To appear in the *Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA, July 31 – August 4, 1994.

Technical Report
R-213-A
April, 1994

Probabilistic evaluation of counterfactual queries

Alexander Balke and **Judea Pearl**
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
<*balke@cs.ucla.edu*> and <*judea@cs.ucla.edu*>

Abstract

Evaluation of counterfactual queries (e.g., "If A were true, would C have been true?") is important to fault diagnosis, planning, and determination of liability. We present a formalism that uses probabilistic causal networks to evaluate one's belief that the counterfactual consequent, C, would have been true if the antecedent, A, were true. The antecedent of the query is interpreted as an external action that forces the proposition A to be true, which is consistent with Lewis' *Miraculous Analysis*. This formalism offers a concrete embodiment of the "closest world" approach which (1) properly reflects common understanding of causal influences, (2) deals with the uncertainties inherent in the world, and (3) is amenable to machine representation.

Introduction

A counterfactual sentence has the form

If A were true, then C would have been true

where A, the counterfactual antecedent, specifies an event that is contrary to one's real-world observations, and C, the counterfactual consequent, specifies a result that is expected to hold in the alternative world where the antecedent is true. A typical instance is "If Oswald were not to have shot Kennedy, then Kennedy would still be alive" which presumes the factual knowledge of Oswald's shooting Kennedy, contrary to the antecedent of the sentence.

The majority of the philosophers who have examined the semantics of counterfactual sentences (Goodman 1983; Harper, Stalnaker, & Pearce 1981; Nute 1980; Meyer & van der Hoek 1993) have resorted to some form of logic based on worlds that are "closest" to the real world yet consistent with the counterfactual's antecedent. Ginsberg (Ginsberg 1986), following a similar strategy, suggested that the logic of counterfactuals could be applied to problems in planning and diagnosis in Artificial Intelligence. The few other papers in AI that have focussed on counterfactual sentences (e.g., (Jackson 1989; Pereira, Aparicio, & Alferes 1991; Boutilier 1992)) have mostly adhered to logics based on the "closest world" approach.

In the real world, we seldom have adequate information for verifying the truth of an indicative sentence, much less the truth of a counterfactual sentence. Except for the small set of relationships between variables which can be modeled by physical laws, most of the relationships in one's knowledge base are nondeterministic. Therefore, it is more practical to ask not for the truth or falsity of a counterfactual, but for one's degree of belief in the counterfactual consequent given the antecedent. To account for such uncertainties, (Lewis 1976) has generalized the notion of "closest world" using the device of "imaging"; namely, the closest worlds are assigned probability scores, and these scores are combined to compute the probability of the consequent.

The drawback of the "closest world" approach is that it leaves the precise specification of the closeness measure almost unconstrained. More specifically, it does not tell us how to encode distances in a way that would (1) conform to our perception of causal influences and (2) lend itself to economical machine representation. This paper can be viewed as a concrete explication of the closest world approach, one that satisfies the two requirements above.

The target of our investigation are counterfactual queries of the form:

If A were true, then what is the probability that C would have been true, given that we know B?

The proposition B stands for the actual observations made in the real world (e.g., that Oswald did shoot Kennedy and that Kennedy is dead) which we make explicit to facilitate the analysis.

Counterfactuals are intertwined with notions of causality: We do not typically express counterfactual sentences without assuming a causal relationship between the counterfactual antecedent and the counterfactual consequent. For example, we can safely state "If the sprinkler were on, the grass would be wet", but the contrapositive form of the same sentence in counterfactual form, "If the grass were dry, then the sprinkler would not be on", strikes us as strange, because we do not think the state of the grass has causal influence on the state of the sprinkler. Likewise, we do not state "All blocks on this table are green, hence, had this white block been on the table, it would have been green". In fact, we could say that people's use of counterfactual statements is aimed precisely at conveying generic causal information, uncontaminated by specific, transitory observations, about the real world. Observed facts often do reflect strange combinations of rare eventualities (e.g., all blocks being green) that have nothing to do with general traits of influence and behavior. The counterfactual sentence, however, emphasizes the law-like, necessary component of the relation considered. It is for this reason, we speculate, that we find such frequent use of counterfactuals in ordinary discourse.

The importance of equipping machines with the capability to answer counterfactual queries lies precisely in this causal reading. By making a counterfactual query, the user intends to extract the generic, necessary connection between the antecedent and consequent, regardless of the contingent factual information available at that moment.

Because of the tight connection between counterfactuals and causal influences, any algorithm for computing counterfactual queries must rely heavily on causal knowledge of the domain. This leads naturally to the use of probabilistic causal networks, since these networks combine causal and probabilistic knowledge and permit reasoning from causes to effects as well as, conversely, from effects to causes.

To emphasize the causal character of counterfactuals, we will adopt the interpretation in (Pearl 1993b), according to which a counterfactual sentence "If it were A, then B would have been" states that B would prevail if A were forced to be true by some unspecified action that is exogenous to the other relationships considered in the analysis. This action-based interpretation does not permit inferences from the counterfactual antecedent towards events that lie in its past. For example, the action-based interpretation would ratify the counterfactual

> If Kennedy were alive today, then the country would have been in a better shape

but not the counterfactual

> If Kennedy were alive today, then Oswald would have been alive as well.

