The Similarity of Causal Structure

Benjamin Eva* Reuben Stern† Stephan Hartmann‡

July 18, 2018

DRAFT – Please quote only with permission.

Abstract

Does y obtain under the counterfactual supposition that x? The answer to this question is famously thought to depend on whether y obtains in the most similar world(s) in which x obtains. What this notion of 'similarity' consists in is controversial, but in recent years, graphical causal models have proved incredibly useful in getting a handle on considerations of similarity between worlds. One limitation of the resulting conception of similarity is that it says nothing about what would obtain were the *causal structure* to be different from what it actually is, or from what we believe it to be. In this paper, we explore the possibility of using graphical causal models to resolve counterfactual queries about causal structure by introducing a notion of similarity between causal graphs. Since there are multiple principled senses in which a graph G^* can be more similar to a graph G than a graph G^{**}, we introduce multiple similarity metrics, as well as multiple ways to prioritize the various metrics when settling counterfactual queries about causal structure.

1 Introduction

Suppose that you had your temperature taken, and that the thermometer correctly indicated a healthy temperature of 37°C (or 98.6°F). What would have happened had the thermometer been broken so that its reading was exactly .5°C higher? Would your body temperature have been different than it actually was? Would you have subsequently taken a pill to reduce your fever?

Inspired by Lewis (1973) and Stalnaker (1968), most philosophers think that the answers to these questions can be determined by checking whether the closest possible world(s) in which the thermometer is broken is/are worlds in which your body temperature is different, or is/are worlds in which you take the pill. What does this closeness consist in? Most philosophers treat worlds as close to the actual world to the extent that they are *similar* to the actual world, where what

*University of Konstanz, 78464 Konstanz (Germany) – http://be0367.wixsite.com/benevaphilosophy – benjamin.eva@uni-konstanz.de.

†Munich Center for Mathematical Philosophy, LMU Munich, 80539 Munich (Germany) – https://sites.google.com/view/reubenstern/home – reuben.stern@gmail.com.

‡Munich Center for Mathematical Philosophy, LMU Munich, 80539 Munich (Germany) – http://www.stephanhartmann.org – s.hartmann@lmu.de.

counts as *similar* varies with context. Although the question of what settles the degree of similarity between possible worlds remains controversial, it is standardly thought that the similarity relation must capture the way in which counterfactual dependence typically tracks causal dependence. For example, given our knowledge of the way in which thermometers work, and of how people respond when thermometers register slight fevers, the relevant notion of similarity should vindicate the claim that the closest possible worlds in which the thermometer is broken may be worlds in which you subsequently take a pill (since the thermometer's appearance has a causal effect on whether you take the pill), but may not be worlds in which your body temperature is different than it actually is (since your body temperature is causally upstream from the thermometer's appearance).

Lewis (1979) famously used a complicated system of weights and priorities to develop an account of 'similarity' that delivers these results, but many philosophers—e.g., Hausman (1998) and Woodward (2005)—have convincingly argued that Lewis's account is inadequate for several reasons, chief among which is its inability to rigorously establish definite similarity orderings.[1] Meanwhile, some other authors—e.g., Briggs (2012), Halpern (2016), Pearl (2009), and Woodward (2005)—have argued that graphical causal models can be used to develop a more precise account that does at least as well at capturing our intuitions about counterfactual dependence. We agree with these authors. Though the details of these accounts vary from one author to the next, we are impressed with their general ability to capture counterfactual reasoning in a rigorous way.

There is, however, at least one important limitation to these authors' accounts. They say nothing about what would obtain were the *causal structure* to be different from what it actually is, or from what we believe it to be. The basic issue is that a graphical causal model specifies how the worlds that the causal model describes are more and less similar to each other, but it says nothing about how the causal model itself is more or less similar to other causal models. Thus these accounts are silent, e.g., with respect to how we should update our beliefs under the counterfactual supposition that your body temperature causally depends on the thermometer's appearance (rather than the other way around). Should we still believe that whether you take a pill is causally downstream from your body temperature? If you want an answer to this question, then you must look elsewhere.

