1.3 Causal Bayesian Networks

The interpretation of directed acyclic graphs as carriers of independence assumptions does not necessarily imply causation; in fact, it will be valid for any set of recursive independencies along any ordering of the variables, not necessarily causal or chronological. However, the ubiquity of DAG models in statistical and AI applications stems (often unwittingly) primarily from their causal interpretation—that is, as a system of processes, one per family, that could account for the generation of the observed data. It is this causal interpretation that explains why DAG models are rarely used in any variable ordering other than those which respect the direction of time and causation.

The advantages of building DAG models around causal rather than associational information are several. First, the judgments required in the construction of the model are more meaningful, more accessible, and hence more reliable. The reader may appreciate this point by attempting to construct a DAG representation for the associations in Figure 1.2 along the ordering $(X_5, X_1, X_3, X_2, X_4)$. Such exercises illustrate not only that some independencies are more vividly accessible to the mind than others but also that conditional independence judgments are accessible (hence reliable) only when they are anchored onto more fundamental building blocks of our knowledge, such as causal relationships. In the example of Figure 1.2, our willingness to assert that X_5 is independent of X_2 and X_3 once we know X_4 (i.e., whether the pavement is wet) is defensible because we can easily translate the assertion into one involving causal relationships: that the *influence* of rain and sprinkler on slipperiness is *mediated* by the wetness of the pavement. Dependencies that are not supported by causal links are considered odd or spurious and are even branded "paradoxical" (see the discussion of Berkson's paradox, Section 1.2.3).

We will have several opportunities throughout this book to demonstrate the primacy of causal over associational knowledge. In extreme cases, we will see that people tend to ignore probabilistic information altogether and attend to causal information instead (see Section 6.1.4).[8] This puts into question the ruling paradigm of graphical models in

[8] Tversky and Kahneman (1980) experiments with causal biases in probability judgment constitute another body of evidence supporting this observation. For

statistics (Wermuth and Lauritzen 1990; Cox and Wermuth 1996), according to which conditional independence assumptions are the primary vehicle for expressing substantive knowledge.[9] It seems that if conditional independence judgments are byproducts of stored causal relationships, then tapping and representing those relationships directly would be a more natural and more reliable way of expressing what we know or believe about the world. This is indeed the philosophy behind causal Bayesian networks.

The second advantage of building Bayesian networks on causal relationships—one that is basic to the understanding of causal organizations—is the ability to represent and respond to external or spontaneous *changes*. Any local reconfiguration of the mechanisms in the environment can be translated, with only minor modification, into an isomorphic reconfiguration of the network topology. For example, to represent a disabled sprinkler in the story of Figure 1.2, we simply delete from the network all links incident to the node Sprinkler. To represent the policy of turning the sprinkler off if it rains, we simply add a link between Rain and Sprinkler and revise $P(x_3|x_1, x_2)$. Such changes would require much greater remodeling efforts if the network were not constructed along the causal direction but instead along (say) the order $(X_5, X_1, X_3, X_2, X_4)$. This remodeling flexibility may well be cited as the ingredient that marks the division between deliberative and reactive agents and that enables the former to manage novel situations instantaneously, without requiring training or adaptation.

1.3.1 Causal Networks as Oracles for Interventions

The source of this flexibility rests on the assumption that each parent-child relationship in the network represents a stable and autonomous physical mechanism—in other words, that it is conceivable to change one such relationship *without* changing the others. Organizing one's knowledge in such modular configurations permits one to predict the

example, most people believe that it is more likely for a girl to have blue eyes, given that her mother has blue eyes, than the other way around; the two probabilities are in fact equal.

[9]The author was as guilty of advocating the centrality of conditional independence as were his colleagues in statistics; see Pearl (1988b, p. 79).

effect of external interventions with minimum of extra information. Indeed, causal models (assuming they are valid) are much more informative than probability models. A joint distribution tells us how probable events are and how probabilities would change with subsequent observations, but a causal model also tells us how these probabilities would change as a result of external interventions—such as those encountered in policy analysis, treatment management, or planning everyday activity. Such changes cannot be deduced from a join distribution, even if fully specified.

The connection between modularity and interventions is as follows. Instead of specifying a new probability function for each of the many possible interventions, we specify merely the immediate change implied by the intervention and, by virtue of autonomy, we assume that the change is local, and does not spread over to mechanisms other than those specified. Once we know the identity of the mechanism altered by an intervention and the nature of the alteration, the overall effect of an intervention can be predicted by modifying the corresponding factors in (1.33) and using the modified product to compute a new probability function. For example, to represent the action "turning the sprinkler On" in the network of Figure 1.2, we delete the link $X_1 \longrightarrow X_3$ and assign X_3 the value On. The graph resulting from this operation is shown in Figure 1.4, and the resulting joint distribution on the remaining variables will be

$$P_{X_3=\text{On}}(x_1, x_2, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_4|x_2, X_3 = \text{On})P(x_5|x_4), \tag{1.36}$$

in which all the factors on the right-hand side (r.h.s.), by virtue of autonomy, are the same as in (1.34).