The former is admitted because the causal influence of Kennedy on the country is presumed to remain valid even if Kennedy became alive by an act of God. The second sentence is disallowed because Kennedy being alive is not perceived as having causal influence on Oswald being alive. The information intended in the second sentence is better expressed in an indicative mood:

> If Kennedy was alive today then he could not have been killed in Dallas, hence, Jack Ruby would not have had a reason to kill Oswald and Oswald would have been alive today.

Our interpretation of counterfactual antecedents, which is similar to Lewis' (Lewis 1979) *Miraculous Analysis*, contrasts with interpretations that require that the counterfactual antecedent be consistent with the world in which the analysis occurs. The set of closest worlds delineated by the action-based interpretation contains all those which coincide with the factual world except on possible consequences of the action taken. The probabilities assigned to these worlds will be determined by the relative likelihood of those consequences as encoded by the causal network.

We will show that causal theories specified in functional form (as in (Pearl & Verma 1991; Druzdzel & Simon 1993; Poole 1993)) are sufficient for evaluating counterfactual queries, whereas the causal information embedded in Bayesian networks is not sufficient for the task. Every Bayes network can be represented by several functional specifications, each yielding different evaluations of a counterfactual. The problem is that, deciding what factual information deserves undoing (by the antecedent of the query) requires a model of temporal persistence, and, as noted in (Pearl 1993c), such a model is not part of static Bayesian networks. Functional specification, however, implicitly contains the temporal persistence information needed.

The next section introduces some useful notation for concisely expressing counterfactual sentences/queries. We then present an example demonstrating the plausibility of the external action interpretation adopted in this paper. We then demonstrate that Bayesian networks are insufficient for uniquely evaluating counterfactual queries whereas the functional model is sufficient. A counterfactual query algorithm is then presented, followed by a re-examination of the earlier example with a quantitative analysis using this algorithm. The final section contains concluding remarks.

Notation

Let the set of variables describing the world be designated by $X = \{X_1, X_2, \ldots, X_n\}$. As part of the complete specification of a counterfactual query, there are real-world observations that make up the background context. These observed values will be represented in the standard form $x_1, x_2, \ldots, x_n$. In addition, we must represent the value of the variables in the counterfactual world. To distinguish between x_i and the value of X_i in the counterfactual world, we will denote the latter with an asterisk; thus, the value of X_i in the counterfactual world will be represented by x_i^*. We will also need a notation to distinguish between events that might be true in the counterfactual world and those referenced explicitly in the counterfactual antecedent. The latter are interpreted as being forced to the counterfactual value by an external action, which will be denoted by a hat (e.g., $\hat{x}$).

Thus, a typical counterfactual query will have the form "What is $P(c^*|\hat{a}^*, a, b)$?" to be read as "Given that we have observed $A = a$ and $B = b$ in the real

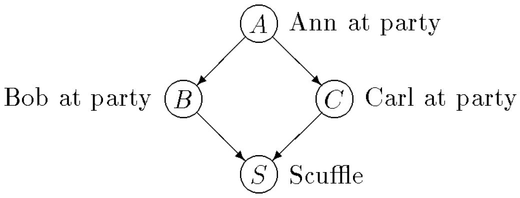

Figure 1: Causal structure reflecting the influence that Ann's attendance has on Bob and Carl's attendance, and the influence that Bob and Carl's attendance has on their scuffling.

world, if A were $\hat{a}^*$, then what is the probability that C would have been c^*?"

Party example

To illustrate the external-force interpretations of counterfactuals, consider the following interpersonal behaviors of Ann, Bob, and Carl:

- Ann sometimes goes to parties.
- Bob likes Ann very much but is not into the party scene. Hence, save for rare circumstances, Bob is at the party if and only if Ann is there.
- Carl tries to avoid contact with Ann since they broke up last month, but he really likes parties. Thus, save for rare occasions, Carl is at the party if and only if Ann is not at the party.
- Bob and Carl truly hate each other and almost always scuffle when they meet.

This situation may be represented by the diamond structure in Figure 1. The four variables A, B, C, and S have the following domains:

$$a \in \left\{ \begin{array}{lcl} a_0 & \equiv & \text{Ann is not at the party.} \\ a_1 & \equiv & \text{Ann is at the party.} \end{array} \right\}$$

$$b \in \left\{ \begin{array}{lcl} b_0 & \equiv & \text{Bob is not at the party.} \\ b_1 & \equiv & \text{Bob is at the party.} \end{array} \right\}$$

$$c \in \left\{ \begin{array}{lcl} c_0 & \equiv & \text{Carl is not at the party.} \\ c_1 & \equiv & \text{Carl is at the party.} \end{array} \right\}$$

$$s \in \left\{ \begin{array}{lcl} s_0 & \equiv & \text{No scuffle between Bob and Carl.} \\ s_1 & \equiv & \text{Scuffle between Bob and Carl.} \end{array} \right\}$$

Now consider the following discussion between two friends (Laura and Scott) who did not go to the party but were called by Bob from his home ($b = b_0$):

Laura: Ann must not be at the party, or Bob would be there instead of at home.

Scott: That must mean that Carl is at the party!

Laura: If Bob were at the party, then Bob and Carl would surely scuffle.

Scott: No. If Bob was there, then Carl would not be there, because Ann would have been at the party.

Laura: True. But if Bob were at the party even though Ann was not, then Bob and Carl would be scuffling.

