GRAPHICAL MODELS, CAUSALITY, AND INTERVENTION

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
judea@cs.ucla.edu

I am grateful for the opportunity to respond to these two excellent papers. Although graphical models are intuitively compelling for conceptualizing statistical associations, the scientific community generally views such models with hesitancy and suspicion. The two papers before us demonstrate the use of graphs – specifically, directed acyclic graphs (DAGs) – as a mathematical tool of great versatility and thus promise to make graphical languages more common in statistical analysis. In fact, I find my own views in such close agreement with those of the authors that any attempt on my part to comment directly on their work would amount to sheer repetition. Instead, as the editor suggested, I would like to provide a personal perspective on current and future developments in the areas of graphical and causal modeling.[1]

I will focus on the connection between graphical models and the notion of causality in statistical analysis. This connection has been treated very cautiously in the papers before us[2] and I would like to supplement the discussion with an account of how causal models and graphical models are related.

It is generally accepted that, because they provide information about the dynamics of the system under study, causal models, regardless of how they are discovered or tested, are more useful than associational models. In other words, whereas the joint distribution tells us how probable events are and how probabilities would change with subsequent observations,

[1]A complementary account of the evolution of belief networks is given in [4].

[2]In [3], the graphs were called "causal networks," for which the authors were criticised; they have agreed to refrain from using the word "causal." In the current paper, Spiegelhalter etal. deemphasize the causal interpretation of the arcs in favor of the "irrelevance" interpretation (page 4). I think this retreat is regrettable for two reasons: first, causal associations are the primary source of judgments about irrelevance and, second, rejecting the causal interpretation of arcs prevents us from using graphical models for making legitimate predictions about the effect of actions. Such predictions are indispensable in applications such as treatment management and patient monitoring.

the causal model also tells us how these probabilities would change as a result of external interventions in the system. For this reason, causal models (or "structural models" as they are often called) have been the target of relentless scientific pursuit and, at the same time, the center of much controversy and speculation. What I would like to discuss in this commentary is how complex information about external interventions can be organized and represented graphically and, conversely, how the graphical representation can be used to facilitate quantitative predictions of the effects of interventions.

The basic idea goes back to Simon [9] and is stated succinctly in his forward to [1]: "The advantage of representing the system by structural equations that describe the direct causal mechanisms is that if we obtain some knowledge that one or more of these mechanisms has been altered, we can use the remaining equations to predict the consequences – the new equilibrium." Here, by "mechanism" Simon means any stable relationship between two or more variables that remains invariant to external influences until it falls directly under such influences.

This mechanism-based model was adapted in [7] for defining probabilistic causal theories; each child-parent family in a DAG Γ represents a deterministic function $X_i = f_i(\mathbf{pa}_i, \epsilon_i)$, where $\mathbf{pa}_i$ are the parents of variable X_i in Γ, and ϵ_i, $0 < i < n$, are mutually independent, arbitrarily distributed random disturbances. Characterizing each child-parent relationship as a deterministic function, instead of the usual conditional probability $P(x_i \mid \mathbf{pa}_i)$, imposes equivalent independence constraints on the resulting distributions and leads to the same recursive decomposition

$$P(x_1, ..., x_n) = \prod_i P(x_i \mid \mathbf{pa}_i) \tag{1}$$

that appears in Eq. (1) of Spiegelhalter etal.'s article. However, the functional characterization also specifies how the resulting distribution would change in response to external interventions, since, by convention, each function is presumed to remain constant unless specifically altered.[3] Moreover, the non-linear character of f_i permits us to treat changes in the function f_i itself as a variable, F_i, by writing

$$X_i = f'_i(\mathbf{pa}_i, F_i, \epsilon_i) \tag{2}$$

where

$$f'_i(a, b, c) = f_i(a, c) \text{ whenever } b = f_i.$$

Thus, any external intervention F_i that alters f_i can be represented graphically as an added parent node of X_i, and the effect of such an intervention can be analyzed by Bayesian conditionalization, that is, by simply setting this added parent variable to the appropriate value f_i.

The simplest type of external intervention is one in which a single variable, say X_i, is forced to take on some fixed value x'_i. Such intervention, which we call *atomic*, amounts to replacing the old functional mechanism $X_i = f_i(\mathbf{pa}_i, \epsilon_i)$ with a new mechanism $X_i = x'_i$ governed by some external force F_i that sets the value x'_i. If we imagine that each variable X_i potentially could be subject to the influence of such an external force F_i, then we can view the causal network Γ as an efficient code for predicting the effects of atomic interventions and of various combinations of such interventions.

[3]This formulation is merely a non-linear generalization of the usual structural equation models, where function constancy (or stability) is implicitly assumed.

