Tracing Actual Causes

Anupam Datta, Depak Garg, Dilsun Kaynar, and Divya Sharma

August 8, 2016

CMU-CyLab-16-004

CyLab
Carnegie Mellon University
Pittsburgh, PA 15213

Tracing Actual Causes*

Anupam Datta (CMU) Deepak Garg (MPI-SWS)
Dilsun Kaynar (CMU) Divya Sharma (CMU)

Abstract

We study the problem of tracing actual causes, i.e. given an event e, we seek to fully *explain* why that event happened. This problem was articulated by David Lewis in his work on causal explanations [Lewis, 1986a]. We address the problem by defining the *causal history* of the explanandum event. Informally, the causal history traces the immediate causes of the explanandum event, the causes of those causes, and so on to produce a full explanation. While related, this problem differs from the problem of determining actual causes where the focus is on identifying individual events that contributed to causing the explanandum event. The central technical contributions of this paper are (a) a formal definition of causal history in the structural equations model; and (b) a theorem that establishes the complexity of the decision problem for causal histories. In addition, we show that our definition cleanly explains a battery of problematic examples from the actual cause literature.

1 Introduction

Actual cause determinations help answer questions of the following form: "Did John's smoking cause his lung cancer?", "What was the cause of the plane crash—a drunken pilot, a system failure, or a maintenance lapse?". This kind of determination is of interest in many fields, ranging from philosophy to law to computer science [Pearl, 2000; Moore, 2009; Spirtes *et al.*, 2000]. The focus on identifying causes of specific events separates actual causation from the related topic of type causation where the focus is on general causal relationships (e.g., "Smoking causes cancer").

Most recent treatments of actual causation involve counterfactuals. The counterfactual tradition goes back to Hume [Hume, 1748] whose position was that an event c is a cause of an event e if had c not occurred (the counterfactual), then e would not have occurred. While this simple idea does not always work, it provides a starting point for a significant body of work [Lewis, 1973; Pearl, 2000; Hitchcock, 2001; Halpern and Pearl, 2005; Halpern, 2008; Hall, 2007; Halpern, 2015]. In particular, an approach based on the structural equations model [Pearl, 2000] developed in the artificial intelligence and philosophy communities has proved to be influential [Hitchcock, 2001; Hall, 2004; Halpern and Pearl, 2005; Halpern, 2015]. Our work also employs the structural equations model but addresses a different (although related) question.

The problem. Our goal is to *trace* actual causes, i.e. given an event e, we seek to fully explain why that event happened. We do so by defining the *causal history* of the explanandum event. Informally, the causal history traces the immediate causes of the explanandum event, the causes of those causes, and so on to produce a full explanation. For example, the causal history of an accident might indicate that the first driver changed lanes suddenly without signaling, the second driver braked immediately, the car skidded because there was ice on the road and the tires were bald, and then hit a third car. We are inspired to pursue this goal, in part, because of a similar goal in Lewis's work on causal explanations [Lewis, 1986a]. Note that this goal is quite different from the goal of determining actual causes where the focus is on identifying individual events that contributed to causing the explanandum event. Thus, in the accident example above, individual events (e.g., changing lanes without signaling, ice on the road, bald tires) will be identified as actual causes; fully explaining why the accident happened would not be a goal of actual cause analysis.

Contributions. The central technical contributions of this paper are (a) a formal definition of causal history in the structural equations model; and (b) a theorem that establishes the complexity of the decision problem for causal histories. In addition, we show that our definition cleanly explains a battery of problematic examples from the actual cause literature.

The structural equations model (SEM) abstracts the world using a set of variables and a set of equations that capture the dependence of each variable on others. An event is simply a variable taking on a specific value. A simplified version of the accident example above can be modeled using variables A, I, C where $A = 1$ denotes the event that the accident occurred, $I = 1$ indicates that the road was icy, $C = 1$ indi-

*This work was partially supported by NSF CNS-1423168 and an AFOSR MURI on Science of Cybersecurity

cates that the driver changed lanes without signaling, and the equation $A \leftarrow I \vee C$ denotes that an accident occurs when the road is icy or the driver changes lanes without signaling. An SEM induces a natural graph (called the causal network) with a vertex for each variable and a directed edge from variable X to Y if the equation for computing X uses Y. Given an SEM M, a context $\vec{u}$ (that supplies the actual values for variables in the SEM), and an event e with $M, \vec{u} \models e$, our definition answers the question: Which paths of the causal network $G(M)$ caused the event e[1]? Our definition answers this question as a set of *causal slices*, where each causal slice is a subgraph of $G(M)$. All paths in each causal slice must act jointly to cause the event. However, each causal slice is sufficient in itself to cause the event. We additionally impose a necessity/minimality constraint: In each causal slice, we include a vertex or edge only if it is necessary to produce the outcome. Thus, each causal slice is a minimal set of paths that together suffice to produce the outcome. The *causal history* of an outcome is the set of all causal slices of the outcome. It represents a complete explanation of how the outcome came to be.