Scott: I agree. It's good that Ann would not have been there to see it.

In the fourth sentence, Scott tries to explain away Laura's conclusion by claiming that Bob's presence would be evidence that Ann was at the party which would imply that Carl was not at the party. Scott, though, analyzes Laura's counterfactual statement as an indicative sentence by imagining that she had observed Bob's presence at the party; this allows her to use the observation for abductive reasoning. But Laura's subjunctive (counterfactual) statement should be interpreted as leaving everything in the past as it was (including conclusions obtained from abductive reasoning from real observations) while forcing variables to their counterfactual values. This is the gist of her last statement.

This example demonstrates the plausibility of interpreting the counterfactual statement in terms of an external force causing Bob to be at the party, regardless of all other prior circumstances. The only variables that we would expect to be impacted by the counterfactual assumption would be the descendants of the counterfactual variable; in other words, the counterfactual value of Bob's attendance does not change the belief in Ann's attendance from the belief prompted by the real-world observation.

Probabilistic vs. functional specification

In this section we will demonstrate that functionally modeled causal theories (Pearl & Verma 1991) are necessary for uniquely evaluating counterfactual queries, while the conditional probabilities used in the standard specification of Bayesian networks are insufficient for obtaining unique solutions.

Reconsider the party example limited to the two variables A and B, representing Ann and Bob's attendance, respectively. Assume that previous behavior shows $P(b_1|a_1) = 0.9$ and $P(b_0|a_0) = 0.9$. We observe that Bob and Ann are absent from the party and we wonder whether Bob would be there if Ann were there $P(b_1^*|\hat{a}_1^*, a_0, b_0)$. The answer depends on the mechanism that accounts for the 10% exception in Bob's behavior. If the reason Bob occasionally misses parties (when Ann goes) is that he is unable to attend (e.g., being sick or having to finish a paper for AAAI), then the answer to our query would be 90%. However, if the only reason for Bob's occasional absence (when Ann goes) is that he becomes angry with Ann (in which case he does exactly the opposite of what she does), then the answer to our query is 100%, because Ann and Bob's current absence from the party proves that Bob is not angry. Thus, we see that the information contained in the conditional probabilities on the observed variables is insufficient for answering counterfactual queries uniquely; some information about the mechanisms responsible for these probabilities is needed as well.

The functional specification, which provides this information, models the influence of A on B by a deterministic function

$$b = F_b(a, \epsilon_b)$$

where ϵ_b stands for all unknown factors that may influence B and the prior probability distribution $P(\epsilon_b)$ quantifies the likelihood of such factors. For example, whether Bob has been grounded by his parents and whether Bob is angry at Ann could make up two possible components of ϵ_b. Given a specific value for ϵ_b, B becomes a deterministic function of A; hence, each value in ϵ_b's domain specifies a *response function* that maps each value of A to some value in B's domain. In general, the domain for ϵ_b could contain many components, but it can always be replaced by an equivalent variable that is minimal, by partitioning the domain into equivalence regions, each corresponding to a single response function (Pearl 1993a). Formally, these equivalence classes can be characterized as a function $r_b : \mathrm{dom}(\epsilon_b) \rightarrow \mathbf{N}$, as follows:

$$r_b(\epsilon_b) = \begin{cases} 0 & \text{if } F_b(a_0, \epsilon_b) = 0 \;\&\; F_b(a_1, \epsilon_b) = 0 \\ 1 & \text{if } F_b(a_0, \epsilon_b) = 0 \;\&\; F_b(a_1, \epsilon_b) = 1 \\ 2 & \text{if } F_b(a_0, \epsilon_b) = 1 \;\&\; F_b(a_1, \epsilon_b) = 0 \\ 3 & \text{if } F_b(a_0, \epsilon_b) = 1 \;\&\; F_b(a_1, \epsilon_b) = 1 \end{cases}$$

Obviously, r_b can be regarded as a random variable that takes on as many values as there are functions between A and B. We will refer to this domain-minimal variable as a *response-function variable*. r_b is closely related to the *potential response variables* in Rubin's model of counterfactuals (Rubin 1974), which was introduced to facilitate causal inference in statistical analysis (Balke & Pearl 1993).

For this example, the response-function variable for B has a four-valued domain $r_b \in \{0, 1, 2, 3\}$ with the following functional specification:

$$b = f_b(a, r_b) = h_{b,r_b}(a) \tag{1}$$

where

$$h_{b,0}(a) = b_0 \tag{2}$$

$$h_{b,1}(a) = \begin{cases} b_0 & \text{if } a = a_0 \\ b_1 & \text{if } a = a_1 \end{cases} \tag{3}$$

$$h_{b,2}(a) = \begin{cases} b_1 & \text{if } a = a_0 \\ b_0 & \text{if } a = a_1 \end{cases} \tag{4}$$

$$h_{b,3}(a) = b_1 \tag{5}$$

specify the mappings of the individual response functions. The prior probability on these response functions $P(r_b)$ in conjunction with $f_b(a, r_b)$ fully parameterizes the model.