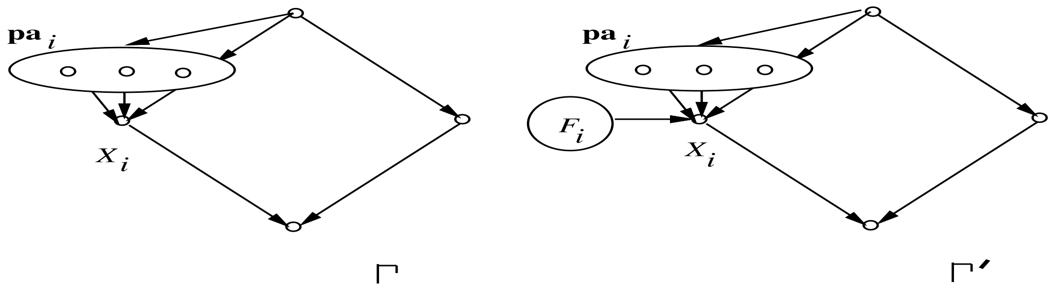

Figure 1: Representing external intervention, F_i, by an augmented network $\Gamma' = \Gamma \cup \{F_i \rightarrow X_i\}$.

The effect of an atomic intervention $set(X_i = x_i')$ is encoded by adding to Γ a link $F_i \longrightarrow X_i$ (see Figure 1), where F_i is a new variable taking values in $\{set(x_i'),\ idle\}$, x_i' ranges over the domain of X_i, and $idle$ represents no intervention. Thus, the new parent set of X_i in the augmented network is $\mathbf{pa}_i' = \mathbf{pa}_i \cup \{F_i\}$, and it is related to X_i by the conditional probability

$$P(x_i \mid \mathbf{pa}_i') = \begin{cases} P(x_i \mid \mathbf{pa}_i) & \text{if } \ F_i = idle \\ 0 & \text{if } \ F_i = set(x_i') \ \text{ and } \ x_i \neq x_i' \\ 1 & \text{if } \ F_i = set(x_i') \ \text{ and } \ x_i = x_i' \end{cases} \tag{3}$$

The effect of the intervention $set(x_i')$ is to transform the original probability function $P(x_1, ..., x_n)$ into a new function $P_{x_i'}(x_1, ..., x_n)$, given by

$$P_{x_i'}(x_1, ..., x_n) = P'(x_1, ..., x_n \mid F_i = set(x_i')), \tag{4}$$

where P' is the directed Markov field dictated by the augmented network $\Gamma' = \Gamma \cup \{F_i \rightarrow X_i\}$ and Eq. (3), with an arbitrary prior distribution on F_i. In general, by adding a hypothetical intervention link $F_i \rightarrow X_i$ to each node in Γ, we can construct an augmented probability function $P'(x_1, ..., x_n; F_1, ..., F_n)$ that contains information about richer types of interventions. Multiple interventions would be represented by conditioning P' on a subset of the F_i's (taking values in their respective $set(x_i')$), while the pre-intervention probability function P would be viewed as the posterior distribution induced by conditioning each F_i in P' on the value $idle$.

This representation yields a simple and direct transformation between the pre-intervention and the post-intervention distributions[4]:

$$P_{x_i'}(x_1, ..., x_n) = \begin{cases} \frac{P(x_1, ..., x_n)}{P(x_i \mid \mathbf{pa}_i)} & \text{if } \ x_i = x_i' \\ 0 & \text{if } \ x_i \neq x_i' \end{cases} \tag{5}$$

The transformation exhibits the following properties:

1. An intervention $set(x_i')$ can affect only the descendants of X_i in Γ.

[4]This transformation reflects the removal of the term $P(x_i \mid \mathbf{pa}_i)$ from the product decomposition of Eq. (1), since $\mathbf{pa}_i$ no longer influence X_i. Transformations involving conjunctive and disjunctive actions can be obtained by straightforward applications of Eq. (4) [10, 2, 5].

2. For any set $\mathbf{S}$ of variables, we have

$$P_{x_i'}(\mathbf{S} \mid \mathbf{pa}_i) = P(\mathbf{S} \mid x_i', \mathbf{pa}_i). \tag{6}$$

In other words, given $X_i = x_i'$ and $\mathbf{pa}_i$, it is superfluous to find out whether $X_i = x_i'$ was established by external intervention or not. This can be seen directly from the augmented network Γ' (see Figure 1), since $\{X_i\} \cup \mathbf{pa}_i$ d-separates F_i from the rest of the network, thus legitimizing the conditional independence $\mathbf{S} \perp\!\!\!\perp F_i \mid (X_i, \mathbf{pa}_i)$.