In the simplified accident example above, the causal history will include two causal slices in a context where $I = C = 1$: a slice with the path from I to A and a slice with a path from C to A. If instead the equation in the example were modified to $A \leftarrow I \wedge C$, then there would be only one causal slice with the two paths. These examples illustrate that our definition can distinguish between joint and independent causes—a distinction that is relevant for joint and several liability in tort cases [Prosser, 1937; Wright, 1987].

We prove that the decision problem for causal slices is $\mathrm{D}_1^{\mathrm{P}}$-complete. $\mathrm{D}_1^{\mathrm{P}}$ is the class of computational problems that can be solved using an NP machine and a co-NP machine simultaneously. Based on this result, we further show that the decision problem for causal histories is in Π_2^{P}.

Closely related work. While Lewis articulates the notion of causal history [Lewis, 1986a], he only discusses it informally. In contrast, we provide a formal definition of this notion and study the complexity of the associated decision problem. Our definition is inspired, in part, by the NESS test of causation (necessary elements of a sufficient set), proposed by Hart and Honoré and examined critically by Wright in the context of tort law [Hart and Honoré, 1985; Wright, 1985]. It also shares some commonalities with actual cause definitions in the structural equation model. The related work section provides a careful comparison.

2 Desiderata for the cause definition

Before describing the formal definition of causal history, we discuss the desideratum for such a definition, using simple examples of boolean circuits.

[1]In the technical section, we deal with a more general case of explaining outcomes that are boolean combinations of events.

Example 1 (Joint causes)**.** Consider a boolean circuit with two inputs connected via an AND gate. This can be modeled using three boolean variables X, A, B and the equation $X \leftarrow A \wedge B$. In this case, when the inputs are $A = 1$ and $B = 1$, the output is $X = 1$ and any reasonable definition of cause should say that both A and B caused this output. Moreover, these causes are *joint* in the sense that they must both be at their actual values 1 for the output $X = 1$: If either $A = 0$ or $B = 0$, then X is 0, not 1.

Example 2 (Causal slices and causal paths)**.** Consider an expansion of the previous example where A is also obtained from a circuit: $A \leftarrow C \wedge D$. Suppose that the inputs B, C, D are all 1, so the output X is also 1. In this case, the causes of $X = 1$ are A, B, C, and D. Again these causes are joint, because if any of A, B, C, D are forced to 0, the output X will become 0. However, this set is not fully descriptive, as it does not explain how these causes combine to result in $X = 1$. A more descriptive cause would also describe dependencies in the form of relevant paths through the circuit. Providing a definition of such a descriptive cause is the main goal of this paper. Specifically, we would want our definition to say that the paths $\{C \rightsquigarrow A \rightsquigarrow X, D \rightsquigarrow A \rightsquigarrow X, B \rightsquigarrow X\}$ caused the outcome $X = 1$, capturing how variables depend on each other in the circuit. We use the term *causal path* for paths like $C \rightsquigarrow A \rightsquigarrow X$ and the term *causal slice* for a set of causal paths like $\{C \rightsquigarrow A \rightsquigarrow X, D \rightsquigarrow A \rightsquigarrow X, B \rightsquigarrow X\}$ that describes an outcome. It will soon become clear that causal slices and causal paths really are slices and paths of a specific graph, namely, the causal network.

Revisiting Example 1, the causal slice would be $\{A \rightsquigarrow X, B \rightsquigarrow X\}$.

Example 3 (Independent causal paths)**.** Consider a revision of Example 1, where we replace the AND gate in X's circuit with an OR gate: $X \leftarrow A \vee B$. If $A = 1$ and $B = 1$, then $X = 1$ and the causes of this output are still A and B. However, in this case the nature of the relationship between these two causes is different: Here, the actual value 1 for either one of A and B will cause $X = 1$, even if the other input is 0. Such causes are called *independent causes*. We would like our definition to distinguish joint causal paths from independent causal paths, say by outputting two different causal slices $\{A \rightsquigarrow X\}$, $\{B \rightsquigarrow X\}$ in this example in place of the single causal slice $\{A \rightsquigarrow X, B \rightsquigarrow X\}$ of Example 1. We use the term *causal history* for all causal slices of an outcome.

