

# Supporting Changes in Structure in Causal Model Construction

Tsai-Ching Lu and Marek J. Druzdzel

Decision Systems Laboratory  
Intelligent Systems Program and School of Information Sciences  
University of Pittsburgh Pittsburgh, PA 15260  
{ching,marek}@sis.pitt.edu

**Abstract.** The term “changes in structure,” originating from work in econometrics, refers to structural modifications invoked by actions on a causal model. In this paper we formalize the representation of reversibility of a mechanism in order to support modeling of changes in structure in systems that contain reversible mechanisms. Causal models built on our formalization can answer two new types of queries: (1) When manipulating a causal model (i.e., making an endogenous variable exogenous), which mechanisms are possibly invalidated and can be removed from the model? (2) Which variables may be manipulated in order to invalidate and, effectively, remove a mechanism from a model?

## 1 Introduction

Graphical probabilistic models, such as Bayesian networks, provide compact and computationally efficient representations of problems involving reasoning under uncertainty. Users can easily update their belief in the states of a modeled system by setting evidence in a model that reflect observations made in the real world. A related formalism of causal models, based on structural equations, in addition to observations, supports prediction of the effects of actions, i.e., external manipulation of modeled systems. Explicit representation of causality in causal models enables users to predict the effects of actions, which in turn allows users to perform counterfactual reasoning [8,12,16].

The problem of predicting the effects of actions was originally referred to in econometrics literatures as predicting the effects of *changes in structure* in simultaneous equation models. Assuming that a modeler has sufficient prior knowledge to predict the effects of changes in structure, researchers in econometrics modeled the effects of actions as “scraping” invalid equations and “replacing” them by new ones [10,13,17,18]. If we assume that the variable manipulated by an action is governed by an *irreversible* mechanism (for example, wearing sunglasses protects our eyes from the sun but it does not make the sun go away), the effect of an action amounts to an arc-cutting operation on the causal graph describing the situation [12,16]. However, there exist a large class of *reversible* mechanisms [4,12,13,15,16, 19,18] that are not amenable to this treatment. For example, a car engine causes the wheels to turn when going up hill, but wheels slow down the engine when going

down hill with transmission being put in a lower gear. An action may reverse the direction of causal relations among variables and consequently have drastic effects on causal graphs.

There have been attempts to assist in predicting the effects of actions on systems containing reversible mechanisms. Bogers [1] developed theorems to support structure modifications when the equation being scraped by an action governs an exogenous variable. Druzdzel and van Leijen [6] studied the conditions under which a conditional probability table in a causal Bayesian network can be reversed when manipulating a reversible mechanism. Dash and Druzdzel [3] demonstrated how various equilibrium systems may violate the arc-cutting operation and further developed *differential causal models* to solve the problem by modeling systems dynamically.

Our approach to supporting changes in structure is based on our representation of reversibility of a mechanism. A mechanism asserts that there exists a relationship among a set of variables. We define the reversibility of a mechanism semantically on the set of possible effect variables of a mechanism. A set of mechanisms is a causal model only if the causal relations among the variables are consistent with the reversibility of its mechanisms. Similarly to STRIPS language [7], we conceptualized an action as consisting of three lists: *PRECONDITION* (a causal model), *ADD* (the set of mechanisms to be added), and *DELETE* (the set of mechanisms to be removed). Consequently, once an action is completely specified, the effect of an action is simply performing the modifications specified in *ADD* and *DELETE* lists on the causal model given in a *PRECONDITION*. Given the *PRECONDITION* and one of the *ADD* or *DELETE* lists of a partially specified action, we proved two theorems to assist modelers in answering two new types of queries: (1) When manipulating a causal model, which mechanisms are possibly invalidated and can be removed from the model? (2) Which variables may be manipulated in order to invalidate and, effectively, remove a mechanism from a model? As an extension of existing approaches [1,3,12,16], we formalize the representation of reversibility of a mechanism and assist modelers in predicting the effects of actions in systems consisting of mixtures of mechanisms.

## 2 Structural Equation Models and Causal Ordering

The work in simultaneous equation models (SEMs) is the root of the work on graphical causal models [8,12,16]. Given an equation  $e$ , we denote the set of variables appearing in  $e$  as  $Vars(e)$ . The set of variables appearing in a set of equations  $E$  is denoted as  $Vars(E) = \bigcup_{e \in E} Vars(e)$ . A structural equation model can be defined as a set of structural equations  $E = \{e_1, e_2, \dots, e_m\}$  on a set of variables  $V = \{v_1, v_2, \dots, v_n\}$  appearing in  $E$ , i.e.,  $V \equiv Vars(E)$ . Each structural equation  $e_i \in E$ , generally written in its implicit form  $e_i(v_1, v_2, \dots, v_n) = 0$ , describes a conceptually distinct mechanism active in a system.<sup>1</sup> A variable  $v_j \in V$  is *exogenous* if it is determined by factors outside the model, i.e., if there exists a

