Probabilistic Evaluation of Counterfactual Queries

Paper by: Alexander Balke and Judea Pearl

Kohei Tsujio

Contents

- Paper Contribution
- Counterfactual Query
- Inference Algorithm
- Conclusion

Paper Contribution

- Provides the formal notation for a counterfactual query such as

  > “If A were true, then what is the probability that C would have been true, given that we know B.”

- Provides an inference algorithm for the probabilistic evaluation of counterfactual queries

Counterfactual Query

- “If A were true, then C would have been true.”
  - A: Counterfactual antecedent
    - Specifies an event that is contrary to one’s real-world observations
  - C: Counterfactual consequent
    - Specifies a result that is expected to hold in the alternative world where the antecedent is true
- e.g.

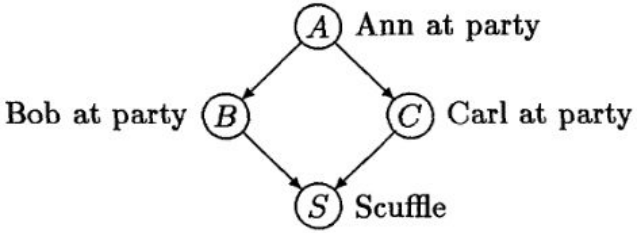
  “If Oswald were not to have shot Kennedy, then Kennedy would still be alive.”

Notation

- Variables describing the world: $X = \{X_1, X_2, \ldots, X_n\}$
- Real-world values: x_i
- Counterfactual world values: x_i^*
- Events that are referenced explicitly in the counterfactual antecedent: $\hat{x}$

- Query: $P(c^* | \hat{a}^*, a, b)$
  - Given that we have observed A = a and B = b in the real-world
  - If A were $\hat{a}^*$, then what is the probability that C would have been c^* ?

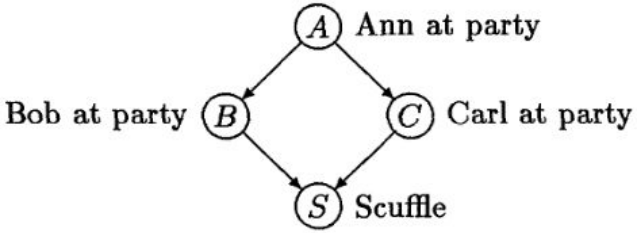
Party Example



- 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.

$$
\begin{array}{lll}
a \in \left\{ \begin{array}{ll} a_0 & \equiv \text{ Ann is not at the party.} \\ a_1 & \equiv \text{ Ann is at the party.} \end{array} \right\} \\
b \in \left\{ \begin{array}{ll} b_0 & \equiv \text{ Bob is not at the party.} \\ b_1 & \equiv \text{ Bob is at the party.} \end{array} \right\} \\
c \in \left\{ \begin{array}{ll} c_0 & \equiv \text{ Carl is not at the party.} \\ c_1 & \equiv \text{ Carl is at the party.} \end{array} \right\} \\
s \in \left\{ \begin{array}{ll} s_0 & \equiv \text{ No scuffle between Bob and Carl.} \\ s_1 & \equiv \text{ Scuffle between Bob and Carl.} \end{array} \right\}
\end{array}
$$

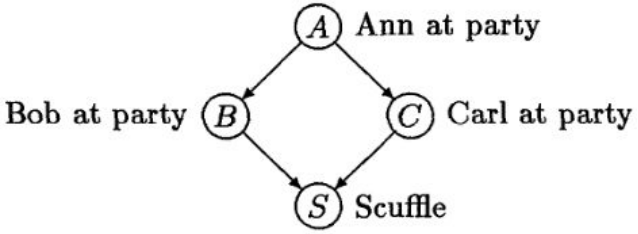
Party Example



- Assume that $P(b_1|a_1) = 0.9$ and $P(b_0|a_0) = 0.9$
- If Ann were at the party, what is the probability that Bob would have been at the party, given that we observe that both of them are absent from the party?

$$P(b_1^*|\hat{a}_1^*, a_0, b_0)$$

Party Example



- Assume that $P(b_1|a_1) = 0.9$ and $P(b_0|a_0) = 0.9$
- If Ann were at the party, what is the probability that Bob would have been at the party, given that we observe that both of them are absent from the party?