In this paper, we explore the possibility of using graphical causal models to resolve counterfactual queries about causal structure by introducing a notion of similarity between causal graphs. Specifically, we aim to answer queries like the above by determining whether the counterfactual graph that is most similar to the original graph is one in which taking a pill is causally downstream from your body temperature. Since the causal structure of the world may be immutable, and since we seldom, if ever, are in a position to manipulate the causal structure of the world, it may seem that this project is valuable only insofar as it informs science fiction, or insofar as it quenches our natural thirst for knowledge about how things would be different were the world to be causally different. (After all, the primary reason that we value knowing what would happen were the thermometer to report inaccurately is that thermometers sometimes *do* break, and we care about the very *real* consequences of their breaking.) We must admit that some of our interest in these counterfactual queries is driven by pure curiosity, but the more pragmatic members of our audience needn't worry.

[1]The ambiguity of Lewis's (1979) account is in part due to his desire to develop an account of similarity that doesn't make any reference to causation (in the hope of reducing causal dependence to counterfactual dependence). Like, e.g., Halpern (2016), Pearl (2009) and Woodward (2005), we have no ambition to reduce causal dependence to counterfactual dependence.

We are also driven by our desire to make progress on the practically important problem of how agents should update their standing beliefs about causal structure when they learn that they are incorrect.

It is clear from work at the intersection of psychology, epistemology, and artificial intelligence that intelligent creatures like us often use qualitative beliefs about causal structure to organize our quantitative beliefs about evidential and counterfactual (ir)relevance.[2] When our beliefs about causal structure play this role, we do not entertain alternate possible causal structures; rather, we take a particular causal structure as given, and let it guide our evidential and practical reasoning. In so doing, we open ourselves up to the possibility of learning something that *conflicts* with our standing qualitative beliefs—e.g., when we initially believe that X causes Y and subsequently learn that Y temporally precedes X—and we need a method of belief revision that is well-suited to this kind of learning. In this setting, we submit that the agent should replace her belief in the original causal graph with a belief in a new causal graph that is among the graphs most similar to the original that are compatible with the new evidence.

Consider our attitudes towards the healthiness of foods. In order to make the world easier to navigate, we often reason as though we're certain that a particular food does (or does not) causally promote heart disease. (This considerably simplifies our deliberation about whether to eat the relevant food.) But when we learn later that we're wrong (e.g., because our best science now tells us that red meat consumption *does* causally promote heart disease), we need to know what causal structure to accept. Our proposal is that we should accept whatever graph is closest to our prior causal graph among the set of graphs that include the learned causal relationship between red meat consumption and heart disease.

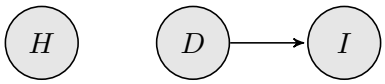
In order to develop a treatment of similarity that is up to the task at hand, we need to think carefully about what information is contained within a causal graph. We will see in what follows that causal graphs are used to represent how an agent's beliefs about causal structure constrain her beliefs about evidential relevance and counterfactual relevance, and that these two kinds of constraints operate somewhat independently from each other. In order to ensure that our treatment of similarity incorporates both considerations, we introduce an evidential similarity relation in Section 2, a counterfactual similarity relation in Section 3, and then consider possible ways of incorporating both kinds of similarity into one master concept in Section 4. We abstain from arguing for any particular master concept of similarity because it seems that there may be some contexts in which one master concept is appropriately deployed, and others where another is appropriately deployed. But in Section 5, we do take stock of when it matters which concept we use—i.e., of when (and when not) counterfactual queries about causal structure have the same answers regardless of which potential master concept is deployed. We argue that the answers to these counterfactual queries may stand on firmer ground than others because, unlike others, they do not depend on whether we give more priority to evidential similarity or counterfactual similarity.

[2]For recent empirical work explicating the role that qualitative beliefs about causal structure play in guiding evidential and counterfactual reasoning, see e.g. Ali, Chater and Oaksford (2011), Lagnado and Sloman (2002).

2 Evidential Similarity

In order to say what accounts for the similarity between causal graphs, we need to first take stock of what information is contained within a causal graph. Our first task is thus to identify the basic properties that distinguish causal structures.