Given $P(r_b)$, we can uniquely evaluate the counterfactual query "What is $P(b_1^* | \hat{a}_1^*, a_0, b_0)$?" (i.e., "Given $A = a_0$ and $B = b_0$, if A were a_1, then what is the probability that B would have been b_1?"). The action-based interpretation of counterfactual antecedents implies that the disturbance ϵ_b, and hence the response-function r_b, is unaffected by the actions that force the counterfactual values[1]; therefore, what we learn about the response-function from the observed evidence is applicable to the evaluation of belief in the counterfactual consequent. If we observe (a_0, b_0), then we are certain that $r_b \in \{0, 1\}$, an event having prior probability $P(r_b = 0) + P(r_b = 1)$. Hence, this evidence leads to an updated posterior probability for r_b (let $\vec{P}(r_b) = \langle P(r_b{=}0), P(r_b{=}1), P(r_b{=}2), P(r_b{=}3) \rangle$)

$$\vec{P}'(r_b) = \vec{P}(r_b | a_0, b_0) = \langle \frac{P(r_b{=}0)}{P(r_b{=}0) + P(r_b{=}1)}, \frac{P(r_b{=}1)}{P(r_b{=}0) + P(r_b{=}1)}, 0, 0 \rangle.$$

According to Eqs. 1-5, if A were forced to a_1, then B would have been b_1 if and only if $r_b \in \{1, 3\}$, which has probability $P'(r_b{=}1) + P'(r_b{=}3) = P'(r_b{=}1)$. This is exactly the solution to the counterfactual query,

$$P(b_1^* | \hat{a}_1^*, a_0, b_0) = P'(r_b{=}1) = \frac{P(r_b{=}1)}{P(r_b{=}0) + P(r_b{=}1)}.$$

This analysis is consistent with the *prior propensity account* of (Skyrms 1980).

What if we are provided only with the conditional probability $(P(b|a))$ instead of a functional model $(f_b(a, r_b)$ and $P(r_b))$? These two specifications are related by:

$$\begin{aligned} P(b_1 | a_0) &= P(r_b{=}2) + P(r_b{=}3) \\ P(b_1 | a_1) &= P(r_b{=}1) + P(r_b{=}3). \end{aligned}$$

which show that $P(r_b)$ is not, in general, uniquely determined by the conditional distribution $P(b|a)$.

Hence, given a counterfactual query, a functional model always leads to a unique solution, while a Bayesian network seldom leads to a unique solution, depending on whether the conditional distributions of the Bayesian network sufficiently constrain the prior distributions of the response-function variables in the corresponding functional model.

In practice, specifying a functional model is not as daunting as one might think from the example above. In fact, it could be argued that the subjective judgments needed for specifying Bayesian networks (i.e., judgments about conditional probabilities) are generated mentally on the basis of a stored model of functional relationships. For example, in the noisy-OR mechanism, which is often used to model causal interactions, the conditional probabilities are derivatives of a functional model involving AND/OR gates, corrupted by independent binary disturbances. This model is used, in fact, to *simplify* the specification of conditional probabilities in Bayesian networks (Pearl 1988).

[1] An observation by D. Heckerman (personal communication)

Evaluating counterfactual queries

From the last section, we see that the algorithm for evaluating counterfactual queries should consist of: (1) compute the posterior probabilities for the disturbance variables, given the observed evidence; (2) remove the observed evidence and enforce the value for the counterfactual antecedent; finally, (3) evaluate the probability of the counterfactual consequent, given the conditions set in the first two steps.

An important point to remember is that it is not enough to compute the posterior distribution of each disturbance variable (ϵ) separately and treat those variables as independent quantities. Although the disturbance variables are initially independent, the evidence observed tends to create dependencies among the parents of the observed variables, and these dependencies need to be represented in the posterior distribution. An efficient way to maintain these dependencies is through the structure of the causal network itself.

Thus, we will represent the variables in the counterfactual world as distinct from the corresponding variables in the real world, by using a separate network for each world. Evidence can then be instantiated on the real-world network, and the solution to the counterfactual query can be determined as the probability of the counterfactual consequent, as computed in the counterfactual network where the counterfactual antecedent is enforced. But, the reader may ask, and this is key, how are the networks for the real and counterfactual worlds linked? Because any exogenous variable, ϵ_a, is not influenced by forcing the value of any endogenous variables in the model, the value of that disturbance will be identical in both the real and counterfactual worlds; therefore, a single variable can represent the disturbance in both worlds. ϵ_a thus becomes a common causal influence of the variables representing A in the real and counterfactual networks, respectively, which allows evidence in the real-world network to propagate to the counterfactual network.

Assume that we are given a *causal theory* $T = \langle D, \Theta_D \rangle$ as defined in (Pearl & Verma 1991). D is a directed acyclic graph (DAG) that specifies the structure of causal influences over a set of variables $X = \{X_1, X_2, \ldots, X_n\}$. Θ_D specifies a functional mapping $x_i = f_i(\mathrm{pa}(x_i), \epsilon_i)$ ($\mathrm{pa}(x_i)$ represents the value of X_i's parents) and a prior probability distribution $P(\epsilon_i)$ for each disturbance ϵ_i (we assume that ϵ_i's domain is discrete; if not, we can always transform it to a discrete domain such as a response-function variable). A counterfactual query "What is $P(c^*|\hat{a}^*, obs)$?" is then posed, where c^* specifies counterfactual values for a set of variables $C \subset X$, $\hat{a}^*$ specifies forced values for the set of variables in the counterfactual antecedent, and *obs* specifies observed evidence. The solution can be evaluated by the following algorithm:

1. From the known causal theory T create a Bayesian network $< G, \mathcal{P} >$ that explicitly models the disturbances as variables and distinguishes the real world variables from their counterparts in the counterfactual world. G is a DAG defined over the set of variables $V = X \cup X^* \cup \epsilon$, where $X = \{X_1, X_2, \ldots, X_n\}$ is the original set of variables modeled by T, $X^* = \{X_1^*, X_2^*, \ldots, X_n^*\}$ is their counterfactual world representation, and $\epsilon = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_n\}$ represents the set of disturbance variables that summarize the common external causal influences acting on the members of X and X^*. $\mathcal{P}$ is the set of conditional probability distributions $P(V_i|\mathrm{pa}(V_i))$ that parameterizes the causal structure G.