<sup>1</sup> Every structural equation normally contains an error term to represent disturbance due to omitted factors. We will leave out error terms for the simplicity of exposition.

structural equation  $e_i(v_j) = 0$  in  $E$ . A variable is *endogenous* if it is determined by solving the model. We denote the set of exogenous and endogenous variables in  $E$  as  $ExVars(E)$  and  $EnVars(E)$  respectively.  $E$  is *independent* if there is no  $e_i \in E$  such that  $e_i$  is satisfied by all simultaneous solutions of any subset of  $E \setminus e_i$ .  $E$  is *consistent* if the solution set of  $E$  is not empty. In order to ensure that  $E$  is independent and consistent, Simon and Rescher [15] defined the concept of *structure*:

**Definition 1.** A structure is a set of equations  $E$  where  $|E| \leq |Vars(E)|$  such that in any subset  $E' \subseteq E$ : (1)  $|E'| \leq |Vars(E')|$ , and (2) If the values of any  $|Vars(E')| - |E'|$  variables in  $Vars(E')$  are chosen arbitrarily, then the values of the remaining  $|E'|$  variables are determined uniquely.

A SEM  $E$  is *self-contained* if  $E$  is a structure and  $|E| = |V|$ .  $E$  is *under-constrained* if  $E$  is a structure and  $|E| < |V|$ .  $E$  is *over-constrained* if  $E$  is not a structure. Whenever  $|E| > |V|$ ,  $E$  is over-constrained. In general, we use a self-contained SEM to describe an equilibrium system since the set of equations is consistent and independent, and the values of variables are determined uniquely. A self-contained structure  $E$  is *minimal* if it does not contain any proper subset of equations in  $E$  which is self-contained. A minimal self-contained structure is a *strongly coupled component* if it contains more than one equation. A set of equations  $E$  can be represented qualitatively as a matrix, called *structure matrix* [5,13,15], with element  $a_{ij} = \mathbf{x}$  if  $v_j \in V$  participates in  $e_i \in E$ , where  $\mathbf{x}$  is a marker, and  $a_{ij} = 0$  otherwise (see Fig. 1).

	$v_1$	$v_2$	$v_3$	$v_4$	$v_5$	$v_6$	$v_7$	$v_8$
$e_1$	$\mathbf{x}$	0	0	0	0	0	0	0
$e_2$	0	$\mathbf{x}$	0	0	0	0	0	0
$e_3$	0	0	$\mathbf{x}$	0	0	0	0	0
$e_4$	0	$\mathbf{x}$	0	$\mathbf{x}$	$\mathbf{x}$	0	0	0
$e_5$	0	$\mathbf{x}$	$\mathbf{x}$	$\mathbf{x}$	$\mathbf{x}$	0	0	0
$e_6$	0	0	$\mathbf{x}$	0	$\mathbf{x}$	$\mathbf{x}$	0	0
$e_7$	0	0	0	$\mathbf{x}$	0	0	$\mathbf{x}$	0
$e_8$	$\mathbf{x}$	0	0	0	0	0	0	$\mathbf{x}$

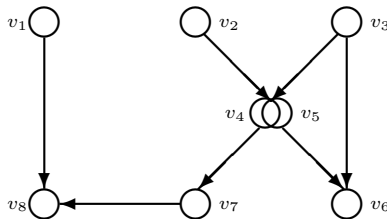


Fig. 1. COA takes a self-contained structure as input and outputs a causal graph.

As shown by Simon [13], a self-contained structure exhibits asymmetries among variables that can be represented by a special type of directed acyclic graph and interpreted causally. He developed a causal ordering algorithm (COA) that takes a self-contained structure  $E$  as input and outputs a *causal graph*  $G_E = \langle N, A \rangle$  where  $N = \{N_1, N_2, \dots, N_r\}$  is a partitioning of  $V$ , consisting of pairwise disjoint sets such that  $\bigcup_{i=1}^r N_i = V$ , and  $A$  is a set of directed arcs  $v \rightarrow N_i$  where  $v \in V$ ,  $N_i \in N$ , and  $v \notin N_i$ . COA starts with *identifying* the minimal self-contained structures in  $E$ . These identified minimal self-contained structures,  $C^0 = \{C_1^0, C_2^0, \dots, C_l^0\}$ , are called *complete structures of 0-th order* and a partition  $N_k^0$  on  $V$  is created for  $Vars(C_k^0)$  for each  $C_k^0 \in C^0$ . For each variable  $v \in N_k^0$ , a corresponding node is created. When a minimal self-contained structure is a strongly coupled component,

