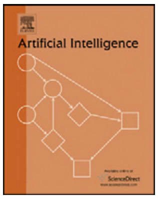
Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint



Causal analysis with Chain Event Graphs

Peter Thwaites [a,*], Jim Q. Smith [a], Eva Riccomagno [b]

[a] Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom
[b] Department of Mathematics, Università degli Studi di Genova, Via Dodecaneso 35, 16146 Genova, Italy

ARTICLE INFO

Article history:
Received 16 January 2009
Received in revised form 13 May 2010
Accepted 13 May 2010
Available online 20 May 2010

Keywords:
Back Door Theorem
Bayesian Network
Causal manipulation
Chain Event Graph
Conditional independence
Event tree
Graphical model

ABSTRACT

As the Chain Event Graph (CEG) has a topology which represents sets of conditional independence statements, it becomes especially useful when problems lie naturally in a discrete asymmetric non-product space domain, or when much context-specific information is present. In this paper we show that it can also be a powerful representational tool for a wide variety of causal hypotheses in such domains. Furthermore, we demonstrate that, as with Causal Bayesian Networks (CBNs), the identifiability of the effects of causal manipulations when observations of the system are incomplete can be verified simply by reference to the topology of the CEG. We close the paper with a proof of a Back Door Theorem for CEGs, analogous to Pearl's Back Door Theorem for CBNs.

1. Introduction

Much recent work in the field of causality has focused on how *cause* relates to control, and the analysis of *controlled* models. Here, with the advocates of this approach we assume the existence of a background *idle* system which is then subjected to some sort of *intervention* or *manipulation*.

The Bayesian Network (BN) is the most commonly used graphical tool for representing complex dependency relationships. Interpreting the directionality of the edges of the BN as *causal* leads to the Causal Bayesian Network (CBN), which uses a non-parametric representation based on structural equation models [12,19,21,30]. CBNs provide a framework for expressing assertions about what might happen when the system under study is externally manipulated and some of its variables are assigned certain values.

BNs and CBNs are ideal for problems which admit a natural product space structure. However many processes do not have this — they are asymmetric in the sense that measurement variables may have different collections of possible outcomes given different vectors of values for sets of ancestral variables. For a variety of examples see [4,1,10,16,2,17,23]. Context-specific variants of BNs exist [2,26,23,18], usually with tree-structured conditional probability tables annexed to the vertices of the BN to allow for the analysis of context-specific independence properties. Although these graphs allow the analyst a greater flexibility than unmodified BNs, they are still inefficient representations of processes (such as treatment regimes) whose unfolding depends on the state of the system at any particular point and the values of specific covariates at that point. Similarly, they are not ideal representations of problems where **no** outcomes of certain variables could logically occur given some vectors of values of ancestral variables.

[*] Corresponding author.
E-mail address: Peter.Thwaites@warwick.ac.uk (P. Thwaites).

doi:10.1016/j.artint.2010.05.004

There have been major advances in CBN theory in the last decade (see [6,14,20,7,35,34], and [21] for a good review of these). The basic *Do* intervention of Pearl [19] has been extended to *functional* manipulations ($Do\, X = g(Z)$), and *stochastic* manipulations which assign a new probability distribution to the state space of the manipulated variable. Nonetheless, at the most primitive level a manipulation of a BN still corresponds to the setting of certain measurement variables to specific values, possibly following some rule or policy. However, whereas the effects of a cause can be reasonably represented by a random variable, at times the specification of a cause as the value of a random variable can be artificial. Causes are more naturally represented as conditioning *events*, and such conditioning is not elegantly expressed in the BN. An analogous case is made by Dawid [5] who argues that causes are decisions and not decision rules.

Although topologically complex, event trees (the elicitation of which often provides the first stage in the development of a model) explicitly acknowledge structural asymmetries — context-specific and sample space information is embedded in the topology of the tree. Their semantics are also often closer to many verbal descriptions of the world, especially when those descriptions revolve around how things happen rather than how the world appears. Event trees however, cannot be readily interrogated for the conditional independence structure of a model.

Trees also have their advocates in the study of causality [27,30,25]. In the related field of decision analysis, French and Insua [11] argue that the advantages of influence diagrams over decision trees are illusory, and point out that asymmetric problems *in which a particular choice of action at a decision node makes available different choices of action at subsequent decision nodes than those available after an alternative choice* are the rule rather than the exception. Using trees we can also choose the level of detail we include in our representation, and this can be dependent on what we intend to *do* to the system. We can incorporate context specific information that is informative about various causal hypotheses (see for example [8]). This is particularly useful in models of biological regulatory mechanisms, which typically contain many noisy *and* and *or* gates [28].

In [28] we introduced an alternative graphical model — the *Chain Event Graph* (CEG), constructed from an event tree together with a set of exchangeability assumptions. It can be seen as a generalisation of a probability graph [3,27], and typically has many fewer nodes than the original tree. The CEG retains those characteristics of the event tree which allow for the representation of asymmetric problems; but they are also more flexible and useful, since their nodes represent intrinsic events in the problem and their edges dependencies between them. They express topologically all the conditional independence structure associated with a problem (this is not bolted on as with context-specific BNs), and also any sample space information generated by the asymmetry of the problem. So in a non-causal context, CEGs provide a more expressive (if somewhat more complicated) topological framework for expressing collections of conditional independence statements than the discrete BN.

We present here an extension of CEG models which provides a framework for **causal** reasoning. We believe this extension to be as transparent and compelling as the extension from BNs to CBNs. In Section 2 we give a brief definition of the CEG and a description of how to read conditional independence properties from it. This section also contains an example of how an asymmetric problem can be depicted using such a graph. Section 3 introduces the manipulation of these graphs, and this theory is developed in Section 4 where we address the identification of the effects of manipulations. Section 5 introduces a Back Door Theorem for CEGs, a generalisation of Pearl's Back Door Theorem for BNs [21].

2. Chain Event Graphs

2.1. Definition

We provide here a brief definition and description of the CEG. A more comprehensive definition can be found in [28].

The CEG is a function of an *event tree* [27], and we begin this section with a brief description of this graph. An event tree T is a directed, rooted tree, with vertex set $V(T)$ and edge set $E(T)$. The non-leaf vertices are called *situations* and the set of situations $S(T)$. The root-to-leaf paths $\{\lambda\}$ of T form the atoms of the event space (called the *path σ-algebra* of T), and label the different possible unfoldings of the described process. Events measurable with respect to this space are unions of these atoms.

Each situation v serves as an index of a random variable $X(v)$ whose values describe the next stage of possible developments of the unfolding process. The state space $\mathbb{X}(v)$ of $X(v)$ can be identified both with the set of directed edges $e(v, v') \in E(T)$ emanating from v in T and the set of end-nodes $v' \in V(T)$ of these edges. For each $X(v)$ $(v \in S(T))$ we let

$$\Pi(v) \equiv \left\{\pi\left(v' \mid v\right) \mid v' \in \mathbb{X}(v)\right\}$$

where $\pi(v' \mid v) \equiv P(X(v) = v')$ are called the *primitive* probabilities of the tree; and

$$\Pi(T) \equiv \left\{\Pi(v)\right\}_{v \in S(T)}$$

A full specification of the probability model is given by $(T, \Pi(T))$.

We extend Shafer's definition of an event tree by the introduction of three further properties — *coloured* edges, *stages* and *positions*.

Definition 1. The *stages, colouring* and *positions* of an event tree are defined as follows:

1. Two situations v^1 and v^2 are in the same *stage* u if $X(v^1)$ and $X(v^2)$ have the same distribution under some bijection ψ between their sample spaces. We label the set of stages of the tree T by $L(T)$.
2. For $v^1, v^2 \in u$ (for some stage u), the edges $e(v^1, v^{1\prime})$ and $e(v^2, v^{2\prime})$ have the same *colour* if $e(v^1, v^{1\prime})$ maps to $e(v^2, v^{2\prime})$ under this bijection ψ, and $\pi(v^{2\prime} \mid v^2) = \pi(v^{1\prime} \mid v^1)$.
3. Two situations v^1 and v^2 are in the same *position* w if for each subpath emanating from v^1, the ordered sequence of colours is the same as that for some subpath emanating from v^2. We label the set of positions of the tree T by $K(T)$.

So two situations are in the same stage when the **immediate** future evolution from both situations is governed by the same probability law. Two situations are in the same position when the **entire** future evolution from both situations is governed by the same probability law.

Definition 2. The Chain Event Graph C (a function of a tree T) is the coloured mixed graph with vertex set $V(C)$, directed edge set $E_d(C)$ and undirected edge set $E_u(C)$ defined by:

1. $V(C) \equiv K(T) \cup \{w_\infty\}$.
2. (a) For $w, w' \in V(C) \setminus \{w_\infty\}$, there exists a directed edge $e(w, w') \in E_d(C)$ *iff* there are situations $v, v' \in S(T)$ such that $v \in w \in K(T)$, $v' \in w' \in K(T)$ and there is an edge from v to v' in $E(T)$.
   (b) For $w \in V(C) \setminus \{w_\infty\}$, there exists a directed edge $e(w, w_\infty) \in E_d(C)$ *iff* there is a situation $v \in S(T)$ and a leaf-node v' of T such that $v \in w \in K(T)$ and there is an edge from v to v' in $E(T)$.
3. For $w^1, w^2 \in V(C)$, there exists an undirected edge $e(w^1, w^2) \in E_u(C)$ *iff* there are situations $v^1, v^2 \in S(T)$ such that $v^1 \in w^1 \in K(T)$, $v^2 \in w^2 \in K(T)$ but v^1, v^2 are members of the same stage u for some $u \in L(T)$. We say that w^1 and w^2 are in the same stage u, and label the set of stages of C by $L(C)$.
4. If $v \in w \in K(T)$, $v' \in w' \in K(T)$ and there is an edge from v to v' in $E(T)$, then the edge $e(w, w') \in E_d(C)$ has the same colour as the edge $e(v, v')$.

There is a one-to-one correspondence between the root-to-leaf paths in T and the root-to-sink paths in C. Each atom of T becomes a path $\lambda(w_0, w_\infty)$ in C, and these paths form the atoms of the σ-algebra of the CEG. Events in C are unions of $w_0 \to w_\infty$ paths. We write $w \prec w'$ when the position w precedes the position w' on a $w_0 \to w_\infty$ path. We call w a *parent* of w' if there exists an edge $e(w, w') \in E_d(C)$. A collection W of positions $w \in V(C)$ is called a *fine cut* of C if all $w_0 \to w_\infty$ paths in C pass through exactly one $w \in W$.

When the set of stages $L(T)$ of an event tree is identical to the set of positions $K(T)$, we call the resultant CEG C *simple*. Simple CEGs have no undirected edges and since the colouring is therefore redundant, they can be treated as directed acyclic graphs. An example of a simple CEG can be found in [33].

Each stage u in our CEG C serves as an index of a random variable $X(u)$ whose values describe the next stage of possible developments of the unfolding process. The state space $\mathbb{X}(u)$ of $X(u)$ can be identified with the set of directed edges $e(w, w') \in E_d(C)$ emanating from any $w \in u$. For each $X(u)$ we let

$$\Pi(u) \equiv \{\pi(e(w, w') \mid w) \mid w \in u\} \quad \text{and} \quad \Pi(C) \equiv \{\Pi(u)\}_{u \in L(C)}$$

A full specification of the probability model is given by $(C, \Pi(C))$.

2.2. Conditional independence

The conditional independence properties of a model can be read rapidly from the topology of a CEG-representation of the model.

For a stage $u \in L(C)$, let the event which is the union of all $w_0 \to w_\infty$ paths passing through some $w \in u$ be labelled $\Lambda(u)$, and let $Z(u)$ be a variable whose state space $\mathbb{Z}(u)$ can be identified with the set of $w_0 \to w \in u$ subpaths. Then as shown in [28] we have that

$$X(u) \amalg Z(u) \mid \Lambda(u)$$

which can be read as — $X(u)$ *is independent of any variable defined* upstream *of u, given the event* $\Lambda(u)$.

So, if we know that a unit has reached some stage u, then we do not need to know how our unit reached u (i.e. along which $w_0 \to w \in u$ subpath) in order to predict how the process is going to unfold in the **immediate** future (i.e. along which edge leaving $w \in u$ our unit is going to proceed). In a BN the analogous result is that we need only the vector of values taken by a variable's parents in order to predict the value taken by that variable.

For a position $w \in V(C)$, let the event which is the union of all $w_0 \to w \to w_\infty$ paths be labelled $\Lambda(w)$, let $Y(w)$ be a variable whose state space $\mathbb{Y}(w)$ can be identified with the set of $w \to w_\infty$ subpaths, and let $Z(w)$ be a variable whose state space $\mathbb{Z}(w)$ can be identified with the set of $w_0 \to w$ subpaths. Then (from [28]) we have that

$$Y(w) \amalg Z(w) \mid \Lambda(w)$$

which can be read as — *variables defined* downstream *of w are independent of variables defined* upstream, *given the event* $\Lambda(w)$.

So, if we know that a unit has reached some position w, then we do not need to know how our unit reached w in order to predict how the process is going to behave during its complete future unfolding (i.e. along which subpath emanating from w our unit is going to pass). In a BN the analogous result is that in order to predict the vector of values taken by a set of variables $\mathbf{X}$, we need to know the vector of values of the set $pa(\mathbf{X}) \setminus \mathbf{X}$.

