For example, when using the car power train, we know that neither the mechanism binding the wheels with the transmission, nor the mechanism binding the transmission with the engine are affected by the manipulation and it is only the engine and the wheels that change between being exogenous and endogenous in different modes of manipulation of the power train. The new self-contained system of equations and its causal graph are

$$\left\{ \begin{array}{l} e_2 : f_2(x_2) = 0 \\ e_3 : f_3(x_1, x_2, y_1) = 0 \\ e_4 : f_4(y_2) = 0 \\ e_5 : f_5(y_1, y_2, z) = 0 \\ e_6 : f_6(z) = 0 \end{array} \right.$$

Please note that the two mechanisms modelled by equations e_3 and e_5 are preserved and are still clearly identifiable in the manipulated system—it is only the direction of causality that has changed within each of the mechanisms. \square

An algebraically necessary condition for causal reversibility is that the resulting model is unambiguous with respect to its solutions. This comes down to the existence of explicit functions for the causal ordering that results from manipulation. The following definitions, and especially the definition of injectivity, formalize the algebraic conditions that are necessary for a structural equation to model a reversible causal mechanism.

Definition 1 (equation system): An equation system is a set $S = \{e_1, \dots, e_n\}$ where each e_i is a relation on a set of discrete finite variables $V(e_i) = \{v_1, \dots, v_m\}$, $e_i \subseteq C(v_1) \times \dots \times C(v_m)$. The union of all variables appearing in S will be denoted by $V(S) = \bigcup_{e_i \in S} V(e_i)$.

Note that an equation can be interpreted simply as the set of configurations that are its solutions. In the sequel, we will take that point of view.

Definition 2 (solution): A solution of an equation system S is a configuration $c \in C(V(S))$ such that for each equation $e \in S$ we have that $c_{V(e)} \in e$.

Definition 3 (compatibility): Configurations c_1 and c_2 are compatible if they assign the same values to the variables in their intersection, that is, if for each $v \in V(c_1) \cap V(c_2)$, we have $c_{1, \{v\}} = c_{2, \{v\}}$.

Definition 4 (substitution): If e is an equation, substituting a configuration c into e means removing from e configurations that are not compatible with c and subsequently removing the variables appearing in c from each configuration in e .

Substitution is useful for obtaining solutions of a system by recursively substituting configurations of variables that precede an equation in causal ordering. Note that nothing will happen if c and e do not have a variable in common.

In the sequel, please note that the order in which variables are specified is not significant and is exploited for notational convenience only.

Definition 5 (injectivity):² Let e be an equation on $V(e) = \{v_1, \dots, v_n\}$. e is said to be injective with respect to v_1 if for each configuration $c \in C(v_2, \dots, v_n)$, there is exactly one configuration on $V(e)$ compatible with c .

Injectivity with respect to a variable x means that given an equation and values for all variables but x , we can find a unique value for x to satisfy the equation.

Example: The equation $x^2 - y = 0$ is injective with respect to variable y (given the value of variable x we can determine uniquely the value of variable y) but it is not injective with respect to variable x (given the value of variable y , we cannot in general determine uniquely the value of x). \square

Finally, two last elements useful in formalizing the algebraic conditions for causal reversibility.

Definition 6 (partial solution set): The partial solution set of a complete subset A of a self-contained system of equations S is its solution set if it is a root in causal ordering; otherwise, it is the set of configurations of $C(V(A))$ that simultaneously satisfy the equations in A and the partial solution sets of A 's parents in causal ordering over S .

A is a complete subset of a self-contained system of equations as defined by Simon's procedure of causal ordering described in Section 2. The following definition formalizes the idea that in order for a system to admit a causal interpretation, it must have a unique solution for each choice of its exogenous variables.

Definition 7 (causal unambiguity): A system is causally unambiguous if the partial solution sets of its complete subsets are singleton for each configuration of its exogenous variables.

