Causal analysis with Chain Event Graphs

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

[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

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.

Keywords

Back Door theorem, Bayesian Network, causal manipulation, Chain Event Graph, conditional independence, event tree, graphical model.

1 Causal Manipulation

Much recent work in the field of causality has focussed 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) has been one of the most successful graphical tools for representing complex dependency relationships, and unsurprisingly researchers have looked to interpret the directionality of the edges of the BN as in some way *causal*. This has led to the development of the Causal Bayesian Network (CBN), using a non-parametric representation based on structural equation models [11][17][18][26]. These provide a framework for expressing assertions

*Corresponding author. Email address Peter.Thwaites@warwick.ac.uk

about what might happen when the system under study is externally manipulated and some of its variables are assigned certain values.

We argue that for many causal problems, the BN is not the most appropriate graphical model. There are two main reasons for this claim:
Firstly, many processes cannot be satisfactorily described by a BN. Examples occur in genomics, epidemiology and multi-agent systems; further examples can be found in [2][15][20]. Such processes tend not to admit a natural product space structure – 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, leading if one uses a BN, to sparse clique probability tables with many zeros or repeated probabilities. Many problems may have some variables which have **no** outcomes given some vectors of values of ancestral variables.
Secondly, even for those problems whose idle settings can satisfactorily be depicted as a BN, there are many manipulations which cannot be described adequately via this representation.
The problems with which we are concerned either exhibit significant asymmetry in the representation of their idle state, or are subject to non-symmetric manipulations. Where problems do not display such asymmetries, the BN remains the appropriate choice for representation and analysis.

The BN provides a simple way of representing the dependence relationships between the measurement variables of a problem, but cannot express graphically all the *context-specific* or sample space information needed for an accurate representation of a more asymmetric problem. Other more primitive tools are needed here, and we note that despite the proliferation of graphical models over the last two decades, the first stage in the development of a model is still often based on the elicitation of an event tree. Although topologically complex, event trees have several advantages over BNs, including (i) they explicitly acknowledge asymmetries embedded in a structure, both in its development and in its sample space structure, (ii) their semantics are much closer to many verbal descriptions of the world, especially when those descriptions revolve around how things happen rather than how the world appears. These advantages are compellingly argued, in for example [23][18][26] in the context of causality. In the related field of decision analysis, French and Insua [10] 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.

The CBN is somewhat of a hybrid structure, retaining a representation of some of the conditional independence structure of the idle system whilst also representing manipulations as the setting of certain measurement variables to specific values. But it may not necessarily be the case that the interventions of interest correspond to setting the original variables in the idle system to specific values – the nature of a manipulation is often dependent on the state of a system and

the values of covariates at a specified time. Pearl [18] touches on this idea, and notes that interventions may involve policies whereby a variable X responds to some other variable(s) Z through either a functional relationship $x = g(z)$, or a stochastic relationship whereby X is set to x with some probability dependent on the value(s) z. We can see that even a symmetric idle system can give rise to a highly asymmetric structure given a particular manipulation rule or policy. Note also that not all vertices in a BN are manipulable in the sense that any manipulation can be given a real interpretation. Although 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*. Such conditioning is not elegantly expressed in the BN, but is simply and intrinsically described in a tree. Analogous arguments are made by Dawid [5] who argues that causes are decisions and not decision rules.

There are partial solutions to some of these problems: Context-specific variants of Bayes Nets exist [2][22][20][16], usually with tree-structured conditional probability tables annexed to the vertices of a BN to allow for the analysis of context-specific independence properties. There is also an art to drawing the appropriate BN of a problem and it is sometimes possible to redefine the variables encoding the problem or add more edges on the graph to aid representation. This may produce a graph consistent with a description of a process, but this graph will still be only a partial representation in general. The context-specific Bayes Net is similarly not a universal panacea – any process (such as a treatment regime) whose unfolding depends on the state of the system at any particular point and the values of specific covariates at that point, cannot be efficiently expressed as a context-specific BN, although it can always be expressed efficiently as a tree. In particular, context-specific BNs do not cope adequately with those problems where some variables have no outcomes given some vectors of values of ancestral variables.

We have already noted that the event tree is a useful tool for **representing** asymmetric problems. It also has its uses for the causal analysis of asymmetric problems and the analysis of the effects of asymmetric causal manipulations. There is a clear link between the analysis of *controlled* models and the field of decision analysis. The common definition that A is a cause of B if the probability of B given a manipulation to A is greater than the probability of B given a manipulation to *not* A (see for example [18]) clearly suggests an event-based (as opposed to variable-based) approach to causal analysis; and an obvious initial candidate among graphical models for such an analysis might be the decision tree. We can think of a causal manipulation as the making of some decision (possibly more than one), and as French and Insua [10] note, such manipulations often induce asymmetry in a problem. Again this suggests that a tree would be a sensible representation.

By using the framework of event trees the definition of manipulative cause is also freed from the shackles of the conditional independence relations imposed

by the more restrictive class of BN models. Causal hypotheses are separated from any direct link with the measurement process. 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 [7]). This is particularly useful in models of biological regulatory mechanisms, which typically contain many noisy *and* and *or* gates [24].

In [24] 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][23], and typically has many fewer nodes than the original event tree. The CEG retains those advantages that event trees have over BNs for the representation of asymmetric problems; but they are also much more flexible and useful than event trees, since their nodes represent intrinsic events in the problem and their edges dependencies between them.

CEGs have two principal advantages over BNs for the representation of (unmanipulated) asymmetric discrete problems. They express **topologically** all the conditional independence structure associated with a problem – this is not bolted on as with context-specific BNs. They also express sample space information generated by the asymmetry of the problem – again this information is expressed in the topology of the graph. See [29] for an example of a very simple problem with an elegant representation as a CEG, but which can only be very clumsily represented by a BN.

We present here a causal extension to CEG models, which we believe to be as transparent and compelling as the extension from BNs to CBNs. Section 2 describes the construction of a CEG and contains an example of how an asymmetric problem can be depicted using such a graph. We have not included a formal definition of the CEG here; such a definition can be found in [24], where also can be found more detail on reading CEGs for conditional independence properties. Section 3 introduces the manipulation of these graphs, and this theory is developed in section 4 where we look at identifying the effects of manipulations. Section 5 introduces a Back Door theorem for CEGs, a generalisation of Pearl's Back Door theorem for BNs [18]. The idealised examples throughout the paper can readily be generalised to more realistic scenarios.

2 Chain Event Graphs

2.1 Derivation

The CEG is a function of an *event tree* [23], and in this section we demonstrate how the CEG is derived from this tree.

An event tree is a directed, rooted tree T, 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 measureable 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

$$\Pi(v) = \{\pi(v' \mid v) \mid v' \in \mathbb{X}(v)\}$$

and

$$\Pi(T) = \{\Pi(v)\}_{v \in S(T)}$$

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

If two situations v and $v^\bullet \in S(T)$ are such that their associated random variables $X(v)$ and $X(v^\bullet)$ have the same distribution then we say that v, $v^\bullet$ are in the same *stage* u – if $v, v^\bullet \in u$, and $v', v^{\bullet\prime}$ label the same outcome given $v, v^\bullet$, then $\pi(v^{\bullet\prime} \mid v^\bullet) = \pi(v' \mid v)$. The set of stages $L(T)$ form a partition of the set $S(T)$.

Two situations v and $v^\bullet$ are therefore in the same stage when the **immediate** future evolution from both v and $v^\bullet$ is governed by the same probability law.

In the conversion of the event tree to the CEG, a useful interim graph is the *staged tree*, defined formally in [24], which is a coloured version of the event tree: If a stage $u \in L(T)$ contains a single vertex $v \in u$, then edges emanating from v are not coloured, but if u contains more than one vertex, then all edges emanating from each $v \in u$ are coloured – two edges $e(v, v'), e(v^\bullet, v^{\bullet\prime})$ emanating from $v, v^\bullet \in u$ have the same colour if these edges label the same outcome (hence $\pi(v^{\bullet\prime} \mid v^\bullet) = \pi(v' \mid v)$).

Two situations v and $v^\bullet$ are said to be in the same *position* w if (i) all edges on all subpaths starting at v or $v^\bullet$ are coloured in the staged tree of T, (ii) for each subpath in the set of subpaths emanating from v, the ordered sequence of colours is the same as that for a subpath in the set of subpaths emanating from $v^\bullet$. The set of positions $K(T)$ forms a partition of the set $S(T)$.

Two situations v and $v^\bullet$ are therefore in the same position when the **entire** future evolution from both v and $v^\bullet$ is governed by the same probability law.

To effect the conversion of the staged tree into a CEG, we start by choosing, for each position $w \in K(T)$, a single representative situation $v \in S(T)$. For each edge $e(v, v')$ leaving v we construct a single edge $e(w, w')$, where $w' = w_\infty$ (a sink-node) if v' is a leaf vertex of T; otherwise w' is the position in $K(T)$ chosen to represent the situation v'.

The colour of the edge $e(w, w')$ is the colour of the edge $e(v, v')$ if this edge has a colour in the staged tree, and if the stage containing the situation v corresponds

to more than just one position $w \in K(T)$. Otherwise the edge is uncoloured. Positions in the same stage are then connected by undirected edges.

The resulting graph $C(T)$ is called a Chain Event Graph – a mixed graph with vertex set $W(C)$ consisting of the positions from $K(T)$ and the sink-node w_∞; directed edge set $E_d(C)$ and undirected edge set $E_u(C)$ as described above. Analogously with the event tree, we call the set of stages of the CEG $L(C)$.

There is a one-to-one correspondence between the root-to-leaf paths in T and the root-to-sink paths in $C(T)$. Each atom of T becomes a path $\lambda(w_0, w_\infty)$ in $C(T)$, and these paths form the atoms of the σ-algebra of the CEG. Events in $C(T)$ are unions of $w_0 \to w_\infty$ paths. For two positions $w, w' \in C(T)$ we write $w \prec w'$ when there is a directed path in $C(T)$ passing through w and w', and w precedes w' on this path.