i.e.,  $|C_k^0| > 1$ , we draw the nodes created for variables in  $N_k^0$  as overlapping circles because their values need to be solved simultaneously. Next, COA *solves* for the values of  $\text{Vars}(C^0)$  and *removes*  $C^0$  from  $E$ . We denote the new structure  $E \setminus C^0$  as  $\widehat{E}^1$ . COA then *substitutes* the solved values of  $\text{Vars}(C^0)$  into  $\widehat{E}^1$  to obtain the *derived structure of the first order*  $E^1$ . COA repeats the process of identifying, removing, solving, and substituting on the derived structure of  $p$ -th order until it is empty. In addition, whenever a partition  $N_k^p$  and corresponding nodes are created for a complete structure  $C_k^p$  in the complete structures of  $p$ -th order, COA refers  $C_k^p$  back to its equations before any substitutions in  $E$ , denoted as  $\widehat{C}_k^p$ , and add arcs from nodes representing variables in  $\text{Vars}(\widehat{C}_k^p) \setminus \text{Vars}(C_k^p)$  to the nodes representing  $N_k^p$ . Notice that COA creates one-to-one mapping, denoted as  $\langle \widehat{C}_k^p, N_k^p \rangle$ , between the set of equations,  $\widehat{C}_k^p$ , and the set of variables,  $N_k^p$ , in a causal graph. We say that  $\widehat{C}_k^p$  is *mapped to*  $N_k^p$  or vice versa in  $G_E$  (see Example 1).

Since the concept of *endogenous* and *exogenous* variables relative to the structure before substitutions of a complete structure of  $p$ -th order [13] plays an important role in the rest of the discussion, we introduce it formally as follows.

**Definition 2.** Let  $C^p$  and  $C^q$  be complete structures of  $p$ -th and  $q$ -th order respectively in a self-contained structure  $E$ . Let  $\widehat{C}_k^p$  be the structure before any substitutions of a complete structure  $C_k^p \in C^p$  in  $E$  and  $v \in \text{Vars}(\widehat{C}_k^p)$ . We say that  $v$  is *endogenous* in  $\widehat{C}_k^p$ , if  $v \notin \text{Vars}(C^q)$  for all  $q < p$ , and  $v$  is *exogenous* in  $\widehat{C}_k^p$ , if  $v \in \text{Vars}(C^q)$  for some  $q < p$ . We denote the sets of endogenous and exogenous variables in  $\widehat{C}_k^p$  by  $\text{EnVars}(\widehat{C}_k^p)$  and  $\text{ExVars}(\widehat{C}_k^p)$  respectively.

From Definition 2, we know that each variable  $v$  in a self-contained  $E$  can appear as an endogenous variable in only one  $\widehat{C}_k^p$ . We define the *necessary structure* for  $v$  in  $E$  to support changes in structure defined in Sect. 4.2.

**Definition 3.** Let  $G_E$  be the causal graph generated by applying COA to a self-contained structure  $E$ . Let  $v \in N_k^p$  and  $\text{Anc}(N_k^p)$  be the ancestral set of  $N_k^p$  in  $G_E$ . The necessary structure for  $v$ , denoted as  $NS_v$ , is the set of equations that are mapped to  $N_k^p \cup \text{Anc}(N_k^p)$  by COA.

It is easy to see that a necessary structure is self-contained. In other words,  $NS_v$  consists of all equations in  $E$  that are necessary to determine  $v$  uniquely.

*Example 1.* In Fig. 1, COA takes the structure matrix as inputs and identifies  $C^0 = \widehat{C}^0 = \{\{e_1\}, \{e_2\}, \{e_3\}\}$ ,  $\widehat{C}^1 = \{\{e_4, e_5\}\}$ ,  $\widehat{C}^2 = \{\{e_6\}, \{e_7\}\}$ , and  $\widehat{C}^3 = \{\{e_8\}\}$  to generate the causal graph. The mapping between equations and variables are  $\langle e_1, v_1 \rangle$ ,  $\langle e_2, v_2 \rangle$ ,  $\langle e_3, v_3 \rangle$ ,  $\langle \{e_4, e_5\}, \{v_4, v_5\} \rangle$ ,  $\langle e_6, v_6 \rangle$ ,  $\langle e_7, v_7 \rangle$  and  $\langle e_8, v_8 \rangle$ . From the causal graph, we may read off the causal relations among sets of variables. For example,  $\{v_4, v_5\}$  is caused by  $v_2$  and  $v_3$ ,  $v_6$  is caused by  $v_3$  and  $v_5$ , and  $v_7$  is caused by  $v_4$ . We may also read off indirect causal relations such as that  $v_3$  is an indirect cause of  $v_7$ . However, the causal relations between  $v_4$  and  $v_5$  are undefined, since they are in a strongly-coupled component. Notice that  $v_4$  is endogenous in  $\widehat{C}_1^1 = \{e_4, e_5\}$  but exogenous in  $\widehat{C}_2^2 = \{e_7\}$ . The necessary structure for  $v_4$  is  $\{e_2, e_3, e_4, e_5\}$ .