Example 4 (Mixed causes)**.** In some cases, joint and independent causal paths may mix with each other. Consider a variant of Example 2 where the AND gate in A's equation is replaced by an OR gate: $A \leftarrow C \vee D$, $X \leftarrow A \wedge B$. Suppose C, D and B are all 1, so $X = 1$. In this case, our definition should yield a causal history containing two causal slices for $X = 1$ — $\{C \rightsquigarrow A \rightsquigarrow X, B \rightsquigarrow X\}$ and $\{D \rightsquigarrow A \rightsquigarrow X, B \rightsquigarrow X\}$ — capturing the fact that both the dependencies $A \rightsquigarrow X$ and $B \rightsquigarrow X$, but only one of the dependencies $C \rightsquigarrow A$ and $D \rightsquigarrow A$ is necessary for the outcome.

Preemption Preemption refers to the situation where the occurrence of an event precludes the possibility of another

event, which was otherwise possible. Modeling preemption and determining actual cause in its presence is a recognized challenge. For instance, a series of examples based on preemption were suggested against Lewis's counterfactual theory [Pearl, 2000; McDermott, 1995; Lewis, 1986b]. Hence, an important desideratum for us is that our definition determine causal slices accurately in models with preemptive events. To this end, we propose a new way of modeling preemption, which is compatible with our definition. We explain this in Section 3.5.

Summary To summarize, the primary desideratum for our definition of causal history is that it should produce causal slices and causal paths, not just causal events. Next, we desire that the definition distinguish joint causal paths from independent causal paths. Additionally, we want our definition to handle preemption-based examples cleanly. In the next section, we describe such a definition.

3 Definition

3.1 Model

Before defining causal paths, slices and histories, we need a language for modeling causal processes (circuits, systems, etc.) that generate caused and causative events. Following prior work on causation [Halpern and Pearl, 2005; Halpern, 2015; Pearl, 2000], we model causal processes as structural equations, which we recap briefly. Variables A, B, C, X, etc. model inputs, outputs and intermediate circuit points and are divided into two disjoint sets: the exogenous variables, whose values are determined by factors outside the model and the endogenous variables whose values are determined by the exogenous variables through the causal process described by the model. A signature $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$ specifies a set of endogenous variables ($\mathcal{V}$), a set of exogenous variables ($\mathcal{U}$) and a range $\mathcal{R}(X)$ for each variable $X \in \mathcal{U} \cup \mathcal{V}$. We assume that $\mathcal{V}$ and $\mathcal{R}(X)$ are finite. For a set of variables $\mathcal{W}$, we define $\mathcal{R}(\mathcal{W})$ as $\Pi_{X \in \mathcal{W}} \mathcal{R}(X)$.

A structural equations model or, simply, a *model* M is a pair $M = \langle \mathcal{S}, \mathcal{F} \rangle$ containing a signature $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$ and a set of equations $\mathcal{F}$. $\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function $F_X : D_X \to \mathcal{R}(X)$, where D_X, the domain of F_X, is a subset of $\mathcal{U} \cup \mathcal{V} \backslash \{X\}$. F_X specifies how the value of X can be computed, given the values of all variables in D_X. The relation between X and F_X is symbolically represented as $X \leftarrow F_X$, also called the *equation* for X. Every model M induces a directed graph $G(M)$, called the *causal network*, whose vertices are elements of $\mathcal{V}$ and which has an edge from Y to X iff $Y \in D_X$ [Halpern and Pearl, 2005]. Such an edge is written $Y \rightsquigarrow X$ ("Y might influence X"). As in prior work [Halpern and Pearl, 2005; Halpern, 2015], we are only interested in models M for which $G(M)$ is *acyclic*. In the sequel, we assume that all models under consideration are acyclic in this sense. A *context* $\vec{u}$ is an assignment of values to all exogenous variables $\mathcal{U}$, i.e., an element of $\mathcal{R}(\mathcal{U})$. It is clear that given an acyclic model and a context $\vec{u}$, the values of all endogenous variables $\mathcal{V}$ are uniquely determined by the equations of the model. These values are called the *actual values* of the variables.

Next, we describe the vocabulary with which we express causative and caused events. A *primitive event* is an assertion of the form $X = x$, where $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$. An *event* or *formula* φ is a boolean combination of primitive events. We say that the primitive event $X = x$ holds in M and $\vec{u}$, written $M, \vec{u} \models X = x$, if the actual value of X determined by M and $\vec{u}$ is x. We lift this to the judgment $M, \vec{u} \models \varphi$ in the obvious way.