  $P(b_1^*|\hat{a}_1^*, a_0, b_0)$

- The answer depends on the 10% exception in Bob's behavior
  - If Bob occasionally misses parties (e.g. sick, study), the answer to the query would be 90%
  - However, for example, if Bob becomes angry with Ann in which case he does exactly the opposite of what she does, then the answer to the query is 100%, because Ann and Bob's absence from the party proves that Bob is not angry
- The conditional probabilities on the observed variables is insufficient for answering counterfactual queries uniquely

Functional Specification

- 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
- Given a specific value for ϵ_b, B becomes a deterministic function of A
- Response function is introduced to fully parameterize the model,

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

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

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

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

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

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

Functional Specification

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

- $P(b_1^*|\hat{a}_1^*, a_0, b_0)$
- If we observe (a_0, b_0) , it means $r_b \in \{0,1\}$, and the prior probability is described as $P(r_b=0) + P(r_b=1)$
- Posterior probability is

$$\vec{P}(r_b) = \langle P(r_b{=}0), P(r_b{=}1), P(r_b{=}2), P(r_b{=}3)\rangle)$$

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

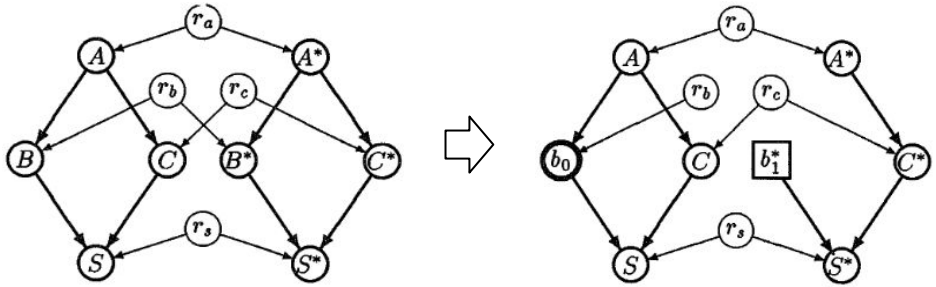
- If A were forced to a_1, then B would have been b_1, iff $r_b \in \{1,3\}$, which has probability $P'(r_b{=}1) + P'(r_b{=}3) = P'(r_b{=}1)$

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

Algorithm

- Assume a causal theory T = $\langle D, \Theta_D \rangle$ is given
  - D: Directed Acyclic Graph (DAG)
  - Θ_D: Functional mapping $x_i = f_i(\mathrm{pa}(x_i), \epsilon_i)$ and prior probability distribution $P(\epsilon_i)$
- The counterfactual query to be solved: $P(c^* | \hat{a}^*, obs)$
  - c^*: Counterfactual values for a set of variables $C \subset X$
  - $\hat{a}^*$: Forced valued for the set of variables in the counterfactual antecedent
  - obs : Observed evidence