When the set of stages $L(T)$ of a staged tree is identical to the set of positions $K(T)$, we call $C(T)$ *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 [29].

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

$$\Pi(u) = \{\pi(e(w, w') \mid w) \mid w \in u\}$$

and

$$\Pi(C) = \{\Pi(u)\}_{u \in L(C)}$$

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

2.2 Conditional independence

The implied conditional independence properties of a staged tree can be read from the topology of a CEG. These properties can appear as a number of different types of statement, and are dealt with in detail in [24] and [27]. These types fall broadly into two categories – *cut*-based properties (developed in [24]), and position-based properties (which appear principally in [27]). By nature of its event-tree-based construction, there may be no intrinsic set of measurement variables for the CEG over which conditional independence is defined. This allows a significant degree of flexibility to our analytical procedures.

We define a collection W of positions $w \in K(T)$ as a *fine cut* of $C(T)$ if all $w_0 \to w_\infty$ paths in $C(T)$ pass through exactly one $w \in W$; and we define a collection U of stages $u \in L(T)$ as a *cut* of $C(T)$ if all $w_0 \to w_\infty$ paths in $C(T)$ pass through exactly one $w \in u \in U$.

The *cut*-based conditional independence properties of a CEG detailed in [24] are of two forms: Firstly, if we know that our process has reached some stage

$u \in U$, then we do not need to know anything about how it reached u in order to predict how the process is going to unfold in the immediate future. Secondly, if we know that our process has reached some position $w \in W$, then we do not need to know anything about how it reached w in order to predict how the process is going to behave during its complete future unfolding.

If our CEG represents a symmetric model which can be perfectly depicted by a BN, then we can produce a sequence of cuts and fine cuts which give us exactly the same set of conditional independence statements that we could deduce from the BN [24]. In practice however, in many applications (for example Bayesian decision analysis [9], risk analysis [1], physics [14], biological regulation [4]) our processes are highly asymmetric, and the first stage of model elicitation produces asymmetric event trees with root-to-leaf paths of unequal lengths and event spaces not admitting a natural product space structure. In such cases a CEG-depiction of the problem allows for the representation of context-specific conditional independence statements that cannot be shown on an unmodified BN, and allows the analyst to deduce other context-specific conditional independence properties that might not be apparent before the elicitation process is undertaken.

2.3 An Example

This section contains an example of a model with the type of asymmetric structure described above. We demonstrate how the model can be represented accurately using a Chain Event Graph, and discuss the difficulties inherent in representing the model via a Bayesian Network.

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 (with proceeding being labelled x_1^1 and not proceeding labelled x_1^0).*

It is uncertain that the suspect was 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).*

It would be perfectly possible to 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, but 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 [25], where we can construct

a *dynamic programming tree* (equivalent to the event tree with the ordering $X_1, X_2, \ldots X_6$) or a *causal tree* (wherein things happen in a temporal order). In section 3 we look at the causal manipulation of CEGs, where the topology of the CEG is altered through a decision of some omniscient decision maker.

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 Figure 1.

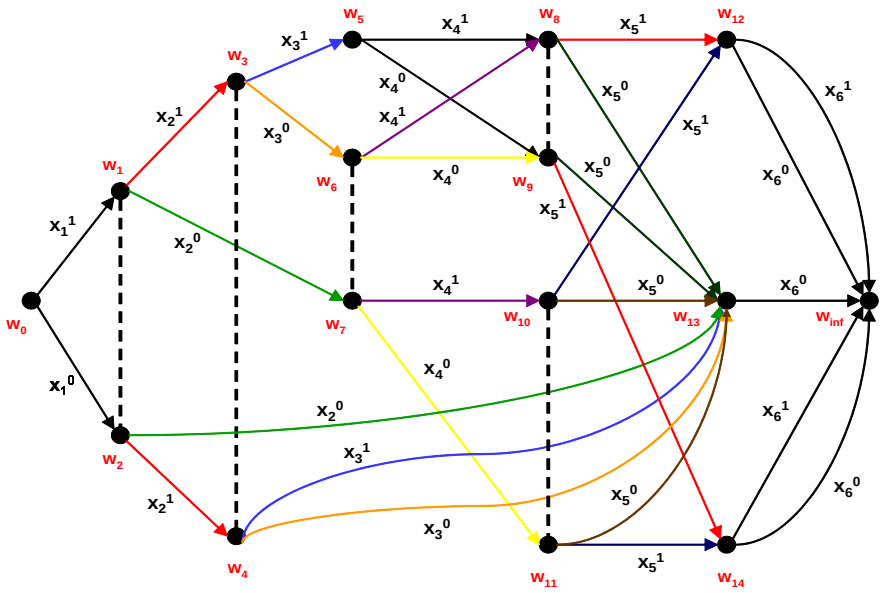

Figure 1: CEG for Example 2.1

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 $x_2^1, x_2^0, x_3^1, x_3^0$ are unaffected by whether they succeed edges labelled x_1^1 or x_1^0. Hence the positions w_1 and w_2 in Figure 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 (*police proceed, S at scene*). S could only have thrown the brick if he was at the scene, so edges labelled x_2^1 are succeeded by edges labelled x_3^1, x_3^0, but edges labelled x_2^0 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 x_4^1, x_4^0, x_5^1 or x_5^0 on $w_0 \to w_\infty$ paths starting with the edge x_1^0.

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. Figure 1 illustrates the fact that we are explicitly using the **topology** of the CEG to express the resulting asymmetric dependency structure.

Could we represent this problem using a BN? Well, of course we could, but our argument is that the CEG is a superior representation as it more accurately describes the problem, and is also a more suitable graph for inference, and for the analysis of causal manipulation.

If we consider the problem contingent on the police proceeding (ie. conditioned on $X_1 = 1$), we can produce a BN on the variables $X_2, X_3, \ldots X_6$ which is consistent with the possible unfoldings of events described above. Such a BN can only be a partial representation as the sample space that includes X_1 is not naturally a product space. Thus (as already noted) if the police do not proceed and S is released, forensic evidence will not be collected, and the witness will not be allowed to testify, so in this sense these variables do not exist under this contingency. This need not stop us trying to draw a BN of the problem, but we can see that such a BN will not be unique. For example, we could make the variables X_4 and X_5 tertiary and label their extra outcomes with the symbol ϕ (to signify that the conditions for X_i taking values corresponding to x_i^1 or x_i^0 have not been met). We could argue that once we know the values of X_4 and X_5 (including $X_4, X_5 = 1_\phi$), we do not need to know the value of X_1 in order to make assessments about X_6. This would suggest a full BN as in Figure 2(a). We could also formally define values of X_4, X_5 conditioned on $X_1 = 0$, in such a way that $X_4 \amalg X_1 \mid (X_2, X_3)$ and $X_5 \amalg X_1 \mid (X_2, X_3, X_4)$; which might lead us to a BN as in Figure 2(b).

Also, if we return to the CEG-representation of the problem in Figure 1, we could insist that every path passes through an edge labelled with outcomes of each of $X_1, X_2, \ldots X_6$ by, whenever we need to add in an edge labelled with outcomes of X_3, X_4, X_5, simply labelling these edges with x_3^0 (S did not throw the brick – here because he wasn't at the crime scene), x_4^0 (no match is found by forensics – here because they didn't do the test), x_5^0 (W did not identify S – here because the case did not go to court). This would also give us a product space structure and allow us to use a BN to model the problem. However the

clique tables for any of our BN-representations of the problem would have a large proportion of zeros (reflecting the actual asymmetry of the problem), with the consequence that our BN would be a very inefficient way of storing the information describing the problem. This would also mean that any attempts to propagate information through the model would be inefficient compared with propagation methods available with CEGs [29].

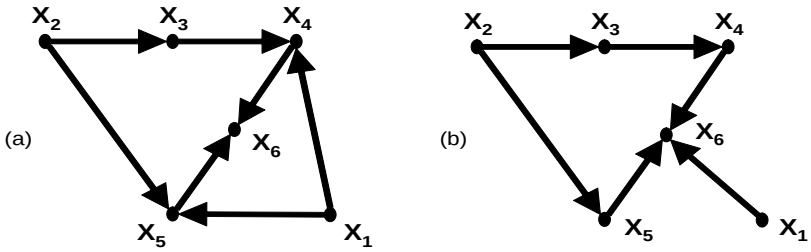

Figure 2: Two possible BNs for Example 2.1

More pertinent perhaps, is that we can only effect this transformation to a BN because our vertex-variables are simple indicators with an outcome *the event of interest does not happen*. Many, if not most, problems in the areas previously mentioned are more complex. For example, in a CEG representing a disease-diagnostic process, the outcomes labelling the edges emanating from a position may be a list of the possible blood types that a patient may have, or a list of their possible combinations of symptoms. To convert such a problem into a BN, additional *dummy* outcomes would need to be added to some vertex-variables whenever we have a set of root-to-sink paths not being all of the same length (for example, if a patient is rhesus +ve, we may not have needed to collect certain information about the patient, as their chances of being affected by some disease is zero [28]). These additions of *dummy* outcomes in order to create our BN result in cumbersome and very inefficient graphical representations.

A more detailed discussion of the difficulties involved in fitting BNs to asymmetric problems appears in [29], wherein we show that even for small problems of this type, the CEG is more efficient than the BN as a means of storing the model structure, but is also more efficient for the propagation of information across the system.

But even if we do create a BN-representation, it will still only convey certain aspects of the story. The fact that S can only have thrown the brick ($X_3 = 1$) if he was present at the crime scene ($X_2 = 1$), or the fact that conviction ($X_6 = 1$) requires positive witness identification ($X_5 = 1$) are not expressed in the BN. We might also be interested in the causal effect of, for example, forcing the witness to identify S as the culprit ($X_5 = 1$) if a match in the glass is found ($X_4 = 1$). This is not represented in the usual semantics of our BNs above. We could of course add an edge between the vertices X_4 and X_5 in our BNs in Figure 2, and then this manipulation could be expressed as a contingent decision, but we would of necessity have lost some information by using this new representation.

