AN ALGORITHMIC ENQUIRY CONCERNING CAUSALITY

by

SAMANTHA KLEINBERG

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Computer Science
New York University
May, 2010

Bhubaneswar Mishra

Have patience with everything unresolved in your heart and try to love *the questions themselves* as if they were locked rooms or books written in a very foreign language. Don't search for the answers, which could not be given to you now, because you would not be able to live them. And the point is, to live everything. *Live* the questions now. Perhaps then, someday far in the future, you will gradually, without even noticing it, live your way into the answer.

—Rainer Maria Rilke, translated by Stephen Mitchell

For Grandma Dorothy

ACKNOWLEDGEMENTS

I must begin by thanking my advisor, Bud Mishra, who has consistently encouraged me to question the status quo and follow my interdisciplinary intuitions. Bud trusted that I could teach myself logic and model checking, and that going all the way back to Hume in my causality reading was a necessary (though not sufficient) condition for understanding. I found myself in Bud's lab as an undergraduate student primarily through luck, but I cannot imagine a better place to have grown as a scholar or another advisor who would have allowed me the freedom and independence I needed along with subtle guidance during periods of difficulty.

I have been fortunate to have a diverse committee who contributed deeply to this thesis in quite varied ways. Ernest Davis has read a multitude of drafts, always with an unbelievable attention to detail. I am grateful for his blunt honesty about my prose, and any unclear sections that remain are due only to me not taking his advice. I thank Petter Kolm for sharing his expertise in finance and working with me over the last few years as I honed my approach. Rohit Parikh has been a thorough reader of this and other works. His depth and breadth of experience has made him invaluable for reminding me of the broader context of my work as I became hyperfocused. Michael Strevens saved me time and again from muddling metaphysics and epistemology and provided critical support and guidance as I delved deeper into philosophy and tried to reconcile it with computer science. I also thank Amnir Pnueli, whose support during

the early stages of this work gave me the confidence to proceed. More fundamentally, without Amir's contributions to temporal logic, this thesis would not have been possible.

I thank Craig Benham, who gave me my first taste of bioinformatics, and Kevin Kelliher, who very patiently introduced me to programming. My deepest thanks to Sebastian Stoenescu who guided me toward the high school internship that fundamentally changed my path. I have always wondered if I would have found this field that I am so passionate about without his influence, and am relieved that I do not have to find out. I was lucky to have someone who knew my interests and abilities better than I did at such a critical moment in my life.

I thank all of the Bioinformatics Group members past and present. In particular I thank Marco Antoniotti, who introduced me to common lisp and changed the way I program.

I thank James, who has read every draft of everything I've written in the last six years.

Finally, I want to thank my mother, who instilled in me a passion for learning and a love of writing. I thank my father for continually reminding me of the importance of art and beauty, and contributing to better living through design. I thank Mark for his unwavering support and encouragement over the last twenty years.

ABSTRACT

In many domains we face the problem of determining the underlying causal structure from time-course observations of a system. Whether we have neural spike trains in neuroscience, gene expression levels in systems biology, or stock price movements in finance, we want to determine why these systems behave the way they do. For this purpose we must assess which of the myriad possible causes are significant while aiming to do so with a feasible computational complexity. At the same time, there has been much work in philosophy on what it means for something to be a cause, but comparatively little attention has been paid to how we can identify these causes. Algorithmic approaches from computer science have provided the first steps in this direction, but fail to capture the complex, probabilistic and temporal nature of the relationships we seek.

This dissertation presents a novel approach to the inference of general (type-level) and singular (token-level) causes. The approach combines philosophical notions of causality with algorithmic approaches built on model checking and statistical techniques for false discovery rate control. By using a probabilistic computation tree logic to describe both cause and effect, we allow for complex relationships and explicit description of the time between cause and effect as well as the probability of this relationship being observed (e.g. "a and b until c, causing d in 10–20 time units"). Using these causal formulas and their associated probabilities, we develop a novel measure for the significance of a cause for its effect, thus allowing

discovery of those that are statistically interesting, determined using the concepts of multiple hypothesis testing and false discovery control. We develop algorithms for testing these properties in time-series observations and for relating the inferred general relationships to token-level events (described as sequences of observations). Finally, we illustrate these ideas with example data from both neuroscience and finance, comparing the results to those found with other inference methods. The results demonstrate that our approach achieves superior control of false discovery rates, due to its ability to appropriately represent and infer temporal information.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

1

INTRODUCTION

> If a man will begin with certainties he shall end in doubts, but if he will be content to begin with doubts he shall end in certainties.
>
> — Francis Bacon

The study of "why" is integral to every facet of science, research and even daily life. When we search for factors that are related to lung cancer or assess fault for a car accident, we seek to predict and explain phenomena or find out who or what is responsible for something. At its most basic, a cause is an answer to a "why" question. Causes tell us not just that two phenomena are related, but *how* they are related: if we ask "why x?" can we respond "because y"? what does knowing y *mean* for knowing x? how does y *explain* x? While correlations can potentially be useful for prediction, they do not have the explanatory power we desire. Knowing, say, that secondhand smoke is correlated with lung cancer doesn't allow us to explain an instance of lung cancer as being due to secondhand smoke nor does it allow us to say that we can prevent lung cancer by avoiding secondhand smoke. If instead we know that secondhand smoke *causes* lung cancer we may be able to make both claims.