Algorithm

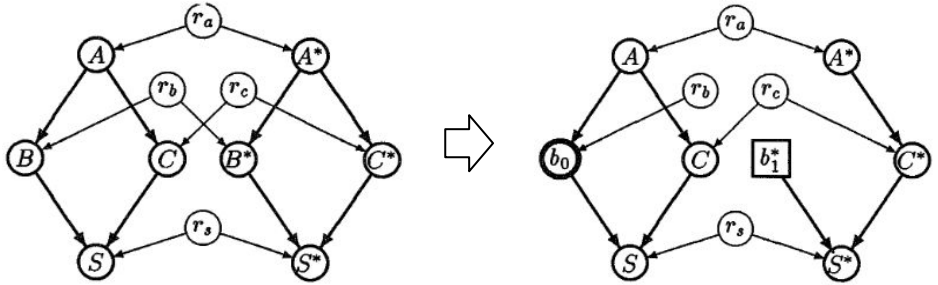


1. Create a Bayesian network $< G, \mathcal{P} >$
   a. G: DAG defined over $V = X \cup X^* \cup \epsilon$
   b. P: Set of conditional probability distributions $P(V_i|\text{pa}(V_i))$
2. Instantiate observed evidence obs on the real world variables X corresponding to obs
3. For every forced value in the counterfactual antecedent specification $\hat{x}_i^* \in \hat{a}^*$, remove the causal edges from $\text{pa}(X_i^*)$ to X_i^* for all $x_i^* \in \hat{a}^*$ and instantiating X_i^* to the value specified in $\hat{a}^*$
4. After instantiating the observations and actions in the network, evaluate the belief in c^* using the standard belief update methods for Bayesian networks.

   The result is the solution to the counterfactual query

Party Example Again

"What is $P(s_1^*|\hat{b}_1^*, b_0)$?"



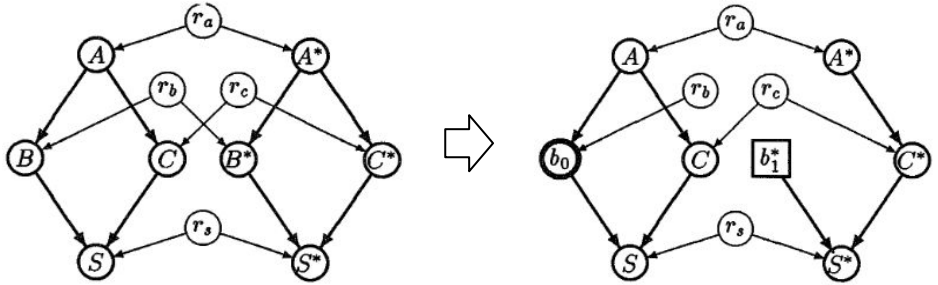
$$
\begin{array}{lll}
a = & f_a(r_a) & = h_{a,r_a}() \\
b = & f_b(a, r_b) & = h_{b,r_b}(a) \\
c = & f_c(a, r_c) & = h_{c,r_c}(a) \\
s = & f_s(b, c, r_s) & = h_{s,r_s}(b, c)
\end{array}
$$

$$
\begin{aligned}
P(r_a) &= \begin{cases} 0.40 & \text{if } r_a = 0 \\ 0.60 & \text{if } r_a = 1 \end{cases} \\
P(r_b) &= \begin{cases} 0.07 & \text{if } r_b = 0 \\ 0.90 & \text{if } r_b = 1 \\ 0.03 & \text{if } r_b = 2 \\ 0 & \text{if } r_b = 3 \end{cases} \\
P(r_c) &= \begin{cases} 0.05 & \text{if } r_c = 0 \\ 0 & \text{if } r_c = 1 \\ 0.85 & \text{if } r_c = 2 \\ 0.10 & \text{if } r_c = 3 \end{cases} \\
P(r_s) &= \begin{cases} 0.05 & \text{if } r_s = 0 \\ 0.90 & \text{if } r_s = 8 \\ 0.05 & \text{if } r_s = 9 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
h_{b,0}(a) &= b_0 \\
h_{b,1}(a) &= \begin{cases} b_0 & \text{if } a = a_0 \\ b_1 & \text{if } a = a_1 \end{cases} \\
h_{b,2}(a) &= \begin{cases} b_1 & \text{if } a = a_0 \\ b_0 & \text{if } a = a_1 \end{cases} \\
h_{b,3}(a) &= b_1
\end{aligned}
$$