Again, it might be argued that a context-specific BN would be perfectly adequate here, but such a representation would still require the addition of dummy outcomes, and conditional probability tables attached to vertices – representing information that is there explicitely in the topology of our CEG. More significantly, since each causal hypothesis may require the addition of extra edges to our BN, we cannot represent all possible hypotheses under consideration with one BN, without a disastrous loss of information – we may well need to create new context-specific BNs for each distinct causal hypothesis we wish to investigate. This is not necessary with our CEG.

3 Manipulating the Chain Event Graph

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 [12][18][23][26]. Here we follow Pearl [18] 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 $p(\cdot \mid\mid \hat{x}_j)$ is defined on $\{X_1, \dots X_n\} \backslash \{X_j\}$ by the formula:

$$p(x_1, \dots x_{j-1}, x_{j+1}, \dots x_n \mid\mid \hat{x}_j) = \prod_{\substack{i=1 \\ i \neq j}}^{n} p(x_i \mid pa_i) \tag{3.1}$$

with $p(x_i \mid pa_i)$ as above, but noting that if X_j is a parent of X_i then X_j takes the value $\hat{x}_j$.

This formula expresses the effect of the manipulation *do* $X_j = \hat{x}_j$. A manipulation of a CEG can be defined in an analogous manner by modifying the distributions of some of the random variables sitting on positions.

Definition 1 *Let $(T, \Pi(T))$ be a tree. Let $D \subset S(T)$ be a subset of the situations of the tree, and $\hat{\Pi}_D = \{\hat{\pi}(v' \mid v) : v \in D,\ v' \in \mathbb{X}(v)\}$ be a new distribution on $v \in D$. Then we define a manipulation of our tree by:*

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

for all $v' \in \mathbb{X}(v),\ v \in S(T)$.

The manipulated tree is the tree so defined, and the manipulated CEG is the CEG of the manipulated tree.

Definition 2 *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, and note that all manipulations of a BN considered by Pearl [17][18][19] are both positioned and staged.

This has a useful consequence for manipulation of CEGs: As manipulations tend to destroy some of the conditional independence structure of a model any way, we can choose to suppress those conditional independence properties encoded by coloured and undirected edges and treat our CEG as *simple*. For *simple* CEGs, each position w is also a stage u, and interventions in the class of *positioned* manipulations of a tree can be enacted on a CEG simply by replacing $(T, \Pi(T))$ by $(C, \Pi(C))$ in Definition 1; $D \subset S(T)$ by $D \subset W(C) \backslash \{w_\infty\}$; $\hat{\Pi}_D = \{\hat{\pi}(v' \mid v) : v \in D,\ v' \in \mathbb{X}(v)\}$ by $\hat{\Pi}_D = \{\hat{\pi}(e(w, w') \mid w) : w \in D\}$, where $\hat{\pi}(e(w, w') \mid w)$ is a new distribution of the random variable $X(u)$ for $u = w$.

The standard manipulations of a BN are those that force some components of the network to take preassigned values, as in expression (3.1). The analogue for CEGs is to consider manipulations which force all paths to pass through a specified set of positions W. This 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).

In a CEG, for a set of positions W, we let $pa(W) = \{w \in W(C) : \exists\ w' \in W \text{ such that } e(w, w') \in E_d(C)\}$ be the set of positions which have an outgoing edge terminating in a position within W. We call the set W a manipulation set if all root-to-sink paths in C pass through exactly one position in $pa(W)$, and each position in $pa(W)$ has exactly one child in 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 Figure 3. As all probabilities after the manipulation remain unchanged, we have *stages* as marked.

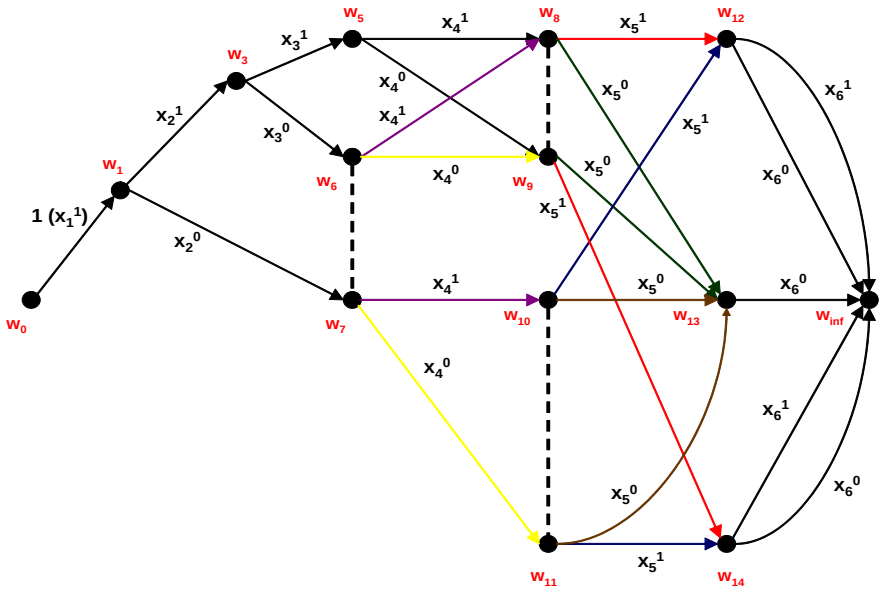

Figure 3: Manipulated CEG $\hat{C}$ for manipulation to w_1

Note that on a BN, we can specify, for example, that a patient is to take some treatment, but we cannot specify how we are going to ensure that the patient takes this treatment. The CEG allows us more flexibility – we can consider a greater range of interventions, many of which may alter the topology of the graph in more complex ways than illustrated here.

We assume that Figure 3 shows a CEG which is valid for our manipulation, but we need to exercise care in making this assumption. If a judge is available, sufficient money has been stolen etc., then the police, believing S to be guilty, will make a **decision** to proceed. In this case our manipulated CEG is almost certainly valid. But suppose the police obtain CCTV footage showing S to be present – then the police will again make a decision to proceed (ensuring 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 acqittal.

If we now consider the manipulation forced to w_{13}, we note that not all root-to-sink paths in C pass through exactly one position in $pa(W)$, the path

$w_0 \to w_2 \to w_4 \to w_{13} \to w_\infty$ path for example passing through two positions in $pa(w_{13})$. This manipulation none-the-less has a straight-forward interpretation (as a contingent manipulation) – if the police proceed, the witness is forced **not** to identify the suspect. A CEG for this interpretation is given in Figure 4.

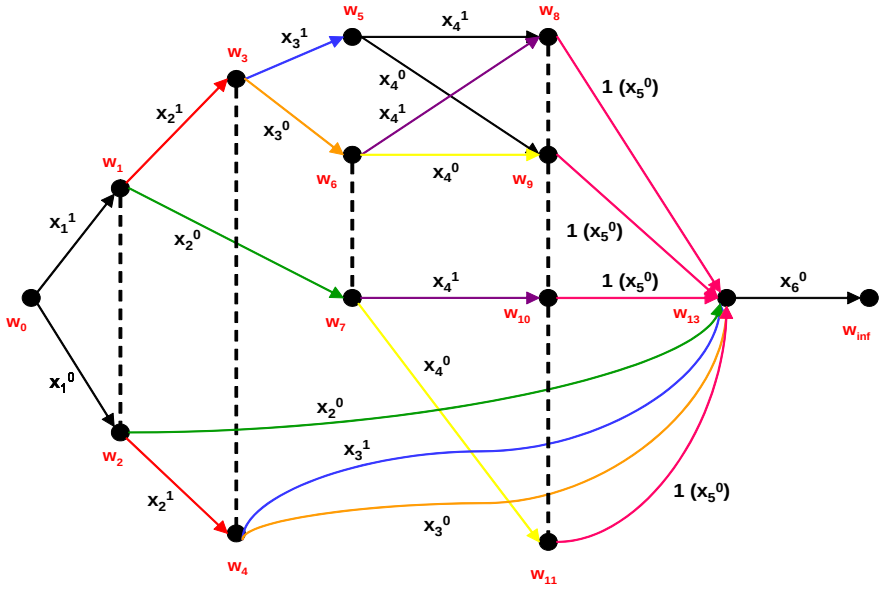

Figure 4: Manipulated CEG $\hat{C}$ for manipulation to w_{13}

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 $e(w_2, w_{14})$ edges, and so the stage structure is as in Figure 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 ($Y = 1$ being *high* satisfaction), we can draw a CEG for this problem as in Figure 5.

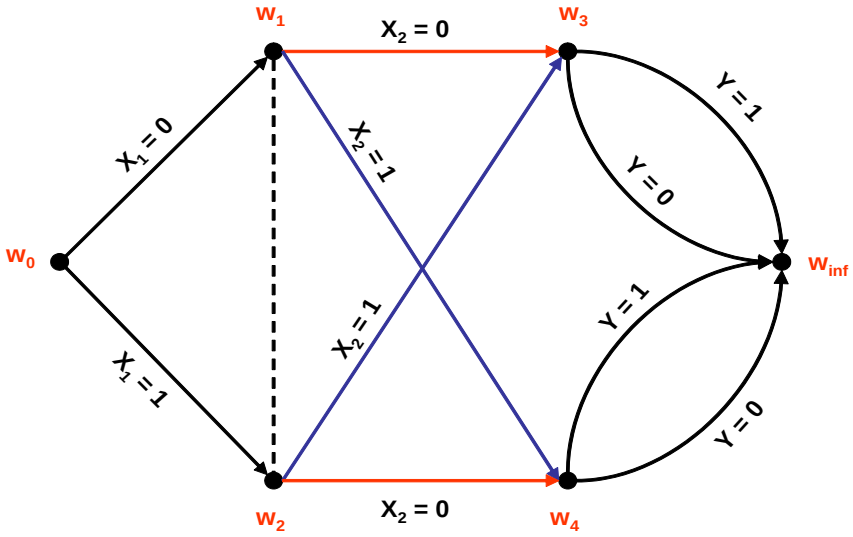

Figure 5: CEG for Example 3.2

The undirected edge between w_1 and w_2 reflects the random allocation of first year students to apartments. A possible BN for this problem, encoding the independence of X_1 and X_2 is given in Figure 6(a).