### 3 Reversible Mechanisms

Like any other scientific modeling, structural equation modeling requires us to clearly relate our definitions of variables and structural equations in a SEM to a system in the real world. In general, we start with identifying entities involved in a system. An entity can be a single object (e.g., a patient), a population of similar objects (e.g., male patients in a hospital), or a group of relevant objects (e.g., patients, doctors, and insurance company in a health system). We then define variables to refer to characteristics of entities (e.g., age of a patient) and define structural equations to describe the linkages among variables (mechanisms) in the system. Our prior domain knowledge serves as a guideline in hypothesizing which mechanisms are involved in a system. Therefore, the definitions of structural equations and variables in a SEM are a-priori [13,18]. Simon [14] suggested three classes of sources for specifying mechanisms: *experimental manipulation*, *temporal ordering*, and “*tangible*” *links*. In [12,18], researchers stressed that mechanisms should be *autonomous* in the sense that the external change on any one of the mechanisms does not imply the change of others. For the purpose of illustration, we define mechanisms as follows.

**Definition 4.** A mechanism  $e$ , represented as a structural equation  $e(v_1, v_2, \dots, v_n) = 0$ , asserts that there exists autonomous linkages among the set of variables  $\{v_1, v_2, \dots, v_n\}$ .

Simon [14] further pointed out that different *a-priori* assumptions for one mechanism may lead to different interpretations of causal relations among variables. For example, schooling helps to increase verbal ability in one experimental context, but verbal ability helps in getting higher schooling in another. He used the term *causal mechanisms* to refer to mechanisms considered under different a-priori assumptions. In other words, each causal mechanism represents a distinct theory that we hypothesized about the observation of a phenomena in the real world and is written as a function to explicitly describe the relation of the effect variable and its causes.

**Definition 5.** Given a mechanism  $e$ , a causal mechanism,  $v = f(Pa(v))$ , describes a function  $f$  between the effect variable  $v \in Vars(e)$  and its direct causes  $Pa(v) = Vars(e) \setminus v$ . We say that  $v = f(Pa(v))$  is instantiated from  $e$ .

Generally, there may be more than one causal mechanism instantiated from a mechanism as long as the functions formalized are consistent with the a-priori assumptions. In practice, we believe that people tend to first express a causal mechanism qualitatively as a specification of the effect variable and its causes, and later give it an explicit function. Assuming that the number of variables appearing in a mechanism is fixed, the number of possible effect variables for a mechanism is finite. Consequently, we can classify mechanisms into four categories according to their *reversibility*: (1) *completely reversible*: every variable in the mechanism can be an effect variable, (2) *partially reversible*: two or more of the variables in the mechanism can be effect variables, (3) *irreversible*: only one of the variables

in the mechanism can be an effect variable, and (4) *unknown*: the reversibility of the mechanism is unspecified, i.e., the modeler only knows that variables in a mechanism are relevant, but does not know how they relate to each other causally.

**Definition 6.** *Given a mechanism  $e$ , let  $EfVars(e) \subseteq Vars(e)$  be the set of all possible effect variables of all causal mechanisms instantiated from  $e$ . We say that  $e$  is (1) completely reversible if  $EfVars(e) = Vars(e)$  and  $|EfVars(e)| > 1$ , (2) partially reversible if  $1 < |EfVars(e)| < |Vars(e)|$ , (3) irreversible if  $|EfVars(e)| = 1$ , and (4) unknown if  $|EfVars(e)| = \emptyset$ .*

We emphasize that the notion of reversibility of a mechanism is a semantic one since it is defined with respect to the set of effect variables of a mechanism. A functional relation may be reversible in *functional* sense (invertible), but may not be reversible in *causal* sense [18, footnote 6]. For example, ideal gas law and Ohm's law are given in [19, pp. 40] and [11, pp. 10] respectively as examples of partially reversible mechanisms, although their functional relations are invertible in general. Traditionally the reversibility of mechanisms is considered mainly applicable to mechanical and physical systems [19, pp. 325], since the concept is defined upon causal mechanisms, i.e., the invertibility of a function is a necessary condition for the reversibility. In our formalization, we define the concept of reversibility on the set of effect variables of a mechanism so that we can apply the reversibility to other domains. For example, it would be a mere coincidence that schooling,  $s$ , and verbal ability,  $a$ , can be described as  $s = f(a)$  in one context and  $a = f^{-1}(s)$  in another. However, it is more likely that  $s = f(a)$  in one context and  $a = g(s)$ , where  $g \neq f^{-1}$ , in another.

Notice that the notion of entity plays an essential role in our modeling. We should not confuse the reversibility of a mechanism with *causal mixtures* [2] in which members of entities may not share the same causal relationships. For example, if the relation between schooling and verbal ability is modeled as a causal mixture, we may find that schooling helps to increase verbal ability in one subpopulation of students but verbal ability helps to getting higher schooling in another. However, reversible mechanisms model the same entities in different contexts. For example, the verbal ability helps some population of students to get higher schooling in one context, but in another context the schooling helps the same students to increase their verbal ability.

Taking the reversibility of mechanisms into account, we can define a *causal model* as follows.