For any function $f : (A_1 \times \ldots \times A_n \times B) \to C$ and any $\vec{a} \in \mathcal{R}(\Pi_{i=1}^n A_i)$, we define the specialization of f to $\vec{a}$, written $f|_{\vec{A} \leftarrow \vec{a}}$, as the function $g : B \to C$ defined by $g(b) = f(\vec{a}, b)$.

3.2 The definition of causal history

We now present our definition of causal history that captures the desideratum from Section 2. Briefly, given a model M, a context $\vec{u}$ and a formula φ with $M, \vec{u} \models \varphi$, our definition answers the question: Which paths of the causal network $G(M)$ caused φ? Our definition answers this question as a *causal history*, which is a set of *causal slices*, where each causal slice is a subgraph of $G(M)$. All paths in each causal slice must act *jointly* to cause the outcome. However, each causal slice is sufficient *in itself* to cause the outcome φ. Thus paths within a causal slice are joint causes and paths in two separate causal slices are independent causes. We additionally impose a necessity/minimality constraint: In each causal slice, we include a vertex or edge only if it is necessary to produce the outcome. Thus, each causal slice is a minimal set of paths that together suffice to produce the outcome.

Definition 5 (Causal slice, causal path)**.** Suppose we are given a model $M = \langle \mathcal{S} = \langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle, \mathcal{F} \rangle$, a context $\vec{u}$ and a formula φ. Let G be a subgraph of $G(M)$, let $\{X_1, \ldots, X_n\}$ be the set of vertices (endogenous variables) in G and for $i \in \{1, \ldots, n\}$, let x_i denote the actual value of X_i over M and u, i.e., suppose that $M, u \models X_i = x_i$.

We call G a *causal slice* of φ over $M, \vec{u}$ if the following hold:

1. (**Outcome**) $\bigwedge_{i=1}^n (X_i = x_i)$ entails φ. (And, hence, $M, \vec{u} \models \varphi$.)
2. (**Sufficiency**) For any vertex $X_i \in \vec{X}$, if $X_i \leftarrow F_{X_i}$ is the equation for X_i and $T = \{X_j \leftarrow x_j \mid (X_j \rightsquigarrow X_i) \in G\}$, then $F_{X_i}|_{T, \mathcal{U} \leftarrow \vec{u}}$ is the constant function that returns x_i.
3. (**Minimality**) No proper subgraph of G satisfies both conditions 1 and 2.

If G is a causal slice of φ over $M, \vec{u}$, then we call each maximal path in G a *causal path*.

Intuitively, G is a causal slice of φ if it is a *minimal* subgraph of $G(M)$ that *suffices* to cause the outcome φ for the specific context $\vec{u}$. This means that even if all dependencies $X \rightsquigarrow Y$ outside of G are broken (by using arbitrary values for X in the evaluation of Y), we still get the φ.

In detail, the Outcome condition checks that we have included enough vertices (endogenous variables) in the causal slice to justify φ: The actual values of included variables must

entail φ. A simple consequence is that, unless φ is a tautology, at least one endogenous variable occurring in φ must also occur in the causal slice.

The Sufficiency condition says that for each variable X_i that occurs in G, the actual values of variables X_j on the incoming edges of X_i in G are sufficient to force X_i to its actual value, irrespective of the values of other endogenous variables. This is exactly what the phrase "$F_{X_i}|_{T,\mathcal{U}\leftarrow\vec{u}}$ is the constant function that returns x_i" means.

The Minimality condition checks that everything included in G is necessary for the outcome φ.

The *causal history* of an outcome is the set of all causal slices of the outcome. It represents a complete explanation of how the outcome came to be.

Definition 6 (Causal history)**.** The causal history of φ over $M, \vec{u}$ is the set of all causal slices of φ over $M, \vec{u}$.

3.3 Examples

We now present several examples that illustrate our definition. We represent causal slices (and graphs in general) as sets of causal paths.

We start by revisiting Examples 1–4 from Section 2. Example 1 can be modeled in structural equations using three endogenous variables X, A, B, two exogenous variables U_A and U_B, which represent the outcomes of the (external) processes that determine the inputs A and B, respectively, and the equations $X \leftarrow A \wedge B$, $A \leftarrow U_A$ and $B \leftarrow U_B$. In the actual scenario, $U_A = U_B = 1$. It is clear that $X = 1$ holds, so we ask what the causal history of $X = 1$ is. Definition 5 yields exactly one causal slice, as expected: $G = \{A \rightsquigarrow X, B \rightsquigarrow X\}$. To see that this is a causal slice, note that because X is in G, the Outcome condition is trivially satisfied. Sufficiency holds trivially at A and B because the right hand sides of the equations of A and B are independent of endogenous variables. Sufficiency holds at X because both the edges $A \rightsquigarrow X$ and $B \rightsquigarrow X$ are in G, so $F_X|_T = F_X|_{A\leftarrow 1,B\leftarrow 1} = (A \wedge B)|_{A\leftarrow 1,B\leftarrow 1} = 1 \wedge 1 = 1$, which is a constant function that returns 1, the actual value of X. Finally, Minimality holds because removing anything from G breaks Sufficiency at X. For instance, if we remove $A \rightsquigarrow X$ from G, we get $F_X|_T = (A \wedge B)|_{B\leftarrow 1} = A$, which is not a constant function.