   If $X_j \in \mathrm{pa}(X_i)$ in D, then $X_j \in \mathrm{pa}(X_i)$ and $X_j^* \in \mathrm{pa}(X_i^*)$ in G ($\mathrm{pa}(X_i)$ is the set of X_i's parents). In addition, $\epsilon_i \in \mathrm{pa}(X_i)$ and $\epsilon_i \in \mathrm{pa}(X_i^*)$ in G. The conditional probability distributions for the Bayesian network are generated from the causal theory:

   $$P(x_i|\mathrm{pa}_X(x_i), \epsilon_i) = \begin{cases} 1 & \text{if } x_i = f_i(\mathrm{pa}_X(x_i), \epsilon_i) \\ 0 & \text{otherwise} \end{cases}$$

   where $\mathrm{pa}_X(x_i)$ is the set of values of the variables in $X \cap \mathrm{pa}(x_i)$.

   $$P(x_i^*|\mathrm{pa}_{X^*}(x_i^*), \epsilon_i) = P(x_i|\mathrm{pa}_X(x_i), \epsilon_i)$$

   whenever $x_i = x_i^*$ and $\mathrm{pa}_{X^*}(x_i^*) = \mathrm{pa}_X(x_i)$. $P(\epsilon_i)$ is the same as specified by the functional causal theory T.

2. Observed evidence. The observed evidence *obs* is instantiated on the real world variables X corresponding to *obs*.

3. Counterfactual antecedent. For every forced value in the counterfactual antecedent specification $\hat{x}_i^* \in \hat{a}^*$, apply the action-based semantics of $set(X_i^* = \hat{x}_i^*)$ (see (Pearl 1993a; Spirtes, Glymour, & Scheines 1993)), which amounts to severing all the causal edges from $\mathrm{pa}(X_i^*)$ to X_i^* for all $x_i^* \in \hat{a}^*$ and instantiating X_i^* to the value specified in $\hat{a}^*$.

4. Belief propagation. After instantiating the observations and actions in the network, evaluate the belief in c^* using the standard belief update methods for Bayesian networks (Pearl 1988). The result is the solution to the counterfactual query.

In the last section, we noted that the conditional distribution $P(x_k|\mathrm{pa}(X_k))$ for each variable $X_k \in X$ constrains, but does not uniquely determine, the prior distribution $P(\epsilon_k)$ of each disturbance variable. Although the composition of the external causal influences are often not precisely known, a subjective distribution over response functions may be assessable. If a reasonable distribution can be selected for each relevant disturbance variable, the implementation of the above algorithm is straightforward and the solution is unique; otherwise, bounds on the solution can be obtained using convex optimization techniques. (Balke & Pearl 1993) demonstrates this optimization task in deriving bounds on causal effects from partially controlled experiments.

A network generated by the above algorithm may often be simplified. If a variable X_j^* in the counterfactual world is not a causal descendant of any of the variables mentioned in the counterfactual antecedent $\hat{a}^*$, then X_j and X_j^* will always have identical distributions, because the causal influences that functionally determine X_j and X_j^* are identical. X_j and X_j^* may therefore be treated as the same variable. In this case, the conditional distribution $P(x_j|\mathrm{pa}(x_j))$ is sufficient, and the disturbance variable ϵ_j and its prior distribution need not be specified.

Party again

Let us revisit the party example. Assuming we have observed that Bob is not at the party ($b = b_0$), we want to know whether Bob and Carl would have scuffled if Bob were at the party (i.e., "What is $P(s_1^*|\hat{b}_1^*, b_0)$?").

Suppose that we are supplied with the following causal theory for the model in Figure 1:

$$\begin{aligned} a = &\quad f_a(r_a) &= h_{a,r_a}() \\ b = &\quad f_b(a, r_b) &= h_{b,r_b}(a) \\ c = &\quad f_c(a, r_c) &= h_{c,r_c}(a) \\ s = &\quad f_s(b, c, r_s) &= h_{s,r_s}(b, c) \end{aligned}$$

where

$$P(r_a) = \begin{cases} 0.40 & \text{if } r_a = 0 \\ 0.60 & \text{if } r_a = 1 \end{cases}$$

$$P(r_b) = \begin{cases} 0.07 & \text{if } r_b = 0 \\ 0.90 & \text{if } r_b = 1 \\ 0.03 & \text{if } r_b = 2 \\ 0 & \text{if } r_b = 3 \end{cases}$$

$$P(r_c) = \begin{cases} 0.05 & \text{if } r_c = 0 \\ 0 & \text{if } r_c = 1 \\ 0.85 & \text{if } r_c = 2 \\ 0.10 & \text{if } r_c = 3 \end{cases}$$