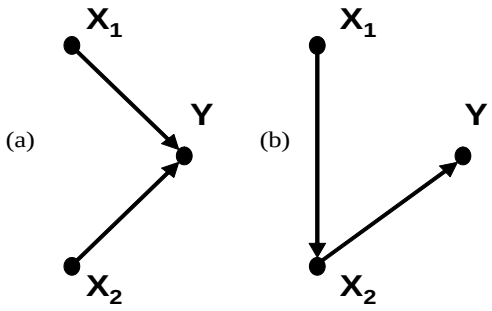

Figure 6: Possible BNs for Example 3.2

A fine cut through the positions $\{w_3, w_4\}$ of the CEG in Figure 5 gives us the conditional independence property that $Y \amalg (X_1, X_2) \mid |X_1 - X_2|$, a property that cannot be deduced from the BN in Figure 6(a). Nor is it possible to determine from this BN (or from the factorisation of the probability mass function of the path events) whether the allocation of the second year students occurs before or after the allocation of the first years. This property of the CEG allows us to consider manipulations where, for example the university places first year students with second years of the same ethnicity. Such an intervention would

correspond to a manipulation of the CEG in Figure 5 to the position w_3. Note that this manipulation would cause the removal of the undirected edge between w_1 and w_2, since $X_1 \not\perp\!\!\!\perp X_2 \mid (X_1 = X_2)$.

We could, of course, 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 Figure 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, this new BN does not fully describe the idle system, as it no longer encodes the property that first year students are allocated at random.

So neither BN is able to describe adequately both the idle and the manipulated systems, whereas the CEG can.

4 Identifying the effects of manipulations

There has been considerable recent interest in causal BN literature [6][18][19] in studying 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. Typically, sufficient conditions on the topology of the BN are given for such identifiability to exist. This allows us to design experiments on the idle system so as to be able 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. Indeed it can be used 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 [18] we prove several sufficient conditions for identifiability, and generalise Pearl's Back Door theorem to CEG models. We first need to provide some notation and a couple of definitions.

Recall that w indicates a position in our CEG. 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 more usually $\mu(w_1, w_2)$, where w_1 and w_2 indicate the start and end positions of the subpath.

A union of atoms constitutes an event, denoted Λ, and M is used to indicate a union of subpaths – usually this is of the form $M(w_1, w_2)$ for positions w_1 and w_2. $\Lambda(w)$ is used to represent the union of all paths passing through the position w, and $\Lambda(e)$ the union of all paths passing through the edge e. Both $\Lambda(w)$ and $\Lambda(e)$ are events. $\Lambda(\mu(w_1, w_2))$ is the event which is the union of all paths utilising the subpath $\mu(w_1, w_2)$.

We use $\pi(w) = \pi(\Lambda(w))$ to denote the probability of passing through the position w, which is also the probability of reaching w from w_0. The probability of reaching w_2 from w_1 is $\pi(\Lambda(w_2) \mid \Lambda(w_1))$, usually simplified to $\pi(w_2 \mid w_1)$. Similarly $\pi_\mu(w_2 \mid w_1) = \pi(\Lambda(\mu(w_1, w_2)) \mid \Lambda(w_1))$ is the probability of utilising

the subpath $\mu(w_1, w_2)$ given that we have reached w_1 – this can be thought of as the probability **of** the subpath $\mu(w_1, w_2)$. By thinking of an edge as a (very short) subpath, we can also define $\pi_e(w_2 \mid w_1)$, the probability of the edge $e(w_1, w_2)$.

We now consider random variables defined on a CEG.

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 3 *A random variable Y is called* observed *if and only if indicators of the events $\{\Lambda_y\}$ are observed for all levels y.*

Definition 4 *Call a manipulation of a CEG $(C, \Pi(C))$* forced to (the position) *w if:*

1. *it assigns probability one to the event $\Lambda(w)$,*
2. *all primitive probabilities in the manipulated CEG $(\hat{C}, \hat{\Pi}(\hat{C}))$ associated with edges downstream of w in C are those of the idle system.*

In Example 3.1, both our manipulations are manipulations forced to a position.

We first consider an effect random variable $\hat{Y}$ defined on the path σ-algebra of $\hat{C}$, where the initial manipulations under consideration are manipulations forced to w. $\hat{Y}$ generates a partition of the root-to-sink paths of $\hat{C}$ with each outcome y corresponding to a union of $w_0 \to w_\infty$ paths Λ_y.

Each $w_0 \to w_\infty$ path in $\hat{C}$ can be thought of as 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 $M(w_0, w)$.

We wish to consider $\hat{Y}$ as perhaps a measurement of an effect after a manipulation forced to w, and so we wish $\hat{Y}$ to be in some sense *after* or *downstream* of w. To do this is straightforward. We require that our partition $\{\Lambda_y\}$ consists of events each of which is $M(w_0, w)$ conjoined to a union of subpaths from $\{\mu(w, w_\infty)\}$ – for outcome y, call this union $M_y(w, w_\infty)$.

We can define a random variable Y on the path σ-algebra of C so that the outcomes of Y partition the root-to-sink paths of C and whenever such a path passes through w and the equivalent path in $\hat{C}$ belongs to the event $\hat{Y} = y$ then in C this path belongs to the event $Y = y$.

In practical situations of course, this defining of Y and $\hat{Y}$ is done the other way around. In a CEG of a BN, sets of edges the same distance from w_0 represent outcomes of the same variable, and we might well label a subset of such edges with the outcome y_0 for example. 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, but those that are left can 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.
Where there is no possibility of ambiguity we can drop the hat from $\hat{Y}$.

Lemma 1 *For all levels y, under a manipulation forced to w*

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

provided that in the unmanipulated system $\pi(w) > 0$.

We have already equated the event $\hat{Y} = y$ with the union of $w_0 \to w_\infty$ paths Λ_y in $\hat{C}$. We can, without ambiguity, when working on C, equate the event $Y = y$ with the union of $w_0 \to w_\infty$ paths Λ_y in C, since those paths that compose Λ_y in $\hat{C}$ are simply those paths in C which satisfy $Y = y$ and pass through the position w. The result of Lemma 1 can hence be expressed as:

$$\hat{\pi}(\Lambda_y) = \pi(\Lambda_y \mid \Lambda(w))$$

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. But this is not however always possible, even in models that can be described by a CBN. When we cannot observe such indicators, we can often 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 5 *A set of positions W of a CEG C is called* C-regular *if no two positions in W lie on the same directed path of C.*

We now construct an effect random variable associated with a manipulation forced to W, where W is a C-regular set. So, as before, 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 as before, we wish $\hat{Y}$ to be *downstream* of W.
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 define Y on C as before, and where there is no possibility of ambiguity we drop the hat from $\hat{Y}$.
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. To aid us here, we construct a graph representing what happens up until we reach a given position w. Let $C^*(w)$ denote the coloured subgraph of C whose vertices and edges are those along the $w_0 \to w$ subpaths of C, and whose edge-colouring (ie. edge-probabilities) is inherited from C also. Usually $C^*(w)$ is not

a CEG. Write $K(C^*(w))$ for the subset of $W(C)$ of positions retained in $C^*(w)$, with the exception of w. Also, for a C-regular set of positions W, let $C^*(W)$ denote the coloured subgraph of C whose vertices and edges are those along the $w_0 \to w \in W$ subpaths of C, and whose edge-colouring is inherited from C. If W contains more than one position $C^*(W)$ is not a CEG. We let

$$K(C^*(W)) = \bigcup_{w \in W} K(C^*(w))$$

Definition 6 *Call a set of positions W* simple *if and only if:*

1. *W is C-regular,*
2. *there exists a partition of the set $K(C^*(W))$ into $K^\alpha(C^*(W))$ and $K^\beta(C^*(W))$, called* active *and* background *positions respectively, such that for $w \in W$, $\pi(\Lambda(w)) = \pi(\Lambda(M(w_0, w)))$ can be decomposed as $A(w) \times B(w)$, where $A(w)$ is a function of the* active *positions and $B(w)$ is a function of the* background *positions,*
3. *$A(w) = A(W)$ is constant $\forall\ w \in W$.*

When W contains a single position, W is clearly *simple*.

Definition 7 *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 $\hat{\Pi}(W) = 1$,*
3. *$\Pi(C)$ and $\hat{\Pi}(\hat{C})$ differ only on edges whose parents lie in $K^\alpha(C^*(W))$.*

Example 4.1 *Consider the binary BN and corresponding CEG in Figure 7. Let the set $W = \{w_7, w_9\}$.*

Here $C^*(W)$ would be the subgraph of the CEG in Figure 7 consisting of the four subpaths joining w_0 to w_7, w_8; retaining the CEG's edge colouring (and labels) but not its undirected edges. We have

$$\pi(\Lambda(w_7)) = \pi(c_0) \sum_d \pi(d)\ \pi(x_0 \mid d) = \pi(c_0)\ \pi(x_0)$$
$$\pi(\Lambda(w_9)) = \pi(c_1) \sum_d \pi(d)\ \pi(x_0 \mid d) = \pi(c_1)\ \pi(x_0)$$

So here the position $w_0 \in K^\beta(C^*(W))$, $B(w_7) = \pi(c_0)$, $B(w_9) = \pi(c_1)$; the positions $w_3, w_4, w_5, w_6 \in K^\alpha(C^*(W))$, $A(w_7) = A(w_9) = \pi(x_0) = A(W)$, and hence W is simple in C.