3. A necessary and sufficient condition for an external intervention $set(X_i = x_i')$ to have the same effect on X_j as the passive observation $X_i = x_i'$ is that X_i d-separates $\mathbf{pa}_i$ from X_j, that is,

$$P_{x_i'}(x_j) = P(x_j \mid x_i') \;\text{ iff }\; X_j \perp\!\!\!\perp \mathbf{pa}_i \mid X_i. \tag{7}$$

Eq. (4) explains why randomized experiments are sufficient for estimating the effect of interventions even when the causal network is not given: because the intervening variable F_i enters the networks as a root node (i.e., independent of all other ancestors of X_i) it is equivalent to a treatment-selection policy governed by a random device.

The immediate implication of Eq. (5) is that, given the structure of the causal network Γ, one can infer post-intervention distributions from pre-intervention distributions; hence, we can reliably estimate the effects of interventions from passive (i.e., non-experimental) observations. Of course, Eq. (5) does not imply that we can always substitute observational studies for experimental studies, as this would require an estimation of $P(x_i \mid \mathbf{pa}_i)$. The mere identification of $\mathbf{pa}_i$ (i.e., the direct causal factors of X_i) requires substantive causal knowledge of the domain which is often unavailable. Moreover, even when we have sufficient substantive knowledge to structure Γ, some members of $\mathbf{pa}_i$ may be unobservable, or *latent*. Fortunately, there are conditions for which an unbiased estimate of $P_{x_i'}(x_j)$ can be obtained even when the $\mathbf{pa}_i$ variables are latent and, moreover, a simple graphical criterion can tell us when these conditions are satisfied.

Assume we are given a causal network Γ together with non-experimental data on a subset $\mathbf{X}_o$ of observed variables in Γ and we wish to estimate what effect the intervention $set(X_i = x_i')$ would have on some response variable X_j. In other words, we seek to estimate $P_{x_i'}(x_j)$ from a sample estimate of $P(\mathbf{X}_o)$. Applying Eq. (4), we can write

$$\begin{aligned} P_{x_i'}(x_j) &= P'(x_j \mid F_i = set(x_i')) \\ &= \sum_{\mathbf{S}} P'(x_j \mid \mathbf{S}, X_i = x_i', F_i = set(x_i'))P'(\mathbf{S} \mid F_i = set(x_i')), \end{aligned} \tag{8}$$

where $\mathbf{S}$ is any set of variables. Clearly, if $\mathbf{S}$ satisfies

$$\mathbf{S} \perp\!\!\!\perp F_i \;\text{ and }\; X_j \perp\!\!\!\perp F_i \mid (X_i, \mathbf{S}), \tag{9}$$

then Eq. (8) can be reduced to

$$\begin{aligned} P_{x_i'}(x_j) &= \sum_{\mathbf{S}} P(x_j \mid \mathbf{S}, x_i')P(\mathbf{S}) \\ &= E_{\mathbf{S}}[P(x_j \mid \mathbf{S}, x_i')]. \end{aligned} \tag{10}$$

Thus, if we find a set $\mathbf{S} \subseteq \mathbf{X}_o$ of observables satisfying Eq. (9), we can estimate $P_{x_i'}(x_j)$ by taking the conditional expectation (over $\mathbf{S}$) of $P(x_j \mid x_i')$, and the latter can easily be estimated from non-experimental data. It is also easy to verify that Eq. (9) is satisfied by any set $\mathbf{S}$ that meets the following *back-door criterion*:

1. No node in $\mathbf{S}$ is a descendant of X_i, and
2. $\mathbf{S}_i$ d-separates X_i from X_j along every path containing an arrow toward X_i.

In Figure 2, for example, the sets $\mathbf{S}_1 = \{X_3, X_4\}$ and $\mathbf{S}_2 = \{X_4, X_5\}$ would qualify under the back-door criterion, but $\mathbf{S}_3 = \{X_4\}$ would not because X_4 does not d-separate X_i from X_j along the path $(X_i, X_3, X_1, X_4, X_2, X_5, X_j)$. Thus, we have obtained a simple graphical criterion for finding a set of observables for estimating (by conditioning) the effect of interventions from purely non-experimental data.

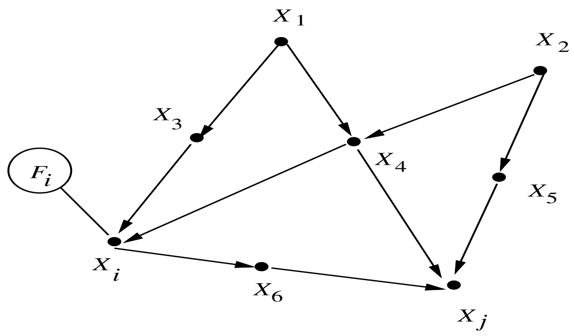

Figure 2

It is interesting that the conditions formulated in Eq. (9) are equivalent to those known as *strongly ignorable treatment assignment* (SITA) conditions in Rubin's model[5] for causal effect [8, 6]. Reducing the SITA conditions to the graphical back-door criterion facilitates the search for an optimal conditioning set $\mathbf{S}$ and significantly simplifies the judgments required for ratifying the validity of such conditions in practical situations.