Example 2 can be analyzed similarly. There, the only causal slice is $\{C \rightsquigarrow A \rightsquigarrow X, D \rightsquigarrow A \rightsquigarrow X, B \rightsquigarrow X\}$.

Example 3 is modeled like Example 1 above, but the equation for X is $X \leftarrow A \vee B$. Now our definition yields two causal slices for $X = 1$: $G_1 = \{A \rightsquigarrow X\}$ and $G_2 = \{B \rightsquigarrow X\}$ because either one of the edges $A \rightsquigarrow X$ and $B \rightsquigarrow X$ suffices to force F_X to 1. For example, $F_X|_{A\leftarrow 1} = (A \vee B)_{A\leftarrow 1} = (1 \vee B) = 1$, which justifies the Sufficiency condition at X for the causal slice $\{A \rightsquigarrow X\}$.

Example 4 requires a more tedious analysis, but the result is the expected one. We get the two causal slices $\{C \rightsquigarrow A \rightsquigarrow X, B \rightsquigarrow X\}$ and $\{D \rightsquigarrow A \rightsquigarrow X, B \rightsquigarrow X\}$, reflecting the fact that both the edges $A \rightsquigarrow X$ and $B \rightsquigarrow X$ but only one of the edges $C \rightsquigarrow A$ and $D \rightsquigarrow A$ is necessary for the outcome $X = 1$.

All the examples so far had tree-shaped causal networks. The following examples demonstrate how our definition handles general acyclic models.

Example 7 (Backup)**.** This example is paraphrased from Hitchcock [Hitchcock, 2001]. A trainee is required to shoot at a target. His supervisor is also present. If the trainee loses his nerve and does not shoot, then the supervisor will shoot. In the actual scenario, the trainee shoots and hits the target. What is the causal history of the target being hit? This example is interesting because the *target is always hit*, independent of whether or not the trainee shoots. Hence, a naive definition may say that any causal slice should only contain the dependency between the trainee and the supervisor, not the trainee's shot itself. However, our definition correctly identifies the expected causal path from the trainee to the target.

To model this example, we use three endogenous boolean variables — T (1 if the trainee shoots, 0 otherwise), S (1 if the supervisor shoots) and H (1 if the target is hit). We also use one exogenous variable, U_T, which models the outcome of the external process that decides whether or not the trainee shoots ($U_T = 1$ when the trainee shoots). The equations are: $T \leftarrow U_T$, $S \leftarrow \neg T$ and $H \leftarrow T \vee S$. We ask for the causal history of $H = 1$, when $U_T = 1$. Note that $H = 1$ independent of the value of U_T. However, intuitively, it is clear that when $U_T = 1$, the trainee's shot is the cause of the target being hit, so the only expected causal slice is $\{T \rightsquigarrow H\}$. Indeed, our definition determines exactly this causal slice. To see this, note that this slice satisfies Sufficiency at H because $F_H|_{T\leftarrow 1} = (T \vee S)|_{T\leftarrow 1} = (1 \vee S) = 1$, which is the actual value of H. Second, note that any graph that does not include $T \rightsquigarrow H$ cannot be a causal slice of $H = 1$. This is because S's actual value is 0, so unless T is restricted to 1, $T \vee S$ cannot be the constant function 1. If, instead, we ask for the causal slice with $U_T = 0$ (the trainee does not shoot), then our definition correctly identifies the causal slice $\{T \rightsquigarrow S \rightsquigarrow H\}$ (trainee does not shoot, therefore the supervisor shoots and, therefore, the target is hit).

Example 8 (Multiple causal slices with same causal events)**.** This example illustrates a situation that has more than one causal slice, but all with the same events (vertexes/variables). Hence, no definition of cause that finds only events as causes can output all the nuances of this example and this example canonically justifies our use of causal slices in place of causal events for the outcome of causal analysis.

Alice works at a firm. She can be fired if her two managers Bob and Charlie, and the human resources all agree to fire her. However, the responsible human resources employee (HR) is lazy and agrees to fire anyone if either Bob or Charlie wish to fire the person. In the actual scenario, both Bob and Charlie agree to fire Alice. Each conveys this to HR, who then also agrees. As a result, Alice is fired. What led to Alice's firing?