**Definition 7.** *A causal model is a set of mechanisms  $E = \{e_1, e_2, \dots, e_m\}$  such that there exists a set of causal mechanisms  $F = \{f_1, f_2, \dots, f_m\}$  instantiated from  $E$ , where each  $f_i \in F$  is an instantiation of  $e_i \in E$ , and  $F$  is a self-contained structure.*

Given a set of mechanisms  $E$ , we can test if  $E$  can form a causal model by checking whether there exists a self-contained  $F$  instantiated from  $E$ . The procedure, denoted as  $IsCausalModel(E)$ , first checks if  $|E| = |Vars(E)|$ . If so, the procedure assumes that  $E$  is a self-contained structure and applies COA qualitatively on  $E$ 's structure matrix to generate the graph  $G_E$ . For each node in  $G_E$ ,

the procedure checks if the mapped mechanisms have valid causal mechanisms to be instantiated, i.e., if there exists a causal mechanism whose effect variable is the same as the one depicted in  $G_E$ . If there exists a set of causal mechanism  $F$ , instantiated from  $E$ , whose effect variables are consistent with  $G_E$ , the procedure verifies that  $E$  is a causal model. In order to assist modelers in hypothesizing causal relations in a mechanism whose reversibility is unknown, the procedure treats its reversibility as completely reversible. Notice that for those  $E$  containing strongly coupled components, we may have several instantiations  $F$  from  $E$ . In other words, an irreversible mechanism cannot participate in a strongly coupled component.

*Example 2.* Assume that the set of mechanisms  $E = \{e_1, e_2, \dots, e_8\}$  for the set of variables  $V = \{v_1, v_2, \dots, v_8\}$  shown in Fig. 1 is stored in a knowledge base along with their causal mechanisms. In the knowledge base,  $e_6$  and  $e_7$  are irreversible where  $EfVars(e_6) = \{v_6\}$  and  $EfVars(e_7) = \{v_7\}$ ,  $e_4$  and  $e_5$  are completely reversible where  $EfVars(e_4) = Vars(e_4)$  and  $EfVars(e_5) = Vars(e_5)$ , and  $e_8$  is partially reversible where  $EfVars(e_8) = \{v_1, v_8\}$ . Consequently,  $E$  is a causal model since there exists a self-contained structure  $F$  that can be instantiated from  $E$ . However, if in the knowledge base we have  $EfVars(e_7) = \{v_4\}$  instead of  $EfVars(e_7) = \{v_7\}$ , then  $E$  is not a causal model since there is no instantiation of  $e_7$  that can make any instantiation  $F$  of  $E$  self-contained.

## 4 Actions in Causal Models

### 4.1 Representation of Actions

Given a causal model that describes a system of interest, we may easily hypothesize different manipulations, such as “raise interest rate” or “reduce tax,” with the intention to influence the values of some target variables. Still, we may not know how other parts of the system may respond to these hypothetical manipulations. In other words, we suspect that our hypothetical manipulations will affect the variables of interest, which are the descendants of the manipulated variables in causal graph, but we are not certain how the equilibrium system will be disturbed by our hypothetical manipulations. Therefore, the process of policy making usually focuses on deliberating the side effects of a manipulation. How should we represent an action in causal modeling to facilitate this deliberation?

Pearl [12, pp. 225] suggested to use the notation  $do(q)$ , where  $q$  is a proposition (variable), to denote an action, since people use the phrases such as “reduce tax” in daily language to express actions. More precisely, an *atomic action*, denoted as  $manipulate(v)$  in [2,16] and  $do(v = v)$  in [12], is invoked by an external force or agent to manipulate the variable  $v$  by imposing on it a probability distribution or holding it at a constant value,  $v = v$ , and replacing the causal mechanism,  $v = f(Pa(v))$ , that directly governs  $v$  in a causal model. The corresponding change in the causal graph is depicted as an arc-cutting operation in which all incoming arcs to the manipulated variable  $v$  are removed [12,16]. Notice that the implicit assumption behind the arc-cutting operation is that the manipulated variable is governed by an *irreversible* mechanism, i.e., only  $v$  can be an effect variable in

mechanism  $e(v, Pa(v)) = 0$ . In order to ensure that the manipulated causal model is self-contained, the irreversible mechanism that governed the manipulated variable has to be removed from the original model. However, when the manipulated variable is governed by a *reversible* mechanism, the arc-cutting operation may lead to inconsistent results. We therefore argue that an action in causal modeling should be defined at the level of mechanisms, not propositions.

In econometric literature (e.g., [10,13,17,18]), a system is represented as a SEM, a set of structural equations, and actions are modeled as “scraping” invalid equations and “replacing” them by new ones. In STRIPS language [7], a situation is represented by a state, conjunctions of function-free ground literals (propositions), and actions are represented as *PRECONDITION*, *ADD*, and *DELETE* lists which are conjunctions of literals. There is a clear analogy between these two modeling formalisms, where the effects of actions are modeled explicitly as adding or deleting fundamental building blocks which are mechanisms in SEM and propositions in STRIPS. We therefore directly translate the “scraping” into *DELETE* and “replacing” into *ADD* and define an action in causal modeling as follows.

