Local Characterizations of Causal Bayesian Networks[⋆]

Elias Bareinboim[1], Carlos Brito[2], and Judea Pearl[1]

[1] Cognitive Systems Laboratory
Computer Science Department
University of California Los Angeles
CA 90095
{eb, judea}@cs.ucla.edu
[2] Computer Science Department
Federal University of Ceará
carlos@lia.ufc.br

Abstract. The standard definition of causal Bayesian networks (CBNs) invokes a global condition according to which the distribution resulting from any intervention can be decomposed into a truncated product dictated by its respective mutilated subgraph. We analyze alternative formulations which emphasizes local aspects of the causal process and can serve therefore as more meaningful criteria for coherence testing and network construction. We first examine a definition based on "modularity" and prove its equivalence to the global definition. We then introduce two new definitions, the first interprets the missing edges in the graph, and the second interprets "zero direct effect" (i.e., *ceteris paribus*). We show that these formulations are equivalent but carry different semantic content.

1 Introduction

Nowadays, graphical models are standard tools for encoding probabilistic and causal information [Pearl, 1988; Spirtes *et al.*, 1993; Heckerman and Shachter, 1995; Lauritzen, 1999; Pearl, 2000; Dawid, 2001; Koller and Friedman, 2009]. One of the most popular representations is a *causal Bayesian network*, namely, a directed acyclic graph (DAG) G which, in addition to the traditional conditional independencies also conveys causal information, and permits one to infer the effects of interventions. Specifically, if an external intervention fixes any set $\mathbf{X}$ of variables to some constant $\mathbf{x}$, the DAG permits us to infer the resulting post-intervention distribution, denoted by $P_{\mathbf{x}}(\mathbf{v})$, [3] from the pre-intervention distribution $P(\mathbf{v})$.

The standard reading of post-interventional probabilities invokes a mutilation of the DAG G, cutting off incoming arrows to the manipulated variables and leads to a "truncated product" formula [Pearl, 1993], also known as "manipulation theorem"

[⋆] This work was supported in parts by National Institutes of Health #1R01 LM009961-01, National Science Foundation #IIS-0914211 and #IIS-1018922, and Office of Naval Research #N000-14-09-1-0665 and #N00014-10-1-0933.

[3] [Pearl, 2000] used the notation $P(\mathbf{v} \mid set(\mathbf{t}))$, $P(\mathbf{v} \mid do(\mathbf{t}))$, or $P(\mathbf{v} \mid \hat{\mathbf{t}})$ for the post-intervention distribution, while [Lauritzen, 1999] used $P(\mathbf{v} \parallel \mathbf{t})$.

[Spirtes *et al.*, 1993] and "G-computation formula" [Robins, 1986]. A local characterization of CBNs invoking the notion of modularity was presented in [Pearl, 2000, p.24] and will be shown here to imply as well as to be implied by the truncated product formula. This characterization requires the network builder to judge whether the conditional probability $P(Y \mid \mathbf{PA_y})$ for each parents-child family remains invariant under interventions outside this family. Whereas the "truncated product" formula computes post-intervention from pre-intervention probabilities, $given$ a correctly specified CBN, the local condition assists the model builder in constructing a correctly specified CBN. It instructs the modeller to focus on each parent-child family separately and judge whether the parent set is sufficiently rich so as to "shield" the child variable from "foreign" interventions.

A second type of local characterization treated in this paper gives causal meaning to individual arrows in the graph or, more accurately, to its missing arrows. These conditions instruct the modeller to focus on non-adjacent pairs of nodes in the DAG and judge whether it is justified to assume that there is no (direct) causal effect between the corresponding variables. Two such conditions are formulated; the first requires that any variable be "shielded" from the combined influence of its non-neighbours once we hold its parents constant; the second requires that for every non-adjacent pair in the graph, one of the variables in the pair to be "shielded" from the influence of the other, holding every other variable constant (*ceteris paribus*).

From a philosophical perspective, these characterizations define the empirical content of a CBN since, in principle, each of these assumptions can be tested by controlled experiments and, if any fails, we know that the DAG structure is not a faithful representation of the causal forces in the domain, and will fail to correctly predict the effects of some interventions. From a practical viewpoint, however, the main utility of the conditions established in this paper lies in their guide to model builders, for they permit the modeller to focus judgement on local aspects of the the model and ensure that the sum total of those judgements be consistent with one's knowledge and all predictions, likewise, will cohere with that knowledge.

In several ways the conditions introduced in this paper echo the global, local, and pairwise conditions that characterize directed Markov random fields [Pearl, 1988; Lauritzen, 1996]. The global condition requires that every d-separation condition in the DAG be confirmed by a corresponding conditional independence condition in the probability distribution. The local Markov condition requires that every variable be independent of its non descendants, conditional on its parents. Finally, the pairwise condition requires that every pair of variables be independent conditional on all other variables in the graph. The equivalence of the three conditions has been established by several authors [Pearl and Verma, 1987; Pearl, 1988; Geiger *et al.*, 1990; Lauritzen, 1996]. Our characterization will differ of course in its semantics, since our notion of "dependence" is causal; it is similar nevertheless in its attempt to replace global with local conditions for the sake of facilitating judgement of coherence.

[Tian and Pearl, 2002] provides another characterization of causal Bayesian networks with respect to three norms of coherence called Effectiveness, Markov and Recursiveness, and showed their use in learning and identification when the causal graph is not known in advance. This characterization relies on equalities among products of probabilities under different interventions and lacks therefore the qualitative guidance needed for constructing the network.

The rest of the paper is organized as follows. In Section 2, we introduce the basic concepts, and present the standard global and local definitions of CBNs together with discussion of their features. In Section 3, we prove the equivalence between these two definitions. In Section 4, we introduce two new definitions which explicitly interpret the missing links in the graph as representing absence of causal influence. In Section 5, we prove the equivalence between these definitions and the previous ones. Finally, we provide concluding remarks in Section 6.

2 Causal Bayesian networks and interventions

The notion of intervention and causality are tightly connected. Interventions are usually interpreted as an external agent setting a variable to a certain level (e.g., treatment), which contrasts with an agent just passively observing variables' levels.

The dichotomy between observing and intervening is extensively studied in the literature [Pearl, 1994; Lindley, 2002; Pearl, 2009, pp. 384-387] , and one example of its utilization is in the context of randomized clinical trials. It is known that performing the trial (intervening), and then collecting the underlying data is equivalent to applying the treatment uniformly over the entire population. This class of experiments is entirely different from simply collecting passive census data, from which no causal information can be obtained.

The concept of intervention precedes any graphical notion. We consider here the most elementary kind of intervention, that is, the atomic one, where a set $\mathbf{X}$ of variables is fixed to some constant $\mathbf{X} = \mathbf{x}$. All the probabilistic and causal information about a set of variables $\mathbf{V}$ is encoded in a collection of interventional distributions over $\mathbf{V}$, of which the distribution associated with no intervention (also called *pre-intervention* or *observational* distribution) is a special case.