Eq. (4) was derived under the assumption that the pre-intervention probability P is given by the product of Eq. (1), which represents a joint distribution prior to making any observations. To predict the effect of action F_i *after* observing C, we must also invoke assumptions about persistence, so as to distinguish properties that will terminate as a result of F_i from those that will persist despite of acting F_i. Such a model of persistence was invoked in [5]; there, it was assumed that only those properties should persist that are not under any causal influence to terminate. This assumption yields formulas for the effect of *conditional interventions* (conditioned on the observation C) which, again, given Γ, can be estimated from non-experimental data.

A more ambitious task has been explored by Spirtes, Glymour, and Scheines [10] – estimation of the effect of intervention when the structure of Γ is not available and must also be inferred from the data. Recent developments in graphical models [7, 10] have produced methods that, under certain conditions, permit us to infer plausible causal structures from

[5] The graphical translation of Rubin's model invokes the mechanism $X_i \rightarrow X_j \leftarrow \mathbf{r}$, where X_i represents the treatment-assignment, X_j the observed response, and $\mathbf{r}$ represents the causal-effect variable. Indeed, following the counterfactual interpretation of $\mathbf{r}$, X_j is a deterministic function of X_i and $\mathbf{r}$, and $\mathbf{r}$ plays the role of f_i in Eq. (2).

non-experimental data, albeit with a weaker set of guarantees than those obtained through controlled randomized experiments. These guarantees fall into two categories: minimality and stability [7]. Minimality guarantees that any other structure compatible with the data is necessarily more redundant, and hence less trustworthy, than the one(s) inferred. Stability ensures that any alternative structure compatible with the data must be less stable than the one(s) inferred; namely, slight fluctuations in the distributions of the disturbances ϵ_i (Eq. (2)) will render that structure no longer compatible with the data.

When the structure of Γ is to be inferred under these guarantees, the formulas governing the effects of interventions and the conditions required for estimating these effects become rather complex [10]. Alternatively, one can produce bounds on the effect of interventions by taking representative samples of inferred structures and estimating $P_{x_i'}(x_j)$ according to Eq. (10) for each such sample.

In summary, I hope my comments convince the reader that DAGs can be used not only for specifying assumptions of conditional independence but also as a formal language for organizing claims about external interventions and their interactions. I hope to have demonstrated as well that DAGs can serve as an analytical tool for predicting, from non-experimental data, the effect of actions (given substantive causal knowledge), for specifying and testing conditions under which randomized experiments are not necessary and for aiding experimental design and model selection.

BIBLIOGRAPHY

[1] Glymour, C., Scheines, R., Spirtes, P., and Kelly, K., *Discovering Causal Structure*, Academic Press, Orlando, FL, 1987.

[2] Goldszmidt, M., and Pearl, J., "Default ranking: A practical framework for evidential reasoning, belief revision and update," in *Proceedings of the Third International Conference on Knowledge Representation and Reasoning*, Cambridge, MA, 661-672, April 1992.

[3] Lauritzen, S.L., and Spiegelhalter, D.J., "Local computations with probabilities on graphical structures and their applications to expert systems," *Proceedings of the Royal Statistical Society, Series B.*, 50, 154-227, 1988.

[4] Pearl, J., "Belief networks revisited," *Artificial Intelligence*, 59, 49-56, 1993.

[5] Pearl, J., "A calculus of pragmatic obligation," *Proceedings of the AAAI Spring Symposium on Reasoning about Mental States*, Stanford, CA, March 1993.

[6] Pearl, J., "Aspects of graphical models connected with causality," *Proceedings of 49th Session, International Statistical Institute: Invited papers*, Florence, Italy, August 1993.

[7] Pearl, J., and Verma, T., "A theory of inferred causation," in Allen, J.A., Fikes, R., and Sandewall, E. (Eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, Morgan Kaufmann, San Mateo, CA, 441-452, April 1991.

[8] Rosenbaum, P., and Rubin, D., "The central role of propensity score in observational studies for causal effects," *Biometrica*, 70, 41-55, 1983.

[9] Simon, H.A., *Models of Discovery: and Other Topics in the Methods of Science*, D. Reidel, Dordrecht, Holland, 1977.

[10] Spirtes, P., Glymour, C., and Schienes, R., *Causation, Prediction, and Search*, Springer-Verlag, New York, 1993.