**Definition 8.** *An action in causal modeling is a triple  $\langle PRECONDITION, ADD, DELETE \rangle$  where *PRECONDITION* is a causal model  $E$  and *ADD* and *DELETE* are the sets of mechanisms to be added and removed from  $E$  respectively when applying action to  $E$ .*

We consider the context and the effects of an action explicitly in Definition 8. This is consistent with our daily dialogue where we talk about an action and its possible effects under a certain context. For example, the phrase “reduce tax” is usually stated in an economic context with some expectations about how economic units would react.

Note that Definition 8 does not constrain us in what types of mechanisms and how many mechanisms can be specified in *ADD* and *DELETE* lists. There is also no guarantee that the manipulated model will be a self-contained structure. However, the atomic action defined in [12,16], which can be expressed explicitly as  $\langle \{E\}, \{v = v\}, \{e(v, Pa(v)) = 0\} \rangle$  using our definition, always derives a self-contained structure. We use the term *atomic addition*, denoted as  $add(v)$ , to refer to the *ADD* list of an action that consists of only one mechanism,  $\{v = v\}$ , which expresses the manipulation on variable  $v$  in  $E$ . We use the term *atomic deletion*, denoted as  $delete(e)$ , to refer to the *DELETE* list of an action that consists of only one mechanism  $e$  in  $E$ . In order to account for systems with mixtures of different mechanisms, we say that an action is *atomic* if it consists of atomic addition and atomic deletion such that the manipulated model is self-contained.

## 4.2 Action Deliberation

Once we chose to represent an action explicitly including its effects and context, we shift the problem of predicting the effects of an action to which mechanisms should be specified in *ADD* and *DELETE* lists. We call the process of deciding which mechanisms should be in *ADD* and *DELETE* lists *action deliberation*. In this section, we develop theorems to facilitate the process of deliberating about

an atomic action. Given a causal model  $E$ , we seek to answer two new types of queries (1) When making an endogenous variable exogenous, which mechanisms are possibly invalidated and can be removed from the model? (2) Which variables may be manipulated in order to invalidate and, effectively, remove a mechanism from a model? In other words, Query (1) assists modelers in modeling the effects of an action considering the manipulation alternatives at hand. Query (2), on the other hand, assists modelers in identifying the set of possible manipulation alternatives. We start by defining the set of *minimal over-constrained* equations that describes the situation where an atomic addition is added into a model.

**Definition 9.** *A set of over-constrained equations is minimal if it does not contain any over-constrained proper subsets itself.*

**Lemma 1.** *Let  $E$  be a self-contained structure and  $add(v) \equiv \{v = v\}$  be an atomic addition where  $v \in EnVars(E)$ . Let  $E'_v = add(v) \cup E$ . The set of equations  $O'_v = NS_v \cup add(v)$  is minimal over-constrained where  $NS_v$  is the necessary structure of  $v$  in  $E$ .*

Lemma 1 states that an atomic addition makes a self-contained structure minimal over-constrained. Next, we prove Lemma 2 to identify the set of equations such that removing any one of them makes the set of minimal over-constrained equations self-contained again.

**Lemma 2.** *Given  $O'_v$  of  $E'_v$ , deleting any equation  $e \in NS_v$  makes  $O_v = O'_v \setminus e$  self-contained and consequently  $E_v = E'_v \setminus e$  self-contained.*

**Corollary 1.** *Given  $E'_v = add(v) \cup E$ ,  $E'_v$  will remain over-constrained if none of equations  $e \in O'_v$  is removed.*

*Example 3.* Consider the self-contained structure  $E$  in Fig. 1. If we manipulate on variable  $v_7$ , i.e.,  $add(v_7)$ , the resulting set of equations  $E'_{v_7} = E \cup add(v_7)$  becomes over-constrained. From Lemma 1, we know that the set of equations  $O'_{v_7} = \{e_2, e_3, e_4, e_5, e_7, add(v_7)\}$  is minimal over-constrained. From Lemma 2, we know that removing any equation  $e \in \{e_2, e_3, e_4, e_5, e_7\}$  makes the remaining set of equations  $E_{v_7} = E'_{v_7} \setminus e$  a self-contained structure. If we instead remove  $e_6$ , the set of equations  $E'_{v_7} \setminus e_6$  remains over-constrained according to Corollary 1.

Notice that Lemmas 1 and 2 hold for sets of equations. As stated in Sect. 3, a self-contained structure is not necessarily a causal model unless it can be instantiated from a set of mechanisms. Therefore, in order to deliberate about an atomic action in a causal model, we need to verify that the manipulated set of mechanisms is a causal model. In general, we can simply enumerate each mechanism  $e \in NS_v$  and use the procedure  $IsCausalModel(E_v)$  outlined in Sect. 3 to check if the manipulated model  $E_v$  is a causal model. However, we observed that the irreversibility of mechanisms allows us to find the set of possible atomic deletions *locally*.