Despite the need for methods for understanding causality, the question of what makes something a cause (let alone how to find one) has plagued philosophers and scientists since at least the time of Aristotle. At the same time, people manage to make causal inferences and judgments in daily life: children learn that touching a hot pot leads to a painful burn and juries weigh evidence and sequences of events to determine guilt or innocence. One of the primary difficulties in the philosophical search for a theory of causality has been the desire for a single theory that accounts for all types and instances of causality. Thus there are arguments against any theory that does not produce expected results in at least one case, leading to a multitude of competing theories, none of which provides the desired perfect approach. At the other end of the spectrum, computer scientists have honed in on one main framework, with little consideration of whether this approach is truly the correct one for all cases. I argue that it is futile to insist on a single unified theory that can handle all counterexamples and all applications. Instead I focus on one particular type of problem and aim to develop the best tool for this job. I will not attempt to capture all intuitions about causality or handle all conceivable problems.

I argue that one of the most critical pieces of information about causality – the time it takes for the cause to produce its effect – has been ignored. If we do not know when the effect will occur, we have little hope of being able to act on this information. We need to know the timing of biological processes in order to disrupt them to prevent disease. We need to know when to take a position in the market if we want to trade based on causes affecting a stock's price. We need to know a patient's sequence of symptoms and related events to determine her diagnosis. Further, policy

and personal decisions may vary enormously with changes in the length of time between cause and effect. The warning that "smoking causes lung cancer" tells us nothing about how long it will take for lung cancer to develop. But while a deterministic relationship that will take 80 years may not change a person's behavior, a relationship with a somewhat lower probability at a time scale of only 10–20 years might be significantly more alarming. In order to clarify such claims, we need to understand both what causality is, and how to represent the finer details of the relationship between cause and effect.

The primary goal of this work is the inference of causal relationships from temporal data. I seek a description of causality that is philosophically sound, a method of inference that is logically rigorous (and allows an automated algorithmic approach), and a statistically thorough procedure of determining which of the causes inferred are genuine. Further, we desire to use these methods in a variety of domains – such as biology, politics, and finance – so the definitions should be applicable in all of those areas and the methods should work with the variety of data currently available. As our primary aim is to infer causal relationships from data, we need to capture the probabilistic nature of the data, and be able to reason about potentially complex relationships as well as the time between cause and effect. It will be argued that the previous methods for causal inference (primarily resulting in the creation of networks or graphs) do not achieve these goals. Instead I present an alternative approach based on the idea of causal relationships as logical statements, which borrows from philosophical notions of probabilistic causality, work in temporal logic and statistical methods for false discovery rate (fdr) control.

In this approach, cause and effect and the conditions for causality are described in terms of logical formulas. By doing this we can capture relationships such as: "smoking causes lung cancer with probability 0.6 in between 10 and 20 years." I show that while we focus only on the case of temporal data, the working definitions allow us to correctly handle many of the difficult cases commonly posed to theories of causality. Further, the use of temporal logic, with clearly defined syntax and semantics, allows us to automatically test any relationship that can be described in the logic.

I will also relate this framework to singular, or token, causality. This problem has great practical importance as a significant use of token causality is in diagnosis of patients, where one wants to find the cause of someone's symptoms but many diseases may share similar symptoms. As electronic health records become more prevalent, it is increasingly desirable to be able to scan these records automatically upon check-in at a doctor's office or hospital. Then one can assess a patient's history and symptoms in order to identify possible causes that will require immediate attention or hospitalization. We can use similar methods to predict events. In the case of patient records, this allows for prognosis determination, given the patient's history and known causes of various illnesses.

1.1 OVERVIEW OF THESIS

This thesis is intended to be accessible to computer scientists and philosophers, as well as interested biologists and researchers in finance and other areas. For that reason, the work is mostly self-contained, and assumes no background in statistics, logic, or philosophy. Included as well is

a glossary containing most of the technical terms used throughout the thesis.

In Chapter 2 I begin with a short introduction to philosophical theories of causality, beginning with historical foundations and then critically discussing probabilistic and counterfactual theories of causality. This chapter introduces the problem of defining and recognizing causal relationships as well as the traditional approaches to this problem. While these theories are not immediately applicable to experimental problems, they provide a foundation on which later methods are based. Further, before we can discuss how to find causes, we must understand what it means for something to be a cause.

In Chapter 3 I review the state of the art in causal inference. I discuss graphical model approaches (which are based on the philosophical literature) and their extensions, as well as commonly used approaches with no philosophical basis, such as Granger causality, a method from finance. I then discuss various approaches to causal reasoning in AI and logic. Most of these do not attempt to relate to philosophical theories about causality, but rather aim to find the effects of actions on a system where a model and causal theories (defining what happens when various actions are taken) are assumed as given. Finally, I discuss experimental approaches in areas of interest (gene expression, neural spike trains, and financial time series).

In the remaining chapters, we turn our attention to formulating a new approach to causal inference and evaluating this approach on various datasets. In Chapter 4 I begin by defining what will be meant by "causes" and what types of causes we will attempt to identify. I introduce a new measure for the significance of causes that is computationally feasible, but

grounded in the philosophical theories discussed in Chapter 2. I relate these definitions to probabilistic temporal logic formulas and discuss how the definitions deal with common counterexamples posed to theories of causality. I show that by using a well-chosen logic we can address the previously ignored problem of representing detailed causal relationships that contain important timing information, while also allowing for automated and computationally feasible testing of these causes in data. Readers unfamiliar with temporal logics should refer to Appendix A, which provides an introduction to logic and model checking.

In Chapter 5 I develop the algorithms needed for testing causal relationships in data. I formalize the methods for testing temporal logic formulas in traces, discussing what it means for a formula to be satisfied relative to such a sequence of observations. I augment PCTL to suit our needs, allowing specification of formulas true within a window of time (See Appendix C.2 for related proofs). I then discuss the problems associated with determining the significance of causes (See Appendix B for an introduction to multiple hypothesis testing and false discovery control). First I describe the computation of significance scores in depth, then discuss how to determine an appropriate threshold for the level at which something is statistically significant. I show that since we are primarily interested in applications that involve a large number of relationships being tested simultaneously, the problem can be treated as one of multiple hypothesis testing and false discovery control, where we infer the null hypothesis from the data. I apply well-established methods from statistics for this purpose. In this chapter I show the correctness of all methods and analyze their computational complexity.