Definition 1 (Set of interventional distributions). *Let $P(\mathbf{v})$ be a probability distribution over a set $\mathbf{V}$ of variables, and let $P_{\mathbf{x}}(\mathbf{v})$ denote the distribution resulting from the intervention $do(\mathbf{X} = \mathbf{x})$ that sets a subset $\mathbf{X}$ of variables to constant $\mathbf{x}$. Denote by $\mathbf{P}_*$ the set of all interventional distributions $P_{\mathbf{x}}(\mathbf{v}), \mathbf{X} \subseteq \mathbf{V}$, including $P(\mathbf{v})$, which represents no intervention (i.e., $\mathbf{X} = \emptyset$). We assume that $\mathbf{P}_*$ satisfies the following condition for all $\mathbf{X} \subseteq \mathbf{V}$:*

i. *[Effectiveness] $P_{\mathbf{x}}(v_i) = 1$, for all $V_i \in \mathbf{X}$ whenever v_i is consistent with $\mathbf{X} = \mathbf{x}$;*

The space of all interventional distributions can be arbitrary large and complex, therefore we seek formal schemes to parsimoniously represent the set of such distributions without being required to explicitly list all of them. It is remarkable that a single graph can represent the sum total all interventional distributions in such a compact and convenient way. This compactness however means that the interventional distributions are not arbitrary but highly structured. In other words, they are constrained by one another through a set of constraints that forces one interventional distribution to share properties with another. Our goal is to find meaningful and economical representations of these constraints by identifying their ”basis”, namely, a minimal set of constraints that imply all the others.

This exercise is similar in many ways to the one conducted in the 1980’s on ordinary Bayes netwroks [Pearl, 1988] where an economical basis was sought for the set of observational distributions represented in a DAG. Moving from probabilistic to causal Bayesian network will entail encoding of both probabilistic and interventional information by a single basis.

Formally, a causal Bayesian network (also known as a *Markovian model*) consists of two mathematical objects: (i) a DAG G, called a causal graph, over a set $\mathbf{V} = \{V_1, ..., V_n\}$ of vertices, and (ii) a probability distribution $P(\mathbf{v})$, over the set $\mathbf{V}$ of discrete variables that correspond to the vertices in G. The interpretation of the underlying graph has two components, one probabilistic and another causal, and we discuss in turn global and local characterizations of these two aspects/components.

2.1 Global characterization

We begin by reviewing the global conditions that provide an interpretation for the causal Bayesian networks. [4]

The probabilistic interpretation specifies that the full joint distribution is given by the product

$$P(\mathbf{v}) = \prod_i P(v_i \mid \mathbf{pa_i}) \tag{1}$$

where $\mathbf{pa_i}$ are (assignments of values to) the parents of variables V_i in G.

The causal interpretation is based on a global compatibility condition, which makes explicit the joint post-intervention distribution under any arbitrary intervention, and makes a parallel to the full factorization of the (pre-interventional) probabilistic interpretation. This condition states that any intervention is associated with the removal of the terms corresponding to the variables under intervention, reducing the product given by the expression in eq. (1) to the so called “truncated product” formula.

This operation is formalized in the following definition.

Definition 2 (Global causal condition [Pearl, 2000]). *A DAG G is said to be globally compatible with a set of interventional distributions $\mathbf{P}_*$ if and only if the distribution $P_\mathbf{x}(\mathbf{v})$ resulting from the intervention $do(\mathbf{X} = \mathbf{x})$ is given by the following expression:*

$$P_\mathbf{x}(\mathbf{v}) = \begin{cases} \prod_{\{i \mid V_i \notin \mathbf{X}\}} P(v_i \mid \mathbf{pa_i}) & \mathbf{v} \textit{ consistent with } \mathbf{x}. \\ 0 & \textit{otherwise.} \end{cases} \tag{2}$$

Eq. (2) is known as the *truncated factorization product*, since it has the factors corresponding to the manipulated variables “removed”. This formula can also be found

[4] A more refined interpretation, called functional, is also common [Pearl, 2000], which, in addition to interventions, supports counterfactual readings. The functional interpretation assumes deterministic functional relationships between variables in the model, some of which may be unobserved. Complete axiomatizations of deterministic counterfactual relations are given in [Galles and Pearl, 1998; Halpern, 1998].

in the literature under the name of “manipulation theorem” [Spirtes *et al.*, 1993] and is implicit in the “G-computation formula” [Robins, 1986]. Even when the graph is not available in its entirety, knowledge of the parents of each manipulated variable is sufficient for computing post-intervention from the preintervention distributions.

2.2 Local characterization

The truncated product is effective in computing post-interventional distributions but offers little help in the process of constructing the causal graph from judgemental knowledge. We next present a characterization that explicates a set of local assumptions leading to the global condition. Since the two definitions are syntactically very different, it is required to prove that they are (logically) equivalent.

The local characterization of causal Bayesian networks also consists of a DAG G and a probability distribution over $\mathbf{V}$, and the probabilistic interpretation [Pearl, 1988] in this characterization views G as representing conditional independence restrictions on P: each variable is independent of all its non-descendants given its parents in the graph. This property is known as the *Markov condition*, and can characterize the Bayesian network absent of any causal reading. Interestingly, the collection of independences assertions formed in this way suffices to derive the global assertion in eq. (1), and vice versa.

Worth to remark that this local characterization is most useful in *constructing* Bayesian networks, because selecting as parents the “direct causes” of a given variable automatically satisfies the local conditional independence conditions. On the other hand, the (probabilistic) global semantics leads directly to a variety of algorithms for reasoning.

More interestingly, the arrows in the graph G can be viewed as representing potential *causal influences* between the corresponding variables, and the factorization of eq. (1) still holds, but now the factors are further assumed to represent *autonomous data-generation processes*. That is, each family conditional probability $P(v_i \mid \mathbf{pa_i})$ represents a stochastic process by which the values of V_i are assigned in response to the values $\mathbf{pa_i}$ (previously chosen for V_i’s parents), and the stochastic variation of this assignment is assumed independent of the variations in all other assignments in the model.

This interpretation implies all conditional independence relations of the graph (dictated by Markov), and follows from two facts: (1) when we fix all parents, the only source of randomness for each variable is the stochastic variation pointing to the nodes [5]; (2) the stochastic variations are independent among themselves, which implies that each variable is independent of all its non-descendents.

This fact together with the additional assumption known as *modularity*, i.e., each assignment process remains *invariant* to possible changes in the assignments processes that govern other variables in the system, enable us to predict the effects of interventions, whenever interventions are described as specific modification of some factors in the product of eq. (1).

Note that the truncated factorization of the global definition follows trivially from this interpretation, because assuming modularity the post-intervention probabilities $P(v_i \mid$

[5] In the structural interpretation, they are represented by the error terms [Pearl, 2000, Ch. 7].