$$P(r_s) = \begin{cases} 0.05 & \text{if } r_s = 0 \\ 0.90 & \text{if } r_s = 8 \\ 0.05 & \text{if } r_s = 9 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\begin{aligned} h_{a,0}() &= a_0 \\ h_{a,1}() &= a_1 \end{aligned}$$

$$h_{s,0}(b,c) = s_0$$

$$h_{s,8}(b,c) = \begin{cases} s_0 & \text{if } (b,c) \neq (b_1, c_1) \\ s_1 & \text{if } (b,c) = (b_1, c_1) \end{cases}$$

$$h_{s,9}(b,c) = \begin{cases} s_0 & \text{if } (b,c) \in \{(b_1,c_0),(b_0,c_1)\} \\ s_1 & \text{if } (b,c) \in \{(b_0,c_0),(b_1,c_1)\} \end{cases}$$

The response functions for B and C (h_{b,r_b} and h_{c,r_c}) both take the same form as that given in Eq. (5).

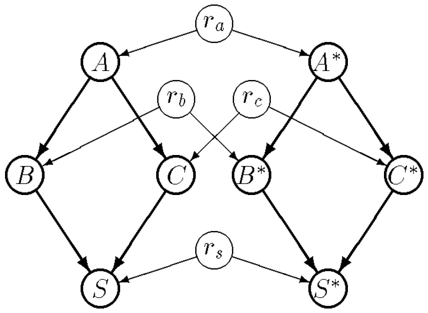

Figure 2: Bayesian model for evaluating counterfactual queries in the party example. The variables marked with $*$ make up the counterfactual world, while those without $*$, the factual world. The r variables index the response functions.

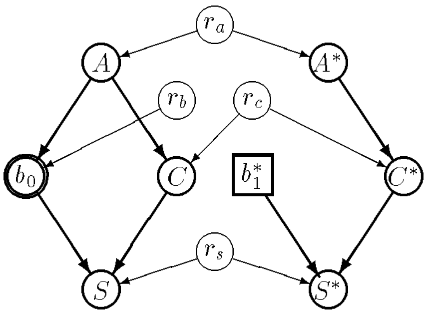

Figure 3: To evaluate the query $P(s_1^*|\hat{b}_1^*, b_0)$, the network of Figure 2 is instantiated with observation b_0 and action $\hat{b}_1^*$ (links pointing to b_1^* are severed).

These numbers reflect the authors' understanding of the characters involved. For example, the choice for $P(r_b)$ represents our belief that Bob usually is at the party if and only if Ann is there ($r_b = 1$). However, we believe that Bob is sometimes ($\sim 7\%$ of the time) unable to go to the party (e.g., sick or grounded by his parents); this exception is represented by $r_b = 0$. In addition, Bob would sometimes ($\sim 3\%$ of the time) go to the party if and only if Ann is not there (e.g., Bob is in a spiteful mood); this exception is represented by $r_b = 2$. Finally, $P(r_s)$ represents our understanding that there is a slight chance (5%) that Bob and Carl would not scuffle regardless of attendance ($r_s = 0$), and the same chance ($P(r_s{=}9) = 5\%$) that a scuffle would take place either outside or inside the party (but not if only one of then shows up).

Figure 2 shows the Bayesian network generated from step 1 of the algorithm. After instantiating the real world observations (b_0) and the actions ($\hat{b}_1^*$) specified by the counterfactual antecedent in accordance with steps 2 and 3, the network takes on the configuration shown in Figure 3.

If we propagate the evidence through this Bayesian network, we will arrive at the solution

$$P(s_1^*|\hat{b}_1^*, b_0) = 0.79.$$

which is consistent with Laura's assertion that Bob and Carl would have scuffled if Bob were at the party, given that Bob actually was not at the party. Compare this to the solution to the indicative query that Scott was thinking of:

$$P(s_1|b_1) \;=\; 0.11.$$

that is, if we had observed that Bob was at the party, then Bob and Carl would probably not have scuffled. This emphasizes the difference between counterfactual and indicative queries and their solutions.

Special Case: Linear-Normal Models

Assume that knowledge is specified by the structural equation model

$$\vec{x} \;=\; B\vec{x} + \vec{\epsilon}$$

where B is a triangular matrix (corresponding to a causal model that is a DAG), and we are given the mean $\vec{\mu}_\epsilon$ and covariance $\Sigma_{\epsilon,\epsilon}$ of the disturbances $\vec{\epsilon}$ (assumed to be normal). The mean and covariance of the observable variables $\vec{x}$ are then given by:

$$\vec{\mu}_x \;=\; S\vec{\mu}_\epsilon \tag{6}$$

$$\Sigma_{x,x} \;=\; S\Sigma_{\epsilon,\epsilon}S^t \tag{7}$$

where $S = (I - B)^{-1}$.