If a model can be depicted by a BN then in our CEG of this model we can combine these position and stage-based expressions to give us exactly the same set of conditional independence statements that we could deduce from the BN (see [28]). However, as noted in Section 1, in many applications our processes are highly asymmetric, and model elicitation produces asymmetric event trees with event spaces not admitting a natural product space structure. In such cases a CEG-depiction of the problem embeds context-specific conditional independence properties within the **topology** of the graph (which is not the case with BNs), and allows the analyst to deduce other context-specific properties that might not be apparent before the elicitation process is undertaken. The examples below illustrate these points.

We believe that the Markov property will prove to be complete with respect to the class of independence properties presented here and in [31,28], but this is a topic for a future paper.

2.3. An example

This section contains an example of a model with the type of asymmetric structure described above. For simplicity the problem variables in this example are all binary and in the form of indicators — something happens or it doesn't.

Example 2.1. The police hold a suspect S whom they believe threw a brick through a shop window and stole a quantity of money. They wish to bring S to court, but there may be reasons for them not proceeding (such as the lack of availability of a judge; police-force policy on the amount of money needing to be stolen before they are prepared to pay for forensic testing, or take suspects to court etc.). Whether they proceed or not can be thought of as outcomes of an indicator X_1.

It is uncertain that the suspect was present at the scene when the money was stolen (indicator X_2), that he was the individual who threw the brick and stole the money (indicator X_3), that the forensic service will find glass matching the window glass on the clothing of S (indicator X_4), that a witness W will identify S (indicator X_5), and whether S will be convicted or released (the *effect* indicator of interest X_6).

We could construct our event tree and hence our CEG in temporal order so that edges representing the outcomes of X_2 and X_3 preceded those associated with X_1. However, if we suppose that we are constructing our tree through eliciting information from members of the police force then X_1 is the first indicator of interest. In this our method is similar to that used in the construction of decision trees in decision analysis [29].

Unless S is identified by the witness W, then S will not be convicted. The glass match is believed only to depend on whether S threw the brick; and the quality of the witness identification is believed to depend only on whether S was at the scene of the crime or not. This is sufficient information for us to construct a CEG for the problem. Our CEG is given in Fig. 1, where for simplicity only a subset of the edges in $E_d(C)$ have been coloured.

As the reasons which might lead to the police not proceeding are not related to their beliefs about S's presence at the crime scene etc., we can see that the probabilities associated with edges labelled *present*, *not present*, *stole money*, *did not steal money* are unaffected by whether they succeed edges labelled *proceed* or *not proceed*. Hence the positions w_1 and w_2 in Fig. 1 are in the same stage (and so connected by an undirected edge), as are the positions w_3 and w_4. The position w_3 represents the history (*proceed, present*). S could only have thrown the brick if he was present at the scene, so edges labelled *present* are succeeded by edges labelled *stole money*, *did not*, but edges labelled *not present* are not.

If the police do not proceed, then forensic evidence is not collected, and as S is not taken to court, W will not be asked to testify. Hence there are no edges labelled *glass match*, *no match*, *identifies S* or *does not* on $w_0 \to w_\infty$ paths starting with the edge *not proceed*.

The success of the forensic test being dependent only on whether or not S threw the brick tells us that the positions w_6 and w_7 are in the same stage (and hence connected by an undirected edge). The quality of identification being dependent only on whether S was at the crime scene or not tells us that the positions w_8 and w_9 are in the same stage, and that the positions w_{10} and w_{11} are in the same stage.

If W does not identify S (position w_{13}), then the probability of conviction is zero, and there is only one edge $e(w_{13}, w_\infty)$. If W does identify S, then the probability of conviction depends on whether the forensic test was successful (position w_{12}) or not (position w_{14}). This last is not explicit in what the police have told us, but is apparent from the fact that the police would not pay for the forensic test if it was not going to be any use to them in the case.

The detailing above of the possible developments of the case amounts to a description of the conditional independence structure of the problem, and clearly most of the information provided is context-specific. Fig. 1 illustrates the fact that we are explicitly using the topology of the CEG to express the resulting asymmetric dependency structure.

We can of course represent this problem using a BN by adding *dummy* outcomes to the sets of possible outcomes of variables X_3, X_4 and X_5, and imposing a product space structure onto the problem. The BN would have to be supplemented by context-specific conditional independence information, but there are methods for doing this [2,23].

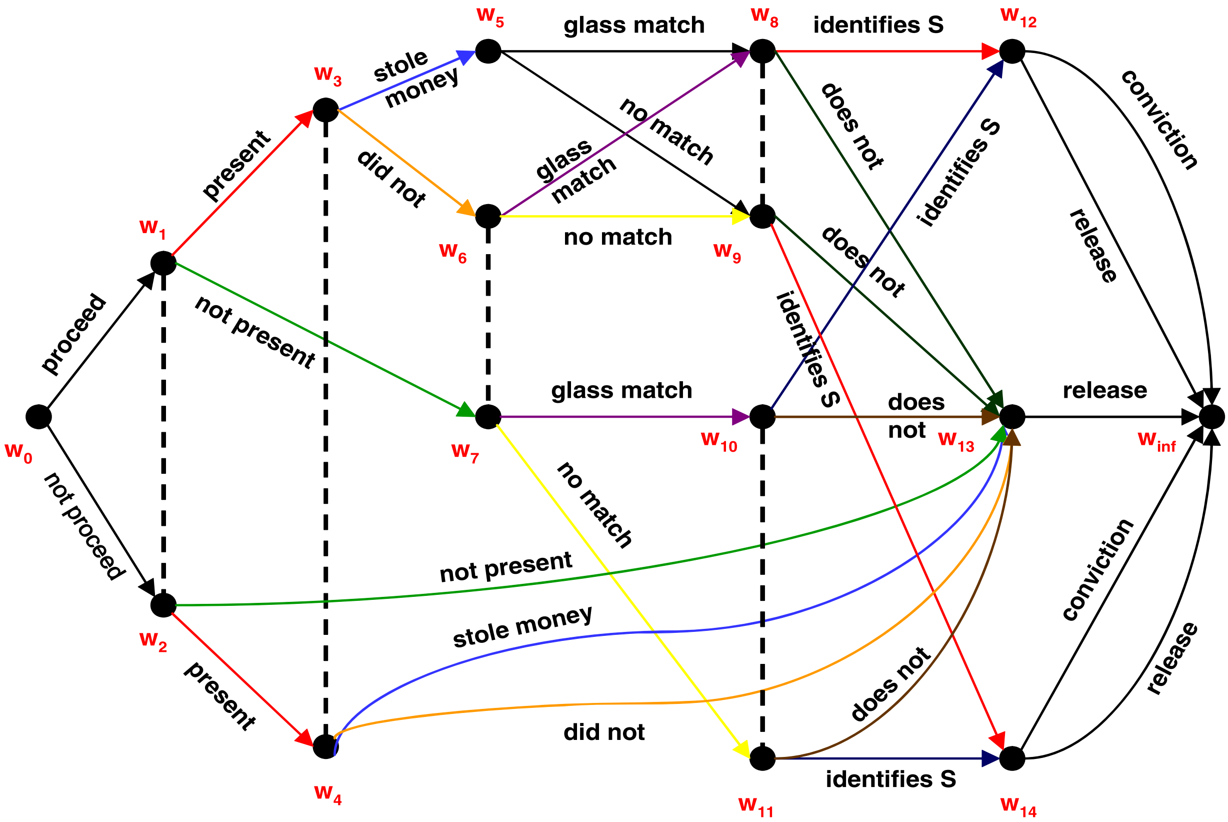


Fig. 1. CEG for Example 2.1.

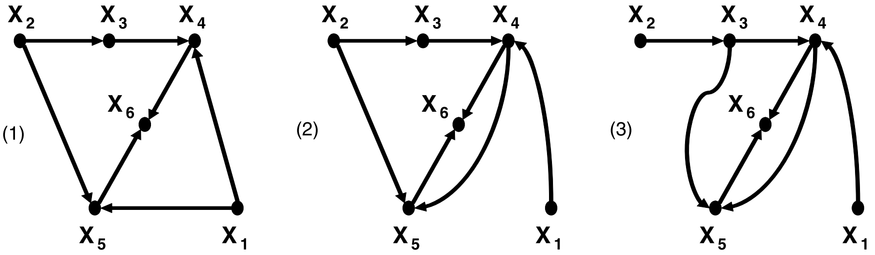


Fig. 2. Three possible BNs for Example 2.1.

The problem with such an approach is that our BN will then be ill-defined. If we add a third possible outcome to the pairs of outcomes already present for X_3, X_4, X_5 (signifying that the conditions for X_i taking either of its current values have not been met), and add edges representing these outcomes to the subpaths leaving w_2 and w_4 and terminating in w_{13}, then we can use the resultant CEG to establish the following conditional independence properties involving the variable X_5:

$$X_5 \amalg (X_3, X_4) \mid (X_1, X_2) \tag{1}$$

$$X_5 \amalg (X_1, X_3) \mid (X_2, X_4) \tag{2}$$

$$X_5 \amalg (X_1, X_2) \mid (X_3, X_4) \tag{3}$$

Coupling each of these in turn with the properties relating to X_1, X_2, X_3, X_4 and X_6, we can draw three different BNs (shown in Fig. 2), but there is no single BN for this problem which depicts all three properties. We could choose one of the BNs in Fig. 2 and supplement it with context-specific information, but this is not ideal.

Now we have already noted that our CEG could be drawn in a different order, so the CEG-representation of a problem is also not unique. But the difference here is that the conditional independence structure of a problem is encoded in the topology of the CEG, and can easily be read from the graph if it is constructed in an order wherein problem variables always appear before their descendants. Furthermore, if the CEG is constructed in such an order, then by supplementing the model with some additional assumptions discussed below, it has a causal interpretation. We are thus able to extend the

CEG's semantics to represent various causal hypotheses in a way analogous to that by which the semantics of the BN are extended to give a CBN.

3. Manipulating the Chain Event Graph

3.1. Principles

In this section we define what we mean by a manipulation of a CEG, and in Section 3.2 we show how such manipulations relate to interventions on BNs.

A CEG provides a flexible framework for expressing what might happen were a model to be manipulated or made subject to some control. Such a manipulation results in a modification (usually a simplification) of the topology of our (*idle*) CEG to produce a *manipulated* CEG. For many manipulations this modification consists simply of the *pruning* (removing) of specified edges and positions and the reassignment of the probabilities on a small subset of the directed edges of the CEG.

Discussions of causal manipulation can be found in [13,21,27,30]. Here we follow Pearl [21] whose *Do* operator describes interventions on directed acyclic graphs (DAGs). The joint density function of a set of random variables $X_1, \dots, X_n$ with sample spaces $\mathbb{X}_1, \dots, \mathbb{X}_n$ factorises according to a DAG as:

$$p(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i \mid pa_i)$$

where $p(x_i \mid pa_i)$ is the probability of X_i taking the value x_i given that its **parents** among $X_1, \dots, X_n$ take values from $x_1, \dots, x_n$.

A random variable is forced to assume a specific value with probability one, say $X_j = \hat{x}_j$ for some $j \in \{1, \dots, n\}$ and $\hat{x}_j \in \mathbb{X}_j$. A new density $p(\cdot \parallel \hat{x}_j)$ (using the notation of [15]) is defined on $\{X_1, \dots, X_n\} \setminus \{X_j\}$ by the formula:

$$p(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n \parallel \hat{x}_j) \equiv \frac{p(x_1, \dots, x_n)}{p(x_j \mid pa_j)} \tag{3.1}$$

This formula expresses the effect of the manipulation *Do* $X_j = \hat{x}_j$. The distribution of the variable X_j has been replaced by a new one which assigns the whole weight to the value x_j. The expression can be readily modified for say a *stochastic* intervention by replacing the distribution of X_j by some other (less crude) new distribution. The manipulation of CEGs is defined in an analogous manner by the replacing of the distributions of some of the random variables sitting on positions by new distributions.

Definition 3. Let $(T, \Pi(T))$ be a tree with corresponding CEG $(C, \Pi(C))$. Let $D \subset S(T)$ be a subset of the situations of the tree, and $\hat{\Pi}_D \equiv \{\hat{\pi}(v' \mid v)\colon\ v \in D,\ v' \in \mathbb{X}(v)\}$ be a new distribution on $v \in D$. Then we define a manipulation of our tree by:

$$\hat{P}\big(X(v) = v'\big) \equiv \begin{cases} \pi(v' \mid v) & v \notin D \\ \hat{\pi}(v' \mid v) & v \in D \end{cases}$$

for all $v' \in \mathbb{X}(v),\ v \in S(T)$. The manipulated tree $(\hat{T}, \hat{\Pi}(\hat{T}))$ is the tree so defined, and the manipulated CEG $(\hat{C}, \hat{\Pi}(\hat{C}))$ is the CEG of the manipulated tree.

Definition 4. A manipulation of a **tree** is called *positioned* if the partition of the positions after the manipulation is equal to or a coarsening of the partition before manipulation. It is called *staged* if the partition of the stages after the manipulation is equal to or a coarsening of the partition before manipulation.

A positioned manipulation of a tree treats all sample units identically when their future development distributions are identical. A staged manipulation treats sample units identically if their **next** development in the idle system is the same. In our experience, it is usually sufficient to restrict study to positioned manipulations. We note that the simple *Do*, *functional* and *stochastic* interventions on a BN considered by Pearl [19,20] are all both positioned and staged.

Useful manipulations of any system tend to be *local* in the sense that only a small number of components are manipulated. This idea is formalised in Definition 3 where only a subset of edges of a tree or CEG have their probabilities reassigned. Where a manipulation of a CEG corresponds to a simple *Do* or a *functional* intervention, one edge only on each root-to-sink path will have its probability altered (to either 0 or 1). This intervention can also be considered as a manipulation to a set of positions W — those positions that terminate the edges which have been assigned a probability of one. Such a manipulation could be, for example, the assignment of patients with particular values of a set of covariates (detailed by their current positions) to a particular treatment regime (a set of subsequent positions W).