$\mathbf{pa_i}$) corresponding to variables in X are either 1 or 0, while those corresponding to unmanipulated variables remain unaltered.[6]

In order to formally capture the idea of invariance of the autonomous mechanism for each family entailed by the local characterization, the following definition encodes such feature facilitating subsequent discussions.

Definition 3 (Conditional invariance (CInv)). *We say that Y is conditionally invariant with respect to $\mathbf{X}$ given $\mathbf{Z}$, denoted $(Y \perp\!\!\!\perp_{ci} \mathbf{X} \mid \mathbf{Z})_{\mathbf{P}_*}$, if intervening on $\mathbf{X}$ does not change the conditional distribution of Y given $\mathbf{Z} = \mathbf{z}$, i.e., $\forall \mathbf{x}, y, \mathbf{z}, P_{\mathbf{x}}(y \mid \mathbf{z}) = P(y \mid \mathbf{z})$.*

We view *CInv* relations as the causal image of conditional independence (or simply *CInd*) relations, and a causal Baysian network as as representing both. Recast in terms of conditional invariance, [Pearl, 2000] proposed the following local definition of causal Bayesian networks:

Definition 4 (Modular causal condition [Pearl, 2000, p.24]). *A DAG G is said to be locally compatible with a set of interventional distributions $\mathbf{P}_*$ if and only if the following conditions hold for every $P_{\mathbf{x}} \in P_*$:*

i. *[Markov] $P_{\mathbf{x}}(\mathbf{v})$ is Markov relative to G;*
ii. *[Modularity] $(V_i \perp\!\!\!\perp_{ci} \mathbf{X} \mid \mathbf{PA_i})_{\mathbf{P}_*}$, for all $V_i \notin \mathbf{X}$ whenever $\mathbf{pa_i}$ is consistent with $\mathbf{X} = \mathbf{x}$.* [7]

In summary, the two definitions of CBNs emphasize different aspects of the causal model; Definition 4 ensures that each conditional probability $P(v_i \mid \mathbf{pa_i})$ (locally) remains invariant under interventions that do not include directly V_i, while Definition 2 ensures that each manipulated variable is not influenced by its previous parents (before the manipulation), and every other variable is governed by its pre-interventional process. Because the latter invokes theoretical conditions on the data-generating process, it is not directly testable, and the question whether a given implemented intervention conforms to an investigator's intention (e.g., no side effects) is discernible only through the testable properties of the truncated product formula (2). Definition 4 provides in essence a series of local tests for Eq. (2), and the equivalence between the two (Theorem 1 below) ensures that *all* empirically testable properties of Eq. (2) are covered by the local tests provided by Definition 4.

2.3 Example

Figure 1 illustrates a simple yet typical causal Bayesian network. It describes the causal relationships among the season of the year (X_1), whether it is raining (X_2), whether the sprinkler is on (X_3), whether the pavement is wet (X_4), and whether the pavement is slippery (X_5).

[6] In the literature, the other side of the implication is implicitly assumed to hold, but it is not immediately obvious, and it is object of our formal analysis in the next section.

[7] Explicitly, modularity states: $P(v_i|\mathbf{pa_i}, do(\mathbf{s})) = P(v_i|\mathbf{pa_i})$ for any set $\mathbf{S}$ of variables disjoint of $\{V_i, \mathbf{PA_i}\}$.

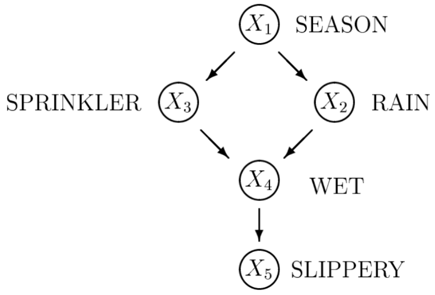

Fig. 1. A causal Bayesian network representing influence among five variables.

In the probabilistic interpretation given by the global definition, we can also use eq. (1) and write the full joint distribution:

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1)P(x_4 \mid x_2, x_3)P(x_5 \mid x_4) \quad (3)$$

Equivalently, the probabilistic interpretation entailed by the modular characterization induces the joint distribution respecting the constraints of conditional independences entailed by the graph through the underlying families. For example, $P(x_4 \mid x_2, x_3)$ is the probability of wetness given the values of sprinkler and rain, and it is independent of the value of season.

Nevertheless, both probabilistic interpretations say nothing about what will happen if a certain *intervention* occurs – i.e., a certain agent interact with the system and externally change the value of a certain variable (also known as action). For example, what if I *turn the sprinkler on*? What effect does that have on the season, or on the connection between wetness and slipperness?

The causal interpretation, intuitively speaking, adds the idea that whenever the sprinkler node is set to $X_3 = on$, so the event $(X_3 = on)$ has all mass of probability, which is clearly equivalent to as if the causal link between the season X_1 and the sprinkler X_3 is removed[8]. Assuming that all other causal links and conditional probabilities remain intact in the model, which is the less intrusive possible assumption to make, the new model that generates the process is given by the equation:

$$P(x_1, x_2, x_4, x_5 \mid do(X_3 = x_3)) = P(x_1)P(x_2 \mid x_1)P(x_4 \mid x_2, X_3 = on)P(x_5 \mid x_4)$$

where we informally demonstrate the semantic content of the *do* operator (also known as the *interventional* operator).

As another point, consider the problem of inferring the causal structure with two variables such that $\mathbf{V} = \{F, S\}$, and in which F stands for "Fire", and S stands for "Smoke". If we consider only the probabilistic interpretation, both structures, $G_1 = \{F \to S\}$ and $G_2 = \{S \to F\}$, are equivalent, and both networks are equally capable

[8] This can be shown more formally without difficulties.

of representing any joint distribution over these two variables. The global interpretation is hard to apply in this construction stage, but the modular interpretation is useful here. To see why, the definition helps one in choosing the causal network G_1 over G_2, because they encode different mechanisms, and so formally different responses under intervention – notice that there is a directed edge from S to F in G_2, but not in G_1. The modular condition as a collection of autonomous mechanisms that may be reconfigured locally by interventions, with the correspondingly local changes in the model, rejects the second network G_2 based on our understanding of the world. (A more transparent reasoning that makes us to prefer structure G_1 over G_2 should be even clearer when we discuss about missing-links in Section 4.)

3 The equivalence between the local and global definitions

We prove next that the local and global definitions of causal Bayesian networks are equivalent. To the best of our knowledge, the proof of equivalence has not been published before.