In Chapter 6 I discuss the problem of token causality in depth. Here we aim to find not general relationships (such as that between smoking and lung cancer) but want to determine the cause on a particular occasion (did Jane's smoking cause her lung cancer?). I begin by discussing why we need a separate treatment of this type of causality, and then review one philosophical theory that will be repurposed. I show how, building on this theory, we can use prior type-level inferences (made using the method developed in the previous chapters) to find the cause of an effect on a particular occasion. We will then examine a number of difficult cases found in the philosophical literature and find that the approach developed can handle these in a manner consistent with intuition about the problems.

Finally, in Chapter 7 I apply the methods developed to data from biological and financial applications. I compare the approach advanced in this work to others (including Granger causality and graphical model-based methods), and demonstrate that I achieve an extremely low false discovery rate, outperforming all other methods by at least one order of magnitude and, in some cases, two. I also show that this performance does not come at the expense of an increase in the false negative rate, as I again have the lowest values of this measure.

2

A BRIEF REVIEW OF CAUSALITY

2.1 PHILOSOPHICAL FOUNDATIONS OF CAUSALITY

The basis for causal inference and the meaning behind causality begins in the philosophical literature. Here, we review the development of "probabilistic causality", particularly in terms of distinguishing between cause and effect. While our study focuses on inferring these relationships, we must have a foundation on which to base these inferences and a vocabulary with which to describe them.

The first modern attempt to frame the question of "why?" came from David Hume in the 18^{th} century. Hume defined a causal relationship between C and E to mean that C is a cause of E if and only if every event of type C is followed by an event of type E. These relations are to be inferred from observations and are subjective due to belief and perception. That is, based on experience, we reason about what will happen, have expectations based on our perceptions, and may establish whether our beliefs are true or false through experimentation and observation. For example, when we hear a noise outside in the morning, we may believe that a garbage truck is outside. Since in the past we heard this noise and saw a garbage truck outside the window, we expect to go to the window and see the same thing this time. This belief may turn out to be false, as perhaps today there is instead a street sweeper causing the noise. The

important point here is that without empirical evidence, we could not have made *any* predictions about the cause of the noise.

Hume examined causality in terms of (1) what is meant when we use the term and (2) what is needed to infer such a relation from empirical evidence. First, addressing the concept of causality, Hume defined three essential relations: contiguity, temporal priority, and necessary connection [57]. The contiguity condition asserts that a cause and its effect must be nearby in time and space. While it may seem this condition does not always hold true, Hume states that any relationships between distant causes and effects can be found to be "linked by a chain of causes, which are contiguous among themselves."[1] The second quality, temporal priority, means that a cause must precede its effect. While Hume traces the chain of events that would occur if we allow cause and effect to be co-temporary, ending with the "utter annihilation of time", it suffices to say that if we do allow cause and effect to be co-temporary we could not distinguish the cause from the effect and would in fact only be able to determine a correlation between the pair. Finally, necessary connection is the defining feature that allows us to make the distinction between causal and non-causal relationships. Here it is stipulated that both the cause and effect must occur. That is, the cause always produces the effect, and the effect is not produced without the cause.

Hume then empirically defines a cause as[2]:

Definition 2.1.1. An object precedent and contiguous to another, and where all the objects resembling the former are placed in a like relation of priority and contiguity to those objects that resemble the latter.

1 [57], 75
2 [57], 172

Necessary connection is replaced here by *constant conjunction*, whereby we may observe two events as being conjoined, but this does not mean that they are necessarily so and nor do we have any basis for being able to make such a statement. One common counterexample to this theory of causality is that "day causes night" satisfies all three criteria, though we would not call day a cause of night. Cases may be made against each of the three criteria; however they represent the first step toward a theory of causality that may be verified through empirical data. The main effect of Hume's work, as stated by Russell, was: "when I assert "Every event of class A causes an event of class B," do I mean merely, "Every event of class A is followed by an event of class B," or do I mean something more? Before Hume, the latter view was always taken; since Hume, most empiricists have taken the former." [3]

2.2 MODERN PHILOSOPHICAL APPROACHES TO CAUSALITY

2.2.1 *Regularity*

Refining Hume's work, in 1974 John Leslie Mackie formalized the ideas of necessity and sufficiency for causes. Here, an event C is a *necessary condition* of an event E if whenever an event of type E occurs, an event of type C also occurs, and C is a *sufficient condition* of E if whenever an event of type C occurs an event of type E also occurs. Thus Mackie states that a cause is an INUS condition: "an *insufficient* but *non-redundant* part of an *unnecessary* but *sufficient* condition" [83]. That is, there are some sets of

3 [111], 454

conditions that result in the effect, E, and the cause, C, is a necessary part of one of those sets.

Definition 2.2.1. $A \wedge B \wedge C$ is a *minimal sufficient condition* for P if no conjunct is redundant (i.e. no part, such as $A \wedge B$, is itself sufficient for P), and $A \wedge B \wedge C$ is sufficient for P.

Definition 2.2.2. C is an *INUS condition* of E iff for some X and for some Y $(C \wedge X) \vee Y$ is a necessary and sufficient condition of E, but C is not a sufficient condition of E and X is not a sufficient condition of E [82, 63].

That is,

1. $C \wedge X$ is sufficient for E,
2. $C \wedge X$ is not necessary since Y could also cause E,
3. C alone is insufficient for E,
4. C is a non-redundant part of $C \wedge X$.

