R-500

REVISED
3:50 pm, Jul 18, 2019

Section in Stuart Russell's Chapter 13

Probabilistic Reasoning[1]

Judea Pearl
Computer Science Department
University of California, Los Angeles
Los Angeles, CA, 90095-1596, USA
judea@cs.ucla.edu

July 18, 2019

[1]There is a general topological criterion called *d-separation* for deciding whether a set of nodes **X** is conditionally independent of another set **Y**, given a third set **Z** (Pearl, 1988; Shachter, 1998). This criterion will be described in the Section 13.5.

13.5 Causality

We have discussed several advantages of keeping node ordering compatible with the direction of causation. In particular, we noted the ease at which conditional probabilities can be assessed if such ordering is maintained, as well as the compactness of the resultant network structure. We noted however that, in principle, any node ordering permits a consistent construction of the network to represent the joint distribution function. This was demonstrated in Figure 13.3(b) which, although bushier and a less natural than Figure 13.3(a), enabled us nevertheless to represent the same distribution on all variables. In this section we will study Causal Bayesian networks, a restricted class of Bayesian networks that forbids all but causally compatible orderings. We will explore how to construct such networks, what is gained by such construction, and how to leverage this gain in decision making tasks.

Consider the simplest Bayesian network imaginable, a single arrow, $\mathit{Fire} \rightarrow \mathit{Smoke}$. It tells as that variables *Fire* and *Smoke* may be dependent, so one needs to specify the prior $P(\mathit{Fire})$ and the conditional probability $P(\mathit{Smoke}|\mathit{Fine})$ in order to specify the joint distribution $P(\mathit{Fire, Smoke})$. However, this distribution can be represented equally well by the reverse arrow $\mathit{Fire} \leftarrow \mathit{Smoke}$, using the appropriate $P(\mathit{Smoke})$ and $P(\mathit{Fire}|\mathit{Smoke})$ computed from Bayes rule. The idea that these two networks are equivalent, hence convey the same information, evokes discomfort and even resistance in most people. How could they convey the same information when we know that *Fire* causes *Smoke* and not the other way around? In other words, we know from our experience and scientific understanding that clearing the smoke would not stop the fire and extinguishing the fire will stop the smoke. We expect therefore to represent this asymmetry through the directionality of the arrow between them. But if arrow reversal only makes things equivalent, how can we hope to represent this important information formally.

Causal Bayesian network, sometimes called Causal Diagrams, were devised to permit us to represent causal asymmetries and to leverage the asymmetries towards reasoning with causal information. The idea is to decide on arrow directionality by considerations that go beyond probabilistic dependence and invoke a totally different type of judgement. Instead of asking an expert whether *Smoke* and *Fire* are "dependent," as we do in ordinary Bayesian networks, we now ask "WHO IS RESPONDING TO WHOM?," *Smoke* to *Fire* or *Fire* to *Smoke*? This may sound a bit metaphorical or even anthropomorphic, but it can be made precise through the notion of "assignment," similar to the assignment operator := in programming languages. If nature assigns a value to *Smoke* on the basis of what Nature learns about *Fire* we draw an arrow from *Smoke* to *Fire*. More importantly, if we judge that Nature assigns *Fire* a truth value that depends on other variables, not *Smoke*, we refrain from drawing the arrow *Fire* $\leftarrow$ *Smoke*. In other words, the value v_i of each variable V_i is determined by a function $f_i(Other\ Variables)$, and an arrow $V_j \rightarrow V_i$ is drawn if and only if V_j is an argument of f_i.