Under such a model, there are well-known formulas (Whittaker 1990, p. 163) for evaluating the conditional mean and covariance of $\vec{x}$ under some observations $\vec{o}$:

$$\vec{\mu}_{x|o} \;=\; \vec{\mu}_x + \Sigma_{x,o}\Sigma_{o,o}^{-1}(\vec{o} - \vec{\mu}_o) \tag{8}$$

$$\Sigma_{x,x|o} \;=\; \Sigma_{x,x} - \Sigma_{x,o}\Sigma_{o,o}^{-1}\Sigma_{o,y} \tag{9}$$

where, for every pair of sub-vectors, $\vec{z}$ and $\vec{w}$, of $\vec{x}$, $\Sigma_{z,w}$ is the sub-matrix of $\Sigma_{x,x}$ with entries corresponding to the components of $\vec{z}$ and $\vec{w}$. Singularities of Σ terms are handled by appropriate means.

Similar formulas apply for the mean and covariance of $\vec{x}$ under an action $\hat{\vec{a}}$. B is replaced by the action-pruned matrix $\hat{B} = [\hat{b}_{ij}]$ defined by:

$$\hat{b}_{ij} \;=\; \begin{cases} 0 & \text{if } X_i \in \hat{\vec{a}} \\ b_{ij} & \text{otherwise} \end{cases} \tag{10}$$

The mean and covariance of $\vec{x}$ under $\hat{B}$ is evaluated using Eqs. (6) and (7), where B is replaced by $\hat{B}$:

$$\hat{\vec{\mu}}_x \;=\; \hat{S}\vec{\mu}_\epsilon \tag{11}$$

$$\hat{\Sigma}_{x,x} \;=\; \hat{S}\Sigma_{\epsilon,\epsilon}\hat{S}^t \tag{12}$$

where $\hat{S} = (I - \hat{B})^{-1}$. We can then evaluate the distribution of $\vec{x}$ under the action $\hat{\vec{a}}$ by conditioning on the value of the action $\hat{\vec{a}}$ according to Eqs. (8) and (9):

$$\vec{\mu}_{x|\hat{a}} \;\triangleq\; \hat{\vec{\mu}}_{x|a} = \hat{\vec{\mu}}_x + \hat{\Sigma}_{x,a}\hat{\Sigma}_{a,a}^{-1}(\hat{\vec{a}} - \hat{\vec{\mu}}_a) \tag{13}$$

$$\Sigma_{x,x|\hat{a}} \;\triangleq\; \hat{\Sigma}_{x,x|a} = \hat{\Sigma}_{x,x} - \hat{\Sigma}_{x,a}\hat{\Sigma}_{a,a}^{-1}\hat{\Sigma}_{a,x} \tag{14}$$

To evaluate the counterfactual query $P(x^*|\hat{a}^*o)$ we first update the prior distribution of the disturbances by the observations $\vec{o}$:

$$\vec{\mu}^o_\epsilon \;\triangleq\; \vec{\mu}_{\epsilon|o} = \vec{\mu}_\epsilon + \Sigma_{\epsilon,\epsilon}S^t(S\Sigma_{\epsilon,\epsilon}S^t)^{-1}(\vec{o} - \vec{\mu}_o)$$

$$\Sigma^o_{\epsilon,\epsilon} \;\triangleq\; \Sigma_{\epsilon,\epsilon|o} = \Sigma_{\epsilon,\epsilon} - \Sigma_{\epsilon,\epsilon}S^t(S\Sigma_{\epsilon,\epsilon}S^t)^{-1}S\Sigma_{\epsilon,\epsilon}$$

We then evaluate the means $\vec{\mu}_{x^*|\hat{a}^*o}$ and variances $\Sigma_{x^*,x^*|\hat{a}^*o}$ of the variables in the counterfactual world (x^*) under the action $\hat{a}^*$ using Eqs. (13) and (14), with Σ^o and μ^o replacing Σ and μ.

$$\vec{\mu}_{x^*|\hat{a}^*o} \;\triangleq\; \vec{\mu}^o_{x|\hat{a}} = \hat{\vec{\mu}}^o_x + \hat{\Sigma}^o_{x,a}(\hat{\Sigma}^o_{a,a})^{-1}(\hat{\vec{a}} - \hat{\vec{\mu}}^o_a)$$

$$\Sigma_{x^*,x^*|\hat{a}^*o} \;\triangleq\; \Sigma^o_{x,x|\hat{a}} = \hat{\Sigma}^o_{x,x} - \hat{\Sigma}^o_{x,a}(\hat{\Sigma}^o_{a,a})^{-1}\hat{\Sigma}^o_{a,x}$$

where, from Eqs. (11) and (12), $\hat{\vec{\mu}}^o_x = \hat{S}\vec{\mu}^o_\epsilon$ and $\hat{\Sigma}^o_{x,x} = \hat{S}\Sigma^o_{\epsilon,\epsilon}\hat{S}^t$.

It is clear that this procedure can be applied to non-triangular matrices, as long as S is non-singular. In fact, the response-function formulation opens the way to incorporate feedback loops within the Bayesian network framework.

Conclusion

The evaluation of counterfactual queries is applicable to many tasks. For example, determining liability of actions (e.g., "If you had not pushed the table, the glass would not have broken; therefore, you are liable"). In diagnostic tasks, counterfactual queries can be used to determine which tests to perform in order to increase the probability that faulty components are identified. In planning, counterfactuals can be used for goal regression or for determining which actions, if performed, could have avoided an observed, unexpected failure. Thus, counterfactual reasoning is an essential component in plan repairing, plan compilation and explanation-based learning.