For interventions of this type, the conditional independence properties characterised by the stage-structure both upstream and downstream of the manipulation are those of the idle system. When a manipulation is to a set of positions W where not all root-to-sink paths pass through positions which are *parents* of positions in W, the stage-structure downstream of W in the manipulated graph is retained from the idle CEG, but the stage-structure upstream of W is often altered. The

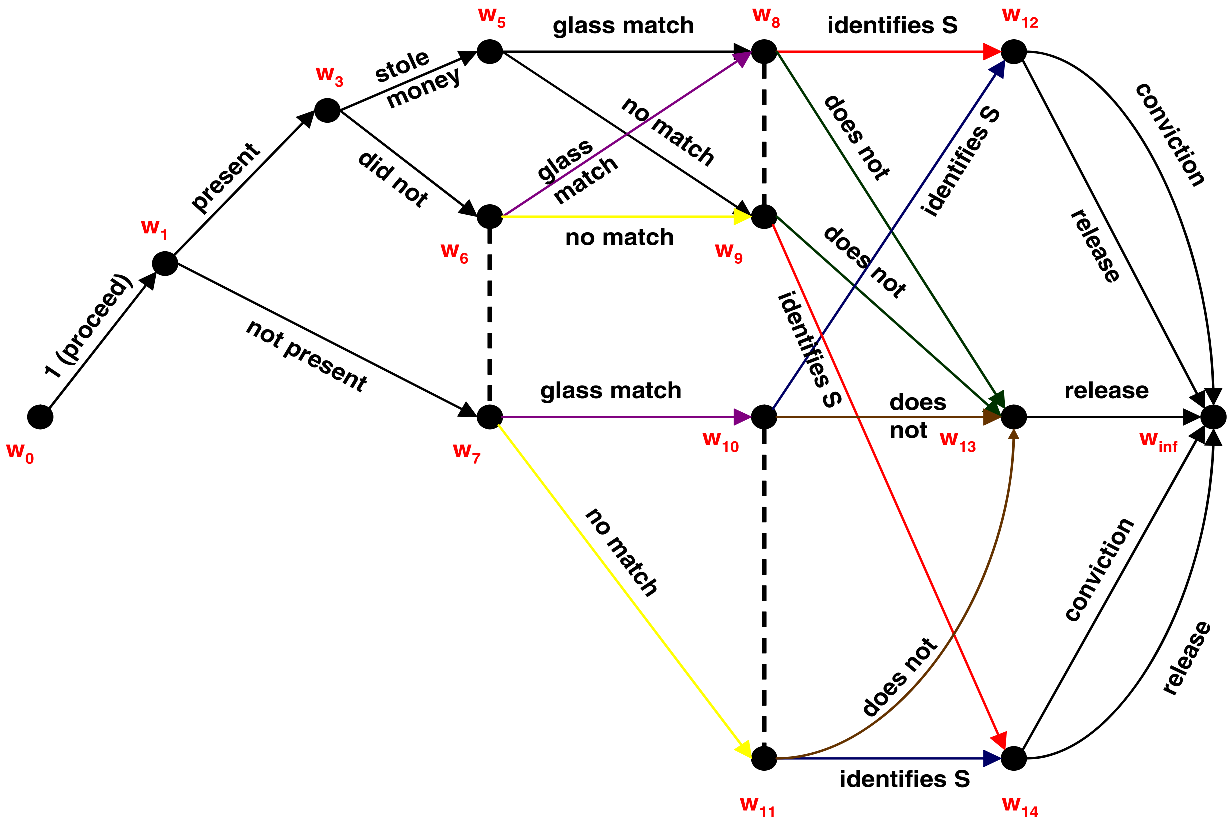


Fig. 3. Manipulated CEG $\hat{C}$ for manipulation to w_1.

conditional independence properties associated with the edges emanating from positions that are parents of positions in W are lost.

This has the useful consequence that for staged manipulations of a CEG we can simply replace $(T, \Pi(T))$ by $(C, \Pi(C))$ in Definition 3; $D \subset S(T)$ by $D \subset L(C)$; $\hat{\Pi}_D \equiv \{\hat{\pi}(v' \mid v)\colon\ v \in D,\ v' \in \mathbb{X}(v)\}$ by $\hat{\Pi}_D \equiv \{\hat{\pi}(e(w, w') \mid w)\colon\ w \in u,\ u \in D\}$, where $\hat{\pi}(e(w, w') \mid w)$ is a new distribution of the random variable $X(u)$ for $w \in u$.

Our focus in this paper is on those manipulations which have analogues for BNs. We start our description with simple interventions which can be characterised as forcing to a manipulation set W (or as in Example 3.1, to a single position w), where each position that is a parent of a position in W has only one child in W. We label the set of parents of positions in W by $pa(W)$.

Example 3.1. In Example 2.1, consider the manipulation forced to w_1 (manipulation set $W = \{w_1\}, pa(W) = \{w_0\}$), which corresponds to ensuring that the suspect goes to court.

This assigns a probability of 1 to the edge $e(w_0, w_1)$, and all vertices and edges not lying on a $w_0 \to w_1 \to w_\infty$ path are deleted. The probabilities on all edges in our manipulated CEG $\hat{C}$ are identical to the corresponding edge-probabilities in C except the probability on the edge $e(w_0, w_1)$. Our manipulated CEG $\hat{C}$ is given in Fig. 3. As all probabilities after the manipulation remain unchanged, we have *stages* as marked.

We assume that Fig. 3 shows a CEG which is valid for our manipulation. However this assumption is a substantive one. If a judge is available, sufficient money has been stolen and so on, then the police, believing S to be guilty, will make a decision to proceed. In this case our manipulated CEG is almost certainly valid. However suppose the police obtain CCTV footage showing S to be present. Then the police will again make a decision to proceed (ensuring that there is a judge available, and ignoring police-force policy if necessary). This can also be interpreted as a manipulation to w_1, but in this case edge-probabilities downstream of the manipulation may well change — the presence of S on CCTV footage may increase the probability of the witness identifying S for example. This manipulation may also alter the topology of the manipulated CEG — the witness failing to identify S may no longer result automatically in an acquittal.

The alternative manipulation, forced to w_{13}, can be interpreted as a contingent manipulation — if the police proceed, the witness is forced **not** to identify the suspect. A CEG for this intervention is given in Fig. 4.

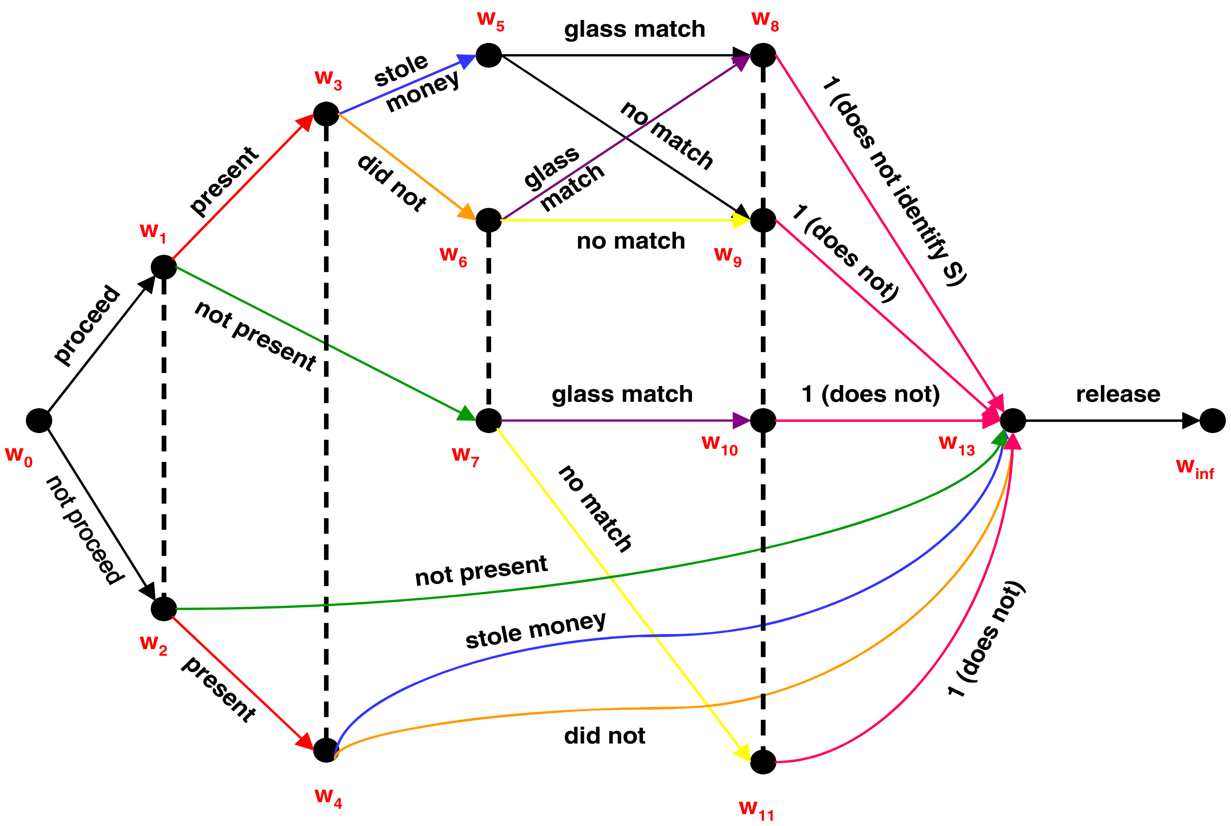


Fig. 4. Manipulated CEG $\hat{C}$ for manipulation to w_{13}.

As the manipulation definition uses the phrase *if the police proceed*, there is no reason here for altering the probabilities on the $e(w_2, w_{13})$ and $e(w_4, w_{13})$ edges, and so the stage structure is as in Fig. 4. Note that this manipulation might be enacted by an *outside* manipulator, such as the suspect's brother!

The manipulation forcing to $\{w_{12}, w_{14}\}$ is considered in Section 5.

Example 3.2. A university has residence blocks of apartments, with two rooms each. It allocates second year students, either English ($X_1 = 0$) or Chinese ($X_1 = 1$), to one of the two rooms in each apartment. The second room is allocated to a first year student, either English ($X_2 = 0$) or Chinese ($X_2 = 1$), and this is done at random. A survey has recorded that the probability of a high satisfaction rating for students placed with another student of the same ethnicity is higher than for students placed with another student of different ethnicity.

Recording student satisfaction via a binary indicator Y, we can draw a CEG for this problem as in Fig. 5. As with Fig. 1, for simplicity only a subset of the edges have been coloured.

The undirected edge between w_1 and w_2 indicates that these positions are in the same stage and hence $X_2 \amalg X_1$, reflecting the random allocation of first year students to apartments. Because w_1 and w_2 are not combined into a single position we can read that $Y \not\amalg X_1$. We can also read the positions w_3 and w_4 to give $Y \amalg (X_1, X_2) \mid X_1 = X_2$ and $Y \amalg (X_1, X_2) \mid X_1 \neq X_2$. These expressions can be combined into the single statement $Y \amalg (X_1, X_2) \mid |X_1 - X_2|$. Since $X_1 \amalg X_2$ and Y depends on both X_1 and X_2, the obvious BN-representation of the problem is as in Fig. 6(a). The BN would need to be supplemented by the extra context-specific information, and if required the information that second year students are allocated before first years.

If we consider the intervention wherein the university places first year students with second years of the same ethnicity, then this would be represented on the CEG in Fig. 5 as a manipulation to the position w_3. Note that this manipulation would cause the removal of the undirected edge between w_1 and w_2 since $X_1 \not\amalg X_2 \mid (X_1 = X_2)$. Using the BN in Fig. 6(a) this would be a functional manipulation $Do\, X_2 = x_1$; or alternatively we could redefine our variable X_2 so that it had outcome space $\{0, 1\}$ corresponding to $\{$*first year student has same ethnicity as second year student*, *first year student has different ethnicity from second year student*$\}$. This would give us the BN as in Fig. 6(b), and our manipulation would correspond to forcing X_2 to the value 0, with the deletion of the arc from X_1 to X_2. However, as with the BN in Fig. 6(a), this BN would

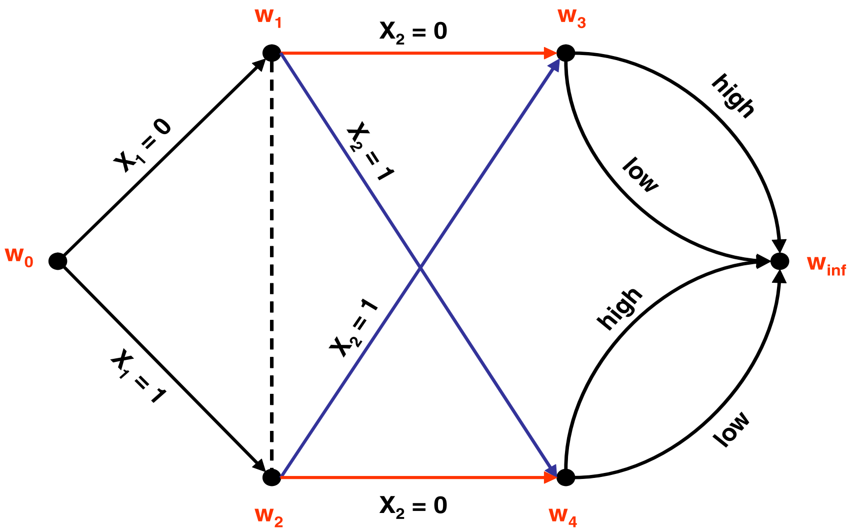


Fig. 5. CEG for Example 3.2.

