THE CAUSAL MANIPULATION OF CHAIN EVENT GRAPHS

By Eva Riccomagno

Department of Mathematics, Università degli Studi di Genova

AND

By Jim Q. Smith

Department of Statistics, The University of Warwick

Discrete Bayesian Networks (BN's) have been very successful as a framework both for inference and for expressing certain causal hypotheses. In this paper we present a class of graphical models called the chain event graph (CEG) models, that generalises the class of discrete BN models. It provides a flexible and expressive framework for representing and analysing the implications of causal hypotheses, expressed in terms of the effects of a manipulation of the generating underlying system. We prove that, as for a BN, identifiability analyses of causal effects can be performed through examining the topology of the CEG graph, leading to theorems analogous to the back-door theorem for the BN.

1. Introduction. Bayesian networks have now been extended to Causal Bayesian Networks (CBN's) using a non-parametric representation based on structural equation models [11, 17, 18, 30]. These 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. Motivated by comments in [25], we develop an alternative graphical

1

representation of a causal model, called *chain event graph model*. This is constructed from an event tree together with a set of ex-changeability assumptions. It can be seen as a generalisation of a probability graph [4, 25] and typically has many less nodes than the original event tree. It was introduced in [28] in parallel with the present paper. In [28] analogues to d-separation theorems that give sufficient conditions for determining whether a conditional independence statement holds, are given for CEG. Here a causal extension of CEG models is discussed which is as transparent and compelling as the extension from BN to CBN is.

We refer to the introduction in [28] for a comparison of CEG models with other frameworks for propagations and for interrogating an elicited model like probability decision graphs [8, 13], context-specific networks [3], cofactors [21] and case-factor diagram [16]. Here briefly we outline some of the reasons why chain event graph models are important.

In some applications e.g. in Bayesian decision analysis [9], risk analysis [2], physics [15], biological regulation [5], often the first stage of the elicitation of a model is based on the elicitation of an event tree. Here is an example of the type of context we have in mind.

EXAMPLE 1.1. The police hold a suspect S who they believe threw a brick through the kitchen window and stole a large quantity of money. The police hope to bring S to court (indicator X_1) but might, for certain legal technicalities, be forced to release him. 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 will identify S (indicator X_5) and whether S will be convicted C or released R (the "effect" indicator of interest X_6). Unless the suspect is identified by the witness as the one who threw the brick, S will not be convicted. The glass match is believed only to depend on whether S was present at the crime scene or not whilst the quality of the witness identification is believed to depend on whether or not S was as at the scene of the

crime as well as on whether or not he threw the brick. The police will later learn whether they have to release S before trial, whether the witness will identify S and the results of the forensic test.

Such a problem has an event tree representation. However this is rather cumbersome. Recently it has been more usual to represent this problem using a BN. Thus given S is not released (i.e. conditional on $X_1 = 1$) the following BN is consistent with the possible unfoldings of events as described above

$$
(1) \qquad
\begin{array}{ccccc}
X_2 & \rightarrow & X_3 & \rightarrow & X_5 \\
& \searrow & \downarrow & & \downarrow \\
& & X_4 & \rightarrow & X_6
\end{array}
$$

EXAMPLE 1.2 (Continuation of Example 1.1).   Although the graphical representation in (1) is illuminating, it is not ideal. First, it is partial because the sample space that includes X_1 is not naturally a product space. Thus if 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. Of course we can formally define X_4 and X_5 conditional on $X_1 = 0$ so that a BN is consistent with this story. One such candidate is given

$$
\begin{array}{ccccccc}
X_2 & \rightarrow & X_3 & \rightarrow & X_5 \\
& \searrow & \downarrow & & \downarrow \\
& & X_4 & \rightarrow & X_6 & \leftarrow & X_1
\end{array}
$$

This stores well probabilities. However, as a qualitative representation of the the possible unfoldings of events—perhaps used as the basis for embodying causal conjectures—it is not ideal. See also Example 3.5 below.

But suppose we use this BN representation of the problem. Second, we note that it only conveys certain aspects of the story. For example the fact that S could only throw the brick $\{X_3 = 1\}$ if he were present at the scene of the crime $\{X_2 = 1\}$ and that conviction $\{X_6 = 1\}$

requires $\{X_5 = 1\}$ are not expressed in the diagram. Moreover we might like to incorporate into the representation context specific information because it is informative about various causal hypotheses (see e.g. [7]). This is particularly so of models of biological regulatory mechanisms which typically contain many noisy "and" and "or" gates [1]. Third, we might well be interested in the causal effect of, for example, forcing the witness to identify S as the culprit $\{X_5 = 1\}$ if a match $\{X_4 = 1\}$ in the glass is found. This is not represented in the usual semantics of the BN above. Furthermore it will not necessarily be the case that the manipulations we want to represent correspond to setting the original variables in the idle system to certain values.

The deficiencies of BN's as expressed in Example 1.2 should not be overemphasised. As stated in [7] there is an art to drawing the appropriate BN of a problem and it is sometimes necessary to redefine the variables defining the problem or add more edges on the graph to aid representation. For example were we to add an edge between X_4 and X_5 then the manipulation described in Example 1.2 can be expressed as a contingent decision (although we then have lost some information in the representation). Furthermore we can often transform the variables in a BN not only to encode more information but also so that the manipulation can be seen as setting this random variable to a value. Nevertheless the BN, whilst being consistent with a story like the one above, will still only be a partial representation in general. The CEG in Figure 1 gives a more expressive graphical representation, able to tell more of the story, which like the BN supports entirely graphical inferences about irrelevances and the potential effects of certain causal manipulations, albeit at the cost of some simplicity.

The partially ordered sequence in which events unfold as expressed by the event tree is retained in the CEG construction. Event trees explicitly acknowledge asymmetries embedded in a structure both in its development and in its sample space. These are retained

in the CEG. Recall that the dimension of the sample space often is critical for determining identifiability especially when there are hidden variables, even in symmetric models [26, 27, 29].

In probability trees conditional independence relations can be embedded through equations linking probabilities labelling the edges of the tree. These are used to construct the vertices of the CEG and its undirected edges. Many conditional independence statements, including all those associated with BN's, many context specific BN's, and also over functions of the variables (such as noisy and/or gates), can be stated as equality of distributions associated with vertices on the underlying tree. This means that all such statements are expressed *explicitly* via the topology of the CEG. See Example 2.7 for complete independence models.

A basic assumption of our definition of causation (see Definition 3.1) is the belief that intervention or manipulation of one or more vertices of the tree/CEG can model external intervention on the underlying process being modelled. This is analogous to the do-operator in [18] and is consistent with the representation of the effect of a cause on a tree model in [25], albeit there not necessarily as a result of a manipulation. It allows us to include naturally information relative to the background idle system in the analysis of causal effects.

An advantage of the use of a framework based on event trees, and hence of CEG's, is that causal hypotheses can be explicitly separated from any direct link with the measurement process as will be illustrated later.

For a good discussion of many of the above points see [25], in particular on the advantages of event trees (and hence CEG's) for coding asymmetrical problems, and as powerful expression of an observer's beliefs especially when those beliefs are based on an underlying conjecture about a causal mechanism

Section 2 contains the basic terminology and definitions. The extension from event trees

and CEG's to causal probability trees and causal CEG is in Section 3. A theorem of identifiability analogous to the back-door theorem is proved in Section 4. Motivating and illustrative examples are presented throughout.

2. Chain event graphs.   From a Bayesian perspective probability trees describe the observer's beliefs about what will happen as events unfold. The edges out of a node v of the probability tree represent the possible unfolding that can occur from the situation labelled by v, or equivalently the event space of a random variable that can be indexed by v. The sample space of the experiment at a node is given by the branches of the probability tree at that point.

Through two equivalence relations on the nodes of the tree, we construct a new model structure, called a chain event graph, which includes Bayesian Networks and which provides a natural framework for defining causality.

We start with some definitions to set up notation and formalise ideas. Some of these definitions are slightly non-standard for reasons that will become apparent later in the paper. Section 2.1 draws strongly from [24, 25, 28] and we refer to those works for further details.

2.1. *Probability trees.*   The model structure upon which a manipulation operation is defined in Section 3 is a graph constructed from a rooted, directed (event) tree $\mathcal{T} = (V(\mathcal{T}), E(\mathcal{T}))$ where $V(\mathcal{T})$ and $E(\mathcal{T})$ are the set of vertices or nodes and of edges, respectively. Write $V = V(\mathcal{T})$ and $E = E(\mathcal{T})$ when there is no ambiguity and assume V and E finite. The single root node is denoted by v_0. Between two vertices there is at most on edge. Each edge $e \in E$ is directed with a parent node v and a child node v'. We write $e = (v, v')$ and note that this edge can be identified with its child vertex v'.

DEFINITION 2.1.   *For $v \in V$ let $\mathbb{X}(v) = \{v' \in V$ such that there exists $e \in E$ such that*

$e = (v, v')\}$ *and call* $\mathbb{X}(v)$ *the set of children of* v.

Thus $\mathbb{X}(v)$ is in one-to-one correspondence with the set of edges out of v. If $\mathbb{X}(v)$ is the empty set, then v is called a *leaf node*, otherwise it is called a *situation*. The set of situations, $S = S(\mathcal{T}) \subset V(\mathcal{T})$, will have particular significance. Note that $\{\mathbb{X}(v) : v \in S\} \cup \{v_0\}$ partitions V.

A *path* between two vertices v and v' is an ordered sequence of nodes $\lambda = \lambda(v, v') = (v_1, \ldots, v_{n[\lambda]+1})$ where $v_1 = v$, $v_{n[\lambda]+1} = v'$ and v_k is the child of v_{k-1} and the parent of v_{k+1} for $k = 2, \ldots, n[\lambda]$. This path can equivalently be identified with the ordered sequence of its edges $\lambda = (e_1, \ldots, e_{n[\lambda]})$ where $e_k = (v_k, v_{k+1})$ for $k = 1, \ldots, n[\lambda]$. The number $n[\lambda]$ of edges in called the *length* of the path. We write $v \in \lambda$ whenever λ contains the vertex v.

DEFINITION 2.2. *Let* $\mathbb{X}(\mathcal{T}) = \mathbb{X} = \{\lambda(v_0, v) : v \in V \backslash S\}$ *be the set of root-to-leaf paths. The elements of* $\mathbb{X}$ *are called* atomic events *of* $\mathcal{T}$.

Clearly, $\mathbb{X}(\mathcal{T})$ is in one-to-one correspondence with the leaves of $\mathcal{T}$. Paths determine a partial order on $V(\mathcal{T})$. A probability tree could be drawn so that the situations along each root-to-leaf path correspond to one of the possible historical developments of the modelled problem. In some cases this directionality is inherent to the problem to be modelled and expresses a conjecture about the ordering in which one situation follows another (see in particular [25, Section 2.8]).

2.2. *Primitive probabilities.* Next we impose a probability structure on $\mathcal{T}$. Consider $\mathbb{X}$ as sample space, the power set $2^{\mathbb{X}}$, and a probability P on $(\mathbb{X}, 2^{\mathbb{X}})$. Thus $\mathrm{P}(\lambda)$ is the probability of the path $\lambda \in \mathbb{X}$. A *probability tree* is a directed tree $\mathcal{T} = (V, E)$ such that to each situation $v \in S(\mathcal{T})$ is associated a discrete random variable $X(v)$ whose sample space is $\mathbb{X}(v)$. The

distribution of $X(v)$, $v \in S(\mathcal{T})$, determines the *primitive probabilities*

$$\pi(e) = \pi(v'|v) = \pi(v') = \mathrm{P}(X(v) = v') \text{ for } v' \in \mathbb{X}(v), \text{ for } e = (v, v')$$

which are fundamental in the paper. Let $\Pi_a(\mathcal{T})$, or simply Π_a, denote the set of all primitive probabilities and $\pi(e)$ is the label or colour of e.

The random variable $X(v)$ can be interpreted as a function on $\mathbb{X}$ which assigns zero probability to the set of paths not through v. Furthermore, for $\lambda \in \mathbb{X}$ the formula $X(v)(\lambda) = v'$ states that $X(v)$ maps λ into the path through v and v', hence the notation $\pi(v'|v)$. This gives an interpretation of $X(v)$ as a function from $\mathbb{X}$ to $\mathbb{X}(v)$. Example 2.5 provides another motivation for this notation. For every situation $v \in \mathcal{S}(\mathcal{T})$ the sum-to-one condition can be written as

$$(2) \qquad \sum_{v' \in \mathbb{X}(v)} \mathrm{P}(X(v) = v') = \sum_{v' \in \mathbb{X}(v)} \pi(v'|v) = 1.$$

Random variables on vertices along a path are required to be mutually independent. It then follows that the probability $\mathrm{P}(\lambda)$ of the atomic event $\lambda = (e_1, \ldots, e_{n[\lambda]}) \in \mathbb{X}$ is the following product of primitive probabilities

$$(3) \qquad \prod_{j=1}^{n[\lambda]} \pi(e_j) = \mathrm{P}(\lambda).$$

Furthermore if $\pi(v'|v) = 0$ for some v, v' then any atomic event including v' has zero probability of occurring. In some circumstances the branch starting at v' could be deleted from the tree.