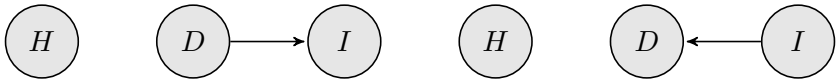
Given some set of variables, $\mathbf{V}$, we define a causal structure over $\mathbf{V}$ as a directed acyclic graph (DAG) over $\mathbf{V}$—i.e., a set of directed edges (or arrows) over $\mathbf{V}$ that are arranged such that no directed path forms a cycle, and where the directed edges are taken to represent direct causal dependencies. For example, if H represents height, D represents diet, and I represents intelligence, then the following DAG represents the causal structure according to which diet has a direct causal influence on intelligence, and no other (direct or indirect) causal relationships obtain between the variables in $\mathbf{V}$.



One way of capturing the characteristic content of this particular causal structure is that it has substantial implications for the *evidential* relationships between H, D, and I. For example, if this really is the true causal structure (and if we haven't omitted any common causes), then H and D should be probabilistically independent. This is because there should be no correlation between the respective values of two variables when there is absolutely no causal relationship between them.

The most powerful and widely accepted way of cataloging the evidential implications of a causal structure is given by the Causal Markov Condition (CMC), a generalization of Reichenbach's common cause principle. The CMC provides a general procedure for inferring probabilistic independencies (such as the independence between height and diet) from DAGs, and also helps us narrow down the set of DAGs that are compatible with a given probability distribution.[3]

One important aspect of the CMC is that it sometimes treats distinct DAGs as compatible with the same probability distributions. To illustrate, consider the following causal structures, G and G^*.



According to the CMC, G and G^* both imply that H is probabilistically independent of both D and I, and allow for the possibility that D is correlated with I. But according to the CMC, the direction of the causal dependence between D and I makes no difference with respect to the probability distribution. Structures like G and G^*—i.e., structures that imply the same probabilistic independencies—are known as *Markov equivalent*.

If we hope to explicate the notion of similarity between causal structures in a way that tracks their role in evidential reasoning, it seems that there should be some dimension of similarity according to which two Markov equivalent structures are always classified as maximally similar. More generally, a natural approach to assessing the similarity between distinct causal structures is to

[3]Note that additional axioms are also commonly employed towards these ends, most notably the causal minimality and causal faithfulness conditions (see Sprites et al., 2000).

measure the extent to which they have the same implications regarding the evidential relationships that obtain among the given variables. This motivates the following definition,

Definition 2.1 *Fix a variable set* $\boldsymbol{V}$ *and let* N_P *denote the cardinality of the set of all possible conditional independencies that can hold among the variables in* $\boldsymbol{V}$*. Given a causal structure* G*, let* P_G *denote the set of conditional independencies entailed by* G *and the CMC. Then we define the 'evidential distance' between two causal structures* G *and* G^* *over* $\boldsymbol{V}$ *to be* $d_E(G, G^*) = \frac{|P_G|\Delta|P_{G^*}|}{N_P}$*, where* Δ *is the symmetric difference.*

$d_E(G, G^*)$, or the *evidential distance* between G and G^*, simply counts the number of conditional independencies that are entailed by one but not both of G and G^*, normalized by the total number of possible conditional independencies that could hold between the variables in $\mathbf{V}$.[4]

Intuitively, $d_E(G, G^*)$ can be thought of as encoding the extent to which G and G^* disagree regarding the evidential relationships between the variables in $\mathbf{V}$. For example, it will be zero if and only if G and G^* are Markov equivalent. To further illustrate, consider the basic structures

$$G_1 : X \to Y \to Z,\ G_2 : X \leftarrow Y \to Z,\ G_3 : X \to Y \leftarrow Z$$

To calculate the evidential distance between them, we simply count how many probabilistic conditional independencies they disagree on and divide by the total number of possible conditional independencies, as below.[5]

	G_1	G_2	G_3
$X \perp Y$	×	×	×
$X \perp Z$	×	×	✓
$Y \perp Z$	×	×	×
$X \perp Y \mid Z$	×	×	×
$X \perp Z \mid Y$	✓	✓	×
$Y \perp Z \mid X$	×	×	×