The deletion of the factor $P(x_3|x_1)$ represents the understanding that, whatever relationship existed between seasons and sprinklers prior to the action, that relationship is no longer in effect while we perform the action. Once we physically turn the sprinkler on and keep it on, a new mechanism (in which the season has no say) determines the state of the sprinkler.

Note the difference between the action $do(X_3 = \text{On})$ and the observation $X_3 = \text{On}$. The effect of the latter is obtained by ordinary Bayesian conditioning, that is, $P(x_1, x_2, x_4, x_5|X_3 = \text{On})$, while that of

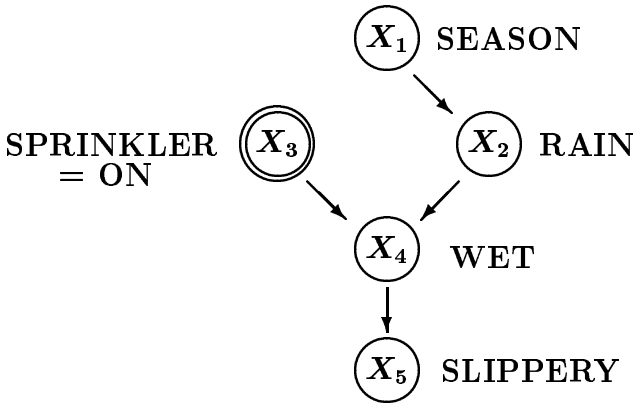

Figure 1.4: Network representation of the action "turning the sprinkler On."

the former by conditioning a mutilated graph, with the link $X_1 \longrightarrow X_3$ removed. This mirrors indeed the difference between seeing and doing: after *observing* that the sprinkler is on, we wish to infer that the season is dry, that it probably did not rain, and so on; no such inferences should be drawn in evaluating the effects of a contemplated *action* "turning the sprinkler On."

The ability of causal networks to predict the effects of actions requires of course a stronger set of assumptions in the construction of those networks, assumptions that rest on causal (not merely associational) knowledge and that ensure the system would respond to interventions in accordance with the principle of autonomy. These assumptions are encapsulated in the following definition of causal Bayesian networks.

Definition 1.3.1 (Causal Bayesian Network)
Let $P(v)$ be a probability distribution on a set V of variables, and let $P_x(v)$ denote the distribution resulting from the intervention $do(X = x)$ that sets a subset X of variables to constants x.[10] *Denote by $\boldsymbol{P}_*$ the set of all interventional distributions $P_x(v)$, $X \subseteq V$, including $P(v)$, which represents no intervention (i.e., $X = \emptyset$). A DAG G is said to be a* causal Bayesian network *compatible with $\boldsymbol{P}_*$ if and only if the following three conditions hold for every $P_x \in \boldsymbol{P}_*$:*

[10] The notation $P_x(v)$ will be replaced in subsequent chapters with $P(v|do(x))$ and $P(v|\hat{x})$ to facilitate algebraic manipulations.

(i) $P_x(v)$ *is Markov relative to* G;

(ii) $P_x(v_i) = 1$ *for all* $V_i \in X$ *whenever* v_i *is consistent with* $X = x$;

(iii) $P_x(v_i|pa_i) = P(v_i|pa_i)$ *for all* $V_i \notin X$ *whenever* pa_i *is consistent with* $X = x$.

Definition 1.3.1 imposes constraints on the interventional space $\boldsymbol{P}_*$ that permit us to encode this vast space economically, in the form of a single Bayesian network G. These constraints enable us to compute the distribution $P_x(v)$ resulting from any intervention $do(X = x)$ as a *truncated factorization*

$$P_x(v) = \prod_{\{i|V_i \notin X\}} P(v_i|pa_i) \quad \text{for all } v \text{ consistent with } x, \tag{1.37}$$

which follows from Definition 1.3.1 and justifies the family deletion procedure on G, as in (1.36). It is not hard to show that, whenever G is a causal Bayes network with respect to $\boldsymbol{P}_*$, the following two properties must hold.