Above we have defined probabilities of the atomic events. Subsequently we defined some random variables over the probability space $(\mathbb{X}, 2^{\mathbb{X}}, \mathrm{P})$ and finally used these random variables to define the primitive probabilities Π_a. In practice when modelling a problem over a probability tree we would usually work the other way round. While constructing the tree from root to leaves the modeller has a (possibly not definite) idea of the distribution of

the random variables sitting on situations. Then Equation (3) is used to determine the probabilities of the atomic events. The P's and π's provide equivalent parametrisations of the probability tree model under the assumption of strict positivity. We do not discuss this further, but give a simple example to show an algebraic reason by which we prefer to parametrise the probability space $(\mathbb{X}, 2^{\mathbb{X}}, P)$ using the primitive probabilities Π_a.

EXAMPLE 2.1. For the tree in Figure 6 Equations (3) are in the left-hand-side (LHS) of the tableau below

$$
\begin{array}{ll|ll}
p_2 = 1 - \pi_1 & p_5 = \pi_1(1 - \pi_3 - \pi_4) & \pi_1 = 1 - p_2 & \pi_4 = \dfrac{p_4}{1 - p_2} \\[2mm]
p_4 = \pi_1\pi_4 & p_7 = \pi_1\pi_3(1 - \pi_6) & \pi_3 = \dfrac{1 - p_2 - p_4 - p_5}{1 - p_2} & \\[2mm]
p_6 = \pi_1\pi_3\pi_6 & & \pi_6 = \dfrac{p_6}{1 - p_2 - p_4 - p_5} &
\end{array}
$$

where p_i is the probability of the path starting at the root vertex and ending in the vertex v_i, $i \in \{1, \ldots, 7\}$ and $\pi_i = \pi(v_i)$. The equations in the LHS of the tableau give the probability of an atomic event λ as a polynomial in the primitive probabilities of degree equal to the length of the path λ.

The RHS of the tableau gives the primitives in terms of the atomic event probabilities. These are ratios of polynomials of degree one in the paths probabilities and are defined if the denominators are not zero.

Note that $p_6 + p_7 = \pi_1\pi_3$ is the probability of the path $\{v_0, v_2, v_3\}$, and that $\pi_4 = \dfrac{p_4}{1 - p_2} = \dfrac{P(X(v_0) = v_1, X(v_1) = v_4)}{P(X(v_0) = v_1)} = P(X(v_1) = v_4 | X(v_0) = v_1)$. Hence the notation $\pi_4 = \pi(v_4|v_1)$ is used to underline its interpretation as the probability of reaching v_4 having arrived in v_1. See also Shafer [25]. We will see later that this directional parametrization is a natural one to use when modelling causal hypothesis just as the directed parametrization of the BN naturally projects into the CBN.

The primitive probabilities, which might be unknown or partially known, satisfy some

logical constraints which often are algebraic. Obvious ones are $0 \leq \pi(e) \leq 1$ and the linear polynomials in Equations (2). Supplementary constraints might be inherent to the logic of the problem been modelled and others might be imposed by the modeller, e.g. in Figure 6 the modeller might know that $X(v_1)$ follows a Binomial distribution with $P(X(v_1) = v_3) = s^2$ and $P(X(v_1) = v_5) = (1-s)^2$ for some $s \in [0, 1]$. The definition of one class of such constraints leads to the notion of chain event graphs below.

EXAMPLE 2.2 (Continuation of Example 2.1). Assume the primitive probabilities and the path probabilities are all unknown indeterminates and consider the polynomial ideal generated by $G = \{p_2 - (1 - \pi_1), p_5 - \pi_1(1 - \pi_3 - \pi_4), p_4 - \pi_1\pi_4, p_7 - \pi_1\pi_3(1 - \pi_6), p_6 - \pi_1\pi_3\pi_6\}$. This is the infinite set of polynomials of the form $\sum_{g \in G} s_g g$ where s_g is any polynomial in $p_2, p_4, p_5, p_6, p_7, \pi_1, \pi_3, \pi_4, \pi_6$. The elimination ideal of the π variables is generated by the polynomial condition $\{p_2 + p_5 + p_4 + p_7 + p_6 - 1\}$, obviously. The constraint $\pi_4 = \pi_3\pi_6$ is considered by adjoining the polynomial $\pi_4 - \pi_3\pi_6$ to G. The elimination ideal contains, obviously, the polynomial $p_4 = p_6$. The propagation of polynomial constraints on the π's and p's is imposed by adjoining the corresponding polynomials to G and computing the relevant elimination ideal. This makes CEG's analysis amenable of techniques from algebraic statistics [20]. Note that the manipulations defined in Section 3.4 are polynomial constraints setting some π's equal to zero.

2.3. *Stages.*

DEFINITION 2.3. *Let $\mathcal{T}$ be a probability tree and $\Pi_a(\mathcal{T})$ an associated set of primitive probabilities. Two situations $v_1, v_2 \in S(\mathcal{T})$ are said to be* stage-equivalent *if and only if $X(v_1)$ and $X(v_2)$ have the same distribution.*

Definition 2.3 requires that there exists a one-to-one map $\mu : \mathbb{X}(v_1) \to \mathbb{X}(v_2)$ and that $\pi(v|v_1) = \pi(\mu(v)|v_2)$ for all $v \in \mathbb{X}(v_1)$. It determines an equivalence relation on $\mathcal{S}(\mathcal{T})$ whose

equivalence classes are called *stages*. For each stage u define

$$\Pi(u) = \{\pi(v'|v) : v' \in \mathbb{X}(v) \text{ for some } v \text{ representative of } u\}$$

and $\Pi = \Pi(\mathcal{T}) = \bigcup_{u \in L(\mathcal{T})} \Pi(u)$. The set of primitive probabilities Π is no larger than the set of all primitive probabilities Π_a and clearly still sufficient for determining the distribution of all random variables measurable with respect to $(\mathbb{X}, 2^{\mathbb{X}})$. The pair $(\mathcal{T}, \Pi(\mathcal{T}))$ is called a *probability tree model*.

2.4. *Examples.*

EXAMPLE 2.3.   Suppose that only three incidents A, B, C can occur and the order of occurrence of B and C is relevant and contingent on whether A or not A ($\bar{A}$) happens. Their history unfolds according to the probability tree in Figure 7. For example the path (v_0, v_1, v_3) represents whether after A has occurred, B occurs. The random variable $X(v_0)$ is the indicator function of the event of the incident A happening; $\pi(v_1)$ is the probability of this event and the primitive $\pi(v_3)$ is the probability that B occurs if A has occurred.

Henceforth we shall assume that the tree fully represents all possible unfoldings of situations. The unfolding "if first A and then B, then C" is simply not part of our story. The probability of the incident of A does not happen and then B does, is given by $\pi(v_2)\pi(v_6) + \pi(v_2)\pi(v_7)\pi(v_9)$ being the sum of the probabilities of the atomic events $\{v_0, v_2, v_6\}$ and $\{v_0, v_2, v_7, v_9\}$. Note that the path σ-algebra is not the σ-algebra generated by $\{A, B, C\}$ where these are thought of as events, since this σ-algebra cannot express the ordering of incidents.

Figure 7 could represent the following situation. A woman with epilepsy happens not want to conceive (A) or wants to conceive ($\overline{A}$). She has three alternatives: to take medicine B, to take medicine C or to follow some other health regime, $\overline{B \cup C}$. Medicine C has less

side effects than B (thus we shall impose $\pi(v_4) >> \pi(v_3)$) but it might have an adverse effect on the formation of the nervous system of the foetus if she becomes pregnant, thus $\pi(v_7) < \pi(v_4)$. If she discovers pregnancy within the first three months from conception then the medicine B can be taken as a supplement to C to reduce greatly this adverse effect. But unless she discovers pregnancy she would never take B after having taken C because of other health risks to herself linked to the combined use of B and C.

EXAMPLE 2.4 (Continuation of Example 2.3). The modeller might want to assert that $\pi(v_3) = \pi(v_7)$ and $\pi(v_4) = \pi(v_6)$. This assigns v_1 and v_2 to the same stage. This assertion implies two things. First if the woman is not taking either B or C then the probability of changing her health regime is the same whether or not she wants to become pregnant. Second if she wants to become pregnant then she will prefer to take C as much as she is likely to take B if she does not want to become pregnant.

EXAMPLE 2.5. Let X and Y be two binary random variables. Let $X = 1$ be the event that a person watches a violent movie on Saturday night and $Y = 1$ the event of that person getting into a fight on Saturday night. In Figure 8 $X(v_0) = X$, $X(v_1) = (Y|X = 1)$ and $X(v_2) = (Y|X = 0)$. The path (v_0, v_1, v_3) corresponds to having watched a violent movie and ending up in a fight and the path (v_0, v_2, v_5) to not having watched a violent movie and ending up in a fight. The three primitive probabilities $\pi(v_1)$, $\pi(v_3)$ and $\pi(v_5)$ are sufficient to parametrise the model.

EXAMPLE 2.6 (Continuation of Example 2.5). In Figure 8 v_1 and v_2 can be in the same stage in two ways: (1) $\pi(v_3) = \pi(v_5)$ and (2) $\pi(v_3) = \pi(v_6)$. In (1) we assume that X and Y are independent, i.e. watching a violent movie has no effect on the probability of subsequent violence; while in (2) we assume that $P(Y = 0|X = 0) = P(Y = 1|X = 1)$, i.e. the probability that violence occurs after seeing violence can be equated with the probability

of non-violence occurring after not seeing violence—a type of conservation of violence law. If the values X and Y take were -1 and 1 then this last equality would imply the independence of X and XY.

2.5. *Chain event graphs.* In this section we assume that the observer is able to express two pieces of qualitative information: the topology of the probability tree and its stages. These two sources of information can be fully represented using a mixed graph called a chain event graph. Mixed means that the graph has directed and undirected edges. Directed edges are labelled with primitive probabilities while undirected edges are not labelled.

Let $(\mathcal{T}, \Pi)$ be a probability tree model. For a situation $v \in S$ let $\mathcal{T}(v)$ be the sub-trees starting at v. In [14] $\mathcal{T}(v)$ is called the subgraph induced by v and its ancestors. Let Π_v be the subset of Π labelling edges in $\mathcal{T}(v)$. Then $(\mathcal{T}(v), \Pi_v)$ is a probability tree model.

DEFINITION 2.4. *Two situations v and v^* in the probability tree $(\mathcal{T}, \Pi)$ are equivalent if and only if*

1. *$\mathcal{T}(v)$ and $\mathcal{T}(v^*)$ are isomorphic. That is, there exists a map μ from the sets of vertices in $\mathcal{T}(v)$ and $\mathcal{T}(v^*)$ such that $(\mu(v_1), \mu(v_2))$ is an edge in $\mathcal{T}(v^*)$ if and only if (v_1, v_2) is an edge in $\mathcal{T}(v)$, and*

2. *for every w situation in $\mathcal{T}(v)$, w and $\mu(w)$ are in the same stage.*

The induced equivalence classes are called positions *and $K(\mathcal{T})$ is the set of positions.*

Item 2. in Definition 2.4 simply means that $\pi(v_2|v_1) = \pi(\mu(v_2)|\mu(v_1))$ for all possible $v_1, v_2 \in \mathcal{T}(v)$. In Definition 2.4 we require that the sub-trees are topologically isomorphic and that their edge probabilities match according to the isomorphism, whether they are known fixed values, unknown fixed values or indeterminates. Clearly the partition of situations into positions is a refinement of the partition into stages: if two situations are in

the same position then they are in the same stage. Two situations are in the same position when the processes governing the stories unfolding from them are believed to be governed by processes with the same distribution. Note that this is a predictive, not a retrospective equivalence. Once a unit reaches the situation v or the situation v^*, all pairs of possible unfoldings from v and v^* occur with the same probabilities. It is in this respect that positions are natural objects on which to describe a causal manipulation.

In broad terms, stages are used for estimation as they reduce the number of parameters (primitive probabilities) and positions are used to express conditional independence statements [28]. Positions are used to form the vertices of a new graph called the chain event graph. Its undirected edges join positions at the same stage and its directed paths correspond to root-to-leaf paths of the probability tree model.

DEFINITION 2.5.   *Let $(\mathcal{T}, \Pi)$ be a probability tree model. Its chain event graph, $\mathcal{C}(\mathcal{T})$, is the mixed graph $(V(\mathcal{C}(\mathcal{T})), E_d(\mathcal{C}(\mathcal{T})), E_u(\mathcal{C}(\mathcal{T})), \Pi(\mathcal{C}(\mathcal{T})))$ where*

1. $V(\mathcal{C}(\mathcal{T})) = K(\mathcal{T}) \cup \{w_\infty\}$ is the vertex set. The vertex w_∞ is called the sink *vertex.*

2. $E_d(\mathcal{C}(\mathcal{T}))$ is a multi-set of directed edges and is partitioned into two sets, $E_1(\mathcal{C}(\mathcal{T}))$ and $E_2(\mathcal{C}(\mathcal{T}))$ constructed as follows. For each $w \in K(\mathcal{T})$ choose $v \in V(\mathcal{T})$ a representative of w. For each (v, v') edge in $E(\mathcal{T})$

    (a) if v' is in position w' then add a directed edge from w to w' to the multi-set $E_1(\mathcal{C}(\mathcal{T}))$,

    (b) if v' is a leaf node then add a directed edge from w to w_∞ to the multi-set $E_2(\mathcal{C}(\mathcal{T}))$.