The function f_i is called *structural equation*, because it describes a stable mechanism in nature which, unlike the probabilities that quantify a Bayesian network, remain invariant to measurements and local changes in the environment

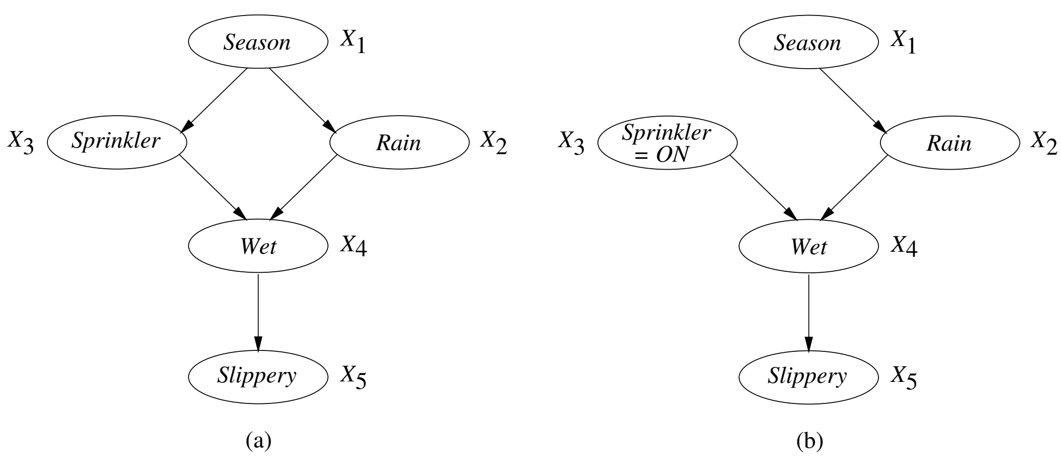

Figure 13.22: (a) A Causal Bayesian network representing cause-effect relations among five variables. (b) Same, after performing the action "*Turn Sprinkler ON*."

To appreciate this stability to local changes, consider Figure 13.22(a), which depicts a slightly modified version of the *Sprinkler* story of Figure 13.14(a). To represent a disabled sprinkler, for example, we simply delete from the network all links incident to the node "*Sprinkler.*" To represent a lawn covered by a tent, we simply delete the arrow *Rain* $\rightarrow$ *Wet Grass.* Any local reconfiguration of the mechanisms in the environment can thus be translated, with only minor modification, into an isomorphic reconfiguration of the network topology. A much more elaborate transformation would be required had the network been constructed contrary to causal ordering. This local stability is particularly important for representing "actions" or "interventions," our next topic of discussion.

13.5.1 Representing actions, the *do*-operator

Consider again the *Sprinkler* story of Figure 13.22(a). We know from Eq. (13.16) that the joint distribution of the five variable is given by a product of five conditional probabilities:

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)\ P(x_2|x_1)\ P(x_3|x_1)\ P(x_4|x_2, x_3)\ P(x_5|x_4) \qquad (13.15)$$

Let us examine now how this distribution would vary if we turn the sprinkler on, namely, if we intervene on the system and impose the condition *Sprinkler* = *ON* by an external force. Since the arrows in the network were judged to represent causal mechanisms, rather than conditional probabilities, the corresponding model would consist of five functions, each representing an autonomous mechanism:

$$\begin{aligned}
X_1 &= U_1 \\
X_2 &= f_2(X_1, U_2) \\
X_3 &= f_3(X_1, U_3) \\
X_4 &= f_4(X_3, X_2, U_4) \\
X_5 &= f_5(X_4, U_5) \qquad (13.16)
\end{aligned}$$

The U's variables (not shown in Figure 13.22(a)) represent unobserved factors, also called "error terms" or "disturbances" which perturb the functional relationships between each variable and its parents. For example, U_4 may represent another potential source of water, in addition to *Sprinkler* and *Rain* (perhaps *Morning Due* or

Firefighting Helicopter). It can be easily shown (Pearl, 2000 (*Causality*)) that, if the U variables are mutually independent, then the probability distribution generated by this system of equations obeys the Markovian property of classical Bayesian networks (Eq. (13.17) and, hence, the joint probability in this example will be given by the product of Eq. (13.15)).

To represent the action "turning the sprinkler ON," denoted $do(X_3 = \mathrm{ON})$, we note that, once the action is enforced the sprinkler will no longer be responding to variations in *Season*, but will permanently be assigned the value ON. We therefore delete the equation $X_3 = f_3(X_1, U_3)$ from the model of Eq. (13.16), and replace it with the equation $X_3 = \mathrm{ON}$.