When reordering a manipulated system of equations, in general the mechanisms will be reversed. This means that the configuration of exogenous and endogenous variables with respect to an equation will change. The following theorem states the implications that Definition 7 has on reversibility; that is, what property equations have to satisfy in order for reordering to render a system causally unambiguous.

For ease of exposition, we will assume each equation has one endogenous variable, or equivalently, we will assume that causal graphs are acyclic, before as well as after reordering.³

Theorem 1 (reversibility): Let S be an unambiguous system of equations and S' be a system of equations that is the result of some intervention on S . S' is causally unambiguous if each equation $e \in S'$ is injective with respect to its endogenous variable.

Proof: For root nodes, the theorem follows trivially since injectivity of an exogenous equation means it has a unique solution so it has a singleton partial solution.

Assume an equation e on $V(e) = v_1, \dots, v_n$ is injective with respect to its endogenous variable v_1 , and assume its parent partial solution sets are singleton. It follows that there is only one configuration $c_{2, \dots, n} \in C(v_2, \dots, v_n)$ compatible with the parent partial solutions.

According to Definition 5, there is exactly one configuration $c_{1, \dots, n}$ compatible with

$c_{2,\dots,n}$ and thus the partial solution is unique. The theorem now follows by induction. \square

Theorem 1 is merely a reformulation of the constraint put on reordering in the previous section. It says that, in order for reordering after an intervention to yield an unambiguous system, equations need to be expressed as an explicit function from the exogenous variables to the endogenous variable.

4. Bayesian networks

Informally, the graph of a Bayesian network encodes qualitative knowledge about relevance relationships in a domain. Its nodes represent discrete, finite stochastic variables and contain numerical information about the probabilities of their values conditional on the values of their parents in the graph. By exploiting the graph as a means of specifying the conditional independencies that must hold between sets of nodes, a Bayesian network can be used to economically represent a joint probability distribution on the nodes of the graph. Efficient algorithms exist for absorbing evidence about the values of nodes into the network and updating marginal probabilities accordingly. Bayesian networks can thus be used for probabilistic inference.

In what follows, we will consider Bayesian networks with discrete variables that can take on any finite number of values.

Definition 8 (Bayesian network): A Bayesian network \mathcal{B} is a tuple (G, F) where

- (1) $G = (V(G), A(G))$ is an acyclic directed graph with vertices $V(G) = \{v_1, \dots, v_n\}$, $n \geq 1$, and arcs $A(G)$, and
- (2) $F = \{f_v | v \in V(G)\}$ is a set of real-valued nonnegative functions $f_v : C(v) \times C(\pi(v)) \rightarrow [0, 1]$, called conditional probability assessment functions, such that for each configuration $c_{\pi(v)}$ of $\pi(v)$, we have

$$\sum_{c_v \in C(v)} f_v(c_v, c_{\pi(v)}) = 1. \quad (3)$$

A joint probability distribution can also be represented by a system of simultaneous equations. Druzdzel and Simon (1993) were able to reproduce the joint distribution over a Bayesian network by representing each conditional probability table by an equation with error terms, and have subsequently proven that causal ordering of this system yields a graph that is isomorphic to the independence graph of the Bayesian network. Interpreting the independence graph as a causal model then amounts to assuming that the corresponding equations are structural.

The representation of probability tables as equations used in Druzdzel and Simon (1993) was for the sake of simplicity of exposition limited to Boolean variables. To allow for a convenient and compact representation of conditional probabilities of multiple-valued variables, we introduce the concept of a probability matrix.

Definition 9 (probability matrix): Let \mathcal{B} be a Bayesian network (G, F) on variables $V = v_1, \dots, v_n$, and let $v \in V$. The probability matrix M_v of v is a two-dimensional matrix, where columns are indexed by the configurations $C(\pi(v))$ of parents of v while the rows are indexed by $C(V)$, the values of v .

Each cell of M_v is assigned an interval of $[0, 1]$ in such a way that

- (1) all intervals in a column of M_v are mutually exclusive,

- (2) the union of all intervals in a column is equal to $[0, 1]$, and
- (3) the length of each interval is equal to the corresponding conditional probability.