3. $E_u(\mathcal{C}(\mathcal{T}))$ is a set of undirected edges joining positions in the same stage, namely

$$
\begin{aligned}
E_u(\mathcal{C}(\mathcal{T})) \quad = \quad & \{\{w, w'\}: \text{ with } w \neq w' \text{ and } \text{ there exist } v, v' \in S(\mathcal{T}), \\
& u \in L(\mathcal{T}) \text{ with } v, v' \in u \text{ and } v \in w, v' \in w'\}.
\end{aligned}
$$

4. $\Pi(\mathcal{C}(\mathcal{T})) = \Pi$ *and if* $e_1, e_2 \in E_d(\mathcal{C}(\mathcal{T}))$ *and* $\pi(e_1) = \pi(e_2)$ *in the original tree, then* e_1 *and* e_2 *have the same label in the CEG.*

When there is no ambiguity we write $(V_{\mathcal{C}}, E_d, E_u, \Pi)$ and $\mathcal{C}$ for the chain event graph. The CEG fully expresses the structure $\mathbb{X}$ of the sample space of a tree because there are as many directed root-to-leaf paths in the original tree as there are root-to-sink paths in its CEG. However often it has many less edges.

If two vertices v and v^* of the original tree are in the same position, then for each path $\lambda(v, v_M)$ in the sub-tree $\mathcal{T}(v)$ there exists a corresponding path $\lambda^*(v^*, v_M^*)$ in $\mathcal{T}(v^*)$ along which the same evolutions occurs. This implies that $P(\lambda) = P(\lambda^*)$. In particular consider the root-to-leaf paths, given in terms of vertices, $\lambda(v_0, \dots, v, \dots, v_M)$ and $\lambda^*(v_0, \dots, v^*, \dots, v_{M^*}^*)$ where v_M and $v_{M^*}^*$ are leaves in $\mathcal{T}$ and v, v^* are in the same position. Then

(4) $\qquad P(\lambda) = P(\lambda(v_0, v)) P(\lambda(v, v_M))$ and $P(\lambda^*) = P(\lambda^*(v_0, v^*)) P(\lambda(v, v_M))$.

The same formula holds when the paths are considered in the CEG. In this case v_0 is substituted by the root node and v_M and $v_{M^*}^*$ become the sink node.

EXAMPLE 2.7 (Continuation of Example 2.5). Figure 9 gives the CEG when v_1 and v_2 are in the same stage. The values of the edge labels indicates whether it is model (1) or (2).

EXAMPLE 2.8. Figures 10 and 11 give a tree and its CEG for the stage partition $\{\{v_0\}, \{v_1, v_3, v_{13}, v_{17}\}, \{v_2, v_7\}, \{v_5, v_9\}, \{v_{19}\}\}$ and the position partition $\{\{v_0\}, \{v_1, v_3\}, \{v_5, v_9\}, \{v_2\}, \{v_7\}, \{v_{13}\}, \{v_{17}\}, \{v_{19}\}, w_\infty\}$.

EXAMPLE 2.9 (Continuation of Example 2.4). The positions are $\{v_0\}, \{v_1\}, \{v_2\}, \{v_7\}, w_\infty$ and the CEG is in Figure 12, where $\pi_i = \pi(v_i)$.

EXAMPLE 2.10 (Bayesian network). In [28] it is proved that any discrete Bayesian network on the random variables $\{X_1, \ldots, X_n\}$ can be expressed by a CEG. Example 2.11 below shows how to retrieve the conditional independence statements of the BN from the topology of any of its CEG's. Like context specific BNs [3] but unlike the probability decision graph [13] or the probability graph [4], the CEG provides a generalisation of the BN. In particular two situations v_{i-1} and v'_{i-1} are in the same stage u_i if, and only if, the values of their parents agree. This fully expresses the conditional independence statement embodied in the BN.

Topological characteristics of a CEG derived from a discrete BN include: *(i)* all the root-to-sink paths have the same length, *(ii)* the stages consist of situations all of whose distances (length of the path from the root to the situation) from the root are the same, and *(iii)* for $2 \leq i \leq n$ all stages u_i associated with different configurations of parents of X_i contain exactly the same number of situations. Examples and d-separation theorems for CEG's are given in [28].

EXAMPLE 2.11. Consider the binary BN $X_2 \leftarrow X_1 \rightarrow X_3$. Its CEG is in Figure 13. Note that the statements that can be read from the topology of this CEG are that the two situations $(\{X_1 = 0, X_2 = 0\}, \{X_1 = 0, X_2 = 1\})$ and the two situations $(\{X_1 = 1, X_2 = 0\}, \{X_1 = 1, X_2 = 1\})$ are in the same stages (respectively $[0,0] \cup [0,1]$ and $[1,0] \cup [1,1]$). From the definition of a stage, this means the two conditional statements

$$
\begin{aligned}
(5) \qquad P(X_3 = 1 | \{x_1 = 0, x_2 = 0\}) &= P(X_3 = 1 | \{x_1 = 0, x_2 = 1\}) \\
P(X_3 = 0 | \{x_1 = 0, x_2 = 0\}) &= P(X_3 = 0 | \{x_1 = 0, x_2 = 1\})
\end{aligned}
$$

and

$$
\begin{aligned}
(6) \qquad P(X_3 = 1 | \{x_1 = 1, x_2 = 0\}) &= P(X_3 = 1 | \{x_1 = 1, x_2 = 1\}) \\
P(X_3 = 0 | \{x_1 = 1, x_2 = 0\}) &= P(X_3 = 0 | \{x_1 = 1, x_2 = 1\})
\end{aligned}
$$

This is synonymous with the statement $X_3 \perp\!\!\!\perp X_2 | X_1$. The CEG in Figure 14 contains the statement (5) but not necessarily (6), which cannot be expressed using a BN.

3. Manipulation and Causality.

3.1. *Manipulations.* A CEG provides a flexible framework for expressing what might happen were a model to be manipulated in certain ways or made subject to some control. Of course as Shafer [25] similarly argues for probability trees, the validity of such a framework is heavily dependent on context. Some discussions of notions of the manipulation of a system and intervention and various applications can be found in [12, 18, 25, 30]. Here we follow Pearl [18]: a model for the manipulation is developed and the issue of suitability of such manipulation for the application under study is left to practical considerations. Recall briefly the standard definition of "do"-operator, which is fundamental in [18]. The joint density function of a set of random variables $X_1, \ldots, X_n$ with sample spaces $\mathbb{X}_1, \ldots, \mathbb{X}_n$, factorises according to a directed acyclic graph (DAG)

$$(7) \qquad p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | pa(x_i))$$

where $pa(x_i)$ are the parents of X_i in DAG language. A random variable is forced to assume a certain value with probability one, say $X_j = \widehat{x}_j$ for some $j \in \{1, \ldots, n\}$ and $\widehat{x}_j \in \mathbb{X}_j$. A new joint density, $p(\cdot || \widehat{x}_j)$, is defined on $\{X_1, \ldots, X_n\} \setminus \{X_j\}$ by the formula

$$(8) \qquad p(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n || \widehat{x}_j) = \prod_{i=1, i \neq j}^{n} p(x_i | \widehat{pa(x_i)})$$

where $\widehat{pa(x_i)}$ is the subset of parents of x_i for which $x_j = \widehat{x}_j$. This formula expresses the effect of the manipulation $X_j = \widehat{x}_j$.

A manipulation of a probability tree [of a CEG] can be defined in an analogous manner by modifying the distributions of some of the random variables sitting on situations [positions].

DEFINITION 3.1. *Let $(\mathcal{T}, \Pi(\mathcal{T}))$ be a probability tree model and $D \subset S$ a subset of situations of the tree. A* manipulation *of the tree is a pair $(D, \widehat{\Pi}_D)$ where $\widehat{\Pi}_D = \{\widehat{\pi}(v'|v) : v \in D$ and $v' \in \mathbb{X}(v)\}$ and $\{\widehat{\pi}(v'|v) : v' \in \mathbb{X}(v)\}$ is a new distribution for $X(v)$.*

The effect *of this manipulation is the transformation $(\mathcal{T}, \Pi) \overset{\widehat{P}}{\rightarrow} (\mathcal{T}, \widehat{\Pi}_D)$ where*

$$
(9) \qquad \widehat{P}(X(v) = v') = \begin{cases} \pi(v'|v) & \text{if } v \notin D \\[2mm] \widehat{\pi}(v'|v) & \text{if } v \in D \end{cases}
$$

for $v' \in \mathbb{X}(v)$. The manipulated tree *is the probability tree model so obtained. The* manipulated CEG *is the CEG of the manipulated tree.*

Definition 3.1 allows large classes of intervention, some of which are illustrated in Example 3.1.

EXAMPLE 3.1 (Continuation of Example 2.1).   (1.) Fix some values of the primitive probabilities or of a function of them, e.g. $\pi_1 = 1$, $\pi_3 + \pi_4 = 0.5$, $p_4 = \pi_1 \pi_4 = 0.5$. (2.) Impose (polynomial) constraints on primitive probabilities, e.g. $\pi_3 = 2\pi_4$, $\pi_1 = \pi_3 = \pi_6$. (3.) Assume that the distribution of the random variable sitting on some situation is from a parametric family, e.g. $X(v_1)$ follows a Binomial distribution with $P(X(v_1) = v_3) = s^2$, $P(X(v_1) = v_4) = 2s(1-s)$ and $P(X(v_1) = v_5) = (1-s)^2$ for $s \in [0,1]$. (4.) In the idle tree $X(v_1)$ follows a Binomial distribution with $\pi_3 = s^2$ and $\pi_5 = (1-s)^2$ and in the manipulated tree $X(v_1)$ has a uniform distribution with $\pi_3 = \pi_4 = \pi_5 = 1/3$.

Here the distinction between intervention and constraint is not a mathematical one. A probability tree model is assigned together with extra information like its stages or some logical constraints (sum-to-one) or experimental regimes (women are randomly assigned to treatment B or C in Example 2.3). These pieces of information impose some constraints on the primitive probabilities. Then, the system is modified by changing one or more of the distribution of the random variables on situation. There might be no values of the

primitive probabilities that satisfy both the manipulated tree and the idle tree, as in the case of Example 3.1 (4). Next we define classes of CEG models and of manipulations for which it is plausible to investigate whether the set of values of the primitive probabilities that satisfy both the idle and the manipulated CEG is not empty.

If two unmanipulated situations were in the same stage of the original tree and are manipulated in the same way, then they remain in the same stage in the manipulated tree.

DEFINITION 3.2. *A manipulation is called* positioned *if the partition of positions after the manipulation is equal to or a coarsening of the partition before manipulation. It is called* staged *if the partition of stages after the manipulation is equal to or a coarsening of the partition before manipulation.*

EXAMPLE 3.2 (Continuation of Example 2.6). Let v_1 and v_2 be in the same stage with v_3 mapping into v_5. The manipulation $D = \{v_1, v_2\}$ and $\widehat{P}(X(v_1) = v_3) = 1$, $\widehat{P}(X(v_2) = v_5) = 1$ is a staged and a positioned manipulation. The CEG of the manipulated tree is in Figure 15 where edges with zero probabilities are not drawn.

EXAMPLE 3.3. In the probability tree in Figure 10 the staged manipulation defined by $\widehat{\pi}(v_5|v_1) = 1$, $\widehat{\pi}(v_9|v_3) = 1$, $\widehat{\pi}(v_{17}|v_{13}) = 1$, $\widehat{\pi}(v_{19}|v_{17}) = 1$, leads to a modification of the CEG in Figure 11 in which the edges into w_∞ from the positions $[v_1, v_3]$, $[v_{13}]$ and $[v_{17}]$ could be not drawn because the associated manipulated probabilities become zero. Indeed this manipulation could be performed directly on the CEG by removing the three aforementioned edges. This idea is developed in Section 3.2.

EXAMPLE 3.4 (Continuation of Example 3.3). The staged and positioned manipulation with $D = \{v_2, v_7\}$ and $\widehat{\pi}(v_6|v_2) = 1$, $\widehat{\pi}(v_{12}|v_7) = 1$ has the effect of cutting off the branch starting at v_2 and going through v_7. Figure 16 gives the CEG of the manipulated tree.

A positioned manipulation manipulates all sample units identically when their future development distributions are identical, using the same (possibly randomising) allocation rule. A staged manipulation will treat sample units identically if their next development in the idle system is the same. Our experience has been that it is often sufficient to restrict study to positioned manipulations. We note for example that all manipulations on a BN considered by Pearl are positioned and also staged. Example 3.5 gives a simple case when a staged manipulation is not appropriate but a positioned manipulation is.

EXAMPLE 3.5.   An English university has residence blocks of apartments with two rooms each. It allocates prospective second year students (either English (E) or Chinese (C)) to one of the two rooms of each apartment. The second room has to be allocated to a prospective first year student. In the past this has been done at random: that is exactly one of the N second year students to go into an apartment and exactly one of the N first year students is allocated the integer $1 \leq i \leq N$ using a randomization devise and students share with the student allocated the same integer. However it has been noticed in a survey that the probability of satisfaction of home students placed with home students is higher and of Chinese students placed with Chinese students is higher than when they are mixed. In order to cause students' satisfaction to increase, the university decides to place first year students with a second year student with the same ethnicity.