The resulting submodel, $M_{X_3=\mathrm{ON}}$, will take the form:

$$\begin{aligned} X_1 &= U_1 \\ X_2 &= f_2(X_1, U_2) \\ X_3 &= ON \\ X_4 &= f_4(X_3, X_2, U_4) \\ X_5 &= f_5(X_4, U_5) \end{aligned} \tag{13.17}$$

and, accordingly, will induce the probability distribution

$$P(x_1, x_2, x_4, x_5 | do(X_3 = ON) = P(x_1)\ P(x_2|x_1)\ P(x_4|x_2, X_3 = ON)\ P(x_5|x_4) \tag{13.18}$$

in which the factor $P(x_3|x_1)$ is absent and variable X_3 (*Sprinkler*) is instantiated to ON. It is easy to see from Eq. (13.17) that the only variables affected by the action are X_4 and X_5, that is, the descendants of the manipulated variable X_3.

Note the difference between the action $do(X_3 = \mathrm{ON})$ and the observation $X_3 = \mathrm{ON}$. The latter is encoded by ordinary Bayesian conditioning on the event '$X_3 = \mathrm{ON}$,' while the former by conditioning a mutilated graph, with the link $X_1 \rightarrow X_3$ 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 the deliberate action "turning the sprinkler ON." The amputation of $X_3 = f_3(X_1, U_3)$ from (13.16) ensures the suppression of all backward inferences from X_3, the recipient of the action.

In general, if the distribution induced by the original model M is given by the product in Eq. (13.16), or

$$P_M(x_1, \ldots, x_n) = \prod_i P(x_i|\mathbf{pa}_i) \tag{13.19}$$

where $\mathbf{pa}_i$ are (values of) the parents of X_i in model M, then the submodel $M_{x'_j}$, representing the action $do(X_j = x'_j)$ induces a product-like distribution, from which the factor $P(x_j|\mathbf{pa}_j)$ is absent:

$$P_{M_{x'_j}}(x_1, \ldots, x_n) \quad = \begin{cases} \prod_{i \neq j} P(x_i|\mathbf{pa}_i) = \frac{P(x_1,\ldots,x_n)}{P(x_j|\mathbf{pa}_j)} & \text{if} \quad x_j = x'_j \\ 0 & \text{if} \quad x_j \neq x'_j \end{cases} \tag{13.20}$$

This *truncated product* reflects the surgical removal of the equation $X_j = f_j(\mathbf{pa}_j, U_j)$ from M, and tying X_j to the value *ON*. If we further interpret the division in Eq. (13.20) as conditioning on $\{x_j, \mathbf{pa}_j\}$, we get a formula for the effect of any variable X_j on another X_k:

$$\begin{aligned} P(X_x = x_k|do(X_j = x_j) \quad &= \quad P_{M_{x_j}}(X_k = x_k) \\ &= \quad \sum_{\mathbf{pa}_j} P(x_k|x_j, \mathbf{pa}_i)P(\mathbf{pa}_j) \end{aligned} \tag{13.21}$$

where $P(x_k|x_j, \mathbf{pa}_i)$ and $P(\mathbf{pa}_j)$ stand for pre-intervention probabilities. Eq. (13.21) is known as an *adjustment formula*. It calls for conditioning $P(x_k|x_j)$ on the parents of X_j and then averaging the result, weighted by the prior probability $P(\mathbf{pa}_j)$. The overall operation of conditioning and summing is called "adjusting" for $\mathbf{pa}_j$.

Note that Eqs. (13.18) through (13.20) are independent of details of the model M. The post-action distribution depends only on observed conditional probabilities $P(x_i|\mathbf{pa}_i)$, not on the particular functional form of $\{f_i\}$ or the distribution $P(\mathbf{u})$ which generate those probabilities. This important feature allows us to estimate the effect of interventions from the observed data and the structure of the graph alone. Moreover, we do not need the structure of the entire graph, the identity of the parents of the manipulated variable suffices. This is shown explicitly in the denominator of Eq. (13.20) which can be estimated from the pre-intervention distribution once we know who the parents are of X_j.