In other words, the intervals in a column of the matrix partition the interval $[0, 1]$ in such a way that we have, for each configuration $c_v \in C(v)$ and $c_{\pi(v)} \in C(\pi(v))$

$$\Pr(\epsilon \in M_v(c_v, c_{\pi(v)})) = f_v(c_v, c_{\pi(v)}). \tag{4}$$

Example: The only difference between a probability table and a probability matrix is the fact that in the latter, probabilities have been replaced by subsets of the intervals $[0, 1]$. Figure 1 shows an example of a probability table and its corresponding probability matrix. The cells of the latter are assigned intervals in one possible way to satisfy the criteria of Definition 9. Note that an interval in a column need not be contiguous and that neither the upper cell need to include 0 nor the lowest cell need to include 1. □

The following theorem, a slightly different framing of a theorem proposed by Druzdel and Simon (1993), describes how the joint probability distribution defined by a Bayesian network can be represented by a system of simultaneous equations with latent variables.

Theorem 2 (representability): Let $\mathcal{B} = (G, F)$ be a Bayesian network on nodes $V = v_1, \dots, v_n$ as in Definition 8, which defines a probability distribution $\Pr_{\mathcal{B}}$. There exists a system of simultaneous equations S on V that defines a distribution \Pr_S such that $\Pr_{\mathcal{B}}(c) = \Pr_S(c)$ for all $c \in C(V)$.

Proof: (Sketch) The proof is by demonstrating a procedure for constructing S . In a Bayesian network, the probability of a configuration $c \in C(V)$ is simply the product of the probabilities for each node v of its configuration c_v conditional on the configuration of its parents, $c_{\pi(v)}$. Therefore, it suffices to replicate the conditional probabilities for each node given its parents.

We will construct one equation for each of the variables. Each equation will include an independent, continuous latent variable ϵ_v , uniformly distributed over the interval $[0, 1]$. Note that $\forall x(0 < x \leq 1) \Pr(\epsilon_v \leq x) = x$.

We first consider a node v without predecessors. Its probability table contains

	v_1	true		false	
$\Pr(v_3 v_1v_2)$	v_2	true	false	true	false
v_3	true	0.9	0.8	0.7	0.1
	false	0.1	0.2	0.3	0.9

	v_1	true		false	
v_2	true	false	true	false	
v_3	true	[0, 0.9]	[0, 0.8]	[0, 0.7]	[0, 0.1]
	false	(0.9, 1]	(0.8, 1]	(0.7, 1]	(0.1, 1]

Figure 1. An example of a probability table $\Pr(v_3|v_1v_2)$ (upper table) and its corresponding probability matrix (lower table).

prior probabilities over its configurations $C(v)$. The following deterministic function reproduces these priors:

$$f_v(\epsilon_v) = c \quad \text{if } \epsilon_v \in M_v(c) \quad \text{for all } c \in C(v). \quad (5)$$

A node v that does have predecessors is treated analogously, except that the function that determines the value of v takes as arguments not just the latent variable but v 's parents $\pi(v)$ as well:

$$\begin{aligned} f_v(c_{\pi(v)}, \epsilon_v) &= c_v \quad \text{if } \epsilon_v \in M_v(c_v, c_{\pi(v)}) \\ \text{for all } c_{\pi(v)} &\in C(\pi(v)) \quad \text{and } c_v \in C(v). \end{aligned} \quad (6)$$

□

From the way the system of simultaneous equations corresponding to a Bayesian network is constructed, it follows immediately that the causal ordering of the system is equivalent to the independence graph of the Bayesian network. This allows us to use the causality framework developed for structural equation models in interpreting the structure of Bayesian networks.

5. Restructuring Bayesian networks

A mechanism-based approach in the context of structural equation models supports true structural changes that can impact the system in such a way that the causal graph changes completely. Knowledge encoded in terms of conditional probability distributions in a directed probabilistic graph, such as a Bayesian network, on the other hand, seems to be valid only in the particular context, in one mode of operation (e.g. going up the hill or going down the hill in the car power train example) in which the system is studied.