The BN and CEG of this problem are given in Figure 17 where X represents the ethnicity of the second year student, Y that of the first year student and Z is a binary index of the satisfaction of two students in the same apartment, taking values U and S. Thus for example $X(v_0) = X$, $X(v_1) = [Y|X = E]$, $X(v_3) = [Z|X = E, Y = E]$ and $\pi(v_5|v_2)$ gives the probability of allocating a Chinese first year student to an apartment with a Chinese second year student. The vertices v_3 and v_5 are in the same stage to indicate a non-mixed apartment, analogous interpretation has the stage $\{v_4, v_6\}$. The undirected edge between

v_1 and v_2 represents the random allocation of the first year student to an apartment.

The relationship between satisfaction and shared race is not depicted in the BN whilst it is in the CEG through the colouring of its edges. More significantly it is impossible to determine, either from the semantics of the BN or the factorisation of the probability mass function of the path events, whether the allocation of the prospective second year student occurs before the allocation of the prospective first year student. The CEG states that second year allocation occurs before first year allocation explicitly, so that "causal" manipulation of the type suggest by the survey above is a possibility. The semantic of a BN is not refined enough to represent the sort of manipulation considered in this example. Note that central to the BN analysis of causal relationships is the absence of edges between vertices, here X and Y (see e.g. [6]). The only way to embody the types of manipulation we consider here is to join X and Y by an edge and so loose this intrinsic information.

A manipulation that forces individuals of the same ethnicity to share an apartment implies a CEG without the directed edge between v_1 and v_2 and without the crossing arrows in the CEG in Figure 17.

3.2. *Manipulating CEG's.* The standard manipulations of a BN are those that force some components of the network to take pre-assigned values, as in Equation (8). The analogue for the CEG is to consider manipulations which force all the paths to pass through an identified set of positions W. For example the assignment of a particular type of unit, here described by their current position, to a particular treatment regime, here described by a set of subsequent positions W (see also Section 4.2).

For a CEG $\mathcal{C}$ and a set of position W in $\mathcal{C}$, let $\mathrm{pa}(W)$ denote the set of all parents of the elements in W, that is $\mathrm{pa}(W) = \{w^* \in V(\mathcal{C}) : \text{there exists } w \in W \text{ such that } (w^*, w) \in E_d(\mathcal{C})\}$. In the analogy above the set $\mathrm{pa}(W)$ corresponds to the positions any unit must reach to be submitted to a treatment forcing them into the positions in W.

DEFINITION 3.3. *A subset W of positions of a CEG $\mathcal{C}$ is called a* manipulation set *if 1. all root-to-sink paths in $\mathcal{C}$ pass through exactly one position in* $\mathrm{pa}(W)$*, and 2. each position in* $\mathrm{pa}(W)$ *has exactly one child in W.*

EXAMPLE 3.6. In the CEG in Figure 15 the position $[v_1, v_2]$ is a manipulation set. Excluding the trivial case of a manipulation set consisting of the root node only, there is no manipulation set in the CEG in Figure 11. In the CEG in Figure 12 $W = \{[v_1], [v_2]\}$ is not a manipulation set. The manipulation described in Example 3.5 is to a manipulation set.

DEFINITION 3.4. *A manipulation $(D, \widehat{\Pi}_D)$ of a CEG is called a* pure manipulation to the positions W *if*

1. *it is a positioned manipulation,*
2. *for each $v \in D$ there exists $w \in W$ such that $\widehat{\mathrm{P}}(X(v|D) = w) = \widehat{\pi}(w|v) = 1$, and*
3. *no $v \notin D$ is manipulated.*

A CEG to be *causal* for an application needs *(i)* to be valid for that application and *(ii)* that for the pure manipulations to any manipulation set the corresponding manipulated CEG is also valid. If a CEG admits a description as a BN and the CEG is causal then the BN is also causal in the sense of [18, Definition 1.3.1]. In this sense a causal CEG is a natural generalisation of a causal BN, applicable to asymmetric models. However it can express a larger variety of manipulations than a causal BN: for example those based on certain functions of preceding variables as in Example 3.5. Next, three manipulation on the CEG for Example 1.2 are discussed.

EXAMPLE 3.7 (Continuation of Example 1.2). **Manipulation 1.** Consider the manipulation forced to w_1—that can be read as ensuring the suspect is taken to court. This would

assign probability 1 to the edge labelled x_1, all vertices other than w_∞ on paths after w_2 and their associated edges are deleted, and the primitive probabilities of the manipulated CEG are like in the unmanipulated CEG except the edge labelled x_1. The manipulated CEG is in Figure 2. Notice here that the hypothesis that this new CEG is valid for the manipulated is a substantive one and in particular will depend upon the how we plan to implement the manipulation, for example if whether the suspect went to court simply depended on whether a judge could be found or on government policy. But if we choose not to proceed with prosecution because, on the basis of their evidence, a third party did not think the case was strong enough to convince a jury, then this causal deduction would almost certainly not be valid. **Manipulation 2.** Consider the manipulation forced to w_{15}, signifying e.g. that the witness does not identify the suspect. The associated causal CEG is given in Figure 2. Note that the inevitable consequence of this manipulation is that suspect is released as there is just the release edge R into w_∞.

4. Identifying effects of a manipulation.

4.1. *Identification of causal effects.* Recent papers on causal BN literature [6, 18, 19] study when the topology of a BN helps to deduce that the effect of a manipulation on a pre-specified node of the BN can be identified from observing a subset of the BN variables that are observed or "manifest" in an unmanipulated system. Experiments on the original "idle" system can then be designed so that the effects of, for example, a proposed new treatment regime, on a manipulated system can be established. Here we demonstrate that the topology of the CEG can also be used to find *functions* of the data, for example subsets of possible measurements, that when observed in the idle system allows us to estimate effects of a given manipulation. We first need some definitions.

4.1.1. *Random variables.* Consider $(\mathcal{T}, \Pi(\mathcal{T}))$ with sample space $\mathbb{X}$. Any function $Y : \mathbb{X} \longrightarrow \mathbb{R}$ is a random variable on $(\mathbb{X}, 2^{\mathbb{X}}, \mathrm{P})$ and defines a partition of the atomic event set, namely

$$(10) \qquad\qquad \Lambda_y = \{\lambda \in \mathbb{X} \text{ such that } Y(\lambda) = y\}$$

where y ranges over the image of Y, written as $\mathrm{Image}(Y)$, which is a finite set. Similarly, given a partition of $\mathbb{X}$, a random variable Y with finite range space can be constructed so that (10) holds.

Now consider a CEG $\mathcal{C}$ constructed from the tree. Let $\mathbb{X}_{\mathcal{C}}$ be the set of root-to-sink paths in the CEG formed by directed edges. As already mentioned, $\mathbb{X}_{\mathcal{C}}$ can be identified with $\mathbb{X}$ and the σ-algebra $2^{\mathbb{X}}$ on the tree is mapped by the CEG construction into the power set of $\mathbb{X}_{\mathcal{C}}$, call it $2^{\mathbb{X}_{\mathcal{C}}}$. Furthermore, the random variable Y corresponds to a random variable on $(\mathbb{X}_{\mathcal{C}}, 2^{\mathbb{X}_{\mathcal{C}}})$ that induces on $\mathbb{X}_{\mathcal{C}}$ the same partition as the one obtained by mapping the Λ_y sets, $y \in \mathrm{Image}(Y)$, on the CEG. Thus, with a slight abuse of notation, for $y \in \mathrm{Image}(Y)$ we can write $\Lambda_y = \{\lambda \in \mathbb{X} \text{ such that } Y(\lambda) = y\} = \{\lambda \in \mathbb{X}_{\mathcal{C}} \text{ such that } Y(\lambda) = y\}$.

DEFINITION 4.1. *A random variable Y on $\mathbb{X}$ is called* observed *(or* manifest*) if, and only if, indicators of the events Λ_y are observed or observable for all $y \in \mathrm{Image}(Y)$.*

Although in practice vectors of manifest random variables are likely to occur, here, without loss of generality, we can work with random variables, because a partition induced by a random vector can be induced by a uni-dimensional random variable.

4.2. *Manipulation forced to a position.*

DEFINITION 4.2. *Call a manipulation of a CEG $\mathcal{C}$ forced to the position w if after manipulation*

1. the probability of the event $\{w\} = \{\lambda \in \mathbb{X} : w \in \lambda\}$ is one, and

2. all primitive probabilities in the manipulated CEG associated with positions at or after w in the original CEG are those in the idle system.

EXAMPLE 4.1. If $\{w\}$ is a manipulation set of $\mathcal{C}$ then the pure manipulation to w is a manipulation forced to w.

EXAMPLE 4.2. In Example 3.5 a manipulation forced to $w = \{v_3, v_5\}$ is obtained by setting $\widehat{\pi}(v_4|v_1) = 0 = \widehat{\pi}(v_6|v_2)$, that is by allocating students with the same ethnicity to the same apartment. This manipulation directly on the CEG is given by $\widehat{\pi}(\{v_4, v_6\}|\{v_1\}) = \widehat{\pi}(\{v_3, v_5\}|\{v_2\}) = 0$.

EXAMPLE 4.3. In Example 2.4 the manipulation $\pi_7 = 1$, forcing a woman who wants to have a child to take medicine C, is trivially a positioned manipulation, but not a staged manipulation. It is also a forced manipulation to $w = \{v_7\}$.

Consider a position w in the CEG $\mathcal{C}$. Every directed root-to-sink path through w can be split into two parts: one from root to w and one from w to the sink node, w_∞. Thus for $\lambda \in \mathbb{X}_{\mathcal{C}}$ if $w \in \lambda$ we can write $\lambda = \lambda(w_0, w) \times \lambda(w, w_\infty)$, where $\times$ indicates concatenation of paths. Note that $\{\lambda(w, w_\infty) : \lambda \in \mathbb{X}_{\mathcal{C}}\}$ is the set of root-to-sink paths in the sub-CEG of $\mathcal{C}$ starting at w, namely the CEG whose root is w, whose vertex set V is formed by positions in $\mathcal{C}$ lying on paths from w to w_∞ and whose edges are those in $\mathcal{C}$ connecting elements in V, likewise its edge labels. Call it $\mathcal{C}(w)$.

Consider a random variable $\widehat{Y}(w)$ on $(\mathcal{C}(w), 2^{\mathcal{C}(w)})$ and let zero be one of the values not taken by $\widehat{Y}(w)$. Let $\{\Lambda_y^+(w) : y \in \text{Image}(\widehat{Y}(w))\}$ be the partition induced by $\widehat{Y}(w)$ on $\{\lambda(w, w_\infty) : \lambda \in \mathbb{X}_{\mathcal{C}}\}$. It can be extended to a partition of $\mathbb{X}_{\mathcal{C}}$ with sets $\Lambda_0(w) = \{\lambda \in \mathbb{X}_{\mathcal{C}} : w \notin \lambda\}$ and $\Lambda_y(w) = \{\lambda = \lambda(w_0, w) \times \lambda(w, w_\infty) : \lambda(w, w_\infty) \in \Lambda_y^+(w)\}$ for $y \in \text{Image}(\widehat{Y}(w))$.

Recall that there exists one, usually many, random variables $Y(w)$ on $(\mathbb{X}_\mathcal{C}, 2^{\mathbb{X}_\mathcal{C}})$ which induce this partition. In Lemma 4.1 we compare the distributions of any such $Y(w)$ and of $\widehat{Y}(w)$ before and after a manipulation forced to w.

LEMMA 4.1.   *Consider a manipulation forced to w in the CEG $(\mathcal{C}, \Pi)$. Let P be the probability measure on the unmanipulated CEG and $\widehat{\mathrm{P}}$ the probability measure on the manipulated CEG, $(\mathcal{C}, \widehat{\Pi})$. Let $\{w\}$ be the event of passing through w in the idle system. Let $Y(w)$ and $\widehat{Y}(w)$ be defined as above. Then for $y \in \mathrm{Image}(\widehat{Y}(w))$*

1. $\mathrm{P}(\widehat{Y}(w) = y) = \widehat{\mathrm{P}}(\widehat{Y}(w) = y)$,

2. $\mathrm{P}(Y(w) = 0) = 1 - \mathrm{P}(\{w\})$ *and* $\widehat{\mathrm{P}}(Y(w) = 0) = 0$,

3. $\mathrm{P}(Y(w) = y) = \mathrm{P}(\{w\})\,\mathrm{P}(\widehat{Y}(w) = y)$, *and*

4. $\widehat{\mathrm{P}}(Y(w) = y) = \mathrm{P}(\widehat{Y}(w) = y)$.

PROOF.   1. This follows from the fact that the primitive probabilities on $\mathcal{C}(w)$ are not changed by the manipulation.

2. The probability of not passing through w in the manipulated system is zero because the manipulation is forced to w while in the manipulated system it is clearly one minus the probability of passing through w.