Theorem 1 (Equivalence between local and global compatibility). *Let G be a DAG and $\mathbf{P}_*$ a set of interventional distributions, the following statements are equivalent:*

i. *G is locally compatible with $\mathbf{P}_*$*
ii. *G is globally compatible with $\mathbf{P}_*$*

Proof. (Definition 4 ⇒ Definition 2)

Given an intervention $do(\mathbf{X} = \mathbf{x})$, $\mathbf{X} \subseteq \mathbf{V}$, assume that conditions 4:(i-ii) are satisfied. For any arbitrary instantiation $\mathbf{v}$ of variables $\mathbf{V}$, consistent with $\mathbf{X} = \mathbf{x}$, we can express $P_{\mathbf{x}}(\mathbf{v})$ as

$$\begin{aligned} P_{\mathbf{x}}(\mathbf{v}) &\overset{\text{def.4:}(i)}{=} \prod_i P_{\mathbf{x}}(v_i \mid \mathbf{pa_i}) \\ &= \prod_{\{i \mid v_i \in \mathbf{X}\}} P_{\mathbf{x}}(v_i \mid \mathbf{pa_i}) \prod_{\{i \mid v_i \notin \mathbf{X}\}} P_{\mathbf{x}}(v_i \mid \mathbf{pa_i}) \\ &\overset{\text{effectiveness}}{=} \prod_{\{i \mid v_i \notin \mathbf{X}\}} P_{\mathbf{x}}(v_i \mid \mathbf{pa_i}) \\ &\overset{\text{def.4:}(ii)}{=} \prod_{\{i \mid v_i \notin \mathbf{X}\}} P(v_i \mid \mathbf{pa_i}) \end{aligned} \tag{4}$$

which is the truncated product as desired.

(Definition 2 ⇒ Definition 4)

We assume that the truncated factorization holds, i.e., the distribution $P_{\mathbf{x}}(\mathbf{v})$ resulting from any intervention $do(\mathbf{X} = \mathbf{x})$ can be computed as eq. (2).

To prove effectiveness, consider an intervention $do(\mathbf{X} = \mathbf{x})$, and let $v_i \in \mathbf{X}$. Let $Dom(v_i) = \{v_{i1}, v_{i2}, ..., v_{im}\}$ be the domain of variable V_i, with only one of those values consistent with $\mathbf{X} = \mathbf{x}$. Since $P_{\mathbf{x}}(\mathbf{v})$ is a probability distribution, we must have

$\sum_j P_{\mathbf{x}}(V_i = v_{ij}) = 1$. According to eq. (2), all terms not consistent with $\mathbf{X} = \mathbf{x}$ have probability zero, and thus we obtain $P_{\mathbf{x}}(v_i) = 1, v_i$ *consistent with* $\mathbf{X} = \mathbf{x}$.

To show Definition 4:(ii), we consider an ordering $\pi : (v_1, ..., v_n)$ of the variables, consistent with the graph G induced by the truncated factorization with no intervention $P(\mathbf{v}) = \prod_i P(v_i \mid \mathbf{pa_i})$. Now, given an intervention $do(\mathbf{X} = \mathbf{x})$

$$P_{\mathbf{x}}(v_i \mid \mathbf{pa_i}) = \frac{P_{\mathbf{x}}(v_i, \mathbf{pa_i})}{P_{\mathbf{x}}(\mathbf{pa_i})}$$

$$\begin{aligned}
&\overset{\text{marginal.}}{=} \frac{\sum_{v_j \notin \{V_i, \mathbf{PA_i}\}} P_{\mathbf{x}}(\mathbf{v})}{\sum_{v_j \notin \{\mathbf{PA_i}\}} P_{\mathbf{x}}(\mathbf{v})} \\
&\overset{\text{eq.(2)}}{=} \frac{\sum_{v_j \notin \{V_i, \mathbf{PA_i}, \mathbf{X}\}} \prod_{v_k \notin \mathbf{X}} P(v_k \mid \mathbf{pa_k})}{\sum_{v_j \notin \{\mathbf{PA_i}, \mathbf{X}\}} \prod_{v_k \notin \mathbf{X}} P(v_k \mid \mathbf{pa_k})} \\
&= P(v_i \mid \mathbf{pa_i}) \times \\
&\quad \frac{\sum_{v_j \notin \{V_i, \mathbf{PA_i}, \mathbf{X}\}} \prod_{v_k \notin \mathbf{X}, k \neq i} P(v_k \mid \mathbf{pa_k})}{\sum_{v_j \notin \{\mathbf{PA_i}, \mathbf{X}\}} \prod_{v_k \notin \mathbf{X}} P(v_k \mid \mathbf{pa_k})}
\end{aligned} \tag{5}$$

The last step is due to the fact that variables in $\{V_i, \mathbf{PA_i}\}$ do not appear in the summations in the numerator. Rewriting the numerator, breaking it in relation to variables before and after v_i, we obtain

$$\begin{gathered}
\sum_{v_j \notin \{V_i, \mathbf{PA_i}, \mathbf{X}\}} \prod_{\substack{v_k \notin \mathbf{X} \\ k \neq i}} P(v_k \mid \mathbf{pa_k}) = \\
\sum_{\substack{v_j \notin \{\mathbf{PA_i}, \mathbf{X}\} \\ j < i}} \prod_{\substack{v_k \notin \mathbf{X} \\ k < i}} P(v_k \mid \mathbf{pa_k}) \sum_{\substack{v_j \notin \mathbf{X} \\ j > i}} \prod_{\substack{v_k \notin \mathbf{X} \\ k > i}} P(v_k \mid \mathbf{pa_k})
\end{gathered} \tag{6}$$

Note that $\sum_{\substack{v_j \notin \mathbf{X} \\ j > i}} \prod_{\substack{v_k \notin \mathbf{X} \\ k > i}} P(v_k \mid \mathbf{pa_k}) = 1$ because all $V_j > V_i$ appear in the summation. Thus, we obtain

$$\sum_{v_j \notin \{V_i, \mathbf{PA_i}, \mathbf{X}\}} \prod_{v_k \notin \mathbf{X}} P(v_k \mid \mathbf{pa_k}) = \sum_{\substack{v_j \notin \{\mathbf{PA_i}, \mathbf{X}\} \\ j < i}} \prod_{\substack{v_k \notin \mathbf{X} \\ k < i}} P(v_k \mid \mathbf{pa_k}) \tag{7}$$

Similarly for the denominator,

$$\sum_{v_j \notin \{\mathbf{PA_i}, \mathbf{X}\}} \prod_{v_k \notin \mathbf{X}} P(v_k \mid \mathbf{pa_k}) = \sum_{\substack{v_j \notin \{\mathbf{PA_i}, \mathbf{X}\} \\ j < i}} \prod_{\substack{v_k \notin \mathbf{X} \\ k < i}} P(v_k \mid \mathbf{pa_k}) \tag{8}$$

Observe that eqs. (7) and (8) are identical, equation (5) reduces to $P_{\mathbf{x}}(v_i \mid \mathbf{pa_i}) = P(v_i \mid \mathbf{pa_i})$ as desired.