A necessary condition for supporting true structural changes in the framework of directed probabilistic graphs, such as Bayesian networks, is their ability to represent reversible causal mechanisms. In order to be able to represent reversible mechanisms, the conditional probability tables have to fulfill algebraic conditions analogous to injectivity in equation-based systems.

We will illustrate this point and the possibility of restructuring a probabilistic graph with an example. Consider the following system of equations:

$$\begin{cases} x = x_0 \\ f(x, y, \mathcal{E}) = 0 \end{cases} \quad (7)$$

From system (7), we can easily derive the conditional probability $\Pr(y|x)$. To accomplish this, we need to replace x in f by x_0 and derive the explicit function for y in terms of \mathcal{E} (note that \mathcal{E} , a random variable, is expressed in terms of a probability distribution).

Let us now imagine that we want to model the effect of acting on y rather than on x without disturbing the mechanism that ties x and y . The new system of structural equations will take the following form:

$$\begin{cases} y = y_0 \\ f(x, y, \mathcal{E}) = 0 \end{cases} \quad (8)$$

Please, note that only the first equation in (7) had to be replaced, as the second equation describes a mechanism that is unaffected by our intervention. If the function f is injective with respect to x , we can derive the conditional probability $\Pr(x|y)$

from it. To accomplish this, we need to replace y in f by y_0 and derive the explicit function for x in terms of \mathcal{E} . We assume that the intervention has left the probability distribution of \mathcal{E} unaffected and, therefore, known as a part of knowledge about the mechanism. This assumption is implied by the assumption that f described a mechanism and \mathcal{E} was exogenous and independent of any other exogenous variable. Note that all that was needed to derive the conditional probability distributions $\Pr(y|x)$ or $\Pr(x|y)$ was the functional form f of the equation binding x , y and \mathcal{E} and the probability distribution of the error variable \mathcal{E} .

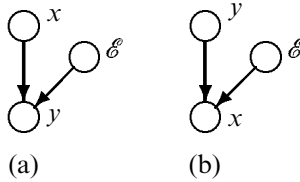


Figure 2. Causal ordering for (a) system (7) and (b) system (8).

The causal ordering applied to (7) and (8) yields the graphs of figure 2(a) and (b) respectively. Imagine a transition from the network (figure 2(a)) to the other network (figure 2(b)), which is similar to the transition between the systems (7) and (8). As explained in section 4, a node in a causal Bayesian network and its direct predecessors can be seen as a mechanism and the conditional probability distribution of this node on its direct predecessors can be seen as a description of the mechanism when the functional form is unknown. From this point of view, construction of the network (figure 2(b)) given that we have a fully quantified network (figure 2(a)) and the fact that the mechanism binding x , y , and \mathcal{E} is reversible, resembles a change in structure in a structural equation model. Under what algebraic constraints on the conditional probability table $\Pr(y|x)$ would we be able to use $\Pr(y|x)$ to derive the conditional probability distribution $\Pr(x|y)$ of the new graph?

Since this may be easily confused with the operation of arc reversal (e.g. Shachter 1986), we would like to stress that they are very distinct. While arc reversal is based on the application of Bayes theorem and could compute the impact of observing y on the probability of x , the operation in question here deals with the manipulation of y and the impact that this manipulation has on the probability distribution of x . Arc reversal requires that we know the prior probability distribution $\Pr(x)$, while manipulation does not.

6. Reversible mechanisms in Bayesian networks

We will consider a conditional probability table with associated equation and investigate what it means to choose a different endogenous variable. The lower table in figure 1 gives an example in the form of a probability matrix as defined in Definition 9.