13.5.2 Causal effects, and the back-door criterion

We have seen in Eq. (13.18) that, once we know the identity of the parents of the manipulated variable X_j, we can estimate the post intervention distribution and, subsequently, the effect of the intervention on any outcome variable, say X_k. For example, if in the model of Figure 13.22(a) we wish to predict the effect of *Sprinkler* (X_3) on *Slippery* (X_5), denoted $P(X_5 = x_5|do(X_3 = x_3))$, we note that X_3 has on one parent, X_1, and Eq. (13.21) gives

$$P(X_5 = x_5|do(X_3 = x_3) = \sum_{x_1} P(x_5|x_3, x_1)P(x_1) \tag{13.22}$$

where $P(x_5|x_3, x_1)$ can be computed by any standard method of computing conditional probabilities in Bayesian networks.

This adjustment however requires that we measure those parent variables, so as to estimate the conditional probability $P(x_j|\mathbf{pa}_j)$. In many problems, however, this information is not available. For example, referring to Figure 13.22, if we do not know what season it is, we will not be able to estimate the conditional probability $P(x_3|x_1)$ required for computing $P(x_5|do(x_3))$ in accordance with Eq. (13.22). In real-life problems, many parent variables will remain unobserved, or hidden, some even nameless (e.g., "genetic factors" or "socio-economic status"), and the question arises whether causal effects can be computed by adjusting for a substitute set S of measured variables. The answer if affirmative, and the criterion for selecting an appropriate set S is known as the *back-door criterion* (Pearl, 1993a). The idea is to select variables that intercept, or "block," every path between the manipulated variables and the outcome variable that contains an arrow into the former. Such paths transmit spurious, or non-causal information and, hence, need be disabled if we are to elicit purely causal dependence from the data.

In our *Sprinkler* model, for example, there is one back-door path between *Sprinkler* and *Slippery*: $X_3 \leftarrow X_1 \rightarrow X_2 \rightarrow X_4 \rightarrow X_5$. This path transmits probabilistic information between X_3 and X_5 and it needs to be blocked if we wish to account only for the causal dependence of X_5 on X_3, through the path $X_3 \rightarrow X_4 \rightarrow X_5$. If we condition on X_2 (*Rain*), the back-door path will be blocked, hence disabled. Therefore, if we do not know the *Season* we have the license to adjust for *Rain* instead, and obtain the effect of *Sprinkler* on *Slippery* via:

$$P(X_5 = x_5 | do(X_3 = x_3) = \sum_{x_2} P(x_5|x_3, x_2)P(x_2) \tag{13.23}$$

with X_2 replacing X_1 in Eq. (13.22).

To formally define of the "blocking" of a path, we now leverage the notion of d-separation, mentioned in footnote 2 (p. 447), which simply tells us when two sets of variables are independent given a third.

Definition 1 (d-separation) *Let a* path *in a DAG be a sequence of consecutive edges, of any directionality. A path p is said to be d-separated (or blocked) by a set of nodes Z iff:*

(i) p contains a chain $i \longrightarrow j \longrightarrow k$ *or a* $fork$ $i \longleftarrow j \longrightarrow k$ *such that the middle node* j *is in* Z*, or,*

(ii) p contains an inverted fork $i \longrightarrow j \longleftarrow k$ *such that neither the middle node* j *nor any of its descendants (in G) are in* Z*.*

If $X, Y,$ *and* Z *are three disjoint subsets of nodes in a DAG G, then Z is said to d-separate X from Y, denoted* $(X \perp\!\!\!\perp Y|Z)_G$*, iff Z d-separates every path from a node in X to a node in Y.*

In Figure 13.22(a), for example, $X = \{X_2\}$ and $Y = \{X_3\}$ are d-separated by $Z = \{X_1\}$; the path $X_2 \leftarrow X_1 \rightarrow X_3$ is blocked by $X_1 \in Z$, while the path $X_2 \rightarrow X_4 \leftarrow X_3$ is blocked because X_4 and all its descendants are outside Z. Thus, $(X_2 \perp\!\!\!\perp X_3|X_1)_G$ holds in Figure 13.22(a). However, X and Y are not d-separated by $Z' = \{X_1, X_5\}$, because the path $X_2 \to X_4 \leftarrow X_3$ is rendered active by virtue of X_5, a descendant of X_4, being in Z'. Consequently, $(X_2 \perp\!\!\!\perp X_3 | \{X_1, X_5\})_G$ does not hold; in words, learning the value of the consequence X_5 renders its causes X_2 and X_3 dependent, as if a pathway were opened along the arrows converging at X_4.