To model this example, we use four endogenous boolean variables – F (1 if Alice is fired), and B, C, H (1 if Bob, Charlie and HR, respectively, agree to fire Alice). There are two exogenous variables U_B and U_C, which are 1 when (external) processes determine that Bob and Charlie should fire Alice. The equations are: $F \leftarrow B \wedge C \wedge H$, $H \leftarrow B \vee C$, $B \leftarrow U_B$ and $C \leftarrow U_C$. In the actual scenario, $U_B = U_C = 1$, and the goal is to find the causal

slice(s) for $F = 1$. It is not difficult to check that there are two causal slices: $\{B \rightsquigarrow F, C \rightsquigarrow F, B \rightsquigarrow H \rightsquigarrow F\}$ and $\{B \rightsquigarrow F, C \rightsquigarrow F, C \rightsquigarrow H \rightsquigarrow F\}$. The edges $B \rightsquigarrow F$, $C \rightsquigarrow F$ and $H \rightsquigarrow F$ are obviously necessary for $F = 1$ due to the equation $F \leftarrow B \wedge C \wedge H$. However, only one of $B \rightsquigarrow H$ and $C \rightsquigarrow H$ is necessary because HR would have agreed to fire Alice at the behest of just one of Bob and Charlie. This justifies the two causal slices informally. Note also that the set of endogenous variables in both the slices is exactly the same, $\{B, C, H, F\}$.

Remark 9 (Normality and defaults)**.** Halpern [Halpern, 2008] observes that in many situations considering all possible counterfactual contingencies for a variable is unreasonable and results in counterintuitive causal determinations. To deal with such situations, he proposes to restrict counterfactual contingencies by augmenting models with information about what is expected or "normal" and what the default values of variables are (in the absence of other information). We note that our definitions of causal slice and causal history are compatible with such restrictions. Specifically, the domains of the non-specialized variables in the constancy test of the Sufficiency condition of Definition 5 can be limited to normal or default values. We omit details due to lack of space.

3.4 Properties and computational complexity

A natural property, rather a sanity check, on our definition of causal slice (Definition 5) is that a causal slice must contain a path from any endogenous variable in it to a variable in the outcome φ, else the former variable obviously cannot influence the outcome. The following lemma captures this property.

Lemma 10 (Relevance of variables in causal slices)**.** If G is a causal slice of φ over $M, \vec{u}$, and X is an endogenous variable that appears in G, then there is a path in G that leads from X to a variable in φ.

Proof. Immediate from the Minimality condition of Definition 5. □

Our next property says that every causal slice G has a closure property: For any variable X in G with actual value x, G also contains a subgraph that is a causal slice of $X = x$. This essentially means that causal slices contain all causes, transitively. For a directed acyclic graph G and a vertex $X \in G$, define $G|_X$ as the subgraph of G containing only those vertices and edges from which X is reachable along some path in G.

Theorem 11 (Causal slices are closed)**.** If G is a causal slice of φ over $M, \vec{u}$, X is an endogenous variable that appears in G and $M, \vec{u} \models X = x$, then some subgraph of $G|_X$ is a causal slice of $X = x$ over $M, \vec{u}$.

Proof. It can be proved that $G|_X$ satisfies Outcome and Sufficiency conditions for $X = x$ over $M, \vec{u}$. Hence, it must contain a minimal subgraph also satisfying these two conditions. □

Finally, we establish the computational complexity of decision problems for causal slices and causal histories. If all functions F_X in the structural equations are computable in polynomial time (i.e., they lie in the complexity class P), then the problem of checking whether a given subgraph of $G(M)$ is a causal slice is $\mathrm{D}_1^{\mathrm{P}}$-complete. To recapitulate briefly, $\mathrm{D}_1^{\mathrm{P}}$ contains a language L iff $L = L_1 \cap L_2$, where $L_1 \in$ NP and $L_2 \in$ co-NP. Note that $\mathrm{D}_1^{\mathrm{P}}$ is *not* the intersection of NP and co-NP. In fact, it contains both these classes. Using this, we can immediately show that the decision problem for causal histories is in Π_2^{P}.

Theorem 12. Assuming that all structural equations are P-time computable, the following language L is $\mathrm{D}_1^{\mathrm{P}}$-complete.

$$L = \{(M, u, \varphi, G) \mid G \text{ is a causal slice of } \varphi \text{ on } (M, u)\}$$

Theorem 13. Assuming that all structural equations are P-time computable, the following language L is in Π_2^{P}.

$$L = \{(M, u, \varphi, H) \mid H \text{ is the causal history of } \varphi \text{ on } (M, u)\}$$

3.5 Handling preemption

Often, the model constrains two or more primitive events to be mutually exclusive and this property is relevant to the determination of cause. Consider the following example, originally due to Hall [Hall, 2004] and quoted here from Halpern and Pearl [Halpern and Pearl, 2005].