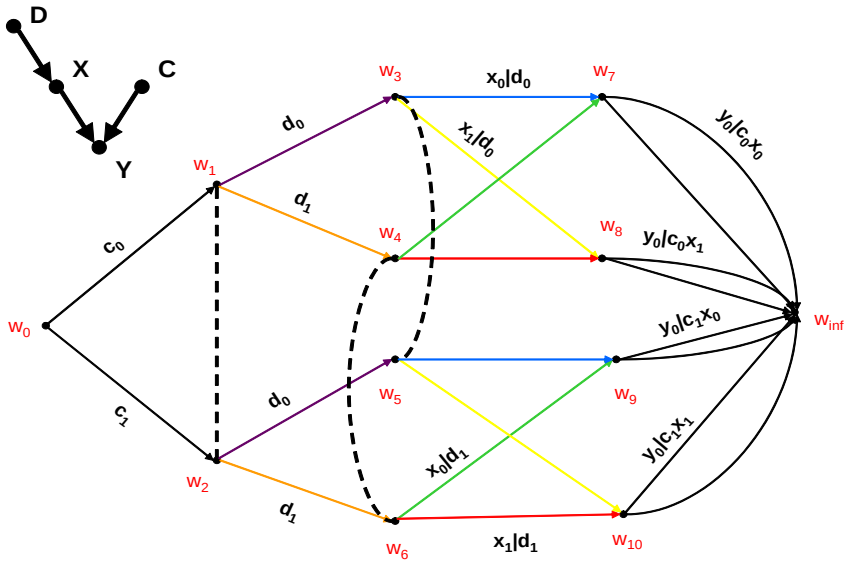

Figure 7: BN and CEG for Example 4.1

If we manipulate C to W (equivalent to the Pearl manipulation $do(X = x_0)$), we get $\hat{C}$ as in Figure 8.

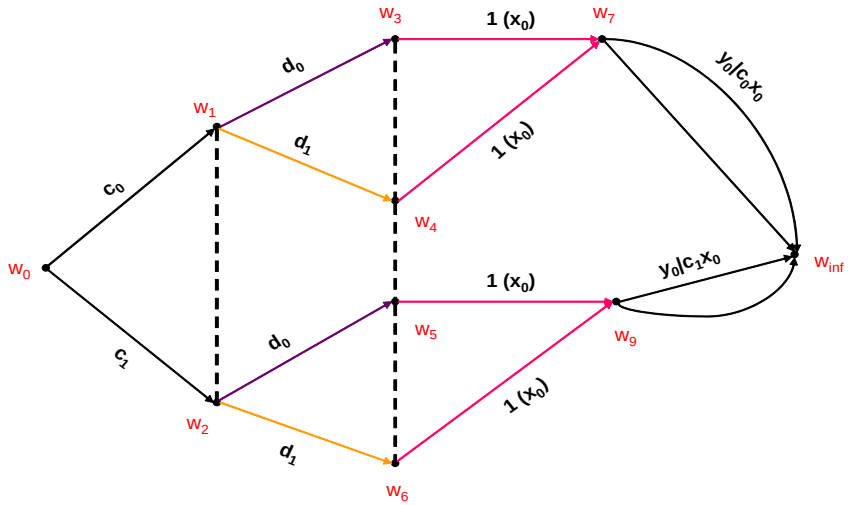

Figure 8: $\hat{C}$ for the manipulation described in Example 4.1

Here

$$\hat{\pi}(\Lambda(w_7)) = \pi(c_0) \sum_d \pi(d) \times 1 = \pi(c_0) \times 1$$
$$\hat{\pi}(\Lambda(w_9)) = \pi(c_1) \sum_d \pi(d) \times 1 = \pi(c_1) \times 1$$

and W is simple in $\hat{C}$.
$\hat{\Pi}$ differs only on the edges leaving w_3, w_4, w_5, w_6, the edges leaving the *active* positions; so this manipulation is *amenable.*

The point of these definitions is that, in a sense to be defined below, the random variables associated with positions lying in $K^\alpha(C^*(W))$ are independent of those associated with positions lying in $K^\beta(C^*(W))$. An amenable manipulation may change probabilities associated with variables labelled by active positions, but will always leave probabilities associated with variables labelled by background positions unchanged.

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 $\pi(W) = \sum_{w \in W} \pi(\Lambda(w)) = \sum_{w \in W} \pi(\Lambda(M(w_0, w)))$, and provided that $\pi(\Lambda(w)) > 0 \ \forall \ w \in W$.

More formally

$$\hat{\pi}(\Lambda_y) = \pi(\Lambda_y \mid \Lambda(W))$$

where $\Lambda(W) = \bigcup_{w \in W} \Lambda(w)$.

Note that the validity of the expression for $\hat{\pi}(\Lambda_y)$ in Lemma 2 depends on the result $\hat{\pi}(\Lambda(w)) = \frac{\pi(\Lambda(w))}{\pi(\Lambda(W))}$ holding, so checking this result is a sensible starting point in our analysis.

Example 4.2 *The manipulation in Example 4.1 is amenable.*

Here we can define Λ_y to be the event $Y = y_0$, and we get, from Figure 8 that

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

From Figure 7 we have that

$$\begin{aligned}\pi(\Lambda_y \mid \Lambda(W)) &= \pi(Y = y_0 \mid x_0) \\ &= \frac{\sum_c \pi(c)\ \sum_d \pi(d)\ \pi(x_0 \mid d)\ \pi(y_0 \mid c, x_0)}{\sum_c \pi(c)\ \sum_d \pi(d)\ \pi(x_0 \mid d)} \\ &= \cdots = \sum_c \pi(c)\ \pi(y_0 \mid c, x_0) = \hat{\pi}(Y = y_0) = \hat{\pi}(\Lambda_y)\end{aligned}$$

Note that $\frac{\pi(\Lambda(w_7))}{\pi(\Lambda(W))} = \frac{\pi(c_0 x_0)}{\pi(x_0)} = \pi(c_0)\ \left(= \hat{\pi}(\Lambda(w_7))\right)$ since $X \amalg C$ (the fine cuts through $\{w_1, w_2\}$ and $\{w_3, w_4, w_5, w_6\}$ give us $C \amalg D,\ X \amalg C \mid D \ \Rightarrow\ X \amalg C$).

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

$$\mathbf{H}_2 \amalg (\mathbf{H}_1, X)$$

and

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

It is straightforward to check that, for a CEG drawn in any order compatible with the ordering of the vertices of such a BN, these are exactly the conditions of Lemma 2. In this correspondence the states of the vector of hidden variables $\mathbf{H}_1$ and X define the values the active positions take, whilst the vector of hidden variables $\mathbf{H}_2$ define the values the background positions take. 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 in Lemma 2 only depend on an appropriate factorisation of probabilities associated with the manipulated set W.

Using Pearl's terminology and the (sets of) variables $X, Y, \mathbf{H}_1, \mathbf{H}_2$ we have that

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

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

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

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

$$\begin{aligned}&= \sum_{h_2} \pi(h_2, y \mid x) \\ &= \pi(y \mid x)\end{aligned}$$

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 active variables are D and X, and our background variable is C.

It is worth pointing out that if we do not use the two conditional independence statements as written, but only the implications

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

then the only derivable expression is

$$\pi(y \mid\mid x) = \sum_{h_2} \pi(h_2)\ \pi(y \mid h_2, x)$$

which is of course the Back Door formula for BNs [18].

Example 4.3 *Note that if we break one of the conditions (eg. $X \not\amalg \mathbf{H}_2$) we no longer have an amenable manipulation. So, consider the binary BN and corresponding CEG in Figure 9.*

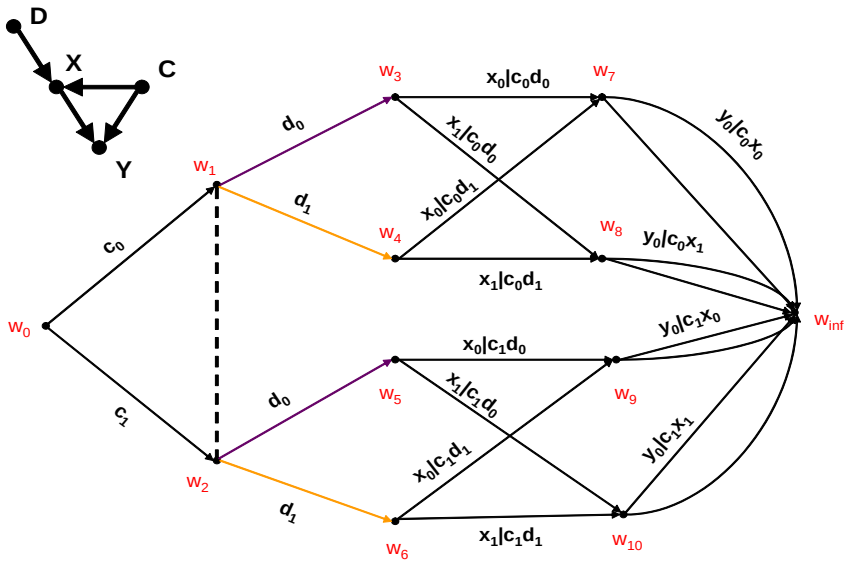

Figure 9: BN and CEG for Example 4.3

If we let the set $W = \{w_7, w_9\}$, then

$$\pi(\Lambda(w_7)) = \pi(c_0)\ \sum_d \pi(d)\ \pi(x_0 \mid c_0, d)$$
$$\pi(\Lambda(w_9)) = \pi(c_1)\ \sum_d \pi(d)\ \pi(x_0 \mid c_1, d)$$

which do not factorise as $A(W)B(w)$.