For example a lit match (C) may be the cause of a house fire. There are, however, many other situations in which a match is lit and does not cause a fire, as well as other situations ($\neg X$) in which a fire occurs without a lit match (Y). In the case of the match causing the fire, there is some set of circumstances (X), each one necessary, which together are sufficient for the fire to occur.

Mackie analyzes an event C as a cause of an event E on a particular occasion (what is also referred to as token, or singular, causality) thusly:

1. C is at least an INUS condition of E,
2. C was present on the occasion in question,

3. The components of X, if there are any, were present on the occasion in question,

4. Every disjunct in Y not containing C as a conjunct was absent on the occasion in question.

Definition 2.2.3. C is *at least an INUS condition* of E iff either C is an INUS condition for E, or C is a minimum sufficient condition for E, or C is a necessary and sufficient condition for E, or C is part of some necessary and sufficient condition for E.

Using the house fire example, a lit match was the cause of a specific fire if it was present, and there was oxygen, flammable material and the other conditions needed for a lit match to create a fire, and there was no unattended cooking, faulty electrical wiring, or other factors that cause fires in the absence of lit matches. That is, the third and fourth conditions above ensure that the other factors necessary for C to cause E are present, while avoiding the problem of overdetermination. For example, if there was a lit match and the house was struck by lightning, we would violate the fourth condition and in fact neither would be deemed the cause of the fire.

2.2.2 *Counterfactuals*

Counterfactuals provide another approach to causality by saying that had the cause not taken place, the effect would not have happened either. This theory is in fact the second part of Hume's definition, where a cause is "*an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second.* Or in other words *where, if the*

first object had not been, the second never had existed" [58]. Though they are supposed restatements of the same theory, the first part, known as the "regularity definition" of causality is quite different from the second part, the "counterfactual definition." If we were to use only the first part of the definition, we would again have the problem of day being a cause of night, as one regularly follows the other. However, the counterfactual definition removes the causal relationship, as had day not been, night would still exist (consider the case of Polar night at the article circle where the sun does not rise at all).

David Lewis developed the primary counterfactual theory of causality, discussing how we can use these conditional statements [24, 121] to distinguish genuine causes from effects and other factors [74]. In this work, Lewis limits causes and effects to events, and looks only at the analysis of causes in terms of particular cases (what is termed token, or singular, causality). He begins by introducing the notion of *possible worlds*, and *comparative similarity* between possible worlds, which may be thought of as maximally consistent sets of propositions true in those worlds. Then, one world is *closer to actuality* than another is if it resembles the actual world more than the other world does. Lewis introduces two constraints on this relation, namely, (1) It involves a weak ordering of the worlds, so any two worlds may be compared, but they may be equal in similarity to the actual world; (2) The actual world is closest to actuality, as it resembles itself more than any other world resembles it [74].

Then, we can take the *counterfactual* of two propositions, A and C. This assertion is represented by $A\Box \rightarrow C$ and means that if A were true, C would be true. Then, the truth condition for this statement is: $A\Box \rightarrow C$ is true (in the actual world w) iff (1) there are no possible A-worlds or (2)

some A-world where C holds is closer (to w) than any A-world where C does not hold. That is, in the non-vacuous case (2), the counterfactual is true iff "it takes less of a departure from actuality to make the consequent true along with the antecedent than it does to make the antecedent true without the consequent" [74].

We switch now to look at events, rather than propositions, and can define causal dependence between them. The dependence defined here means that *whether* e occurs depends on *whether or not* c occurs. This is represented by two counterfactuals, $c\Box \rightarrow e$ and $\neg c\Box \rightarrow \neg e$. After describing causal dependencies, we would like to now describe causation among events. First, take a series of events c_1, c_2. . .c_n and one effect e. If each c_i, with $i > 1$, occurs only with c_{i-1} (each c_i depends causally on the previous c_{i-1}), with e occurring only when c_n occurs, then we say that c_1 is a cause of e, whether or not e might still have taken place without c_1. The causal relationship here is transitive, though the dependence relationship need not be (e need not be dependent on c_1). We define that c is a cause of e if there is some causal chain (i.e. chain of causal dependencies) connecting them.

The main problems facing this approach are transitivity and overdetermination (redundant causation), or preemption. In the first case, we can find situations such that some event a would prevent some event c but in the actual events, a causes another event b, which in turn causes c to occur. Thus the counterfactual account leads to events counterintuitively being labeled causal. McDermott gives one such counterexample [87]. Suppose I give Jones a chest massage (C), without which he would have died. Then, he recovers and flies to New York (F), where he eventually has a violent death (D). Here, C was a cause of F, as without the massage

he would not have been well enough to travel, and F is a cause of D, but C did not cause D. That is, whether or not C occurred, Jones still would have died, but there is a causal chain between C and D.

The second problem for the counterfactual theory of causality is overdetermination, or redundant causation. Consider now that there are two potential causes for an effect (both present) and the effect would have been the result of either, so that the effect depends on neither and the system is overdetermined. This redundant causation may be either symmetrical (each potential cause could equally well be called the cause of the effect, there is nothing to distinguish which was the actual cause) or asymmetrical (there was one cause which *preempted* the other). In the asymmetrical case, if we say c_1 was the preempting cause, c_2 the preempted and e the effect, then had c_1 not occurred, c_2 would still have caused e, and thus c_1 is not the cause of e despite its causing e. This is generally the equivalent of having two causal chains to e, one from c_1 and one from c_2 where something *cuts* the causal chain from c_2 to e, preempting it before it reaches e.[4]

This inconsistency with the counterfactual approach was revisited by Lewis in a more recent paper, where dependencies are not based solely on *whether* events occur, but rather *how*, *when* and *whether* one event occurs depends on *how*, *when* and *whether* the other event occurs [76]. Earlier, Lewis defined that an event is *fragile* "if, or to the extent that, it could not have occurred at a different time, or in a different manner" [75]. Now, we define *alterations* of events (perturbations in time or manner):

4 Lewis later clarifies that there are other cases where both causes occur but one "trumps" the other, preempting it as a cause. However, the case of cutting is more common [76].