Example 14 (Billy-Suzy preemption)**.** Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle had it not been preempted by Suzy's throw.

To model this example, we choose three endogenous boolean variables: ST (1 iff Suzy throws), BT (1 iff Billy throws) and BS (1 iff bottle shatters), and two exogenous variables: U_S (1 when Suzy throws) and U_B (1 when Billy throws). The structural equations are $BS \leftarrow ST \vee BT$, $ST \leftarrow U_S$ and $BT \leftarrow U_B$. We ask for the causal history of $BS = 1$ when $U_S = U_B = 1$. From the textual description of the situation, it is clear that Suzy's throw is the cause. Yet, our definition also yields Billy's throw as a cause. Formally, we get two causal slices: $G_1 = \{ST \rightsquigarrow BS\}$ and $G_2 = \{BT \rightsquigarrow BS\}$. This discrepancy arises because, even though it is clear to us (as humans) that since Suzy's throw reaches the bottle first, Billy's throws cannot reach the bottle first (i.e., the first event preempts the second), the model is symmetric in BT and ST and does not capture this preemption (the model is isomorphic to that of Example 3). Hence, in applying the definition, we consider the spurious contingency that Billy's throw reaches the bottle first, which yields the spurious causal slice.

Hence, the model must be modified to reflect the preemption. There are many ways to do this. We describe here a way that is compatible with our definition and relies on additional *exogenous* variables. We introduce a new exogenous boolean variable R that determines whether Suzy's throw reaches the bottle first ($R = 1$) or Billy's throw reaches the bottle first ($R = 0$). The equation for the bottle's shattering is revised to $BS \leftarrow$ if R then ST else BT. We now ask for the causal history of $BS = 1$ over M and $U_S = U_B = R = 1$. It can

easily be checked that now we get only the expected causal slice $G_1 = \{ST \rightsquigarrow BS\}$.

Note that we introduce the assumption that Suzy's throw reaches the bottle first into the context ($R = 1$) rather than the model M. To obtain the cause when Billy's throw reaches the bottle first, we could simply repose the question with $R = 0$ without having to change the model. This differs from some prior work on actual cause, e.g., [Halpern and Pearl, 2005], where the updated model itself represents the fact that Suzy's throw reaches the bottle first. In such cases, the model must be revised to ask the question about the other circumstance.

This method of modeling preemption through exogenous variables is quite general. We have successfully applied it to other examples from literature that require preemption: early preemption [Hitchcock, 2007; Pearl, 2000] late preemption [Hall, 2004; Hitchcock, 2007] and trumping preemption [Schaffer, 2000].

4 Relationship with other approaches

In this section, we compare our causal history definition to related work.

Lewis' causal history. Our work is most closely related to Lewis' notion of causal histories [Lewis, 1986a]. Lewis describes this notion informally: "The causal history of a particular event includes the event itself...Further it is closed under causal dependence: anything on which an event in the history depends is itself an event in the history...Finally, a causal history includes no more than it must to meet these conditions." His definition treats the notion of causal dependence as a black box. In contrast, we provide a formal definition of causal history with a specific notion of causal dependence. We prove the closure property that Lewis demands (Theorem 11). Indeed this was a criteria that we evaluated our definition against. In addition, we establish the complexity of the associated decision problem.

NESS test of causation. Our definition of causal slice draws inspiration from the NESS test of causation (necessary elements of a sufficient set), proposed by Hart and Honoré and examined critically by Wright in the context of tort law [Hart and Honoré, 1985; Wright, 1985]. The NESS test determines which primitive events (not causal paths or causal slices) cause an outcome. It says that all the *necessary* elements of a set of events that is *sufficient* to cause an outcome are causes of the outcome. An equivalent way to formulate the test is to say that all events in a minimally sufficient set for an outcome are causes of the outcome. Our definition builds on this idea, but generalizes it to define entire causal slices (not just causal events), which is our goal. Specifically, our definition amounts to applying the NESS test to determine the outcome's immediate causes, followed by recursive applications of the NESS test to find the causes of those immediate causes and so on. The resulting subgraph of the causal network is our causal slice.