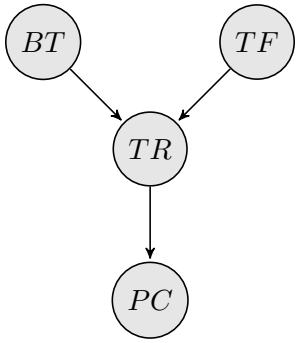
This yields the result that $d_E(G_1, G_2) = 0$ (as expected, since G_1 and G_2 are Markov equivalent) and $d_E(G_1, G_3) = \frac{1}{3} = d_E(G_2, G_3)$ since both G_1 and G_2 disagree with G_3 about two of the six possible independencies.

[4]One might be worried that the proposed definition of the evidential measure involves double counting, since it is always possible to generate the implied probabilistic independencies by considering a subset of independencies that imply the full set via the semi-graphoid axioms for conditional independence. For example, one might consider using only the 'basic independencies' discussed by Forster, Raskutti, Stern and Weinberger (2017). Though this approach will work when one is measuring the similarity between a DAG and one of its subgraphs, it will not work generally. This is because a DAG will generally have multiple different sets of basic independencies, and which set one chooses sometimes makes a difference to the value of d_E. Thus, when evaluating d_E we generally need to consider the full set of probabilistic independencies entailed by the graph, although we do make the harmless simplification of ignoring symmetrized independencies (i.e., we count only one of $X \perp Y|Z$ and $Y \perp X|Z$.)

[5]Given $X, Y, Z \in \mathbf{V}$, we write $X \perp Y|Z$ to indicate that X is probabilistically independent of Y conditional on Z. $X \perp Y$ means that X and Y are unconditionally independent.

3 Counterfactual Similarity

Of course, the characteristic content of a causal structure is not exhausted by its evidential implications. It is widely acknowledged that causal structure also plays a crucial role both in assessing the veracity of counterfactuals and in predicting the outcomes of prospective interventions. To illustrate, let BT represent body temperature, TF represent whether the thermometer properly functions, TR represent the thermometer reading, and PC represent whether a fever-reducing pill is subsequently consumed. The intuitive causal structure in this case is as below.



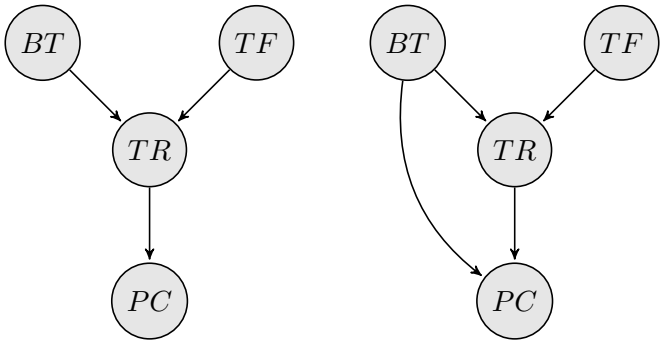
As we mentioned in the introduction, this DAG gives us a good handle on what counterfactually depends on what in most contexts. For example, we know that both TR and PC counterfactually depend on BT, but that TF does not counterfactually depend on BT. Similarly, as mentioned before, we know that BT and TF do not counterfactually depend on TR, and PC plausibly does counterfactually depend on TR.

In the interventionist framework there are two simple ways to account for these dependencies. First, the CMC entails that the *intervention* on X can be associated only with variables that are causally downstream from X, and counterfactual supposition plausibly corresponds to the supposition that one intervenes to set X to x.[6] Second, we can think of a DAG as imposing a partial *counterfactual* ordering on $\mathbf{V}$, where Y comes directly after X in the ordering exactly when X is a direct cause of Y, and where Y possibly counterfactually depends on X exactly when Y (directly or indirectly) follows X in the partial ordering.[7] Thus this graph imposes a counterfactual ordering according to which BT and TF directly precede TR, BT and TF are not ordered, and PC directly follows TR and indirectly follows both BT and TF.