To show Definition 4:(i), we first use the truncated factorization

$$\begin{aligned} P_{\mathbf{x}}(\mathbf{v}) &\overset{\text{eq.(2)}}{=} \prod_{\{i, v_i \notin \mathbf{X}\}} P(v_i \mid \mathbf{pa_i}) \\ &\overset{\text{def.4:}(ii)}{=} \prod_{\{i, v_i \notin \mathbf{X}\}} P_{\mathbf{x}}(v_i \mid \mathbf{pa_i}) \\ &\overset{\text{effectiveness}}{=} \prod_i P_{\mathbf{x}}(v_i \mid \mathbf{pa_i}) \end{aligned} \tag{9}$$

Finally, def. 4:(i) follows from the definition of Markov compatibility (definition 1.2.2 in [Pearl, 2000]).

4 Alternative characterizations of Causal Bayesian Networks

In this section we propose an interpretation of CBNs which focuses on the absence of edges in the causal graph, contrasting with the previous interpretation which focuses on the presence of edges in the causal graph. I.e., now we consider that the missing-links are relevant in the semantic perspective, which encode some sort of absence of causal influence (to be formally defined later on).

Interestingly, the idea that the missing-links carry meaningful information at a higher interpretation can be traced from much earlier. [Pearl, 1988] already discussed, in the pure probabilistic setting, when the absence of edges gives one clue about the inexistence of probabilistic dependence – or more directly, gives insight about the existence of probabilistic irrelevance (usually called *conditional independence*). This has been extensively studied, axiomatized and well understood since 80's.

Specifically, a pair of variables that are not connected by a link in the graph can be in one of the two possible exclusive conditions. First, if both variables are in some ancestral relation (one is ancestor of the other, or they have a common ancestor), then there exists a third set of variables that makes them conditionally independent in the probabilistically sense. Otherwise (i.e. they are not in an ancestral relation), the variables are simply marginally independent.[9] Thus we can see that the absence of an edge between a pair of nodes is conveying some strong qualitative claim about the relationships of the variables. Our goal in this section is to unfold the same kind of relation tailored specifically to the causal domain.

4.1 Missing-link characterization

Let us study the causal intuition analogous to what was just discussed, and consider again a pair of variables X and Y not connected by an edge, and the following two non-exclusive conditions. First, X is not an ancestor of Y, and second X does not have a directed edge going towards Y. In the former case, interventions on X do not affect Y at all, independently of what happens with any other variables in the system. In the

[9] It is possible to formally show both of these claims, but given that they are orthogonal to our goal here, we assume that it suffices for this context to just informally state them.

latter case, an intervention on X may affect Y, but there exists a set of variables that can break this "connection," which in the former case is the empty set.

To formally capture the intuition behind this type of causal invariance, (the idea of) "breaking" causal influence, we introduce the following relation over triplets.

Definition 5 (Interventional invariance (IInv)). *We say that Y is interventionally invariant with respect to $\mathbf{X}$ fixing $\mathbf{Z}$, denoted $(Y \perp\!\!\!\perp_{ii} \mathbf{X} \mid \mathbf{Z})_{\mathbf{P}_*}$, if intervening on $\mathbf{X}$ does not change the interventional distribution of Y given $do(\mathbf{Z} = \mathbf{z})$, i.e., $\forall \mathbf{x}, y, \mathbf{z}, P_{\mathbf{x},\mathbf{z}}(y) = P_{\mathbf{z}}(y)$.*

Now we can start envisioning CBNs through two orthogonal dimensions based on the just defined *IInv* relation. As discussed above, this definition yields an analogous claim of invariance – interventional-causal – contrasting with its probabilistic counterpart, *CInd* relation. Further note that the relations *CInv* and *IInv* (definitions 3 and 5) represent different types of invariance claims, the former relates to irrelevance given a certain observation, while the latter relates to irrelevance given simply another intervention. We are in both cases talking about breaking the causal flow and bringing about a causal invariance claim. (We discuss more about this issue subsequently in the paper.)

We are ready to formally state the definition that explicitly builds on *IInv* relation over the missing edges in the graph.

Definition 6 (Missing-link causal condition). *A DAG G is said to be missing-link compatible with a set of interventional distributions $\mathbf{P}_*$ if and only if the following conditions hold:*

i. *[Markov] $\forall \mathbf{X} \subseteq \mathbf{V}$, $P_{\mathbf{x}}(\mathbf{v})$ is Markov relative to G;*
ii. *[Missing-link] $\forall \mathbf{X} \subset \mathbf{V}, Y \in \mathbf{V}, \mathbf{S} \subset \mathbf{V}$, $(Y \perp\!\!\!\perp_{ii} \mathbf{X} \mid \mathbf{S}, \mathbf{PA_y})$ whenever there is no arrow from $\mathbf{X}$ to Y in G, $\mathbf{pa_y}$ is consistent with $\{\mathbf{X} = \mathbf{x}, \mathbf{S} = \mathbf{s}\}$ and $\mathbf{X}, \{Y\}, \mathbf{S}$ are disjoint.*
iii. *[Parents do/see] $\forall Y \in \mathbf{V}, \mathbf{X} \subset \mathbf{V}$, $P_{\mathbf{x},\mathbf{pa_y}}(y) = P_{\mathbf{x}}(y \mid \mathbf{pa_y})$ whenever $\mathbf{pa_y}$ is consistent with $\mathbf{X} = \mathbf{x}$ and $\mathbf{X}, \{Y\}$ are disjoint.*

Several remarks are worth to make at this point.

Remark 1: The missing-link condition 6:(ii) can be read as when we set $\mathbf{X}$ to some value while keeping the variables $\mathbf{S} \cup \mathbf{PA_y}$ constant, the marginal distribution of Y remains unaltered, independent of the value of $\mathbf{X}$, whenever there is no edge between $\mathbf{X}$ and Y. That is, an intervention on $\mathbf{X}$ does not change Y's distribution while holding constant its parents.

Remark 2: In addition to the missing-link condition, condition 6:(iii) describes the relationship inside each family, i.e., the effect on Y should be the same whether observing (seeing) or intervening (doing) on its parents $\mathbf{PA_y}$. That is, the missing-link condition has to be supplemented to be able to fully characterize causal Bayesian networks – condition 6:(iii) is necessary to describe the relationship between variables when *there exists* a link between them.

To illustrate this fact, consider a DAG G with only two binary variables, A and B, and an edge from A to B. Without condition 6:(iii), the interventional distribution $P_a(b)$ is unconstrained, which allows $P_a(b) \neq P(b \mid a)$. However, Definition 4 implies

$P_a(b) = P(b \mid a)$ since A is the only parent of B, and condition 6:(iii) ensures that this equality will hold.