Recall from the proof of Theorem 2 that this table can be translated into the following equation

$$f_{v_3}(c_{\{v_1, v_2\}}, \epsilon_{v_3}) = c_{v_3}$$

$$\text{if } \epsilon_{v_3} \in M_{v_3}(c_{v_3}, c_{\{v_1, v_2\}})$$

where ϵ_{v_3} is the error term associated with v_3 . Now if, by some intervention, v_1

becomes the endogenous variable and v_3 becomes the exogenous variable, then we will have to compute a new probability table in which v_1 is the dependent variable. This corresponds to calculating the explicit form f'_{v_3} of f_{v_3} for v_1 . Unfortunately, in this case this explicit form does not exist because f_{v_3} is not injective with respect to v_1 which is shown for example by existence of the tuples $f_{v_3}(v_1 = \mathbf{true}, v_2 = \mathbf{true}, \epsilon_{v_3} = 0.5) = \langle v_3 = \mathbf{true} \rangle$ and $f_{v_3}(v_1 = \mathbf{false}, v_2 = \mathbf{true}, \epsilon_{v_3} = 0.5) = \langle v_3 = \mathbf{true} \rangle$.

The following definition provides us with a concept that will express the sufficient and necessary condition for reversibility of a causal mechanism captured by a Bayesian network.

Definition 10 (soundness): *Let e be an equation on v_1, \dots, v_n in an equation system S associated with a Bayesian network \mathcal{B} where v_1 is the endogenous variable. The probability matrix M_{v_1} is sound with respect to v_2 if after transposing the table with respect to v_2 it is still a probability matrix.*

Note that a probability matrix corresponding to a conditional probability table $\text{Pr}(x|\pi(x))$ is by definition sound with respect to x . Table 1 shows an example of a probability matrix that is sound with respect to both variables appearing in it. Consistency of a probability matrix with respect to more than one variable seems like a stern condition because intervals cannot overlap with intervals in the same column as well as in the same row; however, a probability table in which the rows of an exogenous variable v_i add up to 1 can always be translated into a matrix consistent with respect to v_i . Table 2 shows such a ‘doubly normalized’ probability table that can be translated into the probability matrix shown in table 1.

The following theorem proves that soundness is a necessary and sufficient condition for reversibility.

Theorem 3 (reversibility in Bayesian networks): *Let \mathcal{B} be a Bayesian network and S its corresponding equation system. An equation $e \in S$ on variables $V(e) = \{v_1, \dots, v_n\}$ is injective with respect to a variable v_1 if and only if e 's probability matrix M_e is sound with respect to v_1 .*

Table 1. A probability matrix that is sound with respect to both variables v_1 and v_2 appearing in it.

		v_2	true	false	maybe
v_1	true		[0, 0,3]	(0,4, 0,6]	(0,3,0,4] \cup (0,6, 1]
	false		(0,3, 0,5]	(0,6, 1]	[0, 0,3] \cup (0,5, 0,6]
	maybe		(0,5, 1]	[0, 0,4]	(0,4, 0,5]

Table 2. A ‘doubly normalized’ probability table associated with table 1.

$\text{Pr}(v_1 v_2)$		v_2	true	false	maybe
v_1	true		0.3	0.2	0.5
	false		0.2	0.4	0.4
	maybe		0.5	0.4	0.1

Proof: \Rightarrow Assume e is injective with respect to v_1 . Now suppose M_e is not sound with respect to v_1 . This means that if we transpose M_e with respect to v_1 , there must be an $\epsilon \in [0..1]$ such that either (1) a column exists in which ϵ is an element of two cells, or (2) a column exists in which ϵ is not an element of any cell.

In the case of (1), we can conclude from the way equations are created in Theorem 2 that there must be a configuration $c \in C(\{v_2, \dots, v_n\})$ and configurations $c_{v_1,1}$ and $c_{v_1,2}$ in $C(\{v_1\})$ such that both $\langle c_{v_1,1}, c, \epsilon \rangle \in e$ and $\langle c_{v_1,2}, c, \epsilon \rangle \in e$ which makes e non-injective and thus we have a contradiction.

In the case of (2), there must be a configuration $c \in C(\{v_1, \dots, v_n\})$ for which there is no tuple $\langle c, \epsilon \rangle \in e$ again contradicting injectivity.