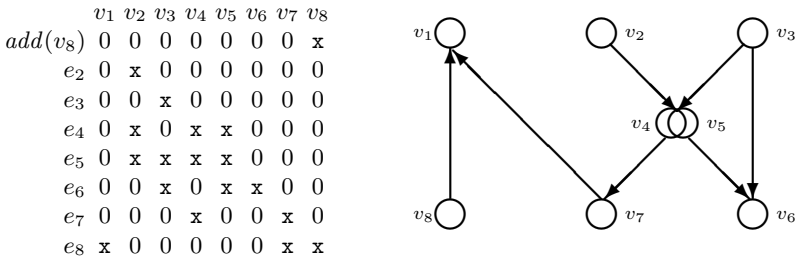
Consider an atomic addition  $add(v)$  on a causal model  $E = \{e_1, e_2, \dots, e_m\}$  and  $v \in EnVars(E)$ . When all mechanisms governing  $EnVars(NS_v)$  in  $NS_v$  are

completely reversible or unknown, we may remove any one of the mechanisms in  $NS_v$  to have a manipulated causal model. When  $v$  is directly governed by an irreversible mechanism  $e$ , we have to remove  $e$  since  $v$  cannot be determined by  $add(v)$  and  $e$  simultaneously in a manipulated model. In other words, the reversibility of mechanism governing the manipulated variable shrinks the set of possible atomic deletions from  $NS_v$  to  $e$ . We therefore learned that propagation of the effects of an atomic addition in a causal model can be blocked by irreversible mechanisms. Now, we prove Theorem 1 to answer Query (1).

**Theorem 1.** *Consider an atomic addition  $add(v)$  in a causal model  $E = \{e_1, e_2, \dots, e_m\}$  and  $v \in EnVars(E)$ . There exists a non-empty set of possible atomic deletions  $D \subseteq NS_v$  such that deleting any mechanism  $d \in D$  derives the causal model  $E_v = E \cup add(v) \setminus d$ .*

Semantically, Theorem 1 identifies the set of manipulated systems that are self-contained. In other words, Theorem 1 assists modelers in hypothesizing a system’s response toward a manipulation. Furthermore, we may find the set of possible atomic deletions *locally* with respect to the order of complete structures in  $NS_v$ . Namely, we perform  $IsCausalModel(E_v)$  checking by enumerating from the mechanisms governing the manipulated variable and recursively up to those governing its ancestors in the causal graph until we reach irreversible mechanisms.

Considering a completely reversible mechanical system, such as the power train described in Sect. 1, a manipulation usually reflects the changes of the operational context as in from driving uphill to driving downhill, for example. The manipulated system normally responds with instantiating different causal mechanisms according to the current operational context. Consequently, the mechanism being removed is usually the one governing the exogenous variable in the system. However, if the mechanism being removed was governing endogenous variables, it means that the linkage among the set of variables is invalid in the manipulated system. For example, transmission or clutch between the engine and the wheels may be broken. Consequently, the link between engine and wheel is no longer valid. We therefore suggest modelers to use different enumeration orders to inspect the set of possible atomic deletions in different applications. When a system consists of irreversible mechanisms, Theorem 1 can further assists modelers in deliberating about the set of possible atomic deletions *locally*.



**Fig. 2.** The structure matrix and its corresponding graph after the atomic action  $\langle E, add(v_8), delete(e_1) \rangle$ .

*Example 4.* Consider the set of mechanisms in Fig. 1 and its reversibility assumed in Example 2. The set of possible atomic deletions for manipulating variable  $v_8$ ,  $add(v_8)$ , is  $\{e_1, e_8\}$  according to Theorem 1. Notice that the irreversibility of mechanisms allows us to find the set of possible atomic deletions in  $\{e_1, e_7, e_8\}$  instead of  $NS_{v_8}$ . Moreover,  $E_{v_8} = E \cup add(v_8) \setminus e_7$  is not a causal model since  $v_7$  cannot be an effect variable in  $e_8$  according to the reversibility of  $e_8$  in the knowledge base. However, if we choose to remove  $e_1$ ,  $delete(e_1)$ , the manipulated model is shown in Fig. 2.

The dual theorem to Theorem 1 is to identify the set of possible atomic additions given an atomic deletion, which answers Query (2).