Remark 3: The *CInd* claims encoded by the CBNs are of the form $(Y \perp\!\!\!\perp \mathbf{ND_Y} \mid \mathbf{PA_y})_{\mathbf{P}_*}$, where $\mathbf{ND_Y}$ represents the set of non-descendants of Y, and the *IInv* claims are of the form $(Y \perp\!\!\!\perp_{ii} X \mid \mathbf{PA_y}, \mathbf{S})_{\mathbf{P}_*}$, $\forall X, \mathbf{S}$. In both cases, $\mathbf{PA_y}$ is required to separate Y from other variables. In the observational case Y is separated from its non-descendants, while in the experimental one it is separated from all other variables. This is so because in the experimental case, an intervention on a descendant of a variable Z cannot influence Z (as is easily shown by d-separation in the mutilated graph).

Interestingly, the *IInv* relation as used in the definition impose strong invariance claims, and can be seen as sort of causal *Markov blankets*, making parallel with its probabilistically counterpart that separates each node from all others in the network.

Remark 4: More importantly, we argue that the missing-link definition is more intuitive than the previous ones because it relies exclusively on causal relationships in terms of which the bulk of scientific knowledge is encoded. This is so because *CInv* as well as *CInd* relations are subject to the phenomenon known as “explaining away” [Pearl, 1988], which can make the analysis of the data much more involved, while this does happen in the *IInv* relation.

For instance, it is quite possible for two unrelated variables to become related upon conditioning on a third variable, a phenomenon which surprises many people and is even viewed as an optical illusion.

More concretely, consider the graph $G = \{A \rightarrow B, C \rightarrow B, C \rightarrow D\}$, and from *modularity* follows that $P_A(D) = P(D)$. But it also follows that $P_A(D \mid B) \neq P_A(D)$, which looks contrived given that A should not be able to “causally affect” D given the topology of the graph, and our very first intuition about causality. It can be cumbersome to judge and systematically perform reasoning in terms of *CInvs*.

4.2 Pairwise characterization

The missing-link definition requires some non-trivial knowledge about the parent sets, which is not always available during the network construction phase. In this Section, we still consider the missing-links but in a different perspective, extending the previous definition towards a pairwise condition based on the more elementary notion of *zero direct effect*, which is even more aligned with our intuition about causal relationships, especially these emanating from typical experiments.

We want to capture the intuition behind the pairwise basic causal principle known as *ceteris paribus*, which says that whenever all variables in the system are fixed but X and Y (the universe is kept constant), if one “wriggles” X and Y does not “feel” these variations emanating from X, it is possible to conclude that there is no “direct causal effect” from X on Y. This notion conveys a different kind of claim that the one encoded in the previous missing-link condition, and now it is required that all variables in the system are held fixed and not only some subset of it. It turns out that both definitions are equivalent, to be shown next. First consider the following definition that formalizes this notion of *zero direct effect*.

Definition 7 (Zero direct effect). *Let $\mathbf{X} \subset \mathbf{V}$, $Y \in \mathbf{V}$ and $\mathbf{S_{XY}} = \mathbf{V} - \{X, Y\}$.* [10] *We say that $\mathbf{X}$ has zero direct effect on Y, denoted $ZDE(\mathbf{X}, Y)$, if*

$$(Y \perp\!\!\!\perp_{ii} \mathbf{X} \mid \mathbf{S_{xy}})$$

Now we are ready to incorporate this concept to CBNs as the following definition purports.

Definition 8 (Pairwise causal condition). *A DAG G is pairwise compatible with a set of interventional distributions $\mathbf{P}_*$ if the following conditions hold:*

i. *[Markov] $\forall \mathbf{X} \subseteq \mathbf{V}, P_{\mathbf{x}}(\mathbf{v})$ is Markov relative to G;*
ii. *[ZDE] $\forall X, Y \in \mathbf{V}$, $ZDE(X, Y)$ whenever there is no arrow from X to Y in G;*
iii. *[Additivity] $\forall \mathbf{X} \subset \mathbf{V}, Z, Y \in \mathbf{V}$, $ZDE(\mathbf{X}, Y)$ and $ZDE(Z, Y) \Rightarrow ZDE(\mathbf{X} \cup \{Z\}, Y)$;*
iv. *[Parents do/see] $\forall Y \in \mathbf{V}, \mathbf{X} \subset \mathbf{V}$, $P_{\mathbf{x},\mathbf{pa_y}}(y) = P_{\mathbf{x}}(y \mid \mathbf{pa_y})$ whenever $\mathbf{pa_y}$ is consistent with $\mathbf{X} = \mathbf{x}$ and $\mathbf{X}, \{Y\}$ are disjoint.*

The main feature of Definition 8 resides in the pairwise condition (ii), which implies that varying X from x to x' while keeping all other variables constant does not change Y's distribution – this corresponds to an ideal experiment in which all variables are kept constant and the scientist "wriggles" one variable (or set) at a time, and contemplates how the target variable reacts (i.e., *ceteris paribus*).

This condition is supplemented by condition 8:(iii), which has also an intuitive appeal: if experiments show that separate interventions on $\mathbf{X}$ and $\mathbf{Z}$ have no direct effect on Y, then a joint intervention on $\mathbf{X}$ and $\mathbf{Z}$ should also have no direct effect on Y (except for a set of measure zero). It is tempting to believe that this condition automatically holds, which usually happens in practice but still have to be formally contemplated for technical reasons. Conditions (i) and (iv) are the same as in the missing-link definition.

One distinct feature of this new definition emerges when we test whether a given pair $< G, \mathbf{P}_* >$ is compatible. First, the modularity condition of Definition 3 requires that each family is invariant to interventions on all subsets of elements "outside" the family, which is combinatorially explosive and rarely feasible to evaluate in practice. So the investigator is by her own without any normative procedure to help in constructing the structure of the network.

In contrast, condition (ii) involves singleton pairwise experiments which are easier to envision and perform. Put another way, when this pairwise condition does not hold, it implies the existence of an edge between the respective pair of nodes thus providing fewer and easier experiments in testing the structure of the graph. Further, one should test the Markov compatibility of P and the new induced graph G.

We finally state our main result that all three local definitions of causal Bayesian networks given so far are equivalent.

Theorem 2. *Let G be a DAG and $\mathbf{P}_*$ a set of interventional distributions, the following statements are equivalent:*

i. *G is locally compatible with $\mathbf{P}_*$*

[10] We use $\{\mathbf{A}, \mathbf{B}\}$ to denote the union of $\mathbf{A}$ and $\mathbf{B}$.

ii. *G is missing-link compatible with $\mathbf{P}_*$*
iii. *G is ZDE compatible with $\mathbf{P}_*$*

Note that even though the notion of “parents set” is less attached to modularity and invariance, it is still invoked by the last two compatibility conditions. Therefore we believe that it is an essential conceptual element in the definition of causal Bayesian networks.

The following result follows directly from Theorems 1 and 2.