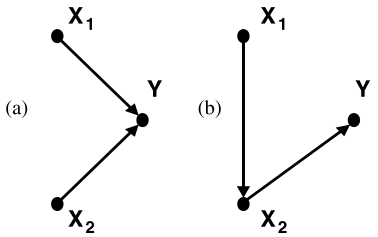


Fig. 6. Possible BNs for Example 3.2.

have to be supplemented by extra information (here the fact that first year students are allocated at random) in order to fully describe the idle system.

The topology of the CEG here fully represents the idle system, and also allows us to both express our manipulation and analyse its effects.

3.2. Interventions on CEGs and BNs

Consider again the BN from Fig. 6(a) and the CEG from Fig. 5. The BN is extended to a CBN by the assumption that the arrows on the edges represent causal directions and mechanisms [21]. So we could for example force X_2 to the value 1, and analyse the effect on the variable Y. In the CEG we cannot make this assumption about edges — X_1 edges immediately precede X_2 edges in Fig. 5, but X_1 is not a *cause* of X_2. But by embedding additional causal hypotheses CEGs can be given a causal interpretation in a very similar manner. Under this interpretation the CEG represents a *controlled* model where some or all of our variables are manipulable.

For a CEG to be causal for a particular manipulation we require that all edges that are to be manipulated lie upstream of any descendants of the variables labelling these edges. We also require that if the CEG were to be manipulated then the assumption that the distribution of variables downstream of the manipulation remains as in the idle system is a valid one. In Example 3.1 we briefly considered a case where this assumption might not be valid. Effectively, a CEG is deemed valid for a manipulation if the assumptions required for Definition 3 are valid.

It is therefore possible for a CEG to have a causal interpretation when not all problem variables appear before their descendants. It is however necessary that the variables to be manipulated satisfy this condition.

If the assumptions required for Definition 3 are valid for a particular CEG C, then a manipulation of this CEG is, in its most general form, the imposition of new probability distributions on the edges leaving one or more positions from $V(C)$. So for example, in Fig. 5 the manipulation $Do\,X_1 = 0$ assigns a probability of 1 to the edge $e(w_0, w_1)$, a probability of 0 to the edge $e(w_0, w_2)$, and leaves all other edge-probabilities unchanged. In practice we would prune the edge $e(w_0, w_2)$ and the edges emanating from w_2 to give a less cluttered diagram.

Clearly the assumptions required for Definition 3 may not be valid for all possible CEG orderings of a problem. We may have two CEGs which both accurately describe the idle state of a problem, but only one of which can be given a causal interpretation for a particular manipulation. Indeed the choice of ordering will be governed by how we wish to use the CEG.

This holds for other graphical representations — in decision analysis for example, a decision tree given a *causal* order will provide an accurate description of how a problem unfolds, but to perform an optimal decision analysis, one would need to have an *extensive form* decision tree.

All the standard interventions on BNs [21,34] are possible on CEGs, and correspond to manipulations of collections of positions. For example the simple *Do* intervention becomes a manipulation of collections of stages, so in models where it is reasonable to talk about manipulating a variable X, the intervention $Do\, X = x_0$ assigns a probability of 1 to all edges labelled x_0, a probability of 0 to all edges labelled x_j $(j \neq 0)$, and leaves all other edge-probabilities unchanged. In practice we prune edges with zero probability and those lying only on zero-probability paths.

Positions in a CEG store vectors of values of preceding variables, so a set of positions whose emanating edges all share the same labels can be partitioned by the values taken by a subset of the preceding variables. A *functional* manipulation $Do\, X = g(Z)$ can then be represented by assigning probabilities to the emanating edges of these positions dependent on which element of the partition the position falls into. A *stochastic Do* is represented simply by assigning a new probability distribution to **all** the edges leaving each position in a set whose members' emanating edges share the same labels.

Definition 3 allows us to look beyond these basic manipulations. So for example they can all be extended so that the manipulated variable (X) no longer corresponds to one of the original measurement variables of the problem. The stochastic *Do* can also be adapted so that for some positions corresponding to X, the distribution imposed on the outgoing edges is identical to that in the *idle* system. This leads to the case where some root-to-sink paths of the CEG have no edges manipulated, corresponding for example to treatment regimes where only patients with certain combinations of symptoms are treated.

We can also consider interventions where some root-to-sink paths are subject to more than one manipulation. Or we could modify our definition of a CEG manipulation to consider interventions which produce possible outcomes at a position which are not possible in the idle system. This would involve not just imposing a new distribution on existing edges, but the adding of extra edges and hence the production of extra paths not present in the original CEG. If we enact the intervention *Build a dam across the valley mouth*, then the event *The village halfway up the valley side gets flooded next year*, which has zero probability in the idle system, now has a probability greater than zero [31].

We have described here how the manipulations of BNs have their counterparts on CEGs. In Section 4 we return to the more general class of manipulations possible with CEGs. The interventions described above can be thought of as special cases of these generic types.

4. Identifying the effects of manipulations

Much of the causal BN literature [6,21,20] studies when the effects of a manipulation on a pre-specified random variable Y can be identified from observing a subset of the BN's variables that are observed or *manifest* in the *idle* system. Necessary and sufficient conditions for causal identifiability (expressed as functions of the topology of the idle BN) have now been proved for most scenarios [22,35,7,34]. These results allow us to use probabilities from the idle system in order to estimate effects on the manipulated system, for example the effects of a proposed new treatment regime.

The topology of the CEG can also be used for this purpose. Pearl [21] states that the causal effect of X on Y is *identifiable* from a graph G if the quantity $p(y \parallel \hat{x})$ can be computed uniquely from any positive probability of the observed variables. We can generalise this for the CEG and state that the causal effect on a variable Y is *identifiable* from a CEG C if the probability of the event $Y = y$ in the manipulated CEG $\hat{C}$ can be expressed solely in terms of observable probabilities from the idle system. We can use the topology of the CEG to find *functions* of the data (not just subsets of possible measurements) that when observed in the idle system allow us to estimate the effect of a given manipulation of a causal CEG. As in [21] we prove several sufficient conditions for identifiability, and generalise Pearl's Back Door Theorem to CEG models. We first provide some notation and a couple of definitions.

We use λ to indicate a root-to-sink $(w_0 \to w_\infty)$ path of our CEG. Each λ is an atom of the path σ-algebra of the CEG, and the set of atoms is denoted Ω. A subpath of a root-to-sink path is denoted μ or $\mu(w_1, w_2)$, where w_1 and w_2 indicate the start and end positions of the subpath.

A union of atoms constitutes an event, denoted Λ. M is used to indicate a union of subpaths, so for example $M(w_1, w_2)$ is the union of all subpaths from w_1 to w_2. Let $\Lambda(w)$ denote the event which is the union of all paths passing through the position w, and $\Lambda(e)$ the union of all paths passing through the edge e. $\Lambda(\mu(w_1, w_2))$ is the event which is the union of all paths utilising the subpath $\mu(w_1, w_2)$.

We use $\pi(w) \equiv \pi(\Lambda(w))$ to denote the probability of passing through the position w. Note that this is also the probability of reaching w from w_0. The probability of reaching w_2 from w_1 is denoted by $\pi(\Lambda(w_2) \mid \Lambda(w_1))$ or more simply $\pi(w_2 \mid w_1)$. Similarly $\pi_\mu(w_2 \mid w_1) \equiv \pi(\Lambda(\mu(w_1, w_2)) \mid \Lambda(w_1))$ is the probability of utilising the subpath $\mu(w_1, w_2)$ given that a unit has reached w_1 — this can be thought of as the probability **of** the subpath $\mu(w_1, w_2)$. Let $\pi_e(w_2 \mid w_1)$ denote the probability of passing from a position w_1 to an adjacent position w_2 along the edge $e(w_1, w_2)$.

Finally we let $Y : \Omega \to \mathbb{R}$ be a random variable measurable with respect to the path σ-algebra of the CEG; and let $\{\Lambda_y\}$ be the partition of Ω generated by Y — namely each Λ_y is the union of those $\lambda \in \Omega$ for which $Y = y$.

Definition 5. For a CEG $(C, \Pi(C))$, and a manipulation of this CEG yielding a manipulated CEG $(\hat{C}, \hat{\Pi}(\hat{C}))$, the manipulation is *forced to the position* w if:

1. $\hat{\pi}(\Lambda(w)) = 1$,
2. under $\hat{\Pi}(\hat{C})$ all primitive probabilities associated with edges downstream of w are those of the idle system.

A manipulation which forces to a position w is one which ensures that at a specified point in a process, all units have the same probability of following each of a set of possible future developments. This is done by arranging that each unit has the same vector of values for a **subset** of the preceding variables (characterised by $\Lambda(w)$). An example of this would be a company preparing employees for possible promotion by ensuring that they had each attended certain training courses or passed certain professional examinations. In Example 3.1, both our manipulations are manipulations forced to a position.

We now consider an effect random variable $\hat{Y}$ defined on the path σ-algebra of $\hat{C}$. $\hat{Y}$ generates a partition of the root-to-sink paths of $\hat{C}$ with each outcome corresponding to a union of $w_0 \to w_\infty$ paths. There then exists a variable Y defined on C such that any path in C which belongs to the event $Y = y$ and which passes through w, has an equivalent path in $\hat{C}$ which belongs to the event $\hat{Y} = y$. Without ambiguity we can denote the union of $w_0 \to w_\infty$ paths in $\hat{C}$ belonging to the event $\hat{Y} = y$ by Λ_y.

Now each $w_0 \to w_\infty$ path in $\hat{C}$ is a conjunction of a $w_0 \to w$ subpath with a $w \to w_\infty$ subpath. We denote these subpaths by $\{\mu(w_0, w)\}$ and $\{\mu(w, w_\infty)\}$ and let the union of **all** $w_0 \to w$ subpaths be $M(w_0, w)$.

The random variable $\hat{Y}$ measures an effect after a manipulation forced to w. So heuristically $\hat{Y}$ needs to be realised after w, i.e. be associated with events downstream of w. Formally we therefore require that our partition $\{\Lambda_y\}$ of $\hat{C}$ consists of events each of which is $M(w_0, w)$ conjoined to a union of subpaths from $\{\mu(w, w_\infty)\}$ – for outcome $\hat{Y} = y$, call this union $M_y(w, w_\infty)$.

Suppose briefly that we are considering a problem which admits a natural product space structure (and could therefore be depicted by a BN). We can then construct a CEG of the problem where all edges can be labelled with the outcomes of the problem variables (although we may sometimes choose to construct our CEG so that these variables are encountered in different orders on different root-to-sink paths). In this case we might well label a subset of edges with for example the outcome y_0. The event $Y = y_0$ would then be the union of all $w_0 \to w_\infty$ paths in C passing through one of these edges. In the manipulated CEG $\hat{C}$ many of these edges will disappear. However those that are left will still be labelled y_0, and the event $\hat{Y} = y_0$ will be the union of all $w_0 \to w_\infty$ paths in $\hat{C}$ passing through one of these edges.

Lemma 1. *Providing that the probability of passing through the position w in the idle system is greater than zero, then for all levels y, under a manipulation forced to w*

$$\hat{\pi}(\hat{Y} = y) = \pi(Y = y \mid w)$$

Proof. The proof of this lemma is presented in Appendix A. □

One consequence of this lemma is that for a manipulation forced to w it may be possible to observe indicators on the events $\{\Lambda_y \cap \Lambda(w)\}$ in the unmanipulated system and to identify the effect on Y of the manipulation, using this expression. However it is not always possible to observe these indicators, even in models that can be described by a CBN. Suppose instead that we can observe indicators for a set of coarser events. We show below that being able to observe indicators on the events $\{\Lambda_y \cap \Lambda(W)\}$ (where W is some **set** of positions) can also be sufficient for identifiability.

Definition 6. A set of positions $W \subset V(C)$ is called *C-regular* (or simply *regular*) if

(i) no two positions in W lie on the same directed path of C, and if
(ii) no two positions in W share a parent in $V(C)$.

For any regular set of positions W, the collection of edges lying on $w_0 \to W$ subpaths can be partitioned into *defining*, *refining* and *passive* edges as follows:

1. the *defining* edges of W are those edges lying on $w_0 \to W$ subpaths which emanate from positions not all of whose outgoing edges lie on $w_0 \to W$ subpaths,
2. the *refining* edges of W are those edges lying on $w_0 \to W$ subpaths which emanate from positions all of whose outgoing edges lie on $w_0 \to W$ subpaths, but not all of whose outgoing edges lie on a $w_0 \to w$ subpath, for any individual $w \in W$,
3. the *passive* edges of W are those edges lying on $w_0 \to W$ subpaths which are neither defining nor refining edges.

Definition 7. For a CEG $(C, \Pi(C))$, and a manipulation of this CEG yielding a manipulated CEG $(\hat{C}, \hat{\Pi}(\hat{C}))$, the manipulation is *forced to the C-regular set W* if:

1. $\sum_{w \in W} \hat{\pi}(\Lambda(w)) = 1$,
2. under $\hat{\Pi}(\hat{C})$ all primitive probabilities associated with edges downstream of any $w \in W$ are those of the idle system.

As noted in Section 3.1, the simple $Do\, X = x$ and functional $Do\, X = g(Z)$ interventions on BNs can be represented on a CEG as manipulations to a regular set of positions.