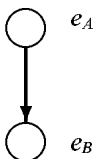
\Leftarrow Assume M_e is sound with respect to v_1 . Now if we transpose M_e with respect to v_1 , again from the way equations are constructed, there must be exactly one tuple for each combination of a configuration for the remaining variables and an $\epsilon \in [0..1]$. \square

Soundness is a condition equivalent to injectivity. When a conditional probability table is ‘multiply normalized’, it can be restructured. Suppose \mathcal{B} is a Bayesian network with node v on which we wish to impose an intervention. We can do this by translating \mathcal{B} to a system S (based on Theorem 2), adding an equation fixing v to one of its values, and subsequently relaxing S by removing an equation from S , to yield S' . The choice of the equation to be removed depends on the real-world meaning of the manipulation, as explained in Section 3. Suitability of this equation depends on semantic reversibility of those equations that are reversed when restructuring S' . Then we can use the procedure of causal ordering to derive the structure of \mathcal{B}' from S' and reuse the conditional probability tables as outlined in this section. In order to be able to translate S' back to a manipulated network \mathcal{B}' , the probability tables of \mathcal{B} should be sound with respect to their endogenous variables in \mathcal{B}' . The probability tables of \mathcal{B}' can then be calculated by transposing those of \mathcal{B} .

7. Example

Consider a communication channel between two locations A and B . Let e_A be the event that there is an error in reading a bit at location A and e_B the event that there is an error in reading a bit at location B . Both events are binary with outcomes **true** and **false**. The error can be caused by the noise in the communication channel. Figure 3 shows a simple Bayesian network fragment modelling the errors e_A and e_B and the probabilistic interactions among them for the case when transmission occurs from location A to location B .

Figure 3. Graph and an associated conditional probability table $\Pr(e_B | e_A)$.

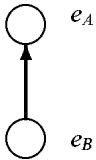


$\Pr(e_B e_A)$	e_A	true	false
		true	0.99
	false	0.01	0.99

There is a clear relationship between error in location A and error in location B : the former causes the latter with high probability. When there is an error at A , the chance that there will be an error at B is 0.99. There is 0.01 chance that the error will restore itself. When there is no error at A , there is a 0.01 chance that sending the bit through the channel will introduce an error at B and 0.99 chance that the transmission will be correct and no error will be observed. The channel is symmetric in the sense of having an equal chance of introducing or restoring an error (i.e. receiving a different bit than the one sent). This is commonly the case for a single bit communication.

Figure 4 shows the same Bayesian network fragment modelling both errors e_A and e_B and the probabilistic interactions among them for the case when transmission occurs from location B to location A .

Figure 4. Graph and an associated conditional probability table $\Pr(e_A | e_B)$.



$\Pr(e_A e_B)$		e_B	true	false
e_A	true		0.99	0.01
	false		0.01	0.99

Please note that the derivation of $\Pr(e_A | e_B)$ from $\Pr(e_B | e_A)$ involved neither $\Pr(e_A)$ nor $\Pr(e_B)$. We want to stress again that the operation of restructuring the network is not the same as the operation of probabilistic arc reversal.

8. Conclusion

Causal manipulation theorems proposed by Spirtes *et al.* (1993) and Pearl (1995) in the context of directed probabilistic graphs, such as Bayesian networks, offer a simple and theoretically sound formalism for predicting the effect of manipulation of a system from its causal model. While the theorems are applicable to a variety of equilibrium causal models, they do not address the issue of reversible causal mechanisms, i.e. mechanisms that are capable of working in several directions, depending on which of their variables are manipulated exogenously. An example involving reversible causal mechanisms is the power train of a car: normally the engine moves the transmission which, in turn, moves the wheels; when the car goes down the hill, however, the driver may want to use the power train to slow down the car, i.e. let the wheels move the transmission, which then moves the engine. Some probabilistic systems can also be symmetric and reversible. For example, the noise introduced by a noisy communication channel does not usually depend on the direction of data transmission.