$$
\begin{aligned}
h_{a,0}() &= a_0 \\
h_{a,1}() &= a_1 \\
h_{s,0}(b, c) &= s_0 \\
h_{s,8}(b, c) &= \begin{cases} s_0 & \text{if } (b, c) \neq (b_1, c_1) \\ s_1 & \text{if } (b, c) = (b_1, c_1) \end{cases} \\
h_{s,9}(b, c) &= \begin{cases} s_0 & \text{if } (b, c) \in \{(b_1, c_0), (b_0, c_1)\} \\ s_1 & \text{if } (b, c) \in \{(b_0, c_0), (b_1, c_1)\} \end{cases}
\end{aligned}
$$

Party Example Again



"What is $P(s_1^*|\hat{b}_1^*, b_0)$?"

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

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

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

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

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

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

$$\begin{array}{lll} h_{b,0}(a) & = & b_0 \\ h_{b,1}(a) & = & \begin{cases} b_0 & \text{if } a = a_0 \\ b_1 & \text{if } a = a_1 \end{cases} \\ h_{b,2}(a) & = & \begin{cases} b_1 & \text{if } a = a_0 \\ b_0 & \text{if } a = a_1 \end{cases} \\ h_{b,3}(a) & = & b_1 \end{array}$$

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

$$\begin{array}{lll} h_{s,0}(b, c) & = & s_0 \\ h_{s,8}(b, c) & = & \begin{cases} s_0 & \text{if } (b, c) \neq (b_1, c_1) \\ s_1 & \text{if } (b, c) = (b_1, c_1) \end{cases} \\ h_{s,9}(b, c) & = & \begin{cases} s_0 & \text{if } (b, c) \in \{(b_1, c_0), (b_0, c_1)\} \\ s_1 & \text{if } (b, c) \in \{(b_0, c_0), (b_1, c_1)\} \end{cases} \end{array}$$

Conclusion

- In this paper, the notation to describe counterfactual queries and the method to answer them are introduced
- Also, this paper provides the algorithm for evaluating the counterfactual queries
- By taking into account the unknown factors (exogenous variables), the authors successfully incorporates their effect into the model and indicates that evaluating counterfactual queries are answered by the proposed method

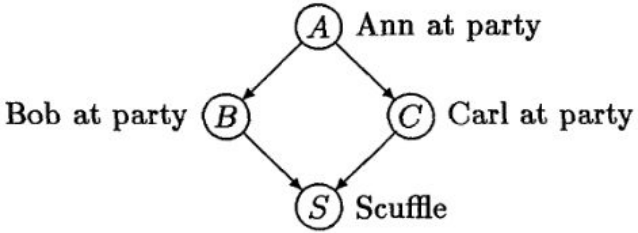
Homework

- Express the solution to the following counterfactual query using $P(r_b = i)$, $i = \{0, 1, 2, 3\}$

  “Given both Bob and Ann attended the party, if Ann were absent, then what is the probability that Bob would have been absent ?”