We now construct an effect random variable associated with a manipulation forced to a C-regular set W. So consider a random variable $\hat{Y}$ defined on the path σ-algebra of $\hat{C}$. Each outcome y of $\hat{Y}$ corresponds to a union of $w_0 \to w_\infty$ paths in $\hat{C}$ (Λ_y), and we wish $\hat{Y}$ to be *downstream* of W. As before, there exists a corresponding variable Y defined on C.

For a position $w \in W$ and outcome y, we can specify an event $M(w_0, w) \times M_y(w, w_\infty)$ provided that the set $\{\mu_y(w, w_\infty)\}$ is not empty. We then define our event $\hat{Y} = y$ (or Λ_y) as the union over all $w \in W$ of the events $\{M(w_0, w) \times M_y(w, w_\infty)\}$.

We wish to be able to state conditions for the effect of a manipulation forced to a C-regular set of positions W being determinable directly from probabilities in the idle system. We do this through the idea of an *amenable* manipulation.

Definition 8. A regular set of positions W is *simple* if:

1. all defining edges of W emanate from positions which have only one outgoing edge lying on a $w_0 \to W$ subpath,
2. all refining edges of W emanate from positions which have only one outgoing edge lying on a $w_0 \to w$ subpath for each $w \in W$,
3. for any $w \in W$, the refining edges on $w_0 \to w$ subpaths are independent of the defining and passive edges on these subpaths in the sense that
   (a) for each $w_1 \in W$ and all $w_2 \in W \setminus w_1$, and for any $\mu(w_0, w_1)$ subpath, there must exist a $\mu(w_0, w_2)$ subpath which differs in colour from the $\mu(w_0, w_1)$ subpath only on refining edges,
   (b) for each $w \in W$ the colouring of the refining edges is the same for each $\mu(w_0, w)$ subpath.

An immediate consequence of Definition 8 is that for each $w \in W$, we can write

$$\pi\big(\Lambda(w)\big) = \pi_w^R \pi\big(\Lambda(W)\big)$$

where π_w^R is the product of probabilities on the refining edges of W lying on the $w_0 \to w$ subpaths. A derivation of this result is given in Appendix A.

Note that in direct analogy with results on causal identifiability in BNs, the conditions of this definition can all be checked with reference to the **topology** of the CEG. Condition 3(b) needs a slight modification if we have chosen to construct our CEG so that different paths pass through the problem variables in different orders.

Suppose briefly that a particular problem is regular enough to admit a natural product space structure, and the edges of our CEG have been labelled with the outcomes of the problem variables. It is then possible to define *passive*, *refining* and *defining* variables. The values labelling the defining edges of W correspond to the state of a vector of defining variables $\mathbf{D}$; the vector of refining variables $\mathbf{R}$ defines the values labelling the refining edges; whilst the vector $\mathbf{P}$ defines the values labelling the passive edges. In this situation we can express condition 3 as $\mathbf{R} \perp\!\!\!\perp (\mathbf{P}, \mathbf{D})$.

Definition 9. A manipulation is called *amenable forcing to a set W* if:

1. the set W is simple in (C, Π),
2. the set W is simple in $(\hat{C}, \hat{\Pi})$, and $\hat{\pi}(\Lambda(W)) = 1$,
3. $\Pi(C)$ and $\hat{\Pi}(\hat{C})$ differ only on the defining edges of W.

Lemma 2. *Consider an amenable manipulation forcing to a simple set W. The distribution of $\hat{Y}$ (as defined above) is identified from the probabilities in the unmanipulated system of the events $\{Y = y, W\}$, and its probabilities are given by the equation*

$$\hat{\pi}(\hat{Y} = y) = \frac{\pi(Y = y, W)}{\pi(W)}$$

where $\pi(W) \equiv \sum_{w \in W} \pi(\Lambda(w))$, and provided that $\pi(\Lambda(w)) > 0\ \forall w \in W$.

Proof. The proof of this lemma is presented in Appendix A. □

Example 4.1. Consider the binary BN and corresponding CEG in Fig. 7. The manipulation to the set $W = \{w_7, w_9\}$ (equivalent to the Pearl manipulation $Do\, X = x_0$) is amenable and satisfies Lemma 2.

For the CEG in Fig. 7 we have

$$\pi\big(\Lambda(w_7)\big) = \pi(c_0) \sum_d \pi(d)\pi(x_0 \mid d) = \pi(c_0)\pi(x_0)$$

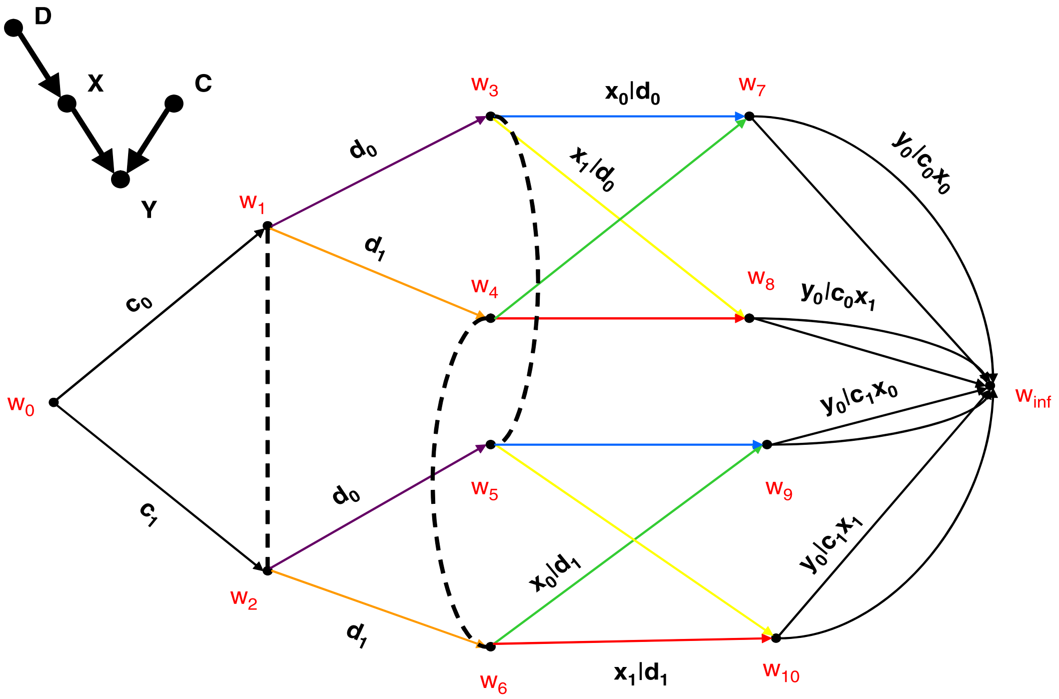


Fig. 7. BN and CEG for Example 4.1.

and similarly for $\pi(\Lambda(w_9))$. W here is simple — the defining edges are those labelled x_0, the refining edges are labelled c_0 and c_1, and the passive edges are those labelled d_0 and d_1. $\pi(\Lambda(W)) = \pi(x_0)$, and $\pi^R_{w_7} = \pi(c_0)$.

The CEG $\hat{C}$ for this manipulation differs from C in that w_3, w_4, w_5 and w_6 are now **all** in the same stage; the edges terminating at w_7 or w_9 all have probability 1; and all edges either terminating at or emanating from w_8 or w_{10} are pruned. Hence

$$\hat{\pi}\big(\Lambda(w_7)\big) = \pi(c_0)\sum_d \pi(d) \times 1 = \pi(c_0) \times 1$$

and similarly for $\hat{\pi}(\Lambda(w_9))$. So W is simple in $\hat{C}$. $\hat{\Pi}$ differs only on the edges labelled x_0, i.e. the defining edges; so this manipulation is *amenable*.

Letting Λ_y be the event $Y = y_0$, we have

$$\hat{\pi}(\Lambda_y) \equiv \hat{\pi}(\hat{Y} = y_0) = \sum_c \pi(c)\sum_d \pi(d) \times 1 \times \pi(y_0 \mid c, x_0) = \sum_c \pi(c)\pi(y_0 \mid c, x_0)$$

From Fig. 7 we have that

$$\pi\big(\Lambda_y \bigm| \Lambda(W)\big) \equiv \pi(Y = y_0 \mid x_0) = \frac{\sum_c \pi(c)\sum_d \pi(d)\pi(x_0 \mid d)\pi(y_0 \mid c, x_0)}{\sum_c \pi(c)\sum_d \pi(d)\pi(x_0 \mid d)}$$
$$= \sum_c \pi(c)\pi(y_0 \mid c, x_0) = \hat{\pi}(Y = y_0)$$

Suppose that our idle CEG can (as in Example 4.1) be represented as a BN, so that in particular defining, refining and passive variables can be defined. In a CBN, the effect of a manipulation of a variable X on a *later* variable Y can be identified from observing the distribution of the unmanipulated pair (X, Y) if and only if the vector of unobserved (hidden) variables $\mathbf{H}$ in the system can be partitioned as $\mathbf{H} = (\mathbf{H}_1, \mathbf{H}_2)$, where

$$\mathbf{H}_2 \amalg (\mathbf{H}_1, X) \quad \text{and} \quad (Y, \mathbf{H}_2) \amalg \mathbf{H}_1 \mid X$$

Returning to the CEG, we learnt in Section 2.2 that for any position w we can write $Y(w) \amalg Z(w) \mid \Lambda(w)$ — i.e. any variable defined downstream of w is independent of variables defined upstream of w conditioned on the event $\Lambda(w)$. Now for $w \in W$ (a simple set), $\Lambda(w)$ can be explicitly characterised simply in terms of the labelling of the defining and refining edges on $w_0 \to w$ subpaths. So in this situation $\Lambda(w)$ can be expressed in terms of the states of the vectors $\mathbf{D}$ and $\mathbf{R}$, and we can write the collection of conditional independence properties $\{Y(w) \amalg Z(w) \mid \Lambda(w)\}_{w \in W}$ as $Y \amalg \mathbf{P} \mid (\mathbf{R}, \mathbf{D})$. Since we already know that $\mathbf{R} \amalg (\mathbf{P}, \mathbf{D})$, we can deduce that $(Y, \mathbf{R}) \amalg \mathbf{P} \mid \mathbf{D}$. Equating $\mathbf{H}_1$ above with $\mathbf{P}$, X with $\mathbf{D}$ and $\mathbf{H}_2$ with $\mathbf{R}$, we can see that the conditions on the CBN are the same as on the CEG. So Lemma 2 is an exact analogue of this well known result for causal BNs for the more general class of CEGs. Moreover, the conditions required by Lemma 2 only depend on an appropriate factorisation of probabilities associated with the manipulated set W.

Using Lauritzen's [15] terminology and the (sets of) variables $X, Y, \mathbf{H}_1, \mathbf{H}_2$, we have from expression (3.1) that

$$\pi(y \parallel x) = \sum_{h_1, h_2} \left[\frac{\pi(x, h_1, h_2, y)}{\pi(x \mid pa(x))} \right]$$

Note that $(X, \mathbf{H}_1) \amalg \mathbf{H}_2 \Rightarrow X \amalg \mathbf{H}_2 \mid \mathbf{H}_1$, so we can equate $\mathbf{PA}(X)$ with $\mathbf{H}_1$, and write

$$\begin{aligned} \pi(y \parallel x) &= \sum_{h_1, h_2} \left[\pi(x \mid h_1)\right]^{-1} \pi(x, h_1) \pi(h_2, y \mid h_1, x) \\ &= \sum_{h_1, h_2} \pi(h_1) \pi(h_2, y \mid x) = \pi(y \mid x) \end{aligned}$$

using $(Y, \mathbf{H}_2) \amalg \mathbf{H}_1 \mid X$.

Under these conditions, manipulating X to x has the same effect on Y as conditioning X to x. Note that in Example 4.1 we can clearly see that $C \amalg (D, X)$ and $(Y, C) \amalg D \mid X$, so our refining variable is C, our defining variable is X, and D is a passive variable.

5. A Back Door Theorem for Chain Event Graphs

A key component of causal analysis on BNs is Pearl's Back Door Theorem [19,21], which owes its derivation in part to the realisation that many manipulations are impossible, unethical or prohibitively expensive in practice, or may be possible to enact but some of their effects may be impossible to observe. The Back Door Theorem gives sufficient conditions for identifying the effect on a variable Y of manipulation of a variable X when we are able to observe the values taken by only a subset Z of the remaining variables in the system. If the set Z is chosen carefully then we can calculate or estimate this effect from a partially observed idle system.

In this section we produce an analogous theorem that applies a graphical and sufficient criterion to a CEG to determine whether we can identify the effect of a manipulation on a random variable Y from the observation of a random variable Z (happening before the manipulation in the partial ordering induced by the paths) in the unmanipulated system. The event-based topology of the CEG allows us to consider a wider class of idle system models, and a wider class of manipulations of these than is generally possible with a standard BN. Similarly, our random variable Z does not need to correspond to any fixed subset of the measurement variables of the problem, giving us more flexibility in our search for an appropriate probability expression.

Before proceeding to this theorem we provide some further notation and a couple of definitions.