Definition 2.2.4. Let an alteration of event E be either a very fragile version of E or else a very fragile alternative event that is similar to E, but numerically different from E.

Then, with distinct actual events C and E, C influences E iff there exist substantial ranges of not-too-distant alterations of C and E (C_1, C_2. . .C_n, E_1, E_2. . .E_n), where at least some differ, and such that for each C_i, if C_i had occurred, E_i would have occurred.[5] Finally, C causes E if there is a chain of influence from C to E (though there is no transitivity of influence and we do not say C influences E, despite the fact that it causes E). Going back to the case of preemption, we can see the advantage of this theory in terms of finding the actual cause of an event. Here, if we alter c_1 while holding c_2 fixed, and then alter c_2 while holding c_1 fixed, we find that in the first case, *e* is altered while in the second case *e* is the same. Since altering c_2 did not influence *e* we find that c_1 was the cause of *e* as its alteration did influence *e*. Note that we may find cases in which an alteration of c_2 would influence *e*, but according to Lewis these may be due to alterations that are too distant. There is also the problem of spurious causation, which Lewis acknowledges is present in both the new and old theories. Here, any event that has a small influence on the time and manner of the effect can be said to be a cause. In this theory, there is no discussion of the *degree* to which the cause influenced the effect.

5 [75], 190

2.3 PROBABILISTIC CAUSALITY

In the prior methods described, causality was generally a deterministic relation. That is, a cause produced its effect without fail. This deterministic view of the world is limiting in a few ways. First, it is not possible to *infer* such deterministic relationships with certainty. There is no amount of events that we could observe that would allow us to pronounce with certainty that one thing causes another with probability one. For example, we may note that in all the years we have known each other, that every time you call me, my phone has rung (let us assume I do not have call waiting, no one else has called, and I do not make calls on my own). We cannot be sure that my phone will *always* ring because you have called, that is, that your call is *the* cause of my phone ringing. What we can infer is that your calling makes it very likely that my phone will ring. In fact, in this case, we can predict with a high probability that when you call, my phone will ring.[6] But, we cannot say this will always, without fail, be the case. Here we must distinguish between the probability due to the actual relationship and the probability due to our lack of knowledge. Just as it is possible for some relationships to be at their core deterministic, it is possible that others are probabilistic. That is, even if we had complete knowledge of the world and all relevant information, we would still find a probabilistic relationship between cause and effect. The other probability is due to our normally incomplete information about the system – but this has no bearing on what the underlying relationship *actually* is. When we

6 Note that there may be other cases where we observe a sequence, such as a fair roulette wheel coming up red 20 times in a row or a fair coin flipped 20 times and coming up heads on each, where these are not indicative of the underlying probabilities. However, note that as the sequence of observations gets longer we will come closer to observing the true probabilities of the system.

observe a causal relationship, we are generally observing a combination of these probabilities.

In order to infer causal relationships, we stipulate that (positive) causes raise the probabilities of their effects and then set about finding which of the causes inferred are the most explanatory. When we say "a cause raises the probability of its effect," we mean that *given* that the cause has occurred, we have a better chance of seeing the effect. That is, with cause C, effect E, and the *conditional probability* $P(E|C) = P(E \wedge C)/P(C)$, we can say that C is a cause of E if:

$$P(E|C) > P(E|\neg C). \tag{2.1}$$

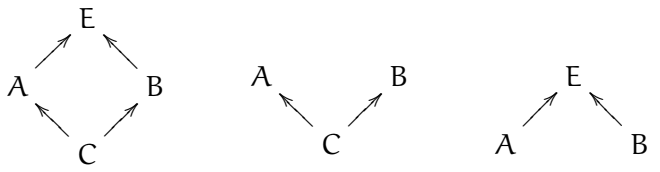
2.3.1 *Screening Off*

One problem for probabilistic theories of causality is that there may be cases where two events are the result of an earlier common cause. In one commonly used example, we may frequently see yellow stained fingers and lung cancer together. We cannot say that yellow stained fingers cause lung cancer, or that lung cancer causes yellow stained fingers. Using more information, we can find an earlier common cause of both: smoking. Here, smoking "screens-off" lung cancer from yellow stained fingers. That is, when we hold fixed that someone is a smoker, the relationship between stained fingers and lung cancer disappears. The idea of earlier "screening off" causes was introduced by Reichenbach [108].

Reichenbach first describes the asymmetry of cause and effect. That is, given that the cause produces the effect, we do not say that the effect produces the cause. However, the probability relations characterizing the causal relationship are symmetric. If C and E are causally related, then the probability of C increases the probability of E and vice versa. Thus Reichenbach attempts to characterize the direction of the temporal relationship with the common cause principle. The *common cause principle* states that, with simultaneous events A and B (where $P(A \wedge B) > P(A)P(B)$) if there is an earlier common cause C of both, C is said to screen off A and B from one another iff:

1. $P(A \wedge B|C) = P(A|C)P(B|C)$,
2. $P(A \wedge B|\neg C) = P(A|\neg C)P(B|\neg C)$,
3. $P(A|C) > P(A|\neg C)$, and
4. $P(B|C) > P(B|\neg C)$.