We see now that DAGs not only have implications with respect to the evidential relationships that obtain between the variables in $\mathbf{V}$, but also imply constraints on the possible counterfactual dependencies that hold between those variables. Just as two distinct DAGs can embody the same evidential implications (e.g., when they are Markov equivalent), two distinct DAGs can embody the same implications regarding the counterfactual dependencies that obtain between the given variables. To illustrate, consider the following two causal structures.

[6] When we intervene to set X to x, we use some exogenous cause of X to set the value of X to x. For a full understanding of interventions, see Pearl (2009) or Spirtes et al. (2000).

[7] We restrict ourselves to talking about *possible* counterfactual dependence because in cases where the causal faithfulness condition fails, Y can follow X in the ordering without being sensitive to interventions on X.



Call the structure on the left G and the structure on the right G^*. G^* is identical to G except for the fact that it includes an additional direct causal influence from BT to PC. This additional causal relationship does not entail any new counterfactual dependencies, since PC already indirectly follows BT in the counterfactual ordering implied by G. Thus, G and G^* encode exactly the same sets of possible counterfactual dependencies, and we can call them *counterfactually equivalent.*[8]

In section 2 we presented a method for evaluating the degree to which different causal structures G and G^* have the same evidential implications. An analogous method can be defined for evaluating the degree to which G and G^* are counterfactually similar. Just as we required that, from the evidential perspective, Markov equivalent DAGs should be maximally similar, we require that counterfactually equivalent DAGs should be maximally similar from the counterfactual perspective.

Definition 3.1 *Fix a variable set* $\boldsymbol{V}$ *and let* N_C *denote the cardinality of the set of all possible counterfactual dependencies that can hold among the variables in* $\boldsymbol{V}$*. Given a causal structure* G*, let* C_G *denote the set of possible counterfactual dependencies allowed by* G *and the CMC. Then we define the 'counterfactual distance' between two causal structures* G *and* G^* *over* $\boldsymbol{V}$ *to be* $d_C(G, G^*) = \frac{|C_G| \Delta |C_{G^*}|}{N_C}$.

Given two causal structures G and G^* over a variable set $\mathbf{V}$, $d_C(G, G^*)$ simply counts all those possible counterfactual dependencies about which G and G^* disagree and normalizes by the cardinality of the set of possible counterfactual dependencies among $\mathbf{V}$. It immediately follows that $d_C(G, G^*) = 0$ if and only if G and G^* are counterfactually equivalent. To illustrate how d_C measures the distance between causal structures and its relationship to d_E, recall the basic structures G_1, G_2, G_3 from the previous section. The following table shows how we calculate the counterfactual distance between these structures (where $X < Y$ means that Y possibly counterfactually depends on X).

	G_1	G_2	G_3
$X < Y$	✓	×	✓
$Y < X$	×	✓	×
$X < Z$	✓	×	×
$Z < X$	×	×	×
$Y < Z$	✓	✓	×
$Z < Y$	×	×	✓

[8]It is easy to see that two DAGs can be counterfactually equivalent only if one is a subgraph of the other.

The evidential and counterfactual distances between these structures is summarized as follows.

	$d_E(-,-)$	$d_C(-,-)$
G_1, G_2	0	1/2
G_1, G_3	1/3	1/2
G_2, G_3	1/3	2/3

This table helps us to see how evidential distance and counterfactual distance can come apart. For example, G_2 and G_3 are equally counterfactually close to G_1, but when it comes to evidential distance, G_2 is closer to G_1 than G_3 is close to G_1. Because these two metrics are independent in this way, if we settle counterfactual queries by minimizing evidential distance, we will not always arrive at the DAG (or set of DAGs) that minimizes counterfactual distance. Likewise, if we minimize counterfactual distance, we will not always arrive at the DAG (or set of DAGs) that minimizes evidential distance. For example, if we start off believing $X \to Y \to Z$ and then learn that, as a matter of fact, $Y \to X$, it's easy to see that the evidentially closest DAG is $X \leftarrow Y \to Z$ (since it's Markov equivalent to the original). But it's also easy to see that $Y \to X \to Z$ is counterfactually closer to the original graph than $X \leftarrow Y \to Z$. This is because $Y \to X \to Z$ permits Z to counterfactually depend on X (as it did in the original graph) while $X \leftarrow Y \to Z$ does not.