Manipulating to W we get the same $\hat{C}$ as before, so

$$\hat{\pi}(y_0) = \sum_c \pi(c)\ \pi(y_0 \mid c, x_0)$$

but we now have

$$\pi(y_0 \mid x_0) = \frac{\sum_c \pi(c)\ \sum_d \pi(d)\ \pi(x_0 \mid c, d)\ \pi(y_0 \mid c, x_0)}{\sum_c \pi(c)\ \sum_d \pi(d)\ \pi(x_0 \mid c, d)}$$

which no longer simplifies to this expression. Note that here

$$\frac{\pi(w_7)}{\pi(W)} = \frac{\pi(c_0, x_0)}{\pi(x_0)} = \pi(c_0 \mid x_0) \neq \pi(c_0)\ \left(= \hat{\pi}(w_7)\right)$$

5 A Back Door theorem for Chain Event Graphs

A key component of causal analysis on BNs is Pearl's Back Door theorem [18][19], 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 we can calculaste 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 an observed 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 not only a wider class of idle system models, but also a wider class of manipulations of these than is possible with a BN. Similarly, our random variable Z no longer needs to correspond to any fixed subset of the measurement variables of the problem, giving us more opportunity of finding an appropriate probability expression.

We have previously considered $C^*(w)$ and $C^*(W)$, graphs replicating the topology of C from w_0 to w or to a set of positions W. The graph $C(w)$ which replicates the topology of C from w to w_∞ is, unlike $C^*(w)$, automatically a CEG, with w as root-node. We can also create a CEG replicating the topology of C from W to w_∞.

Definition 8 *For a set of C-regular positions $W \subset W(C)$, 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 $W(C)$ 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 (ie. 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.

Lemma 3 *For sets of C-regular positions $W^1, W^2 \subset W(C)$, W^2 is* simple *in the CEG $C(W^1)$ if and only if the probability $\pi(\Lambda(w^2) \mid \Lambda(W^1))$ can be decomposed as $A(W^2) \times B(w^2)$ for all $w^2 \in W^2$ (where $A(W^2)$ is constant for all $w^2 \in W^2$).*

This is an important result as it means that whether a set W^2 is *simple* in $C(W^1)$ can be checked on C, without the necessity of drawing the CEG $C(W^1)$.

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

If we let the set of positions to which we intend to manipulate be $W = \{w^2\}$, then for Z to occur before the manipulation we require that every position $w^2 \in W$ lies on a path in C between some position $w^1 \in \Lambda_z$ (for some level z) and w_∞. Note also that as our set $\{w^1\}$ is going to take the role of Z in our Back Door theorem, we need it to be the case 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 $w^2 \in W$ – 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 $Z = z$ (Λ_z) is the same in $\hat{C}$ as in C.

We define our effect variable Y as before, and note that $\{Y = y \mid Z = z\}$ are $C(\Lambda_z)$-measureable events (for each level z) such that

$$\pi(Y = y) = \sum_z \pi(Y = y \mid Z = z)\ \pi(Z = z)$$

since $\{\Lambda_z\}$ partitions the $w_0 \to w_\infty$ paths of C.

Definition 9 *A set of C-regular positions $W \subset W(C)$ is called* simple conditioned on Z *if*

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

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. Note again that such a manipulation does not change any primitive probabilities from the idle system lying on a subpath between w_0 and positions in Λ_z. Letting $\hat{Y}$ be the image of Y in the manipulated CEG, we have that $\{\hat{Y} = y \mid Z = z\}$ are $C(\Lambda_z)$-measureable events such that

$$\hat{\pi}(\hat{Y} = y) = \sum_z \hat{\pi}(\hat{Y} = y \mid Z = z)\ \hat{\pi}(Z = z)$$

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:*

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

or more formally, as

$$\hat{\pi}(\Lambda_y) = \sum_z \pi(\Lambda_y \mid \Lambda(W), \Lambda_z)\ \pi(\Lambda_z)$$

It is worth stressing that the partition $\{\Lambda_z\}$ is constructed so as to help us to calculate $\hat{\pi}(\Lambda_y)$, and that the **choice of positions** within $\{\Lambda_z\}$ will therefore depend on those events which are observable or manifest within the system. If a collection of positions within $\{\Lambda_z\}$ are indistinguishable through observations possible on the idle system, then we would assign these positions to the same Λ_z (ie. assign the same value $Z = z$ to each of these positions). The **further**

assignment of z-values to positions can be completely arbitrary, with our final partition $\{\Lambda_z\}$ being coarser or finer as our brief suggests.

Note that if we construct a CEG of a BN, the ordering of the CEG is constrained to some extent by the topology of the BN (if we label the paths of the CEG in Figure 7 by $CDXY$, we can see that the CEG of the BN in Figure 7 could also be constructed so its paths were labelled by $DCXY$ or $DXCY$). Within these constraints, we find that for a CEG of a BN, and an *atomic* intervention of that BN [18], introducing the *background* variables as early as possible in our CEG makes it more likely that the conditions necessary for Theorem 1 are met. We suspect this is also the case for manipulations of more asymmetric CEGs.

Example 5.1 *In Example 3.1 we considered a manipulation to w_{13}, where if the police proceeded the witness was forced not to identify S. Let us suppose the police* **have** *decided to proceed. It is a very simple matter to modify our CEG to model this contingency. The new CEG, given in Figure 10, is simply the CEG $C(w_1)$, a CEG of the type $C(w)$ as described at the beginning of this section.*

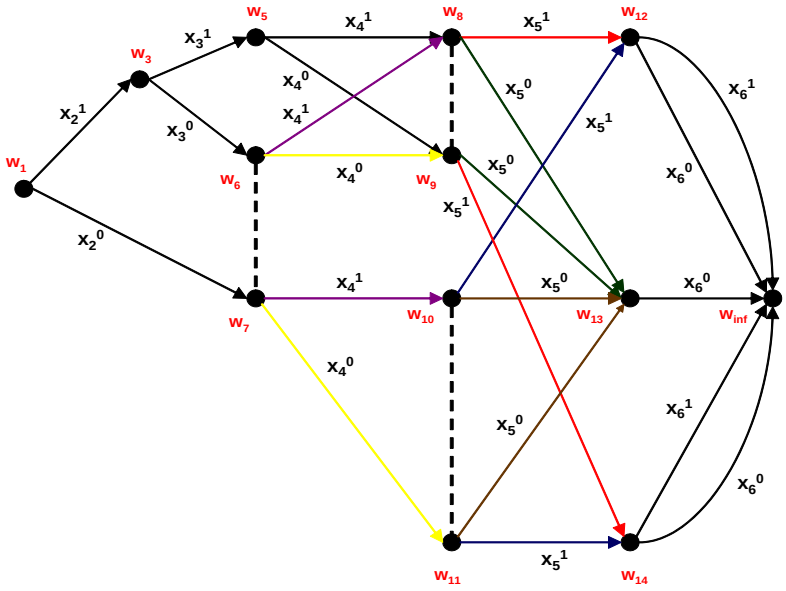

Figure 10: CEG for Example 5.1

Consider the manipulation of our new CEG wherein the witness is forced to identify S. 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. This is a manipulation forced to $W = \{w_{12}, w_{14}\}$. W is a manipulation set under the conditions described in section 3 – $pa(W) = \{w_8, w_9, w_{10}, w_{11}\}$; every root-to-sink path in $C(w_1)$ passes through exactly one position in $pa(W)$; and each posi-

tion in $pa(W)$ has exactly one child in W. The manipulated CEG is given in Figure 11.

The manipulator here would wish to have a very good idea of the effects (on the indicator X_6) of the manipulation, but may not have any means of estimating, for example, the joint distribution of X_3, X_4 (necessary to calculate the full manipulated probability expression for $X_6 = x_6^1$). This need not be a problem, as we demonstrate here. The effect on X_6 of the manipulation can be deduced if we can estimate only a small collection of probabilities from the idle system.

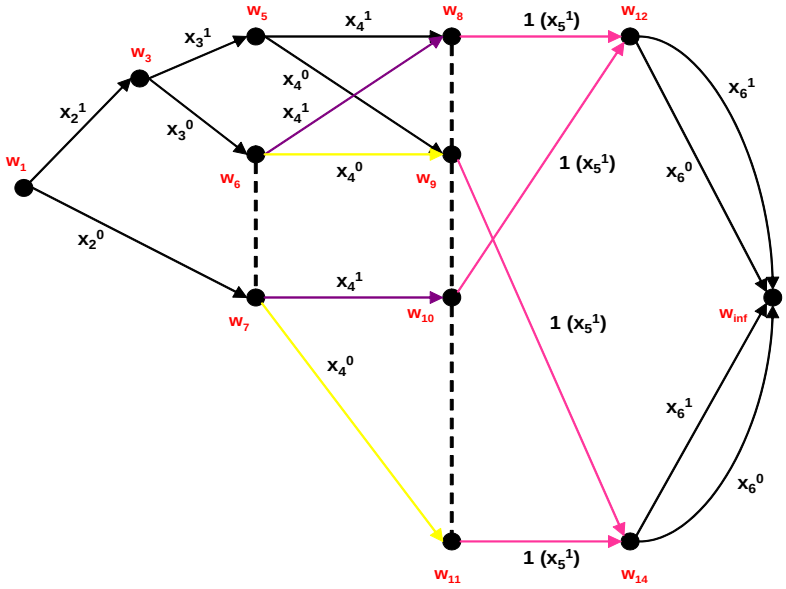

Figure 11: Manipulated CEG for Example 5.1

Suppose we assign the value of $Z = 1$ to paths passing through w_5 (suspect threw brick); $Z = 2$ to paths passing through w_6 (suspect at scene, but did not throw brick): $Z = 3$ to paths passing through w_7 (suspect not at scene). Then if we let $\Lambda_{z(1)} = \{w_5\}$, $\Lambda_{z(2)} = \{w_6\}$, $\Lambda_{z(3)} = \{w_7\}$, we have that $\{\Lambda_z\}$ partitions the root-to-sink paths of our CEG.