3. This is by construction of $Y(w)$ from $\widehat{Y}(w)$. A path λ through w is decomposed as $\lambda = \lambda(w_0, w) \times \lambda(w, w_\infty)$. By Equation (4) its probability in the idle system is $\mathrm{P}(\lambda) = \mathrm{P}(\lambda(w_0, w))\,\mathrm{P}(\lambda(w, w_\infty))$. Now, $\mathrm{P}(\lambda(w_0, w))$ is the probability of reaching w and $\mathrm{P}(\lambda(w, w_\infty)) = \mathrm{P}(\widehat{Y}(w) = y)$ by construction of $Y(w)$ from $\widehat{Y}(w)$ and the fact that the manipulation forced to w does not change the primitives probabilities for edges after w.

4. In the manipulated system $\widehat{\mathrm{P}}(\{w\}) = 1$ and use Item 1.

$\square$

Item 3 in Lemma 4.1 can be re-written as a conditional probability

$$\mathrm{P}(Y(w) = y|\{w\}) = \frac{\mathrm{P}(Y(w) = y)}{\mathrm{P}(\{w\})} = \mathrm{P}(\widehat{Y}(w) = y).$$

Lemma 4.1 can be applied if there is a position w such that, after enacting a manipulation, all paths pass through w and if the fact that $\{w\}$ occurs can be learnt from a set of measurements in the unmanipulated CEG. This can be checked from the CEG topology.

4.3. *Manipulation forced to a set of positions: the sub-CEG* $\mathcal{C}(W)$. It is not always possible, even in models that can be described by a causal BN, to observe indicators on the events $\{\Lambda_y(w) : y \in \mathrm{Image}(Y)\}$ for a suitable choice of w and $\mathrm{Image}(Y)$. Nevertheless being able to observe indicators of the set of coarser events $\Lambda_y(W) = \bigcup_{w \in W} \Lambda_y(w)$ for a set of positions, W, can also be sufficient for identifiability. To show this is less straightforward although the general set-up is a generalisation of Section 4.2.

DEFINITION 4.3. *A set of positions W of a CEG $\mathcal{C}$ is called $\mathcal{C}$-regular if no two positions in W lie on the same directed path of $\mathcal{C}$.*

EXAMPLE 4.4. By definition, a manipulation set of $\mathcal{C}$ is always $\mathcal{C}$-regular.

The analogue of the sub-CEG starting at w for a $\mathcal{C}$-regular set of positions, W, is a new CEG constructed by joining the sub-CEG's starting at each $w \in W$ to a new root-vertex w_0^*. The new edge (w_0^*, w) is labelled $\mathrm{P}(X(w_0^*) = w) = \frac{\mathrm{P}(\{w\})}{\mathrm{P}(W)}$, where, as before, $\mathrm{P}(\{w\}) = \sum_{\lambda \in \mathbb{X}: w \in \lambda} \mathrm{P}(\lambda)$ is the probability of passing through w in the original CEG and $\mathrm{P}(W) = \sum_{w \in W} \sum_{\lambda \in \mathbb{X}: w \in \lambda} \mathrm{P}(\lambda)$ is the probability of passing through a position in the set W in the original CEG. Note that because W is C-regular, $\sum_{w \in W} \mathrm{P}(X(w_0^*) = w) = 1$ Let $\mathcal{C}(W)$ be this new CEG.

EXAMPLE 4.5. The manipulation in Example 3.3 is a manipulation forced to the set of positions $W = \{\{v_1, v_3\}, \{v_2\}\}$ and $w_0^* = \{v_0\}$. A manipulation forced to $W = \{\{v_1, v_3\}, \{v_7\}\}$

requires a new w_0^* and the edge $(w_0^*, \{v_7\})$ has the same probability as the edge $(\{v_0\}, \{v_2\})$ in the original CEG.

It can be shown in analogy to Section 4.2 that the partition induced by a random variable $\widehat{Y}(W)$ on $(\mathbb{X}^{\mathcal{C}(W)}, 2^{\mathbb{X}_{\mathcal{C}(W)}})$ can be extended to a partition on $\mathcal{C}$ and this, in turn, can be interpreted as the range space of a random variable on $(\mathbb{X}^{\mathcal{C}}, 2^{\mathbb{X}_{\mathcal{C}}})$.

4.4. *Manipulation forced to a set of positions: amenable positions.* Let $\mathcal{C}^*(w)$ denote a graph representing what happens until we reach a given position w. Its vertices and edges are those along the root-to-w paths in $\mathcal{C}$ and its edge probabilities are inherited from $\mathcal{C}$ as well. The graph $\mathcal{C}^*(w)$ is not necessarily a CEG, for example in Figure 11 the graph $C^*([v_7])$ is not a CEG because of the undirected edge between $\{v_2\}$ and $\{v_7\}$. Write $K(\mathcal{C}^*(w))$ for the set of positions in $\mathcal{C}$ whose vertices are in $\mathcal{C}^*(w)$ excluding w. For any $\mathcal{C}$-regular set of positions, W, let $K(\mathcal{C}^*(W)) = \bigcup\limits_{w \in W} K(\mathcal{C}^*(w))$.

DEFINITION 4.4. *Call a set of positions, W, simple if*

1. *W is $\mathcal{C}$-regular,*

2. *there exists a partition of $K(\mathcal{C}^*(W))$ into $K^\alpha(\mathcal{C}^*(W))$ and $K^\beta(\mathcal{C}^*(W))$ called* active *and* background *positions respectively such that*

   (a) *for each $w \in W$ an* active *position has exactly one emanating edge along each root-to-w path in $\mathcal{C}^*(w)$ if it lies on that path. Furthermore for any two positions $w_1, w_2 \in W$ every pair of root-to-w_1 path and root-to-w_2 path containing the ordered sequences of active positions $w_1^{\alpha,1}, w_1^{\alpha,2}, \ldots w_1^{\alpha,n}$ and $w_2^{\alpha,1}, w_2^{\alpha,2}, \ldots w_2^{\alpha,n}$ respectively, the pairs of position $\left\{\left(w_1^{\alpha,k}, w_2^{\alpha,k}\right) : 1 \leq k \leq n\right\}$ are in the same stage. If two active positions lie in the same stage in $\mathcal{C}^*(w)$ then their unique emanating edge is labelled by the probability of the same edge in $\mathcal{C}$.*

(b) *Each* background position *in* $\mathcal{C}^*(w)$ *inherits a complete set of emanating edges from* $\mathcal{C}$. *Furthermore for any two positions* w_1, $w_2 \in W$ *every pair of root-to-w_1 path and root-to-w_2 path containing the ordered sequence of background positions* $w_1^{\beta,1}, w_1^{\beta,2}, \ldots w_1^{\beta,n}$ *and* $w_2^1, w_2^2, \ldots w_2^n$ *respectively, the pairs of position* $\left\{ \left(w_1^{\beta,k}, w_2^{\beta,k} \right) : 1 \le k \le n \right\}$ *are in the same stage.*

Note that active positions act as labels for the elements of W because each active position is in at most one path through $w \in W$. Root-to-leaf paths through background positions are governed by the same probability law regardless of the index of $w \in W$.

EXAMPLE 4.6.   A CEG of the binary CBN given by

$$B \;\; \rightarrow \;\; X \;\; \longrightarrow \;\; Y \;\; \leftarrow \;\; A$$

where X is manipulated to a value 1 is given by

$$
\begin{array}{ccccccc}
& & B = 1 & & X = 1 & & \\
A = 1 & w_1^{\beta,1} & \Rightarrow & w_1^{\alpha,2} & \longrightarrow & w_1 & \\
\nearrow & | & B = 0 & | & & \searrow\searrow & \\
w_0 = w^{\alpha,1} & | & & | & & & w_\infty \\
\searrow & | & B = 1 & | & & \nearrow\nearrow & \\
A = 0 & w_2^{\beta,2} & \Rightarrow & w_2^{\alpha,2} & \rightarrow & w_2 & \\
& & B = 0 & & X = 1 & &
\end{array}
$$
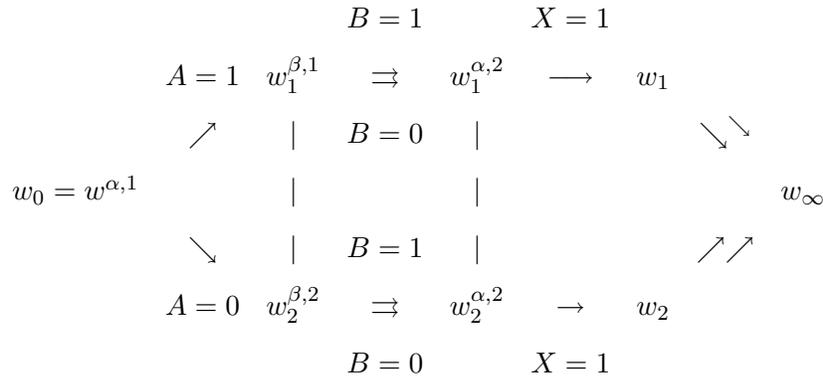
where the manipulation set is $W = \{w_1, w_2\}$. The active and background positions can be identified from $\mathcal{C}(W)$ thinking of $\mathcal{C}(W)$ as a subgraph of $\mathcal{C}$ with edge probabilities inherited from $\mathcal{C}$. In this example active positions are $\left\{ w_0, w_1^{\alpha,2}, w_2^{\alpha,2} \right\}$ where $\left\{ w_0, w_1^{\alpha,2} \right\} = \{A = 1\}$ and $\left\{ w_0, , w_2^{\alpha,2} \right\} = \{A = 0\}$ label the two positions in W and $w_1^{\alpha,2}, w_2^{\alpha,2}$ which lie at the same point in the ordered sequence to w_1 and w_2 lie on the same point in their respective

root to w_1 and root to w_2 sequence and so lie in the same position. The background positions are $w_1^{\beta,1}$ and $w_2^{\beta,1}$, retain all their emanating edges from the original CEG and are also in the same position. The terminology of active and background position here means that active positions might subsequently effect Y if manipulated after X while subsequently manipulating background positions cannot effect Y.

Further examples are in Section 5.

DEFINITION 4.5.   *A manipulation is called* amenable forcing to a set W *if*

1. *the set W is simple in $(\mathcal{C}, \Pi(\mathcal{C}))$,*
2. *the set W is simple in $(\mathcal{C}, \widehat{\Pi}(\mathcal{C}))$ and $\widehat{\mathrm{P}}(W) = 1$, and*
3. *$\Pi(\mathcal{C})$ and $\widehat{\Pi}(\mathcal{C})$ differ only on edges whose parents lie in $K^\alpha(\mathcal{C}^*(W))$.*

Item 2 in Definition 4.5 assumes that the manipulation is forced to W as the probability of passing through a vertex not in W in the manipulated system is zero. Furthermore, an amenable manipulation may change probabilities on edges associated to active positions, but will always leave probabilities associated with background positions unchanged.

EXAMPLE 4.7.   When $W = \{w\}$ is a singleton, the set of active positions will be empty and so all the conditions above are vacuous and W is simple. It follows that a pure manipulation forced to w is amenable.

LEMMA 4.2.   *Consider an amenable manipulation forcing to a simple set W. Then for $w \in W$*

$$\mathrm{P}(\{w\}) = \pi_W^\alpha \ \mathrm{P}^\beta(\{w\}) \quad and \quad \widehat{\mathrm{P}}(\{w\}) = \widehat{\pi}_W^\alpha \ \mathrm{P}^\beta(\{w\})$$

where $\mathrm{P}(\{w\})$ is the probability that a path in $\mathcal{C}$ will pass through the position $w \in W$ in the idle $(\mathcal{C}, \Pi(\mathcal{C}))$ and $\widehat{\mathrm{P}}(\{w\})$ is the analogue probability in the manipulated system; π_W^α and

$P^\beta(\{w\})$ *are products of primitive probabilities in* $\Pi(\mathcal{C})$ *associated with random variables whose positions lie in* $K^\alpha(C^*(W))$ *and* $K^\beta(C^*(W))$, *respectively.*

PROOF. Item 2 in Definition 4.4 means that events associated with background positions and with active positions are independent. Thus, for each $w \in W$ we have $P(\{w\}) = P^\alpha(\{w\}) P^\beta(\{w\})$ where $P^\beta(\{w\})$ is defined above and $P^\alpha(\{w\})$ is the product of primitive probabilities in $\Pi(\mathcal{C})$ associated with random variables whose positions lie in $K^\alpha(C^*(W))$. Furthermore, from the definition of $K^\alpha(C^*(W))$ for any positions $w, w' \in W$ we have $P^\alpha(\{w\}) = P^\alpha(\{w'\}) = \pi_W^\alpha$ (say).