**Theorem 2.** *Consider an atomic deletion  $delete(e)$  for a causal model  $E = \{e_1, e_2, \dots, e_m\}$  where  $e \in E$ . Let  $G_E$  be the causal graph of  $E$ . Let  $e \in \widehat{C}_k^p$  and  $N_k^p$  is mapped to  $\widehat{C}_k^p$  in  $G_E$ . Let  $Des(N_k^p)$  be the descendants of  $N_k^p$  in  $G_E$ . There exists a nonempty set of variables  $A \subseteq (Des(N_k^p) \cup N_k^p)$  such that manipulating any variable  $a \in A$  derives the causal model  $E_a = E \cup add(a) \setminus e$ . The set of mechanisms  $\bigcup_{a \in A} add(a)$  is called the set of possible atomic additions.*

*Example 5.* Consider the set of mechanisms in Fig. 1 and its reversibility assumed in Example 2. The set of possible atomic additions for removing mechanism  $e_4$ ,  $delete(e_4)$ , is  $\{v_4, v_5\}$  according to Theorem 2.

## 5 Discussion

This paper formalizes the representation of reversibility of a mechanism to support modeling of changes in structure. We define the reversibility of a mechanism semantically on the set of possible effect variables. This definition allows us to extend the concept of reversible mechanisms from traditional mechanical and physical systems to other systems. We further draw the analogy between the action represented in SEM and STRIPS languages to argue that the context and the effects of an action should be represented explicitly in causal modeling. Our formalization allows us to answer two new types of queries: (1) When manipulating a causal model, which mechanisms are possibly invalidated and can be removed from the model? (2) Which variables may be manipulated in order to invalidate and, effectively, remove a mechanism from a model? In practical applications, it may be desirable to further encode domain knowledge, such as whether a variable is manipulatable ethically and what is the cost of such manipulation, along with each mechanism.

**Acknowledgments.** This research was supported by the Air Force Office of Scientific Research, grant F49620-00-1-0112 and by the National Science Foundation under Faculty Early Career Development (CAREER) Program, grant IRI-9624629. We thank Denver Dash, Hans van Leijen and Daniel Garcia Sanchez for their helpful comments on the early draft of this paper. We also thank anonymous reviewers for suggestions improving the clarity of the paper. Marek Druzdzel is currently with ReasonEdge Technologies, Pte, Ltd, <http://www.reasonedge.com>, [mjdruzdzel@reasonedge.com](mailto:mjdruzdzel@reasonedge.com).

## References

1. Jeroen J.J. Bogers. Supporting the change in structure in a decision support system based on structural equations. Master's thesis, Department of Technical Mathematics and Informatics, Delft University of Technology, Delft, The Netherlands, August 1997.
2. Gregory F. Cooper. An overview of the representation and discovery of causal relationships using Bayesian networks. In Glymour and Cooper [8], chapter one, pages 3–62.
3. Denver Dash and Marek J. Druzdzel. Caveats for causal reasoning with equilibrium models. In *Sixth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, 2001. In this proceeding.
4. Marek J. Druzdzel. *Probabilistic Reasoning in Decision Support Systems: From Computation to Common Sense*. PhD thesis, Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA, December 1992.
5. Marek J. Druzdzel and Herbert A. Simon. Causality in Bayesian belief networks. In *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI-93)*, pages 3–11, San Francisco, CA, 1993. Morgan Kaufmann Publishers.
6. Marek J. Druzdzel and Hans van Leijen. Causal reversibility in Bayesian networks. *Journal of Experimental and Theoretical Artificial Intelligence*, 13(1):45–62, Jan 2001.
7. Richard E. Fikes and Nils J. Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2(3-4):189–208, 1971.
8. Clark Glymour and Gregory F. Cooper, editors. *Computation, Causation, and Discovery*. AAAI Press, Menlo Park, CA, 1999.
9. William C. Hood and Tjalling C. Koopmans, editors. *Studies in Econometric Method. Cowles Commission for Research in Economics. Monograph No. 14*. John Wiley & Sons, Inc., New York, NY, 1953.
10. Jacob Marschak. Economic measurements for policy and prediction. In Hood and Koopmans [9], chapter I, pages 1–26.
11. P. Pandurang Nayak. Causal approximations. *Artificial Intelligence*, 70(1–2):1–58, 1994.
12. Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2000.
13. Herbert A. Simon. Causal ordering and identifiability. In Hood and Koopmans [9], chapter III, pages 49–74.
14. Herbert A. Simon. The meaning of causal ordering. In Robert K. Merton, James S. Coleman, and Peter H. Rossi, editors, *Qualitative and Quantitative Social Research: Papers in Honor of Paul F. Lazarsfeld*, chapter 8, pages 65–81. The Free Press, A Division of Macmillan Publishing Co., Inc., 1979.
15. Herbert A. Simon and Nicholas Rescher. Cause and counterfactual. *Philosophy of Science*, 33(4):323–340, December 1966.
16. Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer Verlag, New York, 1993.
17. Robert H. Strotz and H.O.A. Wold. Recursive vs. nonrecursive systems: An attempt at synthesis; part I of a triptych on causal chain systems. *Econometrica*, 28(2):417–427, April 1960.
18. Herman Wold. Causality and econometrics. *Econometrica*, 22(2):162–177, April 1954.
19. Herman Wold and Lars Jureen. *Demand Analysis. A Study in Econometrics*. John-Wiley and Sons, Inc., New York, 1953.