If we let $W_{z(1)} = W_{z(2)} = W_{z(3)} = W = \{w_{12}, w_{14}\}$, then by construction, if W is *simple* in $C(\Lambda_z)$ for each Λ_z, the conditions of Definition 9 are satisfied. Lemma 3 enables us to check this on C without the necessity of drawing each $C(\Lambda_z)$. From Figure 10, we see that

$$\pi(\Lambda(w_{12}) \mid \Lambda_{z(1)}) = \pi(w_{12} \mid w_5) = \pi(x_4^1 \mid w_5)\ \pi(x_5^1 \mid x_2^1)$$
$$\pi(\Lambda(w_{14}) \mid \Lambda_{z(1)}) = \pi(x_4^0 \mid w_5)\ \pi(x_5^1 \mid x_2^1)$$

so W is simple in $C(\Lambda_{z(1)})$. It is straightforward to show that W is also simple in $C(\Lambda_{z(2)})$, $C(\Lambda_{z(3)})$. Our variable Z, manipulation set W and effect variable

Y satisfy the conditions for Theorem 1, with $Y = y$ becoming $X_6 = x_6^1$. Hence

$$\hat{\pi}(x_6^1) = \sum_{z=1}^{3} \frac{\pi(x_6^1, W \mid z)}{\pi(W \mid z)} \, \pi(z)$$

Noting that the set W corresponds to the event $X_5 = x_5^1$, we get

$$\hat{\pi}(x_6^1) = \sum_{z=1}^{3} \pi(x_6^1 \mid x_5^1, z) \, \pi(z)$$

It is somewhat laborious, but 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 involving X_4.

Note that checking that our set $\{\Lambda_z\}$ satisfies the conditions required for Theorem 1 is in some ways a similar process to Pearl's method of checking that variables in a BN are d-separated by following possible paths between them [17][18]. This process does not need to be done by hand by the analyst. Similarly, updating the edge-probabilities of our CEG following a manipulation can be done rapidly using algorithms analogous to those described and coded for updating edge-probabilities following an observation in [29].

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 ($X_3 = 0$) or in town L ($X_3 = 1$). Students lodged in towns K and L may have a friendly landlord ($U = 0$) or an unfriendly landlord ($U = 1$), and the friendliness of these landlords is not known to the university.*

When $Z = 0$ it is believed that the CEG in Figure 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 Figure 12.

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 $\hat{\pi}(Y = 1)$ – 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 sort of asymmetries exhibited in this problem make it extremely difficult to represent through a single BN – X_1 is only defined for a student allocated to campus, whilst X_3 and U are only defined for students allocated

to lodgings. Furthermore the manipulation proposed is different for different contingencies.

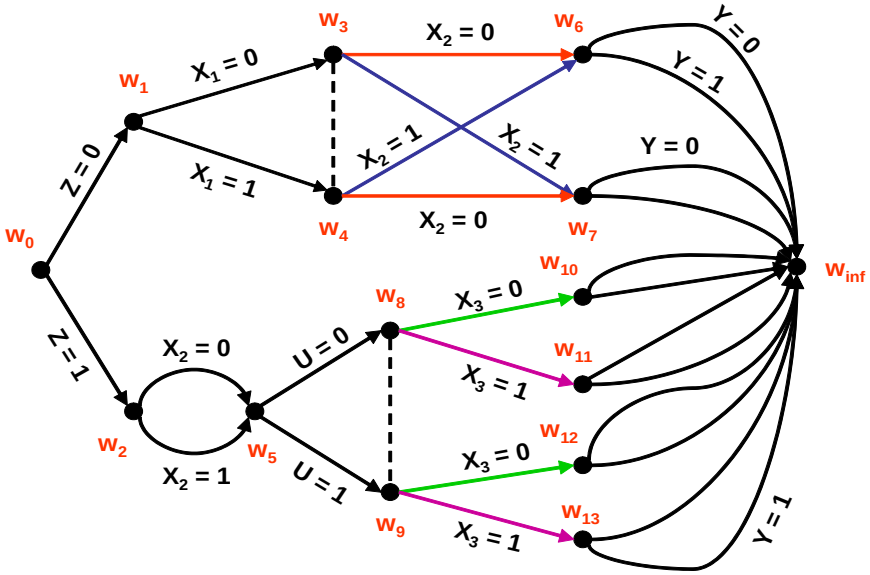

Figure 12: CEG for Example 5.2

The proposal can be considered as a manipulation to $W = \{w_6, w_{11}, w_{13}\}$. If we consider the partition $\{\Lambda_z\} = \{\{w_1\}, \{w_2\}\}$ it is straightforward to check that our variable Z, manipulation set W and effect variable Y satisfy the conditions for Theorem 1, and hence

$$\begin{aligned}\hat{\pi}(Y=1) &= \sum_{z=1}^{2} \pi(Y=1 \mid z, W)\ \pi(z) \\ &= \pi(Y=1 \mid Z=0, X_1 = X_2)\ \pi(Z=0) \\ &\quad + \pi(Y=1 \mid Z=1, X_3 = 1)\ \pi(Z=1)\end{aligned}$$

So $\hat{\pi}(Y=1)$ 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 $\hat{\pi}(Y=1)$ is a function of the event $X_1 = X_2$, or alternatively of the variable $\left|X_1 - X_2\right|$ – a function of X_1 and X_2; a fact that cannot be deduced directly from any BN on the original measurement variables.

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. This is a significant generalisation of Pearl's Back Door theorem:
Firstly, it can be applied to highly asymmetric models as easily as ones exhibiting the strong types of symmetry that can be coded by a BN.
Secondly, the search for an appropriate random variable Z, whose observation ensures identifiability, is not just 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.

6 Discussion

We have demonstrated that the CEG provides a very flexible graphical framework within which to represent and analyse a wide variety of causal hypotheses, even in very asymmetrical domains. Of course the Back Door Theorem presented in this paper is not the only topological crtiterion for determining causal extensions; for example it is possible to produce and prove analogues of Pearl's Front Door Theorem (see [27]). In [8] 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 [13]) – 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 [21]). 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.

Appendix. Proofs

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

$$\begin{aligned}\hat{\pi}(\hat{Y} = y) = \hat{\pi}(\Lambda_y) &= \hat{\pi}(\Lambda(w), \Lambda(M_y(w, w_\infty)) \\ &= \hat{\pi}(\Lambda(w))\ \hat{\pi}(\Lambda(M_y(w, w_\infty)) \mid \Lambda(w)) \\ &= \hat{\pi}(\Lambda(w))\ \hat{\pi}_{M_y}(w_\infty \mid w) \\ &= 1 \times \pi_{M_y}(w_\infty \mid w)\end{aligned}$$

since all paths in $\hat{C}$ pass through w, and using Definition 4 (2).
By definition of Y on C we have

$$\begin{aligned}\pi(Y = y, w) &= \pi(\Lambda_y, \Lambda(w)) \\ &= \pi(\Lambda(w))\ \pi_{M_y}(w_\infty \mid w)\end{aligned}$$

$$\begin{aligned}\Rightarrow\ \hat{\pi}(\hat{Y} = y) = \hat{\pi}(\Lambda_y) &= \frac{\pi(\Lambda_y, \Lambda(w))}{\pi(\Lambda(w))} \\ &= \pi(\Lambda_y \mid \Lambda(w)) \\ &= \pi(Y = y \mid w) \qquad \square\end{aligned}$$

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

$$\begin{aligned}\pi(\Lambda(w)) &= A(W)B(w) \\ \pi(\Lambda(W)) &= \sum_{w \in W} \pi(\Lambda(w)) \qquad \text{as } \{\Lambda(w)\} \text{ partitions } \Lambda(W) \\ &= A(W) \sum_{w \in W} B(w) \\ \hat{\pi}(\Lambda(w)) &= \hat{A}(W)\hat{B}(w) = \hat{A}(W)B(w) \qquad \text{from Definition 7 (2) and (3)} \\ \hat{\pi}(\Lambda(W)) &= \hat{A}(W) \sum_{w \in W} B(w)\end{aligned}$$

But $\hat{\pi}(\Lambda(W)) = 1$ by Definition 7 (2), so

$$\begin{aligned}\hat{\pi}(\Lambda(w)) = \frac{\hat{A}(W)}{A(W)}\ \pi(\Lambda(w)) &= \frac{\hat{\pi}(\Lambda(W))}{\pi(\Lambda(W))}\ \pi(\Lambda(w)) \\ &= \frac{\pi(\Lambda(w))}{\pi(\Lambda(W))}\end{aligned}$$

Recall that in $\hat{C}$, the event of interest $\hat{Y} = y$ (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 $(Y = y, W) = \Lambda_y \cap \Lambda(W)$ (see proof of Lemma 1). So

$$\hat{\pi}(\hat{Y} = y) = \hat{\pi}(\Lambda_y) = \sum_{w \in W} \hat{\pi}(\Lambda(w))\ \hat{\pi}_{M_y}(w_\infty \mid w)$$

since the $\{w\}$ partition W (see proof of Lemma 1)

$$
\begin{aligned}
&= \sum_{w \in W} \hat{\pi}(\Lambda(w))\ \pi_{M_y}(w_\infty \mid w) \quad \text{using Definition 7 (3)} \\
&= \sum_{w \in W} \frac{\pi(\Lambda(w))}{\pi(\Lambda(W))}\ \pi_{M_y}(w_\infty \mid w) \\
&= \frac{\sum_{w \in W} \pi(\Lambda(w))\ \pi_{M_y}(w_\infty \mid w)}{\pi(\Lambda(W))} \\
&= \frac{\pi(\Lambda_y, \Lambda(W))}{\pi(\Lambda(W))} \\
&= \pi(\Lambda_y \mid \Lambda(W)) = \frac{\pi(Y = y, W)}{\pi(W)} \qquad \square
\end{aligned}
$$