This means that if we want to develop a notion of similarity that incorporates *both* kinds of similarity, we need some way of integrating both kinds of consideration into one and the same distance metric. In the next section, we introduce different ways of doing exactly this.

4 One Similarity to Rule Them All?

We are now ready to ask how evidential similarity and counterfactual similarity can be integrated into some procedure for settling counterfactual queries—i.e., that provides a semantics for counterfactuals whose antecedents specify counterfactual facts about causal structure, and that provides agents with instruction who need to revise their qualitative beliefs about causal structure on the basis of new conflicting evidence. This means constructing a procedures that identifies which DAGs that satisfy some constraint (supplied by either the new evidence or the counterfactual antecedent) are most similar to the original DAG. If we let G represent the original graph and let S represent the set of the DAGs that are compatible with the new (counterfactual or evidential) constraint, then the following three procedures are natural candidates.

1: First, find the set $S_E \subseteq S$ of structures in S that are evidentially closest to G, i.e. $S_E = \{G' \in S | d_E(G, G') = \min_{G'' \in S} d_E(G, G'')\}$. Next, find the set $S_{EC} \subseteq S_E$ of structures in S_E that are counterfactually closest to G, i.e. $S_{EC} = \{G' \in S_E | d_C(G, G') = \min_{G'' \in S_E} d_C(G, G'')\}$. Return S_{EC} as the set of structures in S that are most similar to G.

2: First, find the set $S_C \subseteq S$ of structures in S that are counterfactually closest to G, i.e. $S_C = \{G' \in S | d_C(G, G') = \min_{G'' \in S} d_C(G, G'')\}$. Next, find the set $S_{CE} \subseteq S_C$ of structures in S_C that are evidentially closest to G, i.e. $S_{CE} = \{G' \in S_C | d_E(G, G') = \min_{G'' \in S_C} d_E(G, G'')\}$. Return S_{CE} as the set of structures in S that are most similar to G.

3: Define the measure d_α as a weighted average of d_E and d_C, i.e. $d_\alpha(G, G') = (\alpha_E \cdot d_E(G, G')) + (\alpha_C \cdot d_C(G, G'))$. Return the set $S_\alpha \subseteq S$ that minimizes this distance as the set of structures in S that are most similar to G.

Informally, the procedures can be summarized as follows. The first two procedures are lexicographic. They privilege either the evidential or the counterfactual content of causal structures as more fundamental. According to the first procedure, one should first identify those structures whose evidential implications are closest to G's, and then break any ties by going with the structure(s) that is/are counterfactually closest. The second procedure is the exact inverse—i.e., it isolates the set of structures which are most similar to G from the counterfactual perspective and then breaks ties by identifying which elements of that set are evidentially closest to G. While the first and second procedures regard the evidential and counterfactual implications of a causal structure as more fundamental, respectively, the third procedure allows agents to *weigh* counterfactual and evidential considerations against one another in a more nuanced way—i.e., by taking the similarity between two structures to be a weighted average of their evidential and counterfactual similarity.

The relationship between these procedures and the resolution of counterfactual queries is not entirely clear. For example, one can develop the counterfactual semantics such that a given causal-structure-counterfactual is true relative to the standards imposed by a particular procedure only when the structural feature of its consequent is shared by *all* of the graphs returned by the procedure, or, alternatively, only when the structural feature of its consequent is shared by *some* graph returned by the procedure. Similarly, one can develop the norm of belief revision such that it is rationally permissible to accept *any* of the graphs returned by the relevant procedure, or such that it is rationally permissible to accept *no* particular graph when multiple graphs are returned. In this paper, we do not intend to settle questions of this sort, and instead primarily focus on questions about the similarity notion itself.