The importance of d-separation lies in its implications for the probability distribution that governs the data. It states that, regardless of the functions $\{f_i\}$, and for every three subsets (X, Y, Z) of variables, if Z d-separates X and Y in G, then X is independent of Y conditioned on Z (Pearl, 1988). Thus, this criterion can be viewed as the bridge that connects the causal assumptions embodied in M with the probabilisitic properties of our data.

The d-separation criterion has been shown to be complete, in the sense of revealing ALL independencies implied by the topology of a Bayesian network. Furthermore, the criterion can be tested in time linear in the number of edges in the graph (Geiger et al., 1990). Fast algorithms have also been developed for finding a minimal set of variables that d-separates one given set from another (Tian et al., 1998), Thus, a researcher who wishes to test the compatibility of a given model with the data, can obtain such tests by listing the set of d-separation conditions embodied in the graph. This list can easily be identified by inspection: Every pair of non-adjacent variables must be independent given the set of variables that d-separate the two. In the model of Figure 13.22(a) for example we find $X_1 \perp\!\!\!\perp X_4 | X_2, X_3, X_2 \perp\!\!\!\perp X_3 | X_1, X_3 \perp\!\!\!\perp X_5 | X_4$, and more.

Additional applications of the d-separation criterion in causal inference tasks are described in Pearl (2000), Pearl et al. (2016), Pearl and Mackenzie (2018), and Spirtes et al. (2000). This criterion and its back-door corollary, are the basic building block for an elaborate theory of causal reasoning that has emerged in the past two decades and impacted almost all data-intensive disciplines. The main inference tools developed include:

1. Methods of deciding, prior to taking any data, what measurements ought to be taken, whether one set of measurements is as good as to another, and which measurements tend to bias our estimates of the target quantities (Pearl, 1993b,

1995; Tian and Pearl, 2002; Pearl and Paz, 2010).

2. Methods for devising critical statistical tests by which two competing theories can be distinguished (Pearl and Paz, 2010).

3. Methods of deciding mathematically if the causal relationships of interest are estimable from non-experimental or quasi-experimental data and, if not, what additional assumptions, measurements or experiments would render them estimable (Pearl, 1993b, 1995; Kyono, 2010; Tian and Pearl, 2002; Shpitser and Pearl, 2006).

4. Methods of recognizing instrumental variables and auxiliary instrumental variables in structural models (Pearl, 2009, p. 257–8; Brito and Pearl, 2002).

5. Methods of computing probabilities on counterfactual statements, including "cause of effects." For example, how likely it is that Joe's swimming exercise was a necessary (or sufficient) cause of Joe's death (Halpern and Pearl, 2005a,b; Pearl, 2015; Balke and Pearl, 1994).

6. A solution to the so called "Mediation Problem" (Baron and Kenny, 1986), taking the form of estimable formulas for direct and indirect effects that are applicable to both continuous and categorical variables, linear as well as nonlinear interactions (Pearl, 2001, 2010a,b).

7. A formal solution to the "external validity" problem (Campbell and Stanley, 1963), which uses knowledge about differences and commonalities between populations to decide whether and how causal effects in one population can be inferred from experimental studies on another (Pearl and Bareinboim, 2010; Bareinboim and Pearl, 2016).

8. Tools for recovering from missing data, based on causal models of the missingness process (Mohan and Pearl, 2019).

These tools and results are making their way towards enriching machine learning with transparency, explainability, transferability, and robustness (Pearl, 2019).

Bibliography

BALKE, A. and PEARL, J. (1994). Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, vol. I. MIT Press, Menlo Park, CA, 230–237.

BAREINBOIM, E. and PEARL, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences* **113** 7345–7352.

BARON, R. and KENNY, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* **51** 1173–1182.