Proof of Lemma 3. Consider the CEG $(C(W^1), \tilde{\Pi}(C(W^1)))$. As $W^1 = \{w_1\}$ partitions $C(W^1)$, we have

$$
\begin{aligned}
\tilde{\pi}(\Lambda(w^2)) &= \sum_{w^1 \in W^1} \tilde{\pi}(\Lambda(w^1))\ \tilde{\pi}(\Lambda(w^2) \mid \Lambda(w^1)) \\
&= \sum_{w^1 \in W^1} \frac{\pi(\Lambda(w^1))}{\pi(\Lambda(W^1))}\ \pi(\Lambda(w^2) \mid \Lambda(w^1))
\end{aligned}
$$

by Definition 8 (3) and (4)

$$
\begin{aligned}
&= \frac{\sum_{w^1 \in W^1} \pi(\Lambda(w^1), \Lambda(w^2))}{\Lambda(W^1)} \\
&= \frac{\pi(\Lambda(W^1), \Lambda(w^2))}{\Lambda(W^1)} \\
&= \pi(\Lambda(w^2) \mid \Lambda(W^1))
\end{aligned}
$$

As W^2 is C-regular, it is also $C(W^1)$-regular, so W^2 is *simple* in $C(W^1)$ if and only if $\tilde{\pi}(\Lambda(w^2))$ can be decomposed as $A(W^2)B(w^2)$ for all $w^2 \in W^2$ (where $A(W^2)$ is constant for all $w^2 \in W^2$), by Definition 6. The result then follows. $\square$

Proof of Theorem 1. W is simple conditioned on Z, so we can express $W = \bigcup_{z \in \{z\}} W_z$ where W_z is simple in $C(\Lambda_z)$. So by Lemma 3, for a position $w^2 \in W_z$ we have $\pi(\Lambda(w^2) \mid \Lambda_z) = A(W_z)B(w^2)$. Following the argument of the opening lines of the proof of Lemma 2, we have that

$$
\hat{\pi}(\Lambda(w^2) \mid \Lambda_z) = \frac{\pi(\Lambda(w^2) \mid \Lambda_z)}{\pi(\Lambda(W_z) \mid \Lambda_z)}
$$

Consider the event in $\hat{C}$ $(Z = z, W_z, \hat{Y} = y) \equiv \Lambda_z \cap \Lambda(W_z) \cap \lambda_y$. Analogously

with Lemma 2, we can express this as

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

where $M(w_0, w^1)$ is the union of all $\mu(w_0, w^1)$ subpaths, $M(w^1, w^2)$ is the union of all $\mu(w^1, w^2)$ 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

$$\begin{aligned} \hat{\pi}(Z = z, W_z, \hat{Y} = y) &= \hat{\pi}(\Lambda_z, \Lambda(W_z), \Lambda_y) \\ &= \sum_{w^1 \in \Lambda_z} \sum_{w^2 \in W_z} \hat{\pi}(\Lambda(w^1))\ \hat{\pi}(w^2 \mid w^1)\ \hat{\pi}_{M_y}(w_\infty \mid w^2) \end{aligned}$$

since $\{\Lambda_z\}$ and W are C-regular (see proofs of Lemmas 1 and 2). So

$$\hat{\pi}(\hat{Y} = y,\ W_z \mid Z = z) = \hat{\pi}(\Lambda_y, \Lambda(W_z) \mid \Lambda_z) \tag{A.1}$$

$$\begin{aligned} &= \frac{\sum_{w^1} \sum_{w^2} \hat{\pi}(\Lambda(w^1))\ \hat{\pi}(w^2 \mid w^1)\ \hat{\pi}_{M_y}(w_\infty \mid w^2)}{\hat{\pi}(\Lambda_z)} \\ &= \sum_{w^2} \left[\frac{\sum_{w^1} \hat{\pi}(\Lambda(w^1))\ \hat{\pi}(\Lambda(w^2) \mid \Lambda(w^1))}{\hat{\pi}(\Lambda_z)} \right] \hat{\pi}_{M_y}(w_\infty \mid w^2) \\ &= \sum_{w^2} \left[\frac{\sum_{w^1} \hat{\pi}(\Lambda(w^1), \Lambda(w^2))}{\hat{\pi}(\Lambda_z)} \right] \hat{\pi}_{M_y}(w_\infty \mid w^2) \\ &= \sum_{w^2} \left[\frac{\hat{\pi}(\Lambda_z, \Lambda(w^2))}{\hat{\pi}(\Lambda_z)} \right] \hat{\pi}_{M_y}(w_\infty \mid w^2) \\ &= \sum_{w^2} \hat{\pi}(\Lambda(w^2) \mid \Lambda_z)\ \hat{\pi}_{M_y}(w_\infty \mid w^2) \end{aligned} \tag{A.2}$$

$$\begin{aligned} &\sum_{w^2} \left[\frac{\pi(\Lambda(w^2) \mid \Lambda_z)}{\pi(\Lambda(W_z) \mid \Lambda_z)} \right] \hat{\pi}_{M_y}(w_\infty \mid w^2) \\ &= \frac{\sum_{w^2} \pi(\Lambda(w^2) \mid \Lambda_z)\ \pi_{M_y}(w_\infty \mid w^2)}{\pi(\Lambda(W_z) \mid \Lambda_z)} \end{aligned}$$

since our manipulation is amenable (using Definition 7 (3))

$$= \cdots = \frac{\pi(\Lambda_y, \Lambda(W_z) \mid \Lambda_z)}{\pi(\Lambda(W_z) \mid \Lambda_z)} \tag{A.3}$$

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 so far used no aspect of the topology of $\hat{C}$ which is not also true for C.

Now, by Definition 9 (2) $\Lambda_z \cap \Lambda(W_z) = \Lambda_z \cap \Lambda(W)$ in both C and $\hat{C}$. And, as in $\hat{C}$ all paths pass through W, we must have that $\Lambda_z \cap \Lambda(W) = \Lambda_z$ in $\hat{C}$. So

$$\begin{aligned}\hat{\pi}(\Lambda_y, \Lambda(W_z) \mid \Lambda_z) &= \hat{\pi}(\Lambda_y \mid \Lambda_z)\\ \pi(\Lambda_y, \Lambda(W_z) \mid \Lambda_z) &= \pi(\Lambda_y, \Lambda(W) \mid \Lambda_z)\\ \pi(\Lambda(W_z) \mid \Lambda_z) &= \pi(\Lambda(W) \mid \Lambda_z)\end{aligned}$$

and equation (A.3) becomes

$$\begin{aligned}\hat{\pi}(\Lambda_y \mid \Lambda_z) &= \frac{\pi(\Lambda_y, \Lambda(W) \mid \Lambda_z)}{\pi(\Lambda(W) \mid \Lambda_z)}\\ \Rightarrow\ \hat{\pi}(\hat{Y} = y) = \hat{\pi}(\Lambda_y) &= \sum_z \hat{\pi}(\Lambda_y \mid \Lambda_z)\ \pi(\Lambda_z)\\ &= \sum_z \left[\frac{\pi(\Lambda_y, \Lambda(W) \mid \Lambda_z)}{\pi(\Lambda(W) \mid \Lambda_z)}\ \pi(\Lambda_z)\right]\end{aligned}$$

$$= \sum_z \left[\frac{\pi(Y = y, W \mid Z = z)}{\pi(W \mid Z = z)}\ \pi(Z = z)\right] \qquad \square$$

Acknowledgements

This research is being supported by the EPSRC, grant no. EP/F036752/1.

References

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

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

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

[4] G. A. Churchill. Accurate restoration of DNA sequences. In C. Gatsaris et al., editors, *Case Studies in Bayesian Statistics*, volume 2, pages 90–148. Springer-Verlag, 1995.

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

[6] A. P. Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70:161–89, 2002.

[7] A. P. Dawid, J. Moertera, V. L. Pascali, and D. Van Boxel. Probabilistic Expert Systems for Forensic Inference from Genetic Markers. *Scandinavian Journal of Statistics*, 29:577–595, 2002.

[8] G. Freeman and J. Q. Smith. Bayesian model selection of Chain Event Graphs. Research Report, CRiSM, 2008.

[9] S. French, editor. *Readings in Decision Analysis.* Chapman and Hall / CRC, 1989.

[10] S. French and D. R. Insua. *Statistical Decision Theory.* Arnold, 2000.

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

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

[13] D. Heckerman. A Bayesian approach to Learning Causal Networks. In W. Edwards et al., editors, *Advances in Decision Analysis*, pages 202–220. CUP, 2007.

[14] R. Lyons. Random walks and percolation on trees. *Annals of Probability*, 18:931–958, 1990.

[15] A. M. Madrigal and J. Q. Smith. Causal identification in Design Networks. In L. E. Sucar et al., editors, *Advances in Artificial Intelligence 2.* Springer Verlag, 2004.

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

[17] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82:669–710, 1995.

[18] J. Pearl. *Causality: Models, Reasoning and Inference.* Cambridge, 2000.

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

[20] D. Poole and N. L. Zhang. Exploiting contextual independence in probabilistic inference. *Journal of Artificial Intelligence Research*, 18:263–313, 2003.

[21] E. M. Riccomagno and J. Q. Smith. The geometry of Causal probability trees that are algebraically constrained. In L. Pronzato and A. Zhigljavsky, editors, *Optimal Design and related areas in Optimization and Statistics*, pages 131–152. Springer, 2008.

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

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

[24] J. Q. Smith and P. E. Anderson. Conditional independence and Chain Event Graphs. *Artificial Intelligence*, 172:42–68, 2008.

[25] J. Q. Smith and P. A. Thwaites. Decision trees. In E. L. Melnick and B. S. Everitt, editors, *Encyclopedia of Quantative Risk Analysis and Assessment*. Wiley, 2008.

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

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

[28] P. A. Thwaites and 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*, pages 2339–2347, Paris, 2006.

[29] P. A. Thwaites, J. Q. Smith, and R. G. Cowell. Propagation using Chain Event Graphs. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, Helsinki, 2008.