Upon putting these procedures on the table, it is immediately clear that what constitutes the closest graph in which some constraint is satisfied will depend on which procedure is used, and on how the weights are set when using the weighted procedure. Just consider our earlier example, where an agent initially accepts $X \to Y \to Z$ and then learns that $Y \to Z$. If the standards are set by the first procedure, the agent should come to accept that $X \leftarrow Y \to Z$. This is because it's Markov equivalent to the original graph and satisfies the constraint, therefore leaving no ties to be broken. On the other hand, if the standards are set by the second procedure, it can be shown that there are multiple DAGs that are strictly counterfactually closer than this one—e.g., $Y \to X \to Z$—and the evidential tie-breaker therefore cannot provide reason to favor $X \leftarrow Y \to Z$. Similarly, if considerably more weight is given to evidential similarity than counterfactual similarity, then it will be rational to accept $X \leftarrow Y \to Z$, but if considerably more weight is given to counterfactual similarity, then it will be rational to accept some other graph.

Since there may be some contexts where it is reasonable to prioritize evidential similarity (e.g., when beliefs about causal structure function primarily to constrain an agent's evidential probabilistic judgments) and other contexts where it is reasonable to prioritize counterfactual similarity (e.g., when beliefs about causal structure function primarily to impose order on things), the truth-values of causal-structure-counterfactuals, as well as rational belief updates about causal structure, may

depend on the context at hand.[9] Does this mean that there is no objective fact of the matter about how the causal structure would be were some local feature of it different from what it actually is?

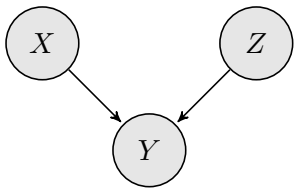
5 Whither Objectivity?

Even if each of the procedures from the last section is sometimes admissible, there may be an objective answer to how any counterfactual query should be answered in a given context. In order to establish this answer, we just determine which procedure should be used in the context at hand, and then the objective fact of the matter is settled by this procedure. But still, it may be surprising that the solutions to counterfactual queries about causal structure appear to depend on what we value—i.e., whether we prioritize evidential similarity or counterfactual similarity—in a way that "normal" counterfactual queries (or queries about the values of variables) do not.

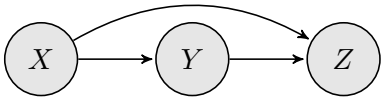
In this section, we consider whether there are any specific types of counterfactual queries about causal structure where the same results are returned no matter which procedure is used. When this happens, just as with "normal" counterfactual queries, there is a context-invariant objective fact of the matter about which causal-structure-counterfactuals are true, and about how agents should revise their qualitative beliefs about causal structure.

We focus primarily on when the two lexicographic procedures agree because it can be easily shown that the lexicographic procedures agree exactly when it doesn't matter what non-extreme weights are used in the weighted procedure.[10] Thus we can check whether context plays a role just by checking whether it matters which of the two lexicographic procedures is used. It is beyond the scope of this paper to prove general results about when the lexicographic procedures agree, but we identify three examples where they do agree that are interesting in their own right, and that may themselves be suggestive about what can be proved in the future.

Case 1 (Collider Conflict): Let G_1 be the basic collider structure



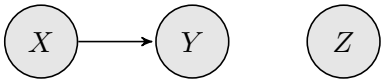
and suppose that we want to find the most similar causal structure in which Y has a direct causal influence on Z (rather than vice-versa). According to either of our lexicographic procedures, the following structure is uniquely most similar to G_1 amongst the set of DAG's that satisfy the given constraint.



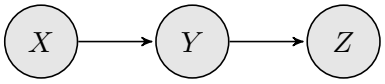
[9] We are officially silent with respect to whether the solutions to counterfactual queries vary with context. Indeed, one of us is sympathetic to the claim that counterfactual similarity should receive priority in every context.

[10] This is true because the lexicographic procedures return the same structures exactly when the intersection of evidentially and counterfactually closest structures is non-empty, and any structure that minimizes both the counterfactual and evidential distances will also minimize any weighted average of those distances.