The fact that W is also simple in $(\mathcal{C}, \widehat{\Pi}(\mathcal{C}))$ for the amenable manipulation implies that $\widehat{P}(\{w\}) = \widehat{\pi}_W^\alpha \widehat{P}^\beta(\{w\})$ for all $w \in W$. Finally from Item 1 in Definition 4.5 we have that $P(\{w\}) = \pi_W^\alpha \ P^\beta(\{w\})$ and from Item 2 $\widehat{P}(\{w\}) = \widehat{\pi}_W^\alpha \ P^\beta(\{w\})$. $\qquad\square$

LEMMA 4.3. *Consider an amenable manipulation forcing to a simple set* W. *The distribution of a random variable* $\widehat{Y}(W)$ *defined on the sub-CEG* $\mathcal{C}(W)$ *is identified from the probabilities in the unmanipulated system of the events* $\{Y(W) = y, W\}$ *for* $y \in \mathrm{Image}(Y)$ *where* $Y(W)$ *is constructed as above and its probabilities are given by the equation*

$$\widehat{P}(\widehat{Y}(W) = y | W) = \frac{P(Y(W) = y | W)}{P(W)}$$

provided that $P(\{w\}) > 0$ *for all* $w \in W$.

PROOF. Clearly for $y \in \mathrm{Image}(Y)$

$$P(Y(W) = y | W) = \frac{P(Y(W) = y, W)}{P(W)}$$

where $P(W) = \sum_{w \in W} P(\{w\}) = \pi_W^\alpha \sum_{w \in W} P^\beta(\{w\})$ by Lemma 4.2. Analogously

$$\widehat{P}(\widehat{Y}(W) = y | W) = \frac{\widehat{P}(\widehat{Y}(W) = y)}{\widehat{P}(W)}$$

where $\widehat{P}(W) = 1$ as the manipulation is forced to W and by Lemma 4.2 $\widehat{P}(W) = \sum_{w \in W} \widehat{P}(\{w\}) = \widehat{\pi}_W^\alpha \sum_{w \in W} P^\beta(\{w\})$. Furthermore

$$
\begin{aligned}
\widehat{P}(\widehat{Y}(W) = y) &= \sum_{w \in W} \widehat{P}(\widehat{Y}(W) = y|W)\widehat{P}(\{w\}) \\
&= \sum_{w \in W} \widehat{\pi}_W^\alpha\, P^\beta(\{w\})\widehat{P}(\widehat{Y}(W) = y|\{w\}) \text{ by Lemma 4.2 (8)} \\
&= \widehat{\pi}_W^\alpha \sum_{w \in W} P^\beta(\{w\})\widehat{P}(\widehat{Y}(W) = y|\{w\}) \text{ multiply and divide by } \pi_W^\alpha \\
&= \frac{\widehat{\pi}_W^\alpha}{\pi_W^\alpha} \sum_{w \in W} P^\beta(\{w\})\pi_W^\alpha \widehat{P}(\widehat{Y}(W) = y|\{w\}) \\
&= \frac{\widehat{\pi}_W^\alpha}{\pi_W^\alpha} \sum_{w \in W} P(\{w\})\widehat{P}(\widehat{Y}(W) = y|\{w\}).
\end{aligned}
$$

Set $\{\widehat{Y}(W) = y|\{w\}\} = \{\widehat{X}(w) = y\}$ and use Lemma 4.1 Item 1 to obtain

$$
\widehat{P}(\widehat{Y}(W) = y) = \frac{\widehat{\pi}_W^\alpha}{\pi_W^\alpha} \sum_{w \in W} P(\{w\})P(\widehat{X}(W) = y).
$$

As for $w \in W$ it holds $P(\{w\}) = P(\{w\}|W)\,P(W)$ then

$$
\widehat{P}(\widehat{Y}(W) = y) = \frac{\widehat{\pi}_W^\alpha}{\pi_W^\alpha} \sum_{w \in W} P(\{w\}|W)P(\widehat{Y}(W) = y|\{w\}) = \frac{\widehat{\pi}_W^\alpha}{\pi_W^\alpha}P(\widehat{Y}(W) = y|W)
$$

Thus we have that $\widehat{P}(\widehat{Y}(W) = y)$ is proportional to $P(\widehat{Y}(W) = y|W)$ and, being probabilities, they must be equal. □

5. A back-door theorem for CEG's. In [18] sufficient topological conditions (see Definition 3.3.1 page 79) on a causal BN are given for when the probability of a random variable Y on the BN is governed by the formula

$$
\widehat{P}_x(Y) = \sum_{z \in \mathbb{Z}} P(Y = y|X = x, \mathbf{Z} = \mathbf{z})\, P(\mathbf{Z} = \mathbf{z})
$$

after X has been manipulated to take a value x. Here $\mathbf{Z}$ is a random vector of variables appearing as vertices in the BN and $\mathbb{Z}$ is its sample space. A particular consequence of

this formula is that $\widehat{\mathrm{P}}_x(Y)$ can be calculated as a function of the marginal probability distribution of $(X, Y, \mathbf{Z})$ and we do not need to observe any other variable in the system.

Next an analogue formula is derived for CEG's. Let (Y, Z) be a random vector taking values $(y, z) \in \mathrm{Image}(Y) \times \mathrm{Image}(Z)$ and let W be a set of positions in a CEG $\mathcal{C}$. We will show that it is sufficient to know the probabilities of the events $\{Y = y, Z = z, W\}$. As above let $\widehat{\mathrm{P}}(Y(W) = y)$ denote the probability of Y after a manipulation to a set of positions W.

For every $z \in \mathrm{Image}(Z)$ let Ω_z be the set of root-to-leaf paths corresponding to the event $\{Z = z\}$. Assume that there exists a regular set of positions W_z such that the set of root-to-leaf paths through W_z is equal to Ω_z. Let $\mathcal{C}(W_z)$ be the sub-CEG defined as in Section 4.3 with new root vertex $w_0(z)$.

THEOREM 5.1. *Consider a random variable Z on a CEG $\mathcal{C}$ and a manipulation forced to a set of positions W in $\mathcal{C}$.*

(i) For $z \in \mathrm{Image}(Z)$ let $W(z)$ be the set of manipulated positions on root-to-leaf paths passing through a position in W_z.

(ii) Assume that if in $\mathcal{C}(W_z)$ $w' \in W(z)$ and $w \in W_z$ are on the same root-to-leaf path then w' comes after w in the path ordering.

(iii) Assume that for each $z \in \mathrm{Image}(Z)$ the manipulation in $\mathcal{C}(W_z)$ is amenable forcing to $W(z)$.

Let $\widehat{Y}(W)$ be a random variable on the manipulated CEG. For $y \in \mathrm{Image}(\widehat{Y}(W))$ and $z \in \mathrm{Image}(Z)$ let $\{\widehat{Y}(W_z) = y\}$ be the set of root-to-leaf paths for the event $\{\widehat{Y}(W) = y\}$ intersected with the set of root-to-leaf paths through W_z. As before, let $Y(W)$ be the extension of $\widehat{Y}(W)$ to the unmanipulated CEG. Then

(11) $$\widehat{\mathrm{P}}(Y(W) = y) = \sum_{z \in \mathrm{Image}(Z)} \mathrm{P}(Y(W) = y | Z = z)\, \mathrm{P}(Z = z).$$

PROOF. From *(i)* for each $z \in \text{Image}(Z)$ the conditional events $\{Y = y, W | Z = z\} = \{Y = y, W(z) | Z = z\}$, $y \in \text{Image}(Y)$ are all measurable with respect to the power set of $\mathcal{C}(W_z)$. From *(ii)* $\widehat{\text{P}}(Z = z) = \text{P}(Z = z)$, and by Lemma 4.3 and condition *(iii)* we have $\widehat{\text{P}}(Y(W_z) = y | Z = z) = \text{P}(Y(W_z = y | Z = z)$. So

$$
\begin{aligned}
\widehat{\text{P}}(Y(W_z) = y) &= \widehat{\text{P}}(Y(W_z) = y | Z = z)\widehat{\text{P}}(Z = z) \text{ by the definition of } W(z) \\
&= \text{P}(Y(W_z) = y | Z = z)\text{P}(Z = z) \text{ by the two identities above} \\
&= \text{P}(Y(W) = y | Z = z)\text{P}(Z = z) \text{ by the definition of } W(z).
\end{aligned}
$$

$\square$

The topology of each graph $\mathcal{C}(W_z)$ is inherited from $\mathcal{C}$ except for the new root vertex $w_0(z)$ and its connecting edges. So the topological conditions on $\mathcal{C}(W_z)$ given in Theorem 5.1 are inherited as topological conditions on the idle CEG $\mathcal{C}$. Theorem 5.1 is particularly useful when the topologies of $\mathcal{C}(W_z)$, $z \in \text{Image}(Z)$, are different, evoking different ways of satisfying the criteria of Theorem 5.1 for different configurations z. Of course, if it is possible to express a model using a BN, then $\mathcal{C}(W_z)$, $z \in \text{Image}(Z)$ are all identical.

EXAMPLE 5.1 (Continuation of Example 3.5). First year students making the university first choice $\{Z = 0\}$ will be allocated a shared apartment on campus whilst others $\{Z = 1\}$ will be lodged either in town W, namely $\{X_3 = 0\}$, or in town C, namely $\{X_3 = 1\}$ and have a friendly landlord $\{U = 0\}$ or an unfriendly landlord $\{U = 1\}$. When $\{Z = 0\}$ it is believed that the CEG of Figure 17 is valid. If $\{Z = 1\}$ the town is chosen independently of the race X_2 of the first year student, the friendliness of the landlord does not depend on the town or race of the student. However the satisfaction Y depend both on friendliness of the landlord and the allocated town C: C having a higher probability of higher satisfaction $\{Y = 1\}$ than W, conditional on $\{Z = 1\}$. This scenario can be expressed as a BN on four

random variables in Figure 12

$$(12) \qquad\qquad X_2 \quad U \ \longrightarrow Y \ \longleftarrow \ X_3$$

We want to consider a proposed manipulation of our allocation policy for next year. We plan to match campus students so that those sharing an apartment are of the same race and to allocate off campus students only to lodgings in town C. Our interest is in $\widehat{P}(Y(W) = 1)$ i.e. the overall predicted probability of high satisfaction were we to implement this policy. We plan to estimate this probability with a small data set, collected from earlier years. The sort of asymmetries exhibited by this problem makes extremely awkward to represent it through a single BN. Thus X_1 is only defined for a student allocated to campus whilst (X_3, U) only to students allocated to lodgings. Furthermore the manipulation proposed is different for different contingencies.

The whole problem can be represented by the CEG in Figure 18 where the set of positions manipulated are coloured in black. Note that $\{Z = 0\}$ and $\{Z = 1\}$ define a cut and for our proposed manipulation condition (i) in Theorem 5.1 is satisfied.

The topology of $\mathcal{C}_{Z=0}$ is identical to the sub-graph of $\mathcal{C}$ consisting of all edges and vertices on root-to-leaf paths containing edge $\{Z = 0\}$. Similarly, $\mathcal{C}_{Z=1}$ can be identified with the sub-graph of $\mathcal{C}$ consisting of all edges and vertices on root-to-leaf paths containing the edge $\{Z = 1\}$.

Since $W(0)$ is a singleton, applying Lemma 4.1 to $\mathcal{C}_{Z=1}$ gives us that

$$\widehat{P}(Y(W(0)) = 1 | Z = 0) = P(Y = 1 | X = 0, Z = 0).$$

A manipulation to $W(1)$ on $\mathcal{C}_{Z=1}$ is clearly amenable so that

$$\widehat{P}(Y(W(0)) = 1 | Z = 1) = P(Y = 1 | X = 1, Z = 0).$$

It follows that

$$\widehat{P}(Y(W(0)) = 1) = P(Y = 1 | X = 0, Z = 0)\, P(Z = 0) + P(Y = 1 | X = 1, Z = 1)\, P(Z = 1).$$

Thus $\widehat{\mathrm{P}}(Y(W) = 1)$ is expressed as a function of only three probabilities from the idle system: the probability that a student is on campus, the probability that a campus student sharing with someone of the same race makes a favourable return given she shares the race of her room-mate and the probability that a non campus student residing in town C makes a favourable return. It follows that we have been able to deduce from the topology of $\mathcal{C}$ that the probability of the ethnicity of match pairs of campus students and the conditional distribution of returns of unmatched students are irrelevant to $\widehat{P}$. Furthermore, the race and probability of friendliness of the landlord for lodged students is also irrelevant to this calculation and need not be estimated. Note here that we have deduced what function of variables is sufficient to discover $\widehat{P}$: here $X = |X_1 - X_2|/2$, a feature that cannot be deduced directly from any BN on the original measurement variables.

COROLLARY 5.1. *Consider a causal CEG, Z and Y as in Theorem 5.1 and a pure manipulation to a manipulation set W. If all the events $\{Y = y, W, Z = z\}$, $y \in \mathrm{Image}(Y)$ and $z \in \mathrm{Image}(Z)$ are manifest, then the effect of the manipulation is identified and given by Equation (11) whenever W is simple, conditioned on Z.*

PROOF. It follows directly from Theorem 5.1.   □

EXAMPLE 5.2. Note that if a CEG of a BN is constructed so that the back-door variables are introduced as early as possible compatibly with the ordering of the BN, then the conditions of Theorem 5.1 are satisfied for atomic interventions on a causal BN.