BRITO, C. and PEARL, J. (2002). Generalized instrumental variables. In *Uncertainty in Artificial Intelligence, Proceedings of the Eighteenth Conference* (A. Darwiche and N. Friedman, eds.). Morgan Kaufmann, San Francisco, 85–93.

CAMPBELL, D. and STANLEY, J. (1963). *Experimental and Quasi-Experimental Designs for Research*. Wadsworth Publishing, Chicago.

GEIGER, D., PAZ, A. and PEARL, J. (1990). Learning causal trees from dependence information. In *Proceedings, AAAI-90*. Boston, MA.

HALPERN, J. and PEARL, J. (2005a). Causes and explanations: A structural-model approach—Part I: Causes. *British Journal of Philosophy of Science* **56** 843–887.

HALPERN, J. and PEARL, J. (2005b). Causes and explanations: A structural-model approach—Part II: Explanations. *British Journal of Philosophy of Science* **56** 843–887.

Kyono, T. (2010). Commentator: A front-end user-interface module for graphical and structural equation modeling. Tech. Rep. R-364, <http://ftp.cs.ucla.edu/pub/stat_ser/r364.pdf>, Master Thesis, Department of Computer Science, University of California, Los Angeles, CA.

Mohan, K. and Pearl, J. (2019). Graphical models for processing missing data. Tech. Rep. R-473-L, <http://ftp.cs.ucla.edu/pub/stat_ser/r473-L.pdf>, Department of Computer Science, University of California, Los Angeles, CA. Forthcoming, Journal of American Statistical Association (JASA).

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

Pearl, J. (1993a). Comment: Graphical models, causality, and intervention. *Statistical Science* **8** 266–269.

Pearl, J. (1993b). From Bayesian networks to causal networks. In *Proceedings of the Adaptive Computing and Information Processing Seminar*. Brunel Conference Centre, London. See also *Statistical Science*, 8(3):266–269, 1993.

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

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.

Pearl, J. (2001). Direct and indirect effects. In *Uncertainty in Artificial Intelligence, Proceedings of the Seventeenth Conference*. Morgan Kaufmann, San Francisco, CA, 411–420.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.

Pearl, J. (2010a). The foundations of causal inference. *Sociological Methodology* Early view, Oct 14 DOI: 10.1111/j.1467–9531.2010.01228.x.

Pearl, J. (2010b). The mediation formula: A guide to the assessment of causal pathways in non-linear models. Tech. Rep. R-363, <http://ftp.cs.ucla.edu/pub/stat_ser/r363.pdf>, Department of Computer Science, University of California, Los Angeles, CA.

Pearl, J. (2015). Causes of effects and effects of causes. *Journal of Sociological Methods and Research* **44** 149–164.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Commun. ACM* **62** 54–60.
URL `http://doi.acm.org/10.1145/3241036`

Pearl, J. and Bareinboim, E. (2010). Transportability across studies: A formal approach. Tech. Rep. R-372, <http://ftp.cs.ucla.edu/pub/stat_ser/r372.pdf>, Department of Computer Science, University of California, Los Angeles, CA.

Pearl, J., Glymour, M. and Jewell, N. (2016). *Causal Inference in Statistics: A Primer*. Wiley, New York.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York.

Pearl, J. and Paz, A. (2010). Confounding equivalence in causal equivalence. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. AUAI, Corvallis, OR, 433–441.

Shachter, R. D. (1998). Bayes-ball: The rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In *Uncertainty in Artificial Intelligence: Proc. Fourteenth Conference*. Morgan Kaufmann, Madison, Wisconsin.

Shpitser, I. and Pearl, J. (2006). Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (R. Dechter and T. Richardson, eds.). AUAI Press, Corvallis, OR, 437–444.

Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, Prediction, and Search*. 2nd ed. MIT Press, Cambridge, MA.

Tian, J., Paz, A. and Pearl, J. (1998). Finding a minimal *d*-separator. Tech. Rep. R-254, <http://ftp.cs.ucla.edu/pub/stat_ser/r254.pdf>, Department of Computer Science, University of California, Los Angeles, CA.

Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*. AAAI Press/The MIT Press, Menlo Park, CA, 567–573.