Property 1 *For all* i,

$$P(v_i|pa_i) = P_{pa_i}(v_i). \tag{1.38}$$

Property 2 *For all* i *and for every subset* S *of variables disjoint of* $\{V_i, PA_i\}$, *we have*

$$P_{pa_i,s}(v_i) = P_{pa_i}(v_i). \tag{1.39}$$

Property 1 renders every parent set PA_i *exogenous* relative to its child V_i, ensuring that the conditional probability $P(v_i|pa_i)$ coincides with the effect (on V_i) of setting PA_i to pa_i by external control. Property 2 expresses the notion of invariance; once we control its direct causes PA_i, no other interventions will affect the probability of V_i.

1.3.2 Causal Relationships and Their Stability

This mechanism-based conception of interventions provides a semantical basis for notions such as "causal effects" or "causal influence," to be defined formally and analyzed in Chapters 3 and 4. For example, to test whether a variable X_i has a causal influence on another variable X_j, we compute (using the truncated factorization formula of (1.37)) the (marginal) distribution of X_j under the actions $do(X_i = x_i)$—namely, $P_{x_i}(x_j)$ for all values x_i of X_i—and test whether that distribution is sensitive to x_i. It is easy to see from our previous examples that only variables that are descendants of X_i in the causal network can be influenced by X_i; deleting the factor $P(x_i|pa_i)$ from the joint distribution turns X_i into a root node in the mutilated graph, and root variables (as the d-separation criterion dictates) are independent of all other variables except their descendants.

This understanding of causal influence permits us to see precisely why, and in what way, causal relationships are more "stable" than probabilistic relationships. We expect such difference in stability because causal relationships are *ontological*, describing objective physical constraints in our world, whereas probabilistic relationships are *epistemic*, reflecting what we know or believe about the world. Therefore, causal relationships should remain unaltered as long as no change has taken place in the environment, even when our knowledge about the environment undergoes changes. To demonstrate, consider the causal relationship S_1, "Turning the sprinkler on would not affect the rain," and compare it to its probabilistic counterpart S_2, "The state of the sprinkler is independent of (or unassociated with) the state of the rain." Figure 1.2 illustrates two obvious ways in which S_2 will change while S_1 remains intact. First, S_2 changes from false to true when we learn what season it is (X_1). Second, given that we know the season, S_2 changes from true to false once we observe that the pavement is wet ($X_4 =$ true). On the other hand, S_1 remains true regardless of what we learn or know about the season or about the pavement.

The example reveals a stronger sense in which causal relationships are more stable than the corresponding probabilistic relationships, a sense that goes beyond their basic ontological-epistemological difference. The relationship S_1 will remain invariant to changes in the mech-

anism that regulates how seasons affect sprinklers. In fact, it remains invariant to changes in *all* mechanisms shown in this causal graph. We thus see that causal relationships exhibit greater robustness to ontological changes as well; they are sensitive to a smaller set of mechanisms. More specifically, and in marked contrast to probabilistic relationships, causal relationships remain invariant to changes in the mechanism that governs the causal variables (X_3 in our example).

In view of this stability, it is no wonder that people prefer to encode knowledge in causal rather than probabilistic structures. Probabilistic relationships, such as marginal and conditional independencies, may be helpful in hypothesizing initial causal structures from uncontrolled observations. However, once knowledge is cast in causal structure, those probabilistic relationships tend to be forgotten; whatever judgments people express about conditional independencies in a given domain are derived from the causal structure acquired. This explains why people feel confident asserting certain conditional independencies (e.g., that the price of beans in China is independent on the traffic in Los Angeles) having no idea whatsoever about the numerical probabilities involved (e.g., whether the price of beans will exceed $10 per bushel).

The element of stability (of mechanisms) is also at the heart of the so-called explanatory accounts of causality, according to which causal models need not encode behavior under intervention but instead aim primarily to provide an "explanation" or "understanding" of how data are generated.[11] Regardless of what use is eventually made of our "understanding" of things, we surely would prefer an understanding in terms of durable relationships, transportable across situations, over those based on transitory relationships. The sense of "comprehensibility" that accompanies an adequate explanation is a natural byproduct of the transportability of (and hence of our familiarity with) the causal relationships used in the explanation. It is for reasons of stability that we regard the falling barometer as predicting but not explaining the rain; those predictions are not transportable to situations where the pressure surrounding the barometer is controlled by artificial means. True understanding enables predictions in such novel situations, where

[11] Elements of this explanatory account can be found in the writings of Dempster (1990), Cox (1992), and Shafer (1996a); see also King et al. (1994, p. 75).

some mechanisms change and others are added. It thus seems reasonable to suggest that, in the final analysis, the explanatory account of causation is merely a variant of the manipulative account, albeit one where interventions are dormant. Accordingly, we may as well view our unsatiated quest for understanding "how data is generated" or "how things work" as a quest for acquiring the ability to make predictions under wider range of circumstances, including circumstances in which things are taken apart, reconfigured, or undergo spontaneous change.