Actual cause definitions. As mentioned in the introduction, actual cause definitions are motivated by a different question: Identifying individual events that contributed to causing the explanandum event [Pearl, 2000; Hitchcock, 2001; Halpern and Pearl, 2005; Hall, 2007; Halpern, 2015]. In contrast, we seek to define the causal history to fully explain an event. The additional information available in a causal history is useful to provide a range of explanations that go significantly beyond what actual causes provide. This includes distinguishing joint and independent causal paths, mixed causes, and multiple causal slices with the same events – as illustrated by the examples in the last two sections of the paper. This comment is not meant as a criticism of the actual cause literature but as an elaboration on how the differences in goals leads to differences in explanatory power. Indeed Lewis starts his paper on causal explanations with exactly this point [Lewis, 1986a].

The actual cause literature offers arguments that actual causes are not transitive [Hitchcock, 2001] — a criticism of Lewis' theory of actual causation [Lewis, 1973] (not to be confused with his work on causal explanation [Lewis, 1986a]). A reader might wonder if the intransitivity of actual cause is inconsistent with the transitive closure property of our causal slices. We remark that there is no inconsistency here: We are forcing a causal slice to be transitively closed, but retaining the separation of direct from indirect causes by keeping the entire structure of causal paths.

5 Conclusion and future work

We present a new take on the old problem of tracing actual causes articulated by David Lewis in his work on causal explanations [Lewis, 1986a]. We address the problem by defining the *causal history* of the explanandum event. Informally, the causal history traces the immediate causes of the explanandum event, the causes of those causes, and so on to produce a full explanation. The central technical contributions of this paper are (a) a formal definition of causal history in the structural equations model; and (b) a theorem that establishes the complexity of the decision problem for causal histories. In addition, we show that our definition cleanly explains a battery of problematic examples from the actual cause literature.

In future work, we will dig deeper into causal explanations and explore the space of useful causal explanations that can be derived from (parts of) causal histories. We will also explore applications of these methods to debugging computer systems.

References

[Hall, 2004] Ned Hall. Two concepts of causation. In John Collins and Ned Hall and Laurie Paul, editor, *Causation and Counterfactuals*, pages 225–276. The MIT Press, 2004.

[Hall, 2007] N. Hall. Structural equations and causation. *Philosophical Studies*, 132(1):109–136, 2007.

[Halpern and Pearl, 2005] Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach: Part I: Causes. *British Journal for the Philosophy of Science*, 56(4):843–887, 2005.

[Halpern, 2008] Joseph Y Halpern. Defaults and Normality in Causal Structures. *Artificial Intelligence*, 30:198–208, 2008.

[Halpern, 2015] Joseph Halpern. A modification of the halpern-pearl definition of causality. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[Hart and Honoré, 1985] H. L. A. Hart and Tony Honoré. *Causation in the Law (Second Edition)*. Oxford University Press, Oxford, UK, 1985.

[Hitchcock, 2001] Christopher Hitchcock. The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy*, 98(6):273–299, 2001.

[Hitchcock, 2007] Christopher Hitchcock. Prevention, preemption, and the principle of sufficient reason. *Philosophical Review*, 116(4):495–532, 2007.

[Hume, 1748] D. Hume. An Enquiry Concerning Human Understanding. *Reprinted Open Court Press, LaSalle, IL, 1958*, 1748.

[Lewis, 1973] David Lewis. Causation. *Journal of Philosophy*, 70(17):556–567, 1973.

[Lewis, 1986a] David Lewis. Causal explanation. In David Lewis, editor, *Philosophical Papers Vol. Ii*, pages 214–240. Oxford University Press, 1986.

[Lewis, 1986b] David Lewis. Events. In *Philosophical Papers Vol. II*, volume 2, pages 241–269. OUP, 1986.

[McDermott, 1995] Michael McDermott. Redundant causation. *British Journal for the Philosophy of Science*, 46(4):523–544, 1995.

[Moore, 2009] Michael S. Moore. *Causation and Responsibility: An Essay in Law, Morals and Metaphysics*. Oxford University Press, 2009.

[Pearl, 2000] Judea Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, New York, NY, USA, 2000.

[Prosser, 1937] William L Prosser. Joint torts and several liability. *California Law Review*, pages 413–443, 1937.

[Schaffer, 2000] Jonathan Schaffer. Trumping preemption. *Journal of Philosophy*, 97(4):165–181, 2000.

[Spirtes *et al.*, 2000] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*, volume 81. MIT press, 2000.

[Wright, 1985] R.W. Wright. Causation in tort law. *California Law Review 73*, pages 1735–1828, 1985.

[Wright, 1987] Richard W Wright. Allocating liability among multiple responsible causes: A principled defense of joint and several liability for actual harm and risk exposure. *UC Davis L. Rev.*, 21:1141, 1987.