In this paper, we addressed the question whether Bayesian networks are capable of representing reversible causal mechanisms. Building on the result of Druzdzel and Simon (1993), which shows that conditional probability tables in Bayesian networks can be viewed as descriptions of causal mechanisms, we demonstrated that representing reversible causal mechanisms in Bayesian networks is possible. Our analysis has shown that in order to model reversible causal mechanisms, conditional

probability tables have to fulfill the condition of *soundness*, which is equivalent to *injectivity* in equations. While this limits the applicability of Bayesian networks, there exist systems where our finding and the resulting framework are directly applicable.

Acknowledgments

This research was supported by the Air Force Office of Scientific Research under grant number F49620-00-1-0112 to University of Pittsburgh, and by the National Science Foundation under Faculty Early Career Development (CAREER) Program, grant IRI-9624629, to Dr. Druzdzel. Anonymous reviewers for and later the participants of the *AAAI 1998 Spring Symposium on Prospects for a Commonsense Theory of Causation*, where an early draft of this paper was presented, prompted us for more clarity in our presentation. An anonymous JETAI reviewer offered us several valuable suggestions that improved the final version of the paper. While we are taking full responsibility for any errors in this paper, we would like to thank Herb Simon for enlightening discussions on the topic of causality and causal reversibility.

Notes

1. Druzdzel and Simon (1993) list and discuss several important reasons for an explicit representation of causality in intelligent systems.
2. This definition differs from the traditional notion of injectivity (e.g. Rudeanu 1974) in that it implies that e represents a function that is complete on its domain. We made this simplification for the sake of clarity of exposition.
3. Though we will not prove it here, it is easy to see that reordering an acyclic system S is guaranteed to yield an acyclic system S' if S is a polytree. On the other hand, manipulating an acyclic system \mathcal{S} introduces cyclicity if there are multiple paths from the node containing the equation to be removed to the node containing the variable to become exogenous.

References

- Balke, A., and Pearl, J., 1995, Counterfactuals and policy analysis in structural models. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, San Francisco, CA (San Mateo: Morgan Kaufmann Publishers), pp. 11–18.
- Davidson, R., and Fehling, M. R., 1994, A structured, probabilistic representation of action. In *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-94)*, San Francisco, CA (San Mateo: Morgan Kaufmann Publishers), pp. 154–161.
- Druzdzel, M. J., and Simon, H. A., 1993, Causality in Bayesian belief networks. In *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI-93)*, San Francisco, CA (San Mateo: Morgan Kaufmann Publishers), pp. 3–11.
- Heckerman, D., and Shachter, R., 1995, A definition and graphical representation for causality. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, San Francisco, CA (San Mateo: Morgan Kaufmann Publishers), pp. 262–273.
- Pearl, J., 1988, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (San Mateo: Morgan Kaufmann Publishers).
- Pearl, J., 1995, Causal diagrams for empirical research. *Biometrika*, **82**: 669–710.
- Pearl, J., and Verma, T. S., 1991, A theory of inferred causation. In *KR-91, Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, Cambridge, MA, edited by J. A. Allen, R. Fikes and E. Sandewall (San Mateo: Morgan Kaufmann Publishers), pp. 441–452.
- Rudeanu, S., 1974, *Boolean Functions and Equations*. (Amsterdam: North-Holland Publishing).
- Shachter, R. D., 1986, Evaluating influence diagrams. *Operations Research*, **34**: 871–882.

- Simon, H. A., 1953, Causal ordering and identifiability. In *Studies in Econometric Method. Cowles Commission for Research in Economics. Monograph No. 14*, edited by W. C. Hood and T. C. Koopmans (New York: John Wiley & Sons, Inc.), chapter III, pp. 49–74.
- Spirtes, P., Glymour, C., and Scheines, R., 1993, *Causation, Prediction, and Search* (New York: Springer Verlag).
- Wold, H., 1954, Causality and econometrics. *Econometrica*, **22**: 162–177.