This says that C raises the probability of A and of B and that if we know that C or that $\neg C$, there is no longer a correlation between A and B. This corresponds to the fork shown in figure 2.1b. The idea here is that if we have such a fork, with some particular a, b, and c, and they satisfy the probability relations given above, it means that c is the common cause of a and b and thus it is also earlier than a and b. Note that the fork open to the past, shown in figure 2.1c., would not account for this relationship. For example, if two lamps burn out simultaneously and the room goes dark, the dark room does not account for the lamps burning out. Rather,



(a) Double fork, common cause and common effect

(b) Open to the future, common cause

(c) Open to the past, common effect

Figure 2.1.: Forks as described by Reichenbach [108].

some earlier common cause such as a burned fuse or problem with a common power supply would account for this.[7]

Definition 2.3.1. Two events, A and B are *causally connected* if either A is a cause of B, B is a cause of A, or there exists an event C such that C is a common cause of both A and B.

Then, an event C is *causally relevant* to another event E iff:

1. C is earlier than E
2. $P(E|C) > P(E)$, and
3. There does not exist a set of events S, earlier than or simultaneous with C, such that S screens off C from E.

That is, there is no other cause screening off C from E and C raises the probability of E.[8]

One difficulty for this as well as other probabilistic definitions of causality is posed by Simpson's Paradox [117]. That is, if C is a cause of E in the general population, we can reverse this relationship, by finding

7 [108], 157.
8 [108], 204.

sub-populations such that in every such sub-population C is a negative cause of E. This situation arises because C is correlated with another factor that prevents E. One common example is based on the case of sex bias in graduate admissions at Berkeley [6]. In that study they found that while in general (looking at the school as a whole), men had a higher rate of admission to the university, within each department there was no correlation between sex and admission rate. Thus, being female did not cause applicants to be rejected, but rather women likely apply to more competitive departments that had lower admissions rates, leading to their overall lower rate of acceptance.

Another common example is given by Brian Skyrms [118]. In general, smoking is a positive cause of lung cancer. Consider now what happens if due to air pollution (which we assume here can cause lung cancer), city-dwellers tend to stop smoking in order to not further jeopardize their lungs. Also suppose that due to cleaner air in the country, people there feel freer to smoke given the lack of air pollution harming their lungs. Then, smoking (C) is a positive cause of lung cancer (E), living in the country (V) is a positive cause of smoking, and living in the country is a negative cause of lung cancer. Then, because V is a positive cause of C and a negative cause of E, depending on the ratio of smokers to non-smokers and the city air quality, since C is correlated with an actual negative cause of E (V), it can be negatively correlated with E despite the fact that it is a positive cause of it (see figure 2.2.). As in the previous case, where being female was associated with a higher probability of admission in each individual department, but a lower probability overall, we find that smoking seems to lower the probability of lung cancer when looking at smokers versus non-smokers, even though it is a positive cause

Resolutions for and more in depth discussion of this issue can be found in [25, 97].

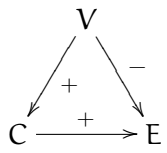


Figure 2.2.: Illustration of Simpson's paradox example.

of lung cancer in general. That is, smoking was correlated with living in the country (and exposure to fresh air), which made it seem to be a negative cause of lung cancer, while non-smoking was associated with living in the city (and exposure to pollution). Similar examples can be constructed where a drug seems to be ineffective when looking at just men and just women, but is effective for people as a whole.

2.3.2 *Suppes*

The primary explanation and development of the theory of probabilistic causation comes from Patrick Suppes [124]. In this work, Suppes defines several types of causes. All of these, aside from negative causes, raise the probability of their effects and the direction of the causal relationship is characterized by temporal priority between cause and effect. Suppes defines probabilities and events in terms of sets, using the notation A_t and $B_{t'}$ to denote event of kind A occurring at time t, and B at time t'. Thus if we are interested in whether smoking causes lung cancer and X is the set of all events, then $C_{t'}$ is the subset of X consisting of all events involving smoking at any time t'(where smoking is followed at any later time t by cancer or no cancer), and E_t is the subset of X consisting of

all events where people have lung cancer at some time t (preceded by smoking or not smoking at any earlier time t'). The temporal subscripts refer only to the "later" and "earlier" conditions. In the computation of the conditional probability $P(E_t|C_{t'})$, $E_t \wedge C_{t'}$ is the intersection of these sets, consisting of the set of events where smoking is followed by lung cancer.[9] The probability of a set of events is the sum of the probabilities of the individual events comprising the set, as each is considered to be a mutually exclusive outcome.[10] Thus we should think of this as the sum over all ways C can precede E and interpret these temporal subscripts as being used only to describe the relationship between times t and t' (e.g. that one is strictly earlier than the other is), not as denoting actual times of occurrence.[11]

The first type of causes, *prima facie* causes, are the simplest causes described. These are *potential* genuine causes.

Definition 2.3.2. An event $B_{t'}$ is a *prima facie* cause of event A_t iff:

1. $t' < t$,

2. $P(B_{t'}) > 0$, and

9 Depending on how finely the events are specified (with this being up to the experimenter), and denoting lung cancer by L and smoking by S, the sets may be as follows. All events (X) could be $\{S_1L_2, S_1\overline{L_2}, \overline{S_1}L_2, \overline{S_1L_2}\}$, where the event space is all combinations of smoking/not-smoking preceding lung cancer/not lung cancer. Then, testing whether $S_{t'}$ causes L_t, the sets are: $C = C_{t'} = \{S_1L_2, S_1\overline{L_t}\}$, $E = E_t = \{S_1L_2, \overline{S_1}L_2\}$. Then, $C_{t'} \wedge E_t = \{S_1L_2\}$. Another X could specify events more finely, such that some event might denote whether lung cancer occurs after ten years but not after five years, and another lung cancer five years but not ten years after smoking. Then E would be comprised of all of these types of events such that lung cancer happens – regardless of when it happens. In another case, outcomes could be of the form $S_1L_1S_2L_2$. Then, C will contain all events that include S_1 or S_2 while E will contain all those with L_1 or L_2 but the intersection of E_t and $C_{t'}$ should only include those where S is prior to L, such as $S_1\overline{L_1}S_2L_2$.