EXAMPLE 5.3 (Continuation of Example 3.7). **Manipulation 3.** Consider a third manipulation forced to the set of positions $W = \{w_{14}, w_{16}\}$—i.e. the witness is forced to positively identify the suspect. The CEG $\mathcal{C}(W)$ is given in Figure 4. Suppose we learn the values of Z are 1 if a path passes through w_7, 2 if through w_8 and 3 if w_4, meaning respectively that the suspect threw the brick, the suspect was present but did not throw the

brick, the suspect not present. The conditions of Theorem 5.1 are satisfied. For example conditioning on $\{Z = 1\}$ gives us the graph in Figure 5. In this sub-graph $\{w_7)$ is the only active position and $\{w_0, w_1, w_3, w_{10}, w_{11}\}$ are the background positions. Similarly it holds for the sub-graphs associated with $\{Z = 2\}$ and $\{Z = 3\}$. Therefore we can conclude that probability of conviction $\{X_6 = 1\}$ after this manipulation is

$$\widehat{P}(X_6 = 1) = \sum_{z=1}^{3} P(Z = z)\, P(X_6 = 1 | Z = z, X_5 = 1)$$

This probability can be estimated from data on similar cases, provided that in these cases the joint distribution of $\{Z, X_6 | X_5\}$ is known and recorded.

6. Discussion. Some problems, which are not satisfactorily expressed in terms of the exchangeable relationships in a BN (see [22, 23] and Example 3.5), can be well represented by a CEG model. This applies when the sample space is asymmetric as in the CEG in Figure 18 or the order of factors may be different for different settings of the factors as in Example 2.3.

Sometimes as in Example 3.5, causality is naturally expressed through predictions concerning the manipulation of unfolding situations rather than through assertions about the effects of manipulations on dependence relationships between measurements. Example 5.1 shows that to determine the effect of a cause function of the measurements, namely $|X_1 - X_2|/2$, may be sufficient instead of resorting only to subsets of measurements like in BN's. BN technology encourages to express causal hypotheses in terms of the random variables and parametrization in which data are conveyed to us within a certain parametrisation. It is now well appreciated that it is often necessary to separate causal structure from the dependence structures introduced into measurements through a particular sampling mechanism specific to the acquisition of information for a particular study. CEG modelling allows that.

Note that the high number of vertices in a probability tree is reduced in a CEG by the sink node that collects the leaves and by the modelling constraints imposed by the position equivalence relation. However in some cases it can remain much larger than the number of factors in the modelled problem. For this reason we do not recommend the use of a CEG model where a simpler structure, like a BN, can be usefully employed. Furthermore the BN gives a representation of dependence structure that does not depend on the sample space of random variables whilst the CEG needs this sample space to be specified and be finite. Sometimes when addressing causal ideas we do not want the size of the nature or size of the sample space to intrude in which case we are again forced back on to a BN representation.

Another difficulty with CEG's is that to our knowledge, the characterization of the equivalence classes both associated with CEG's whose probabilistic structure is identical and also those associated with graphs which can causally be identified are as yet only partially understood. This is in contrast to BN's where such equivalence classes can be determined—respectively by the pattern (or essential graph) or the BN itself. For tasks like model selection a good understanding of such equivalence classes is important to develop for CEG's. We plan to report on this topic in a future paper. Finally, whilst much more varied dependence relationships can be expressed using a CEG rather than a BN, CEG is also limited in the number of factorization formulae it can express simultaneously, being a graph. When such structure is present the CEG no longer provides a fully topological framework expressing all the contingent independence relationships and our criticisms of BN's apply equally to CEG's. We are then thrown on to causal analyses using more algebraic techniques [24]. Despite these caveats we believe that the CEG's provide a powerful graphical tool for the study of implicit causal relationships derivable from its qualitative structure.

We now turn to briefly discuss some generalisations. In the paper we considered the power set σ-algebra to limit technicalities. But generalisations of our results can be considered for

less refined σ-algebras. Searching over functions of measurements to find the cheapest way of identifying a quantity of interest will often be of much great value. This will be particularly useful if those measurements have not yet been collected, or their parametrisations have been chosen by convention rather than because they reflect in some natural way the mechanism by which things happen. Search algorithms need to be developed. Example 2.1 shows that techniques from algebraic geometry can be usefully employed on CEG's as they have been on BN's [10] and for identification of causal effects on BN's [22]. This is only in part being explored and could be combined with search algorithms and design issues.

REFERENCES

[1] P. Anderson and J.Q. Smith (2005). A graphical framework for representing the semantics of asymmetric models. *Technical report* 05-12, CRiSM, *Department of Statistics*, The University of Warwick.

[2] T. Bedford and R. Cooke (2001). *Probabilistic risk analysis: foundations and methods*. Cambridge University Press, Cambridge.

[3] C. Boutilier, N. Freidman, M. Goldszmidt and D. Koller (1996). Context-specific independence in Bayesian networks. In *Proceedings of UAI - 96* 115-123.

[4] R.E. Bryant (1986). Graphical algorithms for Boolean function manipulation. *IEEE Transactions of Computers* C-35 677-691.

[5] G.A. Churchill (1995). Accurate restoration of DNA sequences. In C. Gatsaris et al. (eds.) *Case Studies in Bayesian Statistics vol. II*, Springer-Verlag, New-York, 90-148.

[6] A.P. Dawid (2002). Infuence Diagrams for Causal Modelling and Inference. *International Statistical Reviews*, 70:161 - 89.

[7] A.P. Dawid, J. Mortera, V.L. Pascali and D. W. van Boxel (2002). Probabilistic expert systems for forensic inference from genetic markers. *Scand. J. Statist.*, 29:577595.

[8] N. Freidman and M. Goldszmidt (1999). Learning Bayesian networks with local structure. In M.I. Jordan (ed.) *Learning in Graphical Models* MIT Press, 421-459.

[9] S. French (ed.) (1989). *Readings in Decision analysis*. Chapman and Hall/CRC, London.

[10] L.D. Garcia, M. Stillman and B. Sturmfels (2005). Algebraic geometry of Bayesian networks, *Journal of Symbolic Computation*, 39(3-4):331-355.

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

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

[13] M. Jaeger (2004). Probabilistic decision graphs - combining verification and AI techniques for probabilistic inference. *Int.J. of Uncertainty, Fuzziness and Knowledge-based Systems*, 12:19-42.

[14] S.L. Lauritzen (1996). *Graphical models.* Oxford Science Press, Oxford, 1st edition.

[15] R. Lyons (1990). Random walks and percolation on trees. *Annals of Probability*, 18:931-958.

[16] D. McAllester, M. Collins and F. Pereira (2004). Case-factor diagrams for structured probability models. In *Proceedings of UAI - 2004* 382-391.

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

[18] J. Pearl (2000). *Causality. models, reasoning and inference.* Cambridge University Press, Cambridge.

[19] J. Pearl (2003). Statistics and Causal Inference: A Review (with discussion). *Test*, 12(2):281-345.

[20] G. Pistone, E. Riccomagno and H.P. Wynn (2001). *Algebraic Statistics.* Chapman & Hall/CRC, Boca Raton.

[21] D. Poole and N.L. Zhang (2003). Exploiting contextual independence in probabilistic inference. *Journal of Artificial Intelligence research*, 18:263-313.

[22] E. Riccomagno and J.Q. Smith (2003). Non-Graphical Causality: a generalisation of the concept of a total cause. *Research report series No. 394, Dept of Statistics,* The University of Warwick.

[23] E. Riccomagno and J.Q. Smith (2004). Identifying a cause in models which are not simple Bayesian networks. In *Proceedings of the 10th Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 1315-1322.

[24] E. Riccomagno and J.Q. Smith (2007). The geometry of causal probability trees that are algebraically constrained. In L Pronzato and A A Zigljavsky (eds.) *In Search for Optimality in Design and Statistics: Algebraic and Dynamical System Methods* Springer 131-152.

[25] G. Shafer (1996). *The Art of Causal Conjecture.* MIT Press, Cambridge, MA.

[26] R. Settimi and J.Q. Smith (1998). On the geometry of Bayesian graphical models with hidden variables. In G. Cooper and S. Moral (eds.) *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, S. Francisco, 472–479.

[27] R. Settimi and J.Q. Smith (2000). Geometry, moments and conditional independence trees with hidden variables. *The Annals of Statistics*, 28(4):1179-1205.

[28] J.Q. Smith and E. E. Anderson (2006). Conditional independence and Chain Event Graphs. *Artificial*
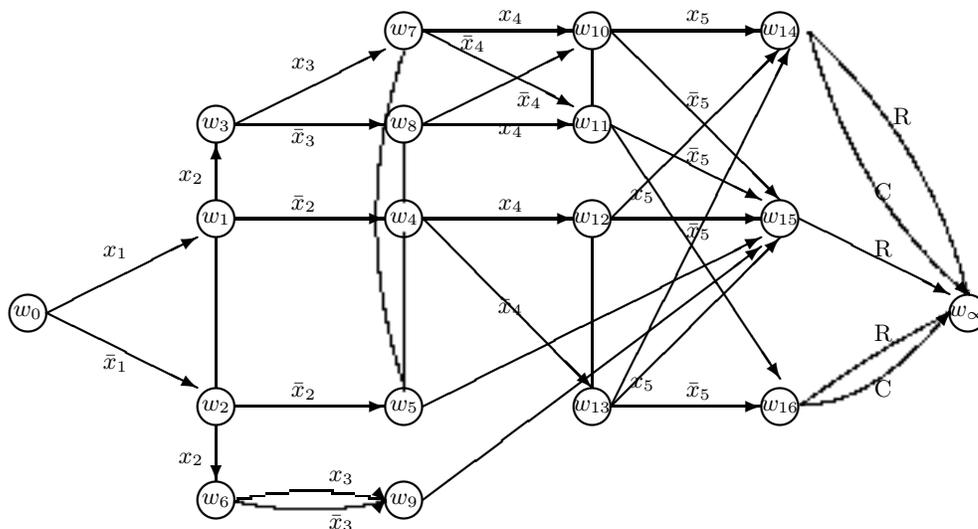
FIG 1. *CEG for Example 1.1*

Intelligence, to appear.

[29] J.Q. Smith and J. Croft (2003). Bayesian networks for discrete multivariate data: An algebraic approach to inference. *J. of Multivariate Analysis*, 84(2):387-402.

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

VIA DODECANESO, 35
GENOVA 16149, ITALY
PRINTEADE1

GIBBET HILL ROAD
COVENTRY CV4 7AL, UK
PRINTEADE1
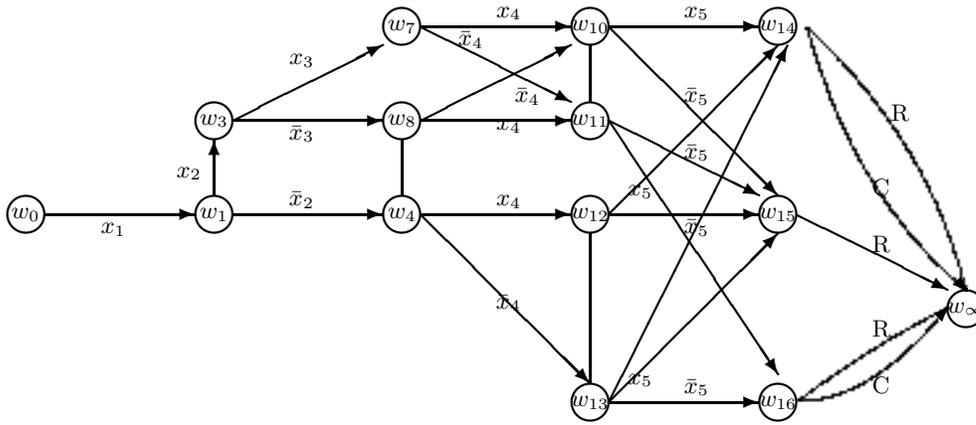
FIG 2. *CEG for Manipulation 1 in Example 3.7*
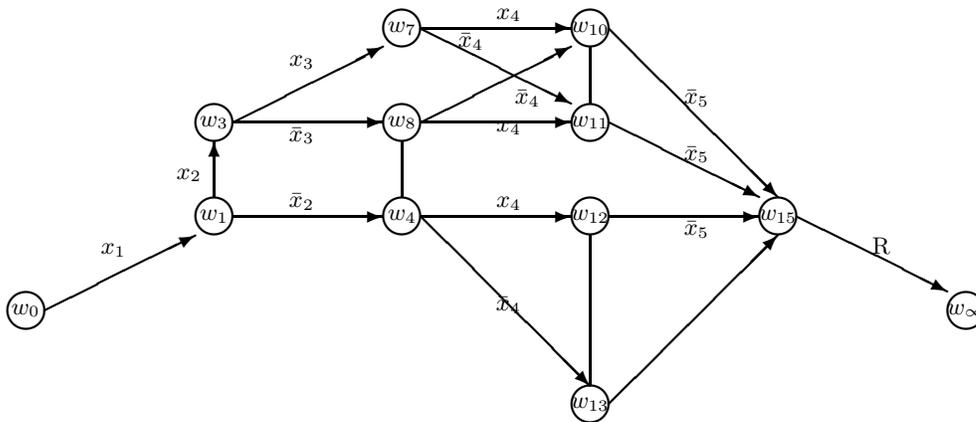
FIG 3. *CEG for Manipulation 2 in Example 3.7*
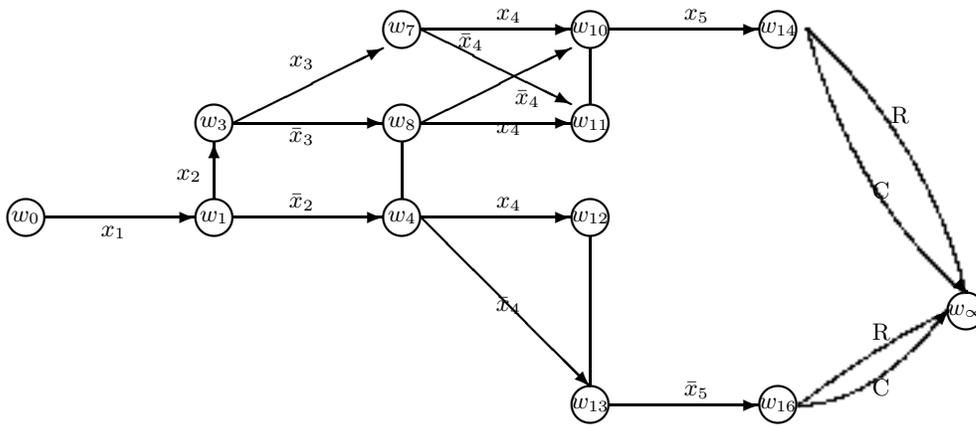
FIG 4. *CEG for Manipulation 3 in Example 3.7*
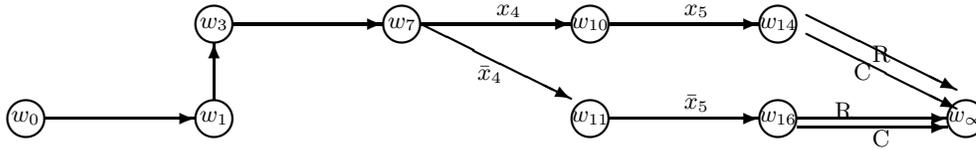
FIG 5. *Backdoor theorem for Example 5.3*
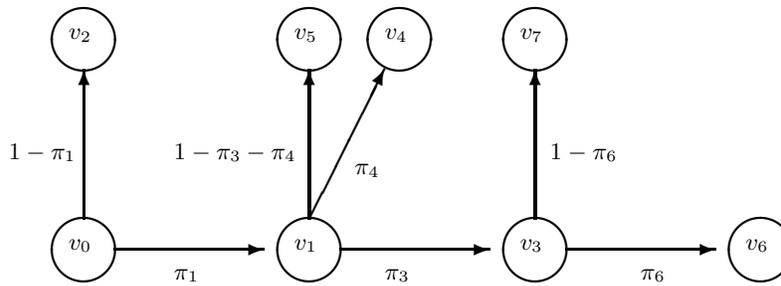
FIG 6. *Primitive probabilities and atomic event probabilities*