Corollary 1. *All local and the global definitions of causal Bayesian networks are equivalent.*

5 Equivalence between the local definitions of causal Bayesian network

Definition 9 (Strong Markov Condition). *Each variable is interventionally independent of every other variable after fixing its parents. That is, for all $Y \in \mathbf{V}$ and $\mathbf{X} \subseteq \mathbf{V} - \{Y, \mathbf{PA_Y}\}$ we have*

$$P_{\mathbf{x},\mathbf{pa_y}}(y) = P_{\mathbf{pa_y}}(y), \textit{ for all } \mathbf{x}, y, \mathbf{pa_y} \tag{10}$$

5.1 [ZDE-CBN] ⇒ [local-CBN]

In this subsection, we assume that the four conditions in the definition of the Zero direct effect causal Bayesian network are valid for a given graph G and set $\mathbf{P}_*$.

The first result simply extends the Zero direct effect semantics to subset of variables:

Lemma 1. *$Zde(\mathbf{W}, Y)$ holds for every $\mathbf{W} \subseteq \mathbf{V} - \{Y, \mathbf{PA_Y}\}$.*

Proof. Note that $\mathbf{W}$ does not contain parents of Y. Then, [ZDE] gives that, for every U in $\mathbf{W}$, we have $Zde(U, Y)$. But then, it follows directly by [Additivity], that $Zde(\mathbf{W}, Y)$ holds.

The next Lemma shows that the strong Markov condition is also valid for G and $\mathbf{P}_*$.

Lemma 2. *For all $Y \in \mathbf{V}$ and $\mathbf{X} \subset \mathbf{V} - \{Y, \mathbf{PA_Y}\}$, the relation $(Y \perp\!\!\!\perp_{ii} \mathbf{X} \mid \mathbf{PA_Y})$ holds.*

Proof. Let $\mathbf{T_1} = \mathbf{V} - \{Y, \mathbf{PA_Y}\}$, and note that $S_{Y\mathbf{T_1}} = \mathbf{PA_Y}$. Since $\mathbf{T_1}$ does not have parents of Y, by Lemma 1, we have $Zde(\mathbf{T_1}, Y)$, that is

$$P_{\mathbf{t_1}, s_{y\mathbf{t_1}}}(y) = P_{s_{y\mathbf{t_1}}}(y) = P_{\mathbf{pa_y}}(y)$$

Now, let $\mathbf{T_2} = V - \{Y, \mathbf{X}, \mathbf{PA_Y}\}$, and note that $S_{Y\mathbf{T_2}} = \{\mathbf{X}, \mathbf{PA_Y}\}$. Since $\mathbf{T_2}$ does not have parents of Y, by Lemma 1, we have $Zde(\mathbf{T_2}, Y)$, that is

$$P_{\mathbf{t_2}, s_{y\mathbf{t_2}}}(y) = P_{s_{y\mathbf{t_2}}}(y) = P_{x,\mathbf{pa_y}}(y)$$

Since $(\mathbf{T_1} \cup S_{Y\mathbf{T_1}}) = (\mathbf{T_2} \cup S_{Y\mathbf{T_2}})$, we obtain

$$P_{\mathbf{x},\mathbf{pa_y}}(y) = P_{\mathbf{pa_y}}(y)$$

Lemma 3. *The condition of [Modularity] is valid for G and $\mathbf{P}_*$.*

Proof. Fix a variable Y and $\mathbf{X} \subset \mathbf{V} - \{Y\}$. We need to show that

$$P_{\mathbf{x}}(y \mid \mathbf{pa_y}) = P(y \mid \mathbf{pa_y})$$

Applying the condition [Parents do/see] to both sides in the equation above, we obtain

$$P_{x,\mathbf{pa_y}}(y) = P_{\mathbf{pa_y}}(y)$$

and we immediately recognize here a claim of the strong Markov condition.

Finally, the observation that the condition [Markov] is present in both definitions, we complete the proof that G is a local causal Bayesian network for $\mathbf{P}_*$.

5.2 [local-CBN] ⇒ [ZDE-CBN]

In this subsection, we assume that the two conditions in the definition of the local causal Bayesian network are valid for a given graph G and set $\mathbf{P}_*$.

Lemma 4. *For all $Y \in \mathbf{V}$ and $\mathbf{X} \subset \mathbf{V} - \{Y, \mathbf{PA_Y}\}$ we have*

$$P_{\mathbf{x},\mathbf{pa_y}}(\mathbf{pa_y} \mid y) = 1$$

whenever $P_{\mathbf{x},\mathbf{pa_y}}(y) > 0$, and $\mathbf{pa_y}$ is compatible with $\mathbf{x}$.

Proof. This is an immediate consequence of the property of [Effectiveness], in the definition of $\mathbf{P}_*$.

Lemma 5. *The condition [Parents do/see] is valid for G and $\mathbf{P}_*$.*

Proof. Fix a variable $\mathbf{X} \subset \mathbf{V}$ and consider an arbitrary instantiation $\mathbf{v}$ of variables $\mathbf{V}$, and $\mathbf{pa_y}$ consistent with $\mathbf{x}$.

Consider the intervention $do(\mathbf{X} = \mathbf{x})$, and given the condition [Modularity], $P_{\mathbf{x}}(y \mid \mathbf{pa_y}) = P(y \mid \mathbf{pa_y}), Y \notin \mathbf{X}$. Now consider the intervention $do(\mathbf{X} = \mathbf{x}, \mathbf{PA_Y} = \mathbf{pa_y})$, and again by the condition [Modularity] $P_{\mathbf{x},\mathbf{pa_y}}(y \mid \mathbf{pa_y}) = P(y \mid \mathbf{pa_y})$. The r.h.s. coincide, therefore

$$\begin{aligned} P_{\mathbf{x}}(y \mid \mathbf{pa_y}) &= P_{\mathbf{x},\mathbf{pa_y}}(y \mid \mathbf{pa_y}) \\ &\overset{\text{Bayes thm.}}{=} \frac{P_{\mathbf{x},\mathbf{pa_y}}(\mathbf{pa_y} \mid y) P_{\mathbf{x},\mathbf{pa_y}}(y)}{P_{\mathbf{x},\mathbf{pa_y}}(\mathbf{pa_y})} \\ &\overset{\text{effectiveness}}{=} P_{\mathbf{x},\mathbf{pa_y}}(\mathbf{pa_y} \mid y) P_{\mathbf{x},\mathbf{pa_y}}(y) \end{aligned} \tag{11}$$

We consider two cases. If $P_{\mathbf{x},\mathbf{pa_y}}(y) > 0$, by lemma 4 $P_{\mathbf{x},\mathbf{pa_y}}(\mathbf{pa_y} \mid y) = 1$, and then substituting back in eq. (11) we obtain $P_{\mathbf{x}}(y \mid \mathbf{pa_y}) = P_{\mathbf{x},\mathbf{pa_y}}(y)$. If $P_{\mathbf{x},\mathbf{pa_y}}(y) = 0$, substituting back in eq. (11) we obtain $P_{\mathbf{x}}(y \mid \mathbf{pa_y}) = P_{\mathbf{x},\mathbf{pa_y}}(\mathbf{pa_y} \mid y) * 0 = 0$, and then $P_{\mathbf{x}}(y \mid \mathbf{pa_y}) = P_{\mathbf{x},\mathbf{pa_y}}(y)$.