Definition 10. For a C-regular set of positions W, the graph C_W with vertex set $V(C_W)$, directed edge set $E_d(C_W)$ and undirected edge set $E_u(C_W)$, is defined by

1. $V(C_W)$ consists of the union of $\{w_0^\bullet\}$, a new root-node, with the set of precisely those positions from $V(C)$ which lie on a $w \to w_\infty$ subpath in C, for some $w \in W$.
2. The root-node $w_0^\bullet$ is connected by an edge to each $w \in W$. $E_d(C_W)$ consists of the union of the set $\{e(w_0^\bullet, w)\}_{w \in W}$ with the set of precisely those edges from $E_d(C)$ which lie on a $w \to w_\infty$ subpath in C, for some $w \in W$.
3. Edge-colourings (i.e. edge-probabilities) on $w \to w_\infty$ subpaths of C_W (for $w \in W$) are retained from C.
4. The edge $e(w_0^\bullet, w)$ $(w \in W)$ is given the probability $\frac{\pi(\Lambda(w))}{\pi(\Lambda(W))}$.
5. If two positions in $V(C_W)$ were connected by an undirected edge in C, then they are connected in C_W. $E_u(C_W)$ is the set of undirected edges in C_W.

It is straightforward to show that C_W is a CEG.

We now let Z be a random variable observed on C, whose events $\{Z = z\}$ partition the set of $w_0 \to w_\infty$ paths of C; and consider W^1, a fine cut of C such that each event $Z = z$ is precisely the set of $w_0 \to w_\infty$ paths in C passing through a (specified) subset of positions from this cut. We can then, without ambiguity, identify each event $Z = z$ with this set of positions — say W_z^1.

Let the set of positions to which we intend to manipulate be W^2. Then for Z to occur before the manipulation we require that every position $w^2 \in W^2$ lies on a path in C between some position $w^1 \in W_z^1$ (for some level z) and w_∞. Note that our fine cut W^1 is going to take the role of Z in our Back Door Theorem. We therefore require that the manipulation does not change any primitive probabilities from the idle system lying on a subpath between w_0 and the positions in W^1.

To ensure this we need to stipulate that for each $w^1 \in W^1$, there must exist a $w_0 \to w^1 \to w^2 \to w_\infty$ path for some $w^2 \in W^2$. If there existed $w^1 \in W^1$ for which there was no such w^2, then $\hat{\pi}(\Lambda(w^1))$ would equal zero, and hence would not equal $\pi(\Lambda(w^1))$. Having imposed this condition, we can ensure that the probability of $Z = z$ is the same in $\hat{C}$ as in C.

Definition 11. A set of C-regular positions $W^2 \subset V(C)$ is called *simple conditioned on* Z if

1. $W^2 \equiv \bigcup_z W_z^2$ where W_z^2 is simple in $C_{W_z^1}$.
2. There is a directed path in C from each position $w_z^1 \in W_z^1$ through a position $w^2 \in W^2$, and W_z^2 is the set of precisely those positions in W^2 which lie on a $w_0 \to w_z^1 \to w_\infty$ path for some $w_z^1 \in W_z^1$.

Note that the union in item 1 *is* **not** *a disjoint union.*

Consider an amenable manipulation to a set W, and let W be simple conditioned on Z. Then Z is called a *Back Door variable* to the manipulation. Let our effect variable $\hat{Y}$ be the image of Y in the manipulated CEG.

Theorem 1. *If a set W is simple conditioned on Z (a Back Door variable), then the distribution of Y after an amenable manipulation to W is identified from the probabilities (in the idle system) of the events $\{Y = y,\ W, Z = z\}$, and its probabilities are given by:*

$$\hat{\pi}(\hat{Y} = y) = \sum_z \frac{\pi(Y = y, W, Z = z)}{\pi(W, Z = z)} \pi(Z = z)$$

Proof. The proof of this theorem is presented in Appendix A. □

Example 5.1. In Example 3.1 we considered a manipulation of the CEG in Fig. 1 to the position w_{13}, where if the police proceeded the witness was forced not to identify S. Consider now the manipulation wherein the witness is forced to identify S. This is a manipulation forced to $W = \{w_{12}, w_{14}\}$. The manipulated CEG is given in Fig. 8.

Whereas the previous manipulation might have been enacted by an *outside* manipulator, such as the suspect's brother, this intervention is likely to have been enacted by someone within the police force, probably acting in an unethical manner. They would wish to have a good idea of the effects (on the indicator X_6) of this manipulation, but as a consequence of the improper nature of their intervention, they might not have any means of obtaining reliable estimates of certain necessary joint distributions of the problem variables. In particular, they would probably have to treat X_4 (indicating whether or not the forensic service will find a glass match) as an unobservable variable. Can we produce a manipulated probability expression which does not depend on X_4?

Consider the BNs in Fig. 2 and the use of Pearl's Back Door Theorem [19,21]. For BN (1) we could use a Back Door blocking set consisting of X_1 with X_2 and/or X_3, whereas for BNs (2) and (3), any blocking set **must** include X_4. So only BN (1) is of use to us here. If this BN is supplemented with the relevant context-specific information then we will be able to produce an identifiable expression. But as noted in Section 2.3, none of the BNs in Fig. 2 are well-defined — none of them encodes the full set of conditional independence properties of the problem, whereas the problem has an unambiguous representation as a CEG, and the CEG does not need to be supplemented with extra information. Moreover, we can use Theorem 1 to deduce the effect on X_6 of our manipulation, and produce an expression which is not dependent on X_4.

Notice that the probabilities in $\hat{C}$ differ only on the defining edges of W (a C-regular set), i.e. on those edges labelled *proceed* and *identifies S*. At first sight the manipulation does not seem to satisfy the conditions for Theorem 1 as the manipulated edge *proceed* must necessarily be upstream of any possible Back Door blocking set we propose. But all $w_0 \to W$ paths pass through w_1, so $\hat{\pi}(conviction) = \hat{\pi}(conviction \mid \Lambda(w_1))$. This is simply the manipulated probability of *conviction* in the CEG C_{w_1} (see Definition 10). We can therefore simply consider this CEG C_{w_1} and apply Theorem 1 to it. W is C_{w_1}-regular and the defining, refining and passive edges of W in C_{w_1} are precisely the defining, refining and passive edges of W in C which lie downstream of the position w_1, so checking the conditions of Theorem 1 can be done on the original graph in Fig. 1.

The ideas here can be used to allow us to apply Theorem 1 in many situations where, at first sight, it is apparently not appropriate.

Note that our set W^1 need no longer be a fine cut of C, but just one of C_{w_1}, i.e. a partition of the root-to-sink paths of $\hat{C}$. So consider the set $W^1 = \{w_5, w_6, w_7\}$, upstream of the set $W \equiv W^2 = \{w_{12}, w_{14}\}$. Assign the value $z = 1$ to paths passing through w_5 (*stole money*); $z = 2$ to paths passing through w_6 (*present, did not steal money*); $z = 3$ to paths passing through w_7 (*not present*). Then $W_{z=1}^1 = \{w_5\}$, $W_{z=2}^1 = \{w_6\}$, $W_{z=3}^1 = \{w_7\}$.

If we let $W_{z=1}^2 = W_{z=2}^2 = W_{z=3}^3 = W^2$, then by construction, if W^2 is simple in $C_{W_z^1}$ for each W_z^1, the conditions of Definition 11 are satisfied and W^2 is simple conditioned on Z (in C_{w_1}). We do not need to produce separate graphs for each $C_{W_z^1}$ — we can do all our checking on C in Fig. 1. For each $C_{W_z^1}$ the defining edges of W^2 are those labelled *identifies S* and the refining edges are those labelled *glass match* and *no match*. These edges obey the conditions of Definition 8, so W^2

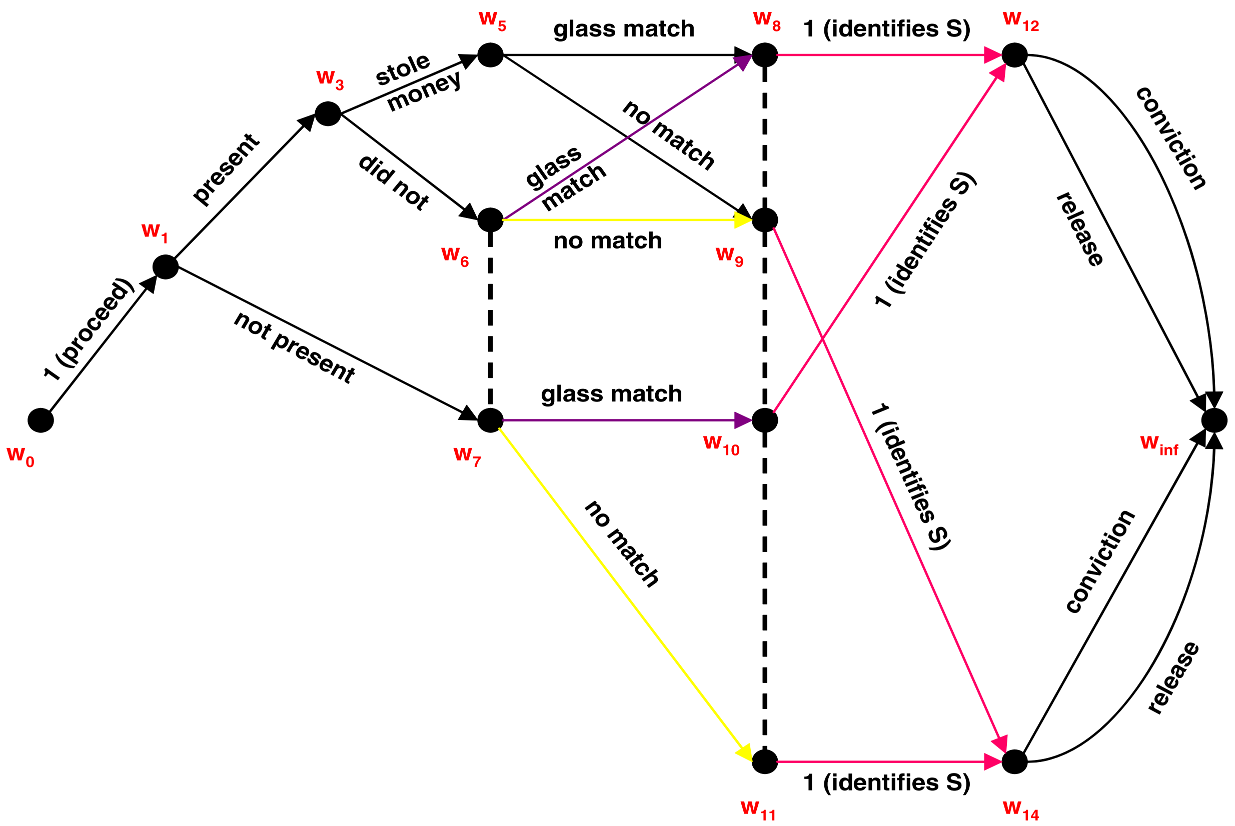


Fig. 8. Manipulated CEG $\hat{C}$ for Example 5.1.

is simple in each $C_{W_z^1}$ and hence simple conditioned on Z in C_{w_1}. Our variable Z, manipulation set $W \equiv W^2$ and effect variable Y therefore satisfy the conditions for Theorem 1 with the X_6 outcome *conviction* equating to $Y = y$.

Hence in the CEG C_{w_1}

$$\hat{\pi}(conviction) = \sum_{z=1}^{3} \frac{\pi(conviction, W, z)}{\pi(W, z)} \pi(z) = \sum_{z=1}^{3} \pi(conviction \mid identifies\ S, z)\pi(z)$$

since the set W corresponds to the event that X_5 takes the outcome *identifies S*.

Note that all primitive probabilities in C_{w_1} are identical to those in C except the probability on the edge $e(w_0, w_1)$, so the probability of z in C_{w_1} is the probability of z in C divided by $\pi(proceed)$. Hence in C our manipulated probability expression is

$$\hat{\pi}(conviction) = \frac{1}{\pi(proceed)} \sum_{z=1}^{3} \pi(conviction \mid identifies\ S, z)\pi(z)$$

It is not difficult to check that this formula correctly expresses the causal effect on X_6 of our manipulation. Moreover, as our three positions w_5, w_6, w_7 can be characterised by values of X_2 and X_3, this expression does not require knowledge of the distribution of X_4 or of joint distributions including X_4.

As with BNs, we conjecture that it will be possible to devise simple automated methods for determining whether there exist variables Z satisfying the conditions for Theorem 1, and procedures for choosing between candidate variables Z. These methods are as yet not fully developed.

If we believe such a variable does exist, then Example 5.1 shows us that the choosing of the positions within our partition can be straightforward. We construct the set $\{W_z^1\}$ to satisfy the conditions of Theorem 1, but also to minimise the amount of work involved. The events $\{Z = z\}$ must be observable or manifest within the system, but it is not necessary that these events are actually observed — in Example 5.1 the police do not observe whether the suspect was at the scene or threw the brick, but they do have what they consider to be reliable estimates for these probabilities. How we assign z values to the positions depends on the information we have available, so for example we could here assign the value $z = 2$ to both w_6 and w_7, and let $W_{z=2}^1 = \{w_6, w_7\}$. W^2 is still simple in this new $C_{W_{z=2}^1}$, so this new partition also satisfies the conditions

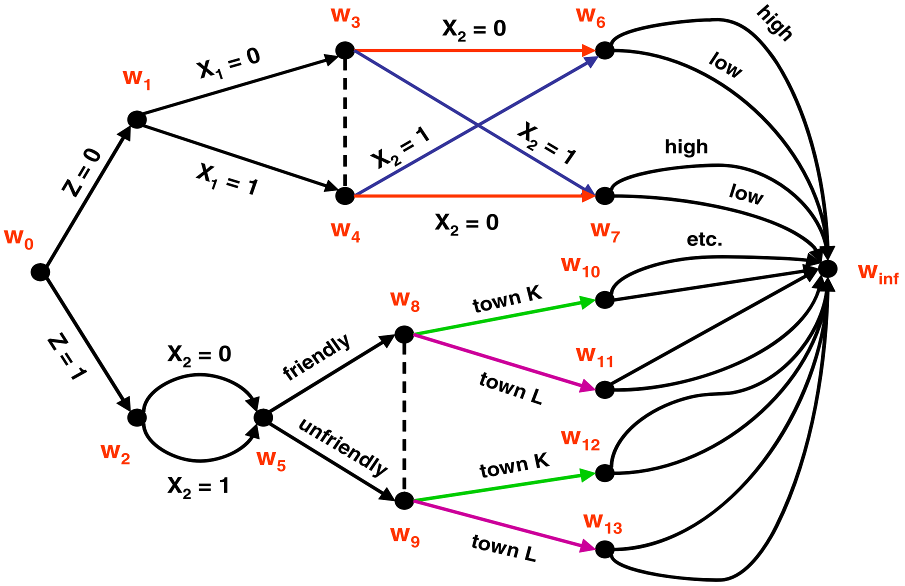


Fig. 9. CEG for Example 5.2.

for Theorem 1. We therefore have a choice of coarser or finer partitions, and which we choose will depend on the structure of the information we hold on the conditional distributions of X_6. If we are able to choose either then we should go for the coarser partition, as the resulting expression for $\hat{\pi}(conviction)$ is simpler.