10 For further details, see the Appendix of [124].

11 Suppes gives an example immediately after the introduction of this notation, of inoculation and incidence of cholera where A_t is the event of contracting cholera, while $B_{t'}$ is the event of being vaccinated against the disease. It is clear that the times refer only to the temporal order, and not to any particular times.([124],12)

3. $P(A_t|B_{t'}) > P(A_t)$.

We interpret this as being for all t and t' where $t' < t$. That is, the probability of A occurring at any time after B is greater than the probability of A occurring at any time. These do not refer to specific t's and t''s, but rather just describe the relationship between t and t'. In some cases, these causes may later turn out to be false. This discussion brings us next to the topic of spurious causes. We may believe that, even though something meets the criterion of being a *prima facie* cause, there is a better explanation for the effect. Then, we need a method to examine whether it is truly a cause. Suppes introduces two ways in which something may be a false, or spurious, cause. In each, the idea is that there is some event earlier than the possible cause that accounts equally well for the effect. That is, the spurious cause does not have any influence (positive or negative) on the effect.[12]

Definition 2.3.3. $B_{t'}$, a prima facie cause of A_t is a *spurious cause* in sense one iff $\exists t'' < t'$ and $C_{t''}$ such that:

1. $P(B_{t'} \wedge C_{t''}) > 0$,
2. $P(A_t|B_{t'} \wedge C_{t''}) = P(A_t|C_{t''})$, and
3. $P(A_t|B_{t'} \wedge C_{t''}) \geqslant P(A_t|B_{t'})$.

The idea here is that $B_{t'}$ is a possible cause of A_t, but there may be another, earlier, event that has more explanatory relevance to A_t. However, condition 2 of the definition above is very strong and perhaps counterintuitive. It means that there exists *an event* that completely eliminates the effectiveness of the cause for predicting the effect. One

12 [124], 24.

[124], 25. Note also that an event that is spurious in sense two is spurious in sense one, but the reverse is not true.

way of relaxing this condition is to look for kinds of events, where given the observation of one of these kinds of events or properties, knowing that the spurious cause occurs is uninformative with regards to whether the effect will occur. Here, a partition, π_t, may be of either the sample space or universe and consists of "pairwise disjoint, nonempty sets whose union is the whole space."

Definition 2.3.4. $B_{t'}$, a prima facie cause of A_t is a *spurious cause* in sense two iff there is a partition, $\pi_{t''}$ where $t'' < t'$ and for every $C_{t''}$ in $\pi_{t''}$

1. $P(B_{t'} \wedge C_{t''}) > 0$, and
2. $P(A_t|B_{t'} \wedge C_{t''}) = P(A_t|C_{t''})$.

One example of this, given by Otte [96], is the case of rain (A), a falling barometer (B) and a decrease in air pressure (C). B is a *prima facie* cause of A, as when it occurs the probability that A will follow is increased. However, $P(A|C \wedge B) = P(A|C)$, that is, given that the air pressure has decreased, the falling barometer does not provide any extra information about the rain. Also, $P(A|B \wedge C) \geqslant P(A|B)$, since the probability of rain given both a decrease in air pressure and a falling barometer is at least as great as the probability given just the falling barometer. So, B is a spurious cause of A in sense one.

We can also show that B is a spurious cause of A in sense two. Taking the partition π being {decreasing air pressure, non-decreasing air pressure} we then find that the probability of A given $(B \wedge C)$ is still equal to the probability of A given C and that the probability of A given $(B \wedge \neg C)$ is equal to the probability of A given $\neg C$. That is, if there is not decreasing air pressure, a falling barometer provides no information about whether it will rain. All causes that are spurious in sense two are also spurious

in sense one, but the reverse is not true in general. Note however that in the limit, where our barometer reports the air pressure perfectly, it will not be spurious in sense two, as $P(B \wedge \neg C) = 0$ (and we require that the probability must be greater than zero) – though it will still be spurious in sense one.

Then, Suppes defines *genuine causes* as non-spurious prima facie causes. These definitions allow us to begin to talk about what it means for something to probabilistically cause another thing. However, they can be rather limiting. Looking at the definition for spurious causes, the stipulation that $P(A_t|B_{t'} \wedge C_{t''}) = P(A_t|C_{t''})$ means that some causes may not be deemed spurious, despite meeting all the conditions, if there is a small difference in the probabilities on either side of this equality. To address this issue, Suppes introduced the notion of an ε-spurious cause.

Definition 2.3.5. An event $B_{t'}$ is an ε-spurious cause of event A_t iff $\exists t'' < t'$ and a partition $\pi_{t''}$ such that for every $C_{t''}$ of $\pi_{t''}$:

1. $t' < t$,
2. $P(B_{t'}) > 0$,
3. $P(A_t|B_{t'}) > P(A_t)$,
4. $P(B_t \wedge C_{t''}) > 0$, and
5. $|P(A_t|B_{t'} \wedge C_{t''}) - P(A_t|C_{t''})| < \varepsilon$.

This definition implies that a genuine cause that has a small effect on the probability of the event being caused will be ruled spurious. The partition, $\pi_{t''}$, separates off the past just prior to the possibly spurious cause, $B_{t'}$.

One issue that arises when using these definitions to determine the true cause of an effect is that we may find earlier and earlier causes that make the later ones spurious. That is, the cause may be quite removed from the effect in time (not to mention space). Suppes does not modify the theory to account for this, rather he introduces the idea of a *direct cause*. This is a concept very similar to screening off and spurious causes, except here we must consider whether there is some event coming between the cause and effect.[13]

Definition 2.3.6. An event $B_{t'}$ is a *direct cause* of A_t iff $B_{t'}$ is a prima facie cause of A_t and there is *no* t'' and *no* partition $\pi_{t''}$ such that for every $C_{t''}$ in $\pi_{t''}$:

1. $t' < t'' < t$,
2. $P(B_{t'} \wedge C_{t''}) > 0$, and
3. $P(A_t | C_{t''} \wedge B_{t'}) = P(A_t | C_{t''})$.