imsart-aos ver. 2005/10/19 file: currentRevAnnals.tex date: July 25, 2007
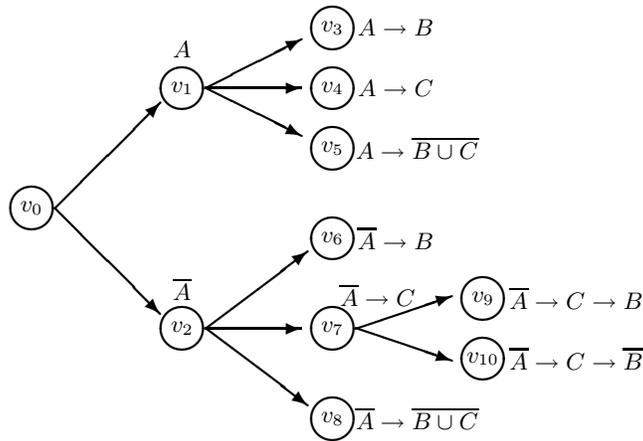
FIG 7. *Stages and independence.*
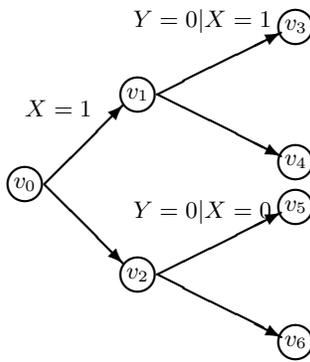
FIG 8. *Probability tree for Example 2.5*
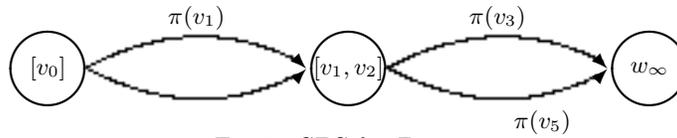
FIG 9. *CEG for Figure 8*
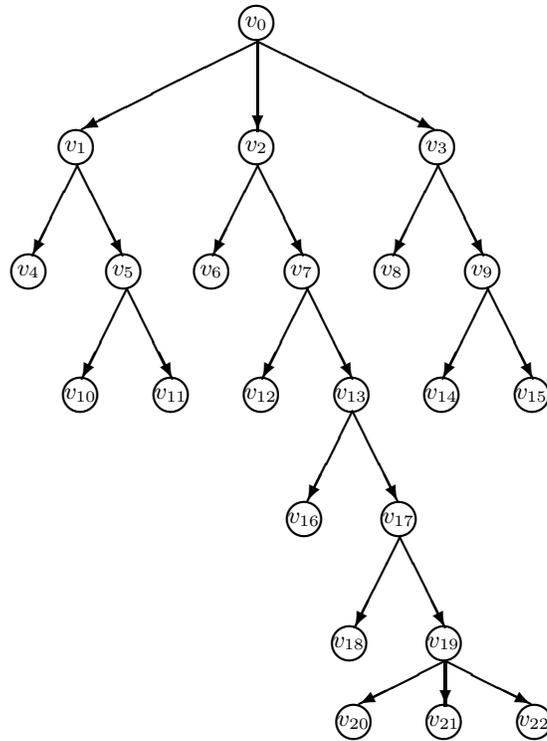
FIG 10. *A tree with unusual stage set*
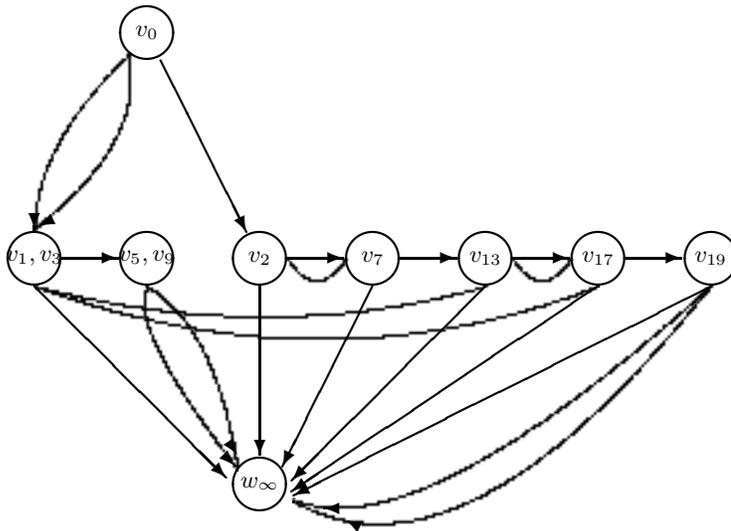
FIG 11. *CEG for the event tree in Figure 10*

imsart-aos ver. 2005/10/19 file: currentRevAnnals.tex date: July 25, 2007
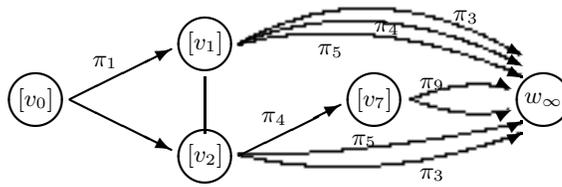
FIG 12. *CEG for the event tree in Figure 7*
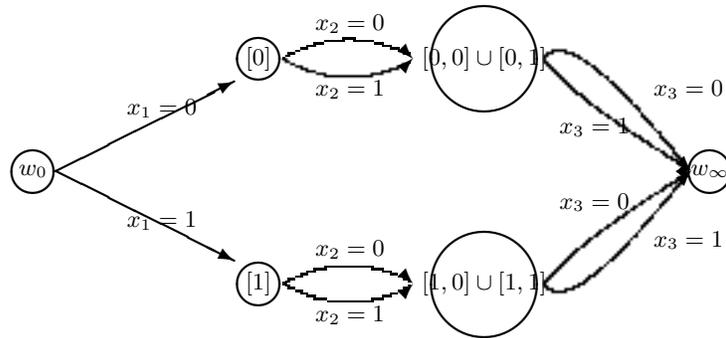
FIG 13. *CEG for binary* $X_2 \leftarrow X_1 \rightarrow X_3$
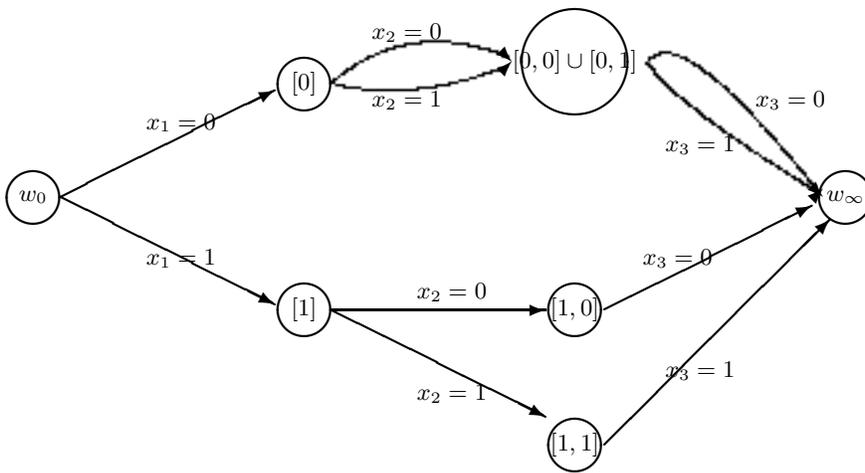
FIG 14. *CEG for condition (5) but not necessarily (6)*
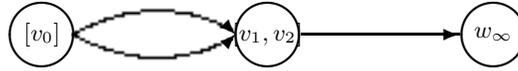
FIG 15. *Manipulated CEG for Example 3.2*
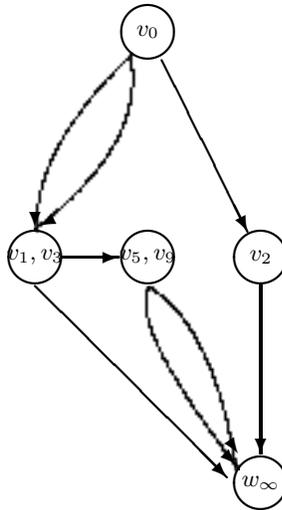


FIG 16. *Manipulated CEG for Example 3.4*
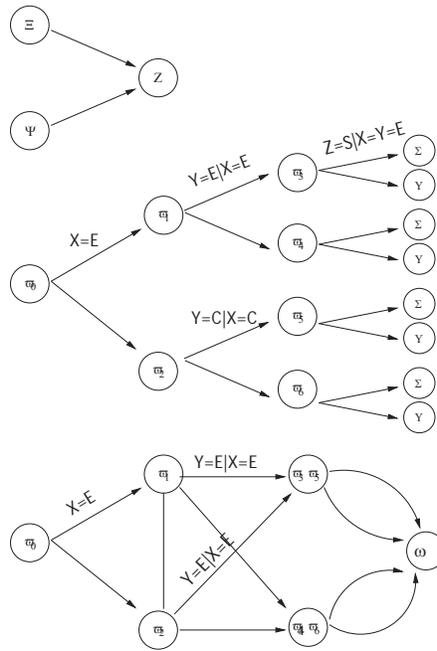
FIG 17. *BN and CEG for the example of university room allocation*
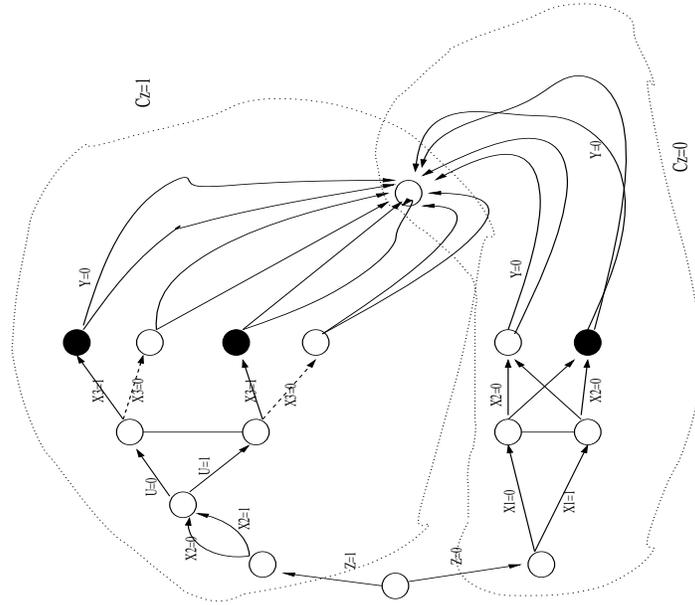
Fig 18. *Modified university example*