The updating of the edge-probabilities of our CEG following a manipulation can already be done rapidly and automatically, using algorithms analogous to those described and coded for updating edge-probabilities following an observation in [33] and [32].

Example 5.2. First year students at the university in Example 3.2 who made the university first choice on their application ($Z = 0$) are allocated a shared apartment on campus, whilst first year students who did not ($Z = 1$) are lodged in either town K or in town L (indicator X_3). Students lodged in towns K and L may have a friendly landlord or an unfriendly landlord (indicator U), and the friendliness of these landlords is not known to the university.

When $Z = 0$ it is believed that the CEG in Fig. 5 is valid (where here Y is explicitly the satisfaction expressed by the first year student). If $Z = 1$ the town in which the student is lodged is chosen independently of the ethnicity X_2 of the student; the friendliness of the landlord does not depend on either the town or the ethnicity of the student; but the satisfaction rating Y expressed by the first year student depends both on the friendliness of the landlord and the allocated town. The problem can be represented by the CEG in Fig. 9.

We wish to consider a proposed manipulation of the allocation policy for next year. The university plans to match campus-based students so that those sharing an apartment are of the same ethnicity, and to allocate off-campus students only to lodgings in town L. Our interest is in $\hat{\pi}(high)$ — the overall predicted probability of high satisfaction were this policy to be implemented. The university intends to estimate this probability with a small data set, collected from earlier years.

The manipulation proposed is different for different contingencies, but this is irrelevant when analysing with a CEG. It can be considered as a manipulation to $W = \{w_6, w_{11}, w_{13}\}$. If we consider the partition $\{W_z^1\} = \{\{w_1\}, \{w_2\}\}$, it is straightforward to check that our variable Z, manipulation set W and effect variable Y satisfy the conditions for Theorem 1, and hence

$$\begin{aligned}\hat{\pi}(high) &= \sum_{z=0}^{1} \pi(high \mid z, W)\pi(z)\\ &= \pi(high \mid Z = 0, X_1 = X_2)\pi(Z = 0) + \pi(Y = 1 \mid Z = 1, town\ L)\pi(Z = 1)\end{aligned}$$

So $\hat{\pi}(high)$ can be expressed as a function of three probabilities from the idle system — that a student resides on campus; that a campus-based student sharing with someone of the same ethnicity gives a high satisfaction rating; and that a student lodging in town L gives a high satisfaction rating. It follows that the probabilities associated with the ethnicity of matched pairs of campus-based students; the satisfaction ratings of unmatched pairs of campus-based students; the ethnicity of non-campus-based students; the friendliness of the landlords of non-campus-based students are all irrelevant to this calculation, and need not be estimated.

Note also that $\hat{\pi}(high)$ is a function of the event $X_1 = X_2$, or alternatively of the variable $|X_1 - X_2|$. This is not one of the original measurement variables of the problem, but appears naturally in the CEG of Fig. 9, where its outcomes correspond to the positions w_6 and w_7.

To summarise — by examining the topology and colouring of an (idle) CEG, it is possible to determine sufficient conditions for whether an effect of a causal manipulation can be identified from a partial set of observations of the system. The CEG is ideally suited to the causal analysis of models which are highly asymmetric. Also, the search for an appropriate random variable Z, whose observation ensures identifiability, is not restricted to subvectors of the original (non-descendant) measurement vectors; we can search over all functions of such measurements. Searching over these functions to find the cheapest way of identifying the quantity of interest will often be of much greater value than simply searching over subsets of measurements. This will be particularly useful if those measurements have not yet been collected, or their parameterisations have been chosen by convention rather than because they reflect some natural description of how a process unfolds.

If an intervention can be identified via a CEG, then it may well be the case that by imposing a product space on the problem we will be able to express it as a BN and find an identifiable expression for the effects of the intervention via this BN. However, this identification will probably require supplementary context-specific information which is not present in the DAG of the BN, but which appears naturally in the CEG-representation of the problem.

6. Discussion

We have demonstrated that the CEG provides a flexible graphical framework within which to represent and analyse a wide variety of causal hypotheses, even in highly asymmetrical domains. There is of course a cost for this flexibility in that CEGs have, in general, more vertices and edges than BNs. For more symmetric problems this favours BN-based procedures, but as problems become less symmetric we have found that CEGs become more efficient both in model-storage space and in the algorithms used for updating probabilities [33]. An analysis of the comparative complexity of CBNs and CEGs is to be the focus of a future paper.

Of course the Back Door Theorem presented in this paper is not the only topological criterion for determining causal extensions; for example it is possible to produce and prove analogues of Pearl's Front Door Theorem (see [31]). In [9] we have shown that CEGs admit conjugate learning and model selection. Currently under investigation are extensions to learning CEGs when underlying experiments can be causally manipulated (similar in approach to [14]) — these also often admit a conjugate analysis. Despite their more complex topology, causal CEGs, being more general and expressive than CBNs, provide a useful complementary technology.

As with the BN, there are limits to the expressiveness of the CEG, and sometimes issues such as whether a cause can be identified can only be addressed algebraically (see [24]). None-the-less, the popularity of the BN has demonstrated the appeal of graphical-based causal inference, as well as how useful such inference can be. CEGs provide a powerful additional graphical tool for the investigation of causal structures which are not easily or fully expressible as CBNs.

Acknowledgements

This research is being supported by the EPSRC, grant no. EP/F036752/1. We would like to thank the referees for their perceptive and helpful comments which have enabled us to markedly improve the clarity of the paper.

Appendix A. Proofs (and Lemma 3)

Proof of Lemma 1. In $\hat{C}$ we can express the event $\hat{Y} = y$ as $\Lambda_y = M(w_0, w) \times M_y(w, w_\infty) = \Lambda(w) \cap \Lambda(M_y(w, w_\infty))$. Hence

$$\hat{\pi}(\hat{Y} = y) \equiv \hat{\pi}(\Lambda_y) = \hat{\pi}\big(\Lambda(w), \Lambda\big(M_y(w, w_\infty)\big)\big) = \hat{\pi}\big(\Lambda(w)\big)\hat{\pi}\big(\Lambda\big(M_y(w, w_\infty)\big) \bigm| \Lambda(w)\big)$$
$$= \hat{\pi}\big(\Lambda(w)\big)\hat{\pi}_{M_y}(w_\infty \mid w) = 1 \times \pi_{M_y}(w_\infty \mid w)$$

using Definition 5(1) and (2).

By definition of Y on C we have

$$\pi(Y = y, w) \equiv \pi\big(\Lambda_y, \Lambda(w)\big) = \pi\big(\Lambda(w)\big)\pi_{M_y}(w_\infty \mid w)$$
$$\Rightarrow \quad \hat{\pi}(\hat{Y} = y) = \big[\pi\big(\Lambda(w)\big)\big]^{-1}\pi\big(\Lambda_y, \Lambda(w)\big) = \pi\big(\Lambda_y \bigm| \Lambda(w)\big) \equiv \pi(Y = y \mid w) \qquad \Box$$

Derivation of result $\boldsymbol{\pi(\Lambda(w)) = \pi_w^R \pi(\Lambda(W))}$. Consider a single subpath $\mu(w_0, w)$ for $w \in W$. This consists of a set of passive, defining and refining edges, so we can write

$$\pi\big(\Lambda\big(\mu(w_0, w)\big)\big) = \pi_\mu^P \pi_\mu^D \pi_\mu^R$$

where π_μ^P is the product of the probabilities on the passive edges of μ (etc.).

For simple W, Definition 8(3)(b) implies that π^R_μ is constant for all $\mu(w_0, w)$. Relabel this π^R_w. Therefore

$$\pi\big(\Lambda(w)\big) = \pi^R_w \sum_{\mu\in\{\mu(w_0,w)\}} \pi^P_\mu \pi^D_\mu$$

Definition 8(3)(a) implies that for each $\mu \in \{\mu(w_0, w_1)\}$, there is a corresponding $\mu \in \{\mu(w_0, w_2)\}$ for which $\pi^P_\mu \pi^D_\mu$ takes the same value, for all $w_2 \in W \setminus \{w_1\}$, and hence that $\sum_{\mu\in\{\mu(w_0,w)\}} \pi^P_\mu \pi^D_\mu$ is constant for all $w \in W$. Relabel this as π'_W. Therefore

$$\pi\big(\Lambda(W)\big) \equiv \sum_{w\in W} \pi\big(\Lambda(w)\big) = \sum_{w\in W} \big[\pi^R_w\, \pi'_W\big] = \Big[\sum_{w\in W} \pi^R_w\Big]\pi'_W$$

By Definition 6(2) and by construction, the possible combinations of refining edges partition the set of $w_0 \to W$ subpaths, so $\sum_{w\in W} \pi^R_w = 1$. Hence $\pi'_W = \pi(\Lambda(W))$ and

$$\pi\big(\Lambda(w)\big) = \pi^R_w \pi\big(\Lambda(W)\big)$$

Proof of Lemma 2. As our manipulation is amenable, for each $w \in W$

$$\pi\big(\Lambda(w)\big) = \pi^R_w \pi\big(\Lambda(W)\big)$$

where π^R_w is the product of probabilities on the refining edges of w. So

$$\hat\pi\big(\Lambda(w)\big) = \hat\pi^R_w \hat\pi\big(\Lambda(W)\big) = \pi^R_w$$

using (i) Definition 9(3), $\Pi(C)$ and $\hat\Pi(\hat C)$ differ only on the defining edges of W (i.e. not on the refining edges of any $w \in W$), and (ii) Definition 9(2). So

$$\hat\pi\big(\Lambda(w)\big) = \big[\pi\big(\Lambda(W)\big)\big]^{-1} \pi\big(\Lambda(w)\big)$$

In $\hat C$, the event $\hat Y = y$ (or Λ_y) is equal to $\bigcup_{w\in W}[M(w_0, w) \times M_y(w, w_\infty)]$. The corresponding event in C is $(Y = y, W) \equiv \Lambda_y \cap \Lambda(W)$. So

$$\begin{aligned}
\hat\pi(\hat Y = y) \equiv \hat\pi(\Lambda_y) &= \sum_{w\in W} \hat\pi\big(\Lambda(w)\big)\hat\pi_{M_y}(w_\infty \mid w) \\
&= \sum_{w\in W} \hat\pi\big(\Lambda(w)\big)\pi_{M_y}(w_\infty \mid w) \quad \text{using Definition 7(2)} \\
&= \big[\pi\big(\Lambda(W)\big)\big]^{-1} \sum_{w\in W} \pi\big(\Lambda(w)\big)\pi_{M_y}(w_\infty \mid w) \\
&= \big[\pi\big(\Lambda(W)\big)\big]^{-1} \pi\big(\Lambda_y, \Lambda(W)\big) \\
&= \pi\big(\Lambda_y \bigm| \Lambda(W)\big) \equiv \frac{\pi(Y = y, W)}{\pi(W)} \qquad \square
\end{aligned}$$

Lemma 3. *If W^1 and W^2 are C-regular sets of positions, and W^2 is simple in the CEG C_{W^1} then $\pi(\Lambda(w^2) \mid \Lambda(W^1), \Lambda(W^2))$ can be written as the product of probabilities on the refining edges of W^2 in C.*

Proof. Let w be a position in C_{W^1} other than its root. By construction of C_{W^1}, the sub-CEG of C_{W^1} rooted in w has precisely the same topology and edge-colouring (i.e. edge-probabilities) as the sub-CEG of C rooted in w.

Suppose the edges leaving w are refining edges of W^2 in C_{W^1}. Then all edges leaving w are on $w^\bullet_0 \to w^2 \in W^2$ subpaths in C_{W^1}, and hence in C. Also, as W^2 is simple in C_{W^1}, only one edge leaving w lies on a $w^\bullet_0 \to w^2$ subpath in C_{W^1} for any individual $w^2 \in W^2$. So only one edge leaving w lies on a $w_0 \to w^2$ subpath in C for any individual $w^2 \in W^2$. Hence the edges leaving w are refining edges in C. The refining edges of W^2 in C_{W^1} are therefore precisely the refining edges of W^2 in C which lie downstream of W^1.