$$P\left(b_0^* \mid \hat{a}_0^*, a_1, b_1\right)$$

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



Thank you for listening

$$r_s(\epsilon_s) = \begin{cases} 0 & \text{if } F_s(b_0, c_0, \epsilon_s) = 0 \& F_s(b_0, c_1, \epsilon_s) = 0 \& F_s(b_1, c_0, \epsilon_s) = 0 \& F_s(b_1, c_1, \epsilon_s) = 0 \\ 1 & \text{if } F_s(b_0, c_0, \epsilon_s) = 0 \& F_s(b_0, c_1, \epsilon_s) = 0 \& F_s(b_1, c_0, \epsilon_s) = 0 \& F_s(b_1, c_1, \epsilon_s) = 1 \\ 2 & \text{if } F_s(b_0, c_0, \epsilon_s) = 0 \& F_s(b_0, c_1, \epsilon_s) = 0 \& F_s(b_1, c_0, \epsilon_s) = 1 \& F_s(b_1, c_1, \epsilon_s) = 0 \\ 3 & \text{if } F_s(b_0, c_0, \epsilon_s) = 0 \& F_s(b_0, c_1, \epsilon_s) = 0 \& F_s(b_1, c_0, \epsilon_s) = 1 \& F_s(b_1, c_1, \epsilon_s) = 1 \\ 4 & \text{if } F_s(b_0, c_0, \epsilon_s) = 0 \& F_s(b_0, c_1, \epsilon_s) = 1 \& F_s(b_1, c_0, \epsilon_s) = 0 \& F_s(b_1, c_1, \epsilon_s) = 0 \\ 4 & \text{if } F_s(b_0, c_0, \epsilon_s) = 0 \& F_s(b_0, c_1, \epsilon_s) = 1 \& F_s(b_1, c_0, \epsilon_s) = 0 \& F_s(b_1, c_1, \epsilon_s) = 1 \\ 6 & \text{if } F_s(b_0, c_0, \epsilon_s) = 0 \& F_s(b_0, c_1, \epsilon_s) = 1 \& F_s(b_1, c_0, \epsilon_s) = 1 \& F_s(b_1, c_1, \epsilon_s) = 0 \\ 7 & \text{if } F_s(b_0, c_0, \epsilon_s) = 0 \& F_s(b_0, c_1, \epsilon_s) = 1 \& F_s(b_1, c_0, \epsilon_s) = 1 \& F_s(b_1, c_1, \epsilon_s) = 1 \\ 8 & \text{if } F_s(b_0, c_0, \epsilon_s) = 1 \& F_s(b_0, c_1, \epsilon_s) = 0 \& F_s(b_1, c_0, \epsilon_s) = 0 \& F_s(b_1, c_1, \epsilon_s) = 1 \\ 9 & \text{if } F_s(b_0, c_0, \epsilon_s) = 1 \& F_s(b_0, c_1, \epsilon_s) = 0 \& F_s(b_1, c_0, \epsilon_s) = 0 \& F_s(b_1, c_1, \epsilon_s) = 0 \\ 10 & \text{if } F_s(b_0, c_0, \epsilon_s) = 1 \& F_s(b_0, c_1, \epsilon_s) = 0 \& F_s(b_1, c_0, \epsilon_s) = 1 \& F_s(b_1, c_1, \epsilon_s) = 1 \\ 11 & \text{if } F_s(b_0, c_0, \epsilon_s) = 1 \& F_s(b_0, c_1, \epsilon_s) = 0 \& F_s(b_1, c_0, \epsilon_s) = 1 \& F_s(b_1, c_1, \epsilon_s) = 0 \\ 12 & \text{if } F_s(b_0, c_0, \epsilon_s) = 1 \& F_s(b_0, c_1, \epsilon_s) = 1 \& F_s(b_1, c_0, \epsilon_s) = 0 \& F_s(b_1, c_1, \epsilon_s) = 1 \\ 13 & \text{if } F_s(b_0, c_0, \epsilon_s) = 1 \& F_s(b_0, c_1, \epsilon_s) = 1 \& F_s(b_1, c_0, \epsilon_s) = 0 \& F_s(b_1, c_1, \epsilon_s) = 0 \\ 14 & \text{if } F_s(b_0, c_0, \epsilon_s) = 1 \& F_s(b_0, c_1, \epsilon_s) = 1 \& F_s(b_1, c_0, \epsilon_s) = 1 \& F_s(b_1, c_1, \epsilon_s) = 1 \\ 15 & \text{if } F_s(b_0, c_0, \epsilon_s) = 1 \& F_s(b_0, c_1, \epsilon_s) = 1 \& F_s(b_1, c_0, \epsilon_s) = 1 \& F_s(b_1, c_1, \epsilon_s) = 0 \end{cases}$$