Lemma 6. *The condition [ZDE] is valid for G and $\mathbf{P}_*$.*

Proof. Fix $Y, X \in \mathbf{V}$ such that there is no arrow pointing from X to Y. Let $\mathbf{S_{XY}} = \mathbf{V} - \{X, Y\}$. We want to show

$$P_{x,\mathbf{s_{xy}}}(y) = P_{\mathbf{s_{xy}}}(y), \text{for all } x, y, \mathbf{s_{xy}}$$

Note that $\mathbf{PA_y} \subseteq \mathbf{S_{xy}}$, and then by the [Parent do/see] condition we have to show

$$P_{x,\mathbf{s'_{xy}}}(y \mid \mathbf{pa_y}) = P_{\mathbf{s'_{xy}}}(y \mid \mathbf{pa_y})$$

where $\mathbf{S'_{xy}} = \mathbf{S_{xy}} - \{\mathbf{PA_y}\}$.

The condition [Modularity] implies that $P_{x,\mathbf{s'_{xy}}}(y \mid \mathbf{pa_y}) = P(y \mid \mathbf{pa_y})$. Again by [Modularity], we obtain $P(y \mid \mathbf{pa_y}) = P_{\mathbf{s'_{xy}}}(y \mid \mathbf{pa_y})$. Applying [Parents do/see], [ZDE] follows.

Lemma 7. *The condition [Additivity] is valid for G and $\mathbf{P}_*$.*

Proof. Fix $\mathbf{X} \subset \mathbf{V}$ and $Z, Y \in \mathbf{V}$. Let $\mathbf{S_{xzy}} = \mathbf{V} - \{\mathbf{X}, Y, Z\}$. Assume $Zde(\mathbf{X}, Y)$ and $Zde(Z, Y)$. For the sake of contradiction, suppose that $Zde(\mathbf{X} \cup \{Z\}, Y)$ is false.

We can rewrite it based on the law of total probability,

$$\sum_{\mathbf{pa_y}} P_{\{\mathbf{x},z\},\mathbf{s_{xzy}}}(y \mid \mathbf{pa_y}) P_{\{\mathbf{x},z\},\mathbf{s_{xzy}}}(\mathbf{pa_y}) \neq \sum_{\mathbf{pa_y}} P_{\mathbf{s_{xzy}}}(y \mid \mathbf{pa_y}) P_{\mathbf{s_{xzy}}}(\mathbf{pa_y})$$

Notice that there is only one configuration of $\mathbf{pa_y}$ consistent with $\mathbf{s_{xzy}}$ in both sides because $\mathbf{PA_y} \subseteq \mathbf{S_{xzy}}$ and [Effectiveness]. Then, this equation reduces to

$$P_{\{\mathbf{x},z\},\mathbf{s_{xzy}}}(y \mid \mathbf{pa_y}) \neq P_{\mathbf{s_{xzy}}}(y \mid \mathbf{pa_y})$$

We reach a contradiction given [Modularity].

The proof for the Missing-link CBN is analogous for the just shown.

6 Conclusions

We first proved the equivalence between two characterizations of Causal Bayesian Networks, one local, based on the modularity condition, and another global, based on the truncated product formula. We then introduced two alternative characterizations of CBNs, proved their equivalence with the previous ones, and showed that some of their features make the tasks of empirically testing the network structure, as well as judgmentally assessing its plausibility more manageable.

Another way to look at the results of our analysis is in terms of the information content of CBNs, that is, what constraints a given CBN imposes on both observational and experimental findings.

For a probabilistic Bayes network the answer is simple and is given by the set of conditional independences that are imposed by the d-separation criterion. For a CBN, the truncated product formula (2) imposes conditional independencies on any interventional distribution $P_x(\mathbf{v})$. But this does not sum up the entire information content of a CBN. The truncated product formula further tells us that the relationship between any two interventional distributions, say $P_x(\mathbf{v})$ and $P_{x'}(\mathbf{v})$, is not entirely arbitrary; the two distributions constrain each other in various ways. For example, the conditional distributions $P_x(v_i|\mathbf{pa_i})$ and $P_{x'}(v_i|\mathbf{pa_i})$ must be the same for any unmanipulated family. Or, as another example, for any CBN we have the inequality: $P_x(y) \geq P(x, y)$ [Tian *et al.*, 2006].

A natural question to ask is whether there exists a representation that encodes all constraints of a given type. The modularity property of Definition 4 constitutes such a representation, and so do the missing-link and the pairwise definitions. Each encodes constraints of a given type and our equivalence Theorems imply that all constraints encoded by one representation can be reconstructed from the other representation without loss of information.

Bibliography

A. P. Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189, 2001.

D. Galles and J. Pearl. An axiomatic characterization of causal counterfactuals. *Foundation of Science*, 3(1):151–182, 1998.

D. Geiger, T.S. Verma, and J. Pearl. Identifying independence in Bayesian networks. In *Networks*, volume 20, pages 507–534. John Wiley, Sussex, England, 1990.

J.Y. Halpern. Axiomatizing causal reasoning. In G.F. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence*, pages 202–210. Morgan Kaufmann, San Francisco, CA, 1998. Also, *Journal of Artificial Intelligence Research* 12:3, 17–37, 2000.

D. Heckerman and R. Shachter. Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, 1995.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

S.L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996.

S.L. Lauritzen. Causal inference from graphical models. In *Complex Stochastic Systems*, pages 63–107. Chapman and Hall/CRC Press, 1999.

D.V. Lindley. Seeing and doing: The concept of causation. *International Statistical Review*, 70:191–214, 2002.

J. Pearl and T. Verma. The logic of representing dependencies by directed acyclic graphs. In *Proceedings of the Sixth National Conference on AI (AAAI-87)*, pages 374–379, Seattle, WA, July 1987.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.

J. Pearl. Belief networks revisited. *Artificial Intelligence*, 59:49–56, 1993.

J. Pearl. A probabilistic calculus of actions. In R. Lopez de Mantaras and D. Poole, editors, *Uncertainty in Artificial Intelligence 10*, pages 454–462. Morgan Kaufmann, San Mateo, CA, 1994.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. Second ed., 2009.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, second edition, 2009.

J.M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

P. Spirtes, C.N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer-Verlag, New York, 1993.

J. Tian and J. Pearl. A new characterization of the experimental implications of causal Bayesian networks. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pages 574–579. AAAI Press/The MIT Press, Menlo Park, CA, 2002.

J. Tian, C. Kang, and J. Pearl. A characterization of interventional distributions in semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pages 1239–1244. AAAI Press, Menlo Park, CA, 2006.