Let probabilities in C_{W^1} be denoted $\tilde\pi$. Then W^2 is simple in C_{W^1} implies that $\tilde\pi(\Lambda(w^2) \mid \Lambda(W^2))$ can be expressed as the product of probabilities on the refining edges of W^2 in C_{W^1}, and from above can therefore be expressed as the product of probabilities on the refining edges of W^2 in C. But

$$
\begin{aligned}
\tilde{\pi}\big(\Lambda(w^2) \mid \Lambda(W^2)\big) &= \big[\tilde{\pi}\big(\Lambda(W^2)\big)\big]^{-1}\tilde{\pi}\big(\Lambda(w^2)\big) \\
&= \Big[\sum_{w^1}\tilde{\pi}\big(\Lambda(w^1)\big)\tilde{\pi}\big(\Lambda(W^2) \mid \Lambda(w^1)\big)\Big]^{-1}\sum_{w^1}\tilde{\pi}\big(\Lambda(w^1)\big)\tilde{\pi}\big(\Lambda(w^2) \mid \Lambda(w^1)\big) \\
&= \Big[\sum_{w^1}\frac{\pi(\Lambda(w^1))}{\pi(\Lambda(W^1))}\pi\big(\Lambda(W^2) \mid \Lambda(w^1)\big)\Big]^{-1}\sum_{w^1}\frac{\pi(\Lambda(w^1))}{\pi(\Lambda(W^1))}\pi\big(\Lambda(w^2) \mid \Lambda(w^1)\big)
\end{aligned}
$$

using Definition 10(3) and (4)

$$
\begin{aligned}
&= \big[\pi\big(\Lambda(W^1), \Lambda(W^2)\big)\big]^{-1}\pi\big(\Lambda(W^1), \Lambda(w^2)\big) \\
&= \pi\big(\Lambda(w^2) \mid \Lambda(W^1), \Lambda(W^2)\big) \qquad \square
\end{aligned}
$$

Proof of Theorem 1. Let $W \equiv W^2$. W^2 is simple conditioned on Z, so we can express $W^2 \equiv \bigcup_z W_z^2$ where W_z^2 is simple in $C_{W_z^1}$. Now

$$
\begin{aligned}
&\pi\big(\Lambda(w^2) \mid \Lambda(W_z^1)\big) = \pi\big(\Lambda(w^2) \mid \Lambda(W_z^1), \Lambda(W^2)\big)\pi\big(\Lambda(W^2) \mid \Lambda(W_z^1)\big) \\
&\quad\Rightarrow\quad \hat{\pi}\big(\Lambda(w^2) \mid \Lambda(W_z^1)\big) = \hat{\pi}\big(\Lambda(w^2) \mid \Lambda(W_z^1), \Lambda(W^2)\big)\hat{\pi}\big(\Lambda(W^2) \mid \Lambda(W_z^1)\big) = \pi\big(\Lambda(w^2) \mid \Lambda(W_z^1), \Lambda(W^2)\big)
\end{aligned}
$$

using (i) Lemma 3, $\pi(\Lambda(w^2) \mid \Lambda(W_z^1), \Lambda(W^2))$ can be written as the product of probabilities on the refining edges of W^2 in C, and Definition 9(3), $\Pi(C)$ and $\hat{\Pi}(\hat{C})$ differ only on the defining edges of W^2; and (ii) the fact that $\hat{\pi}(\Lambda(W^2) \mid \Lambda(W_z^1)) = 1$. Therefore

$$
\hat{\pi}\big(\Lambda(w^2) \mid \Lambda(W_z^1)\big) = \big[\pi\big(\Lambda(W^2) \mid \Lambda(W_z^1)\big)\big]^{-1}\pi\big(\Lambda(w^2) \mid \Lambda(W_z^1)\big)
$$

Consider in $\hat{C}$ the events:

$(Z = z) \equiv \Lambda(W_z^1)$ since every $w^1 \in W_z^1$ exists in $\hat{C}$ by construction,

$(Z = z, \hat{Y} = y) \equiv \Lambda(W_z^1) \cap \Lambda(W^2) \cap \Lambda_y$ since in $\hat{C}$ all paths pass through W^2.

Analogously with Lemma 2, we can express this latter event as

$$
\bigcup_{w^1 \in W_z^1}\ \bigcup_{w^2 \in W^2}\big[M\big(w_0, w^1, w^2\big) \times M_y\big(w^2, w_\infty\big)\big]
$$

where $M(w_0, w^1, w^2)$ is the union of all $\mu(w_0, w^1, w^2)$ subpaths, and $M_y(w^2, w_\infty)$ is the union of all $\mu(w^2, w_\infty)$ subpaths consistent with $\hat{Y} = y$. So

$$
\hat{\pi}(Z = z, \hat{Y} = y) = \sum_{w^1 \in W_z^1}\sum_{w^2 \in W^2}\hat{\pi}\big(\Lambda(w^1), \Lambda(w^2)\big)\hat{\pi}_{M_y}\big(w_\infty \mid w^2\big) \quad \text{and}
$$

$$
\begin{aligned}
\hat{\pi}(\hat{Y} = y \mid Z = z) &= \hat{\pi}\big(\Lambda(W^2), \Lambda_y \mid \Lambda(W_z^1)\big) && \text{(A.1)} \\
&= \big[\hat{\pi}\big(\Lambda(W_z^1)\big)\big]^{-1}\sum_{w^1 \in W_z^1}\sum_{w^2 \in W^2}\hat{\pi}\big(\Lambda(w^1), \Lambda(w^2)\big)\hat{\pi}_{M_y}\big(w_\infty \mid w^2\big) \\
&= \sum_{w^2 \in W^2}\Big[\big[\hat{\pi}\big(\Lambda(W_z^1)\big)\big]^{-1}\sum_{w^1 \in W_z^1}\hat{\pi}\big(\Lambda(w^1), \Lambda(w^2)\big)\Big]\hat{\pi}_{M_y}\big(w_\infty \mid w^2\big) \\
&= \sum_{w^2 \in W^2}\hat{\pi}\big(\Lambda(w^2) \mid \Lambda(W_z^1)\big)\hat{\pi}_{M_y}\big(w_\infty \mid w^2\big) && \text{(A.2)} \\
&= \sum_{w^2 \in W^2}\big[\pi\big(\Lambda(W^2) \mid \Lambda(W_z^1)\big)\big]^{-1}\pi\big(\Lambda(w^2) \mid \Lambda(W_z^1)\big)\pi_{M_y}\big(w_\infty \mid w^2\big)
\end{aligned}
$$

using the above and Definition 9(3) (or Definition 7(2))

$$
= \big[\pi\big(\Lambda(W^2) \mid \Lambda(W_z^1)\big)\big]^{-1}\pi\big(\Lambda(W^2), \Lambda_y \mid \Lambda(W_z^1)\big)
$$

using the equivalence of the entities in expressions (A.1) and (A.2) and removing the *hats*, which we can do as this proof has used no aspect of the topology of $\hat{C}$ which is not also true for C.

Also, since W^1 is a fine cut, all $w_0 \to w_\infty$ paths in C pass through some $w^1 \in W^1$, and by Definition 11(2) each $w^1 \in W^1$ lies on a $w_0 \to W^2$ path. So by Definition 6(1) no edges leaving positions upstream of W^1 are defining edges of

W^2. But our manipulation is amenable to W^2 and by Definition 9(3), $\Pi(C)$ and $\hat{\Pi}(\hat{C})$ differ only on the defining edges of W^2. So the probabilities on all edges upstream of W^1 are the same in $\hat{C}$ as in C, and

$$\hat{\pi}\big(\Lambda\big(w^1\big)\big) = \pi\big(\Lambda\big(w^1\big)\big) \quad \forall w^1 \in W^1$$
$$\Rightarrow \quad \hat{\pi}(Z = z) \equiv \hat{\pi}\big(\Lambda\big(W_z^1\big)\big) = \pi\big(\Lambda\big(W_z^1\big)\big) \equiv \pi(Z = z)$$

So

$$\begin{aligned}\hat{\pi}(\hat{Y} = y) &= \sum_z \hat{\pi}(\hat{Y} = y \mid Z = z)\hat{\pi}(Z = z)\\ &= \sum_z \big[\pi\big(\Lambda\big(W^2\big) \mid \Lambda\big(W_z^1\big)\big)\big]^{-1} \pi\big(\Lambda\big(W^2\big), \Lambda_y \mid \Lambda\big(W_z^1\big)\big)\pi\big(\Lambda\big(W_z^1\big)\big)\\ &= \sum_z \left[\frac{\pi(Y = y, W, Z = z)}{\pi(W, Z = z)}\pi(Z = z)\right] \qquad \square\end{aligned}$$

References

[1] T. Bedford, R. Cooke, Probabilistic Risk Analysis: Foundations and Methods, Cambridge, 2001, pp. 99–151.

[2] C. Boutilier, N. Friedman, M. Goldszmidt, D. Koller, Context-specific independence in Bayesian Networks, in: Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence, Portland, Oregon, 1996, pp. 115–123.

[3] R.E. Bryant, Graphical algorithms for Boolean function manipulation, IEEE Transactions of Computers C 35 (1986) 677–691.

[4] G.A. Churchill, Accurate restoration of DNA sequences, in: C. Gatsaris, et al. (Eds.), Case Studies in Bayesian Statistics, vol. 2, Springer-Verlag, 1995, pp. 90–148.

[5] A.P. Dawid, Causal inference without counterfactuals, Journal of the American Statistical Association 95 (2000) 407–448.

[6] A.P. Dawid, Influence diagrams for causal modelling and inference, International Statistical Review 70 (2002) 161–189.

[7] A.P. Dawid, V. Didelez, Identifying the consequences of dynamic treatment strategies, Research Report 262, University College London, 2005.

[8] A.P. Dawid, J. Moertera, V.L. Pascali, D. Van Boxel, Probabilistic expert systems for forensic inference from genetic markers, Scandinavian Journal of Statistics 29 (2002) 577–595.

[9] G. Freeman, J.Q. Smith, Bayesian MAP model selection of Chain Event Graphs, Research Report 09-06, CRiSM, 2009.

[10] S. French (Ed.), Readings in Decision Analysis, Chapman and Hall/CRC, 1989.

[11] S. French, D.R. Insua, Statistical Decision Theory, Arnold, 2000.

[12] D. Glymour, G.F. Cooper, Computation, Causation and Discovery, MIT Press, 1999.

[13] D. Hausman, Causal Asymmetries, Cambridge University Press, 1998.

[14] D. Heckerman, A Bayesian approach to Learning Causal Networks, in: W. Edwards, et al. (Eds.), Advances in Decision Analysis, CUP, 2007, pp. 202–220.

[15] S.L. Lauritzen, Causal inference from graphical models, in: O.E. Barndorff-Nielsen, et al. (Eds.), Complex Stochastic Systems, Chapman and Hall, 2001.

[16] R. Lyons, Random walks and percolation on trees, Annals of Probability 18 (1990) 931–958.

[17] A.M. Madrigal, J.Q. Smith, Causal identification in Design Networks, in: L.E. Sucar, et al. (Eds.), Advances in Artificial Intelligence, vol. 2, Springer Verlag, 2004.

[18] D. McAllester, M. Collins, F. Periera, Case factor diagrams for structured probabilistic modeling, in: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, 2004, pp. 382–391.

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

[20] J. Pearl, Statistics and causal inference: A review, Sociedad de Estadistica e Investigacion Operativa. Test 12 (2) (2003) 281–345.

[21] J. Pearl, Causality Models, Reasoning and Inference, 2nd edition, Cambridge, 2009.

[22] J. Pearl, J.M. Robins, Probabilistic evaluation of sequential plans from causal models with hidden variables, in: Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, 1995, pp. 444–445.

[23] D. Poole, N.L. Zhang, Exploiting contextual independence in probabilistic inference, Journal of Artificial Intelligence Research 18 (2003) 263–313.

[24] E.M. Riccomagno, J.Q. Smith, The geometry of causal probability trees that are algebraically constrained, in: L. Pronzato, A. Zhigljavsky (Eds.), Optimal Design and Related Areas in Optimization and Statistics, Springer, 2008, pp. 131–152.

[25] J.M. Robins, A new approach to causal inference in mortality studies with sustained exposure period – application to control of the healthy worker survivor effect, Mathematical Modelling 7 (1986) 1393–1512.

[26] A. Salmeron, A. Cano, S. Moral, Importance sampling in Bayesian Networks using probability trees, Computational Statistics and Data Analysis 34 (2000) 387–413.

[27] G. Shafer, The Art of Causal Conjecture, MIT Press, 1996.

[28] J.Q. Smith, P.E. Anderson, Conditional independence and Chain Event Graphs, Artificial Intelligence 172 (2008) 42–68.

[29] J.Q. Smith, P.A. Thwaites, Decision trees, in: E.L. Melnick, B.S. Everitt (Eds.), Encyclopedia of Quantitative Risk Analysis and Assessment, vol. 2, Wiley, 2008, pp. 462–470.

[30] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction and Search, Springer-Verlag, 1993.

[31] P.A. Thwaites, Chain Event Graphs: Theory and application, PhD thesis, University of Warwick, 2008.

[32] P.A. Thwaites, J.Q. Smith, Non-symmetric models, Chain Event Graphs and propagation, in: Proceedings of the 11th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Paris, 2006, pp. 2339–2347.

[33] P.A. Thwaites, J.Q. Smith, R.G. Cowell, Propagation using Chain Event Graphs, in: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence, Helsinki, 2008, pp. 546–553.

[34] J. Tian, Identifying dynamic sequential plans, in: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence, Helsinki, 2008, pp. 554–561.

[35] J. Tian, J. Pearl, A general identification condition for causal effects, in: Proceedings of the 18th National Conference on Artificial Intelligence, AAAI Press, 2002, pp. 567–573.