In this paper we have presented formal notation, semantics, representation scheme, and inference algorithms that facilitate the probabilistic evaluation of counterfactual queries. World knowledge is represented in the language of modified causal networks, whose root nodes are unobserved, and correspond to possible functional mechanisms operating among families of observables. The prior probabilities of these root nodes are updated by the factual information transmitted with the query, and remain fixed thereafter. The antecedent of the query is interpreted as a proposition that is established by an external action, thus pruning the corresponding links from the network and facilitating standard Bayesian-network computation to determine the probability of the consequent.

At this time the algorithm has not been implemented but, given a subjective prior distribution over the response variables, there are no new computational tasks introduced by this formalism, and the inference process follows the standard techniques for computing beliefs in Bayesian networks (Pearl 1988). If prior distributions over the relevant response-function variables cannot be assessed, we have developed methods of using the standard conditional-probability specification of Bayesian networks to compute upper and lower bounds on counterfactual probabilities (Balke & Pearl 1994).

The semantics and methodology introduced in this paper can be adopted to nonprobabilistic formalisms as well, as long as they support two essential components: abduction (to abduce plausible functional mechanisms from the factual observations) and causal projection (to infer the consequences of the action-like antecedent). We should note, though, that the license to keep the response-function variables constant stems from a unique feature of counterfactual queries, where the factual observations are presumed to occur not earlier than the counterfactual action. In general, when an observation takes place before an action, constancy of response functions would be justified if the environment remains relatively static between the observation and the action (e.g., if the disturbance terms ϵ_i) represent unknown pre-action conditions). However, in a dynamic environment subject to stochastic shocks a full temporal analysis using temporally-indexed networks may be warranted or, alternatively, a canonical model of persistence should be invoked (Pearl 1993c).

Acknowledgments

The research was partially supported by Air Force grant #AFOSR 90 0136, NSF grant #IRI-9200918, Northrop Micro grant #92-123, and Rockwell Micro grant #92-122. Alexander Balke was supported by the Fannie and John Hertz Foundation. This work benefitted from discussions with David Heckerman.

References

Balke, A., and Pearl, J. 1993. Nonparametric bounds on causal effects from partial compliance data. Technical Report R-199, UCLA Cognitive Systems Lab.

Balke, A., and Pearl, J. 1994. Bounds on probabilistically evaluated counterfactual queries. Technical report, UCLA Cognitive Systems Lab.

Boutilier, C. 1992. A logic for revision and subjunctive queries. In *Proceedings Tenth National Conference on Artificial Intelligence*, 609–15. Menlo Park, CA: AAAI Press.

Druzdzel, M. J., and Simon, H. A. 1993. Causality in bayesian belief networks. In *Proceedings of the 9th Annual Conference on Uncertainty in Artificial Intelligence (UAI-93)*, 3–11.

Ginsberg, M. L. 1986. Counterfactuals. *Artificial Intelligence* 30:35–79.

Goodman, N. 1983. *Fact, Fiction, and Forecast.* Cambridge, MA: Harvard University Press, 4th edition.

Harper, W. L.; Stalnaker, R.; and Pearce, G., eds. 1981. *Ifs: Conditionals, Belief, Decision, Chance, and Time.* Boston, MA: D. Reidel.

Jackson, P. 1989. On the semantics of counterfactuals. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 1382–7 vol. 2. Palo Alto, CA: Morgan Kaufmann.

Lewis, D. 1976. Probability of conditionals and conditional probabilities. *The Philosophical Review* 85:297–315.

Lewis, D. 1979. Counterfactual dependence and time's arrow. *Noûs* 455–476.

Meyer, J.-J., and van der Hoek, W. 1993. Counterfactual reasoning by (means of) defaults. *Annals of Mathematics and Artificial Intelligence* 9:345–360.

Nute, D. 1980. *Topics in Conditional Logic.* Boston: D. Reidel.

Pearl, J., and Verma, T. 1991. A theory of inferred causation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, 441–452. San Mateo, CA: Morgan Kaufmann.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, CA: Morgan Kaufman.

Pearl, J. 1993a. Aspects of graphical models connected with causality. Technical Report R-195-LLL, UCLA Cognitive Systems Lab. Short version: Statistical Science 8(3):266-269.

Pearl, J. 1993b. From Adams' conditionals to default expressions, causal conditionals, and counterfactuals. Technical Report R-193, UCLA Cognitive Systems Lab. To appear in *Festschrift for Ernest Adams*, Cambridge University Press, 1994.

Pearl, J. 1993c. From conditional oughts to qualitative decision theory. In *Uncertainty in Artificial Intelligence: Proceedings of the Ninth Conference*, 12–20. Morgan Kaufmann.

Pereira, L. M.; Aparicio, J. N.; and Alferes, J. J. 1991. Counterfactual reasoning based on revising assumptions. In *Logic Programming: Proceedings of the 1991 International Symposium*, 566–577. Cambridge, MA: MIT Press.

Poole, D. 1993. Probabilistic Horn abduction and Bayesian networks. *Artificial Intelligence* 64(1):81–130.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688–701.

Skyrms, B. 1980. The prior propensity account of subjunctive conditionals. In Harper, W.; Stalnaker, R.; and Pearce, G., eds., *Ifs*. D. Reidel. 259–265.

Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, Prediction, and Search.* New York: Springer.

Whittaker, J. 1990. *Graphical Models in Applied Multivariate Statistics.* New York: John Wiley & Sons.