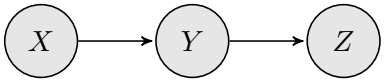
Case 2 (Adding an Arrow): Let G_1 be the structure



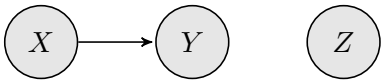
and suppose that we want to find the most similar causal structure in which Y has a direct causal influence on Z (rather than there being no causal relationship between them). Both lexicographic procedures agree on a single closest structure amongst all those that satisfy the given constraint. Specifically, they both return the chain structure



Case 3 (Disconnecting Two Variables): Let G_1 be the chain structure



and suppose that we want to find the most similar causal structure in which Y and Z are completely causally independent of one-another in the sense that neither causally influences the other, and they have no common causes or effects. Both lexicographic procedures agree on a single closest structure amongst all those that satisfy the given constraint:



The fact that these three cases exist shows that there is a noteworthy class of cases for which it does not matter how we prioritize considerations of evidential and counterfactual similarity. In these cases, we can say that there is an objective fact of the matter about what would happen were the causal structure to be different in a sense that does not exist generally—namely, no matter how considerations of evidential similarity and counterfactual similarity are prioritized, the most similar graph(s) is/are the same. Whether there are actually contexts in which each of these procedures should be used goes beyond the scope of this paper, so it is still possible to argue that the resolution of *every* causal-structure-counterfactual does not vary with context (by arguing for one of the lexicographical procedures). But since some practically minded individuals might think there are some contexts that call for one prioritization, and others that call for another, we believe that it is worth pointing out that there are some causal-structure-counterfactual queries that have the same resolution no matter how things are prioritized.

6 Conclusion

We have explicated two ways in which causal structures can be similar, and three ways in which these two notions of similarity can be integrated into a single master conception of similarity. We do not defend any of these master concepts as universally correct, but we have shown that there is a substantial range of cases in which they coincide. In future work we hope to explore

1: When and whether there is principled reason to favor one master conception (or one system of weights) over the alternatives.

2: When and whether any of these potential master conceptions can be used to assess the accuracy of causal search algorithms in terms of similarity to the true causal structure.[11]

3: Whether any of these potential master conceptions can be used to define a procedure for aggregating competing beliefs about causal structure in terms of finding the graph(s) that is on average most similar to the individually accepted graphs (see e.g. Bradley, Dietrich and List (2014)).

References

Ali, N., Chater, N. and Oaksford, M. (2011). The mental representation of causal conditional reasoning: mental models or causal models. *Cognition* 119(3): 403–418.

Bradley, R., Dietrich, F., and List, C. (2014). Aggregating Causal Judgements. *Philosophy of Science* 81(4): 491-515.

Briggs, R. (2012). Interventionist Counterfactuals. *Philosophical Studies* 160: 139–166.

Garant, D. and Jensen, J. (2016). Evaluating Causal Models by Comparing Interventional Distributions. https://arxiv.org/abs/1608.04698.

Hausman, D. (1998). *Causal Asymmetries.* Cambridge: Cambridge University Press.

Halpern, J. (2016). *Actual Causality.* Cambridge MA: MIT Press.

Lagnado, D., and Sloman, S. A. (2002). Learning causal structure. In W. Gray and C. Schunn (Eds.), *Proceedings of the twenty-fourth annual conference of the Cognitive Science Society*: 560–565. Mahwah, New Jersey: Erlbaum.

Lewis, D. (1973). *Counterfactuals.* Cambridge, MA: Harvard University Press. Reissued London: Blackwell, 2001.

Lewis, D. (1979). Counterfactual Dependence and Time's Arrow. *Nous*, 13: 455–476.

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

Stalnaker, R. (1968). A Theory of Conditionals. *Studies in Logical Theory, American Philosophical Quarterly, Monograph*: 2, 98–112

Sprites, P., Glymour, C., and Scheines, R. (2000): *Causation, Prediction and Search*, Cambridge, MA: MIT Press.

Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation.* Oxford Studies in the Philosophy of Science. Oxford: Oxford University Press.

[11]The so called 'graph-edit distance' is sometimes used for exactly this purpose, but it is now widely recognized that a different conception of similarity to the true causal structure is needed (see, e.g., Garant and Jensen (2016)).