It is still possible that we will have a direct cause that is remote in space (and perhaps less possibly, in time), but we may nevertheless use this to rule out indirect remote causes. For example, we could have the case where someone's birth is the cause of them dying (it is an earlier cause screening off any later spurious causes, and it certainly holds with probability 1, though this would require constraining the time of death, since the probability someone will die eventually is 1). However, we can find later causes that are between birth and death that are the direct causes of death. Following the same rationale as for ε-spurious causes, we may define ε-direct causes.

13 Note, however, that there is no link between spurious and indirect causes.

Now we consider the possibility that two prima facie causes may aid one another in producing an effect. Suppes refers to such causes as *supplementary causes*, which are defined as follows:

Definition 2.3.7. Events $B_{t'}$ and $C_{t''}$ are supplementary causes of A_t iff:

1. $B_{t'}$ is a prima facie cause of A_t,
2. $C_{t''}$ is a prima facie cause of A_t,
3. $P(B_{t'} \wedge C_{t''}) > 0$, and
4. $P(A_t|B_{t'} \wedge C_{t''}) > max(P(A_t|B_{t'}), P(A_t|C_{t''}))$.

In this case, t'' may be equal to t'. Analogously to the previous cases, we may also define ε-supplementary causes. With this definition, we can identify combinations of causes that predict effects much better than each cause alone. Causes that result in their effects with probability one, i.e. the limit of prima facie causes where $P(A_t|B_{t'}) = 1$ are referred to as *sufficient* (or *determinate*) causes, using the same terminology as Mackie.

Looking at these definitions, we may identify some potential problems. First, because of the way "spurious" is defined, we run into difficulties with causal chains. For example, if we have a chain of causes that all produce their effects with probability one, every member will be spurious aside from the first member of the chain. Now, if we add another event between the last member of the chain and the final effect, which produces the effect with some probability, $0 < p < 1$, the last member will still be spurious, but it will now be the only direct cause of the effect. In many cases we may find earlier and earlier events to account for the effects, but it is perhaps unsatisfying to say that the only direct cause is spurious and the genuine cause is indirect. Similarly, in the case of the first chain

described, it is unclear whether the first or last link should be the genuine cause.

Secondly, in the case of overdetermination, or redundant causation, where there are two possible causes for an effect and both are present, all causes will turn out to be spurious aside from the earliest. For example, we have Bob and Susie (armed with rocks to throw at the bottle) and a glass bottle. Let us say that Bob is standing a little closer to the bottle than Susie is. So, Susie aims and throws her rock a little earlier than Bob does, but their rocks hit the glass simultaneously, breaking it shortly after impact. In this case, since Susie aimed her rock first, there is an earlier event than Bob aiming his rock and the rocks hitting the glass that accounts for the glass breaking. [14] Here we can see that this does not quite make sense, as Susie's throw did not set off a chain of events leading to the glass breaking any more than Bob's throw did (her throw had no effect on his). Why should one be the genuine cause of the glass breaking, simply because it was earlier? Now, we may also alter this example to look at the case of preemption (this is analogous to the "cutting" of causal chains described by Lewis). If Susie still throws first, but Bob's rock arrives first and thus breaks the glass before Susie's rock hits it, we would think that Bob's throw caused the glass to break. But, since Susie threw her rock first and would have caused the glass to break with probability 1, her throw still caused the glass to break despite the fact that it was already broken when her rock hit it.

To summarize Suppes' theory, a *prima facie* cause raises the probability of its effect and may be a genuine cause if it is not spurious. There are

14 Here we assume that if Susie aims and throws her rock it hits the glass with probability one and the glass breaks with probability one. The same assumption is made for Bob.

two ways in which something may be spurious, which correspond to looking for *particular earlier events* that explain the effect better than the spurious cause versus making a partition and looking at *kinds of events*. It remains to be determined whether the earliest cause should be termed the genuine cause.

2.3.3 *Eells*

Another advancement in probabilistic theories of causality came from Ellery Eells, who described theories of both type and token level causation [27]. *Type* causation refers to relationships between kinds (or *types*) of events, factors, or properties, while, in contrast, *token* causation refers to relationships between particular events that actually occur.

Type level causation

First, Eells states:

Definition 2.3.8. C is a *positive causal factor* for E iff for each i:

$$P(E|K_i \wedge C) > P(E|K_i \wedge \neg C), \tag{2.2}$$

where the K_i's are causal background contexts. By causal background contexts, we mean that if there are n factors other than C that are relevant to E there are 2^n ways of holding these fixed and we are interested in the subset of these that occur with nonzero probability in conjunction with C as well as ¬C [i.e. $P(C \wedge K_i) > 0$ and $P(\neg C \wedge K_i) > 0$] constitute a background context.[15] For example, if we have three factors – $x_1, x_2,$

15 [27], 86.

and x_3 – one among the eight possible background contexts would be $K_i = \neg x_1 \wedge x_2 \wedge \neg x_3$.

We may also define negative as well as neutral causal factors by changing the $>$ in equation 2.2 to $<$ and $=$ respectively. This idea of requiring the causal relationship to hold in all background contents is referred to as *context unanimity*. There is ongoing debate on this requirement, but for Eells's theory we will assume it as a given.[16] Lastly, C may also have *mixed* relevance for E, it may not be negative, positive or neutral. This corresponds to C's role varying depending on the context. Eells defines that C is *causally relevant* to E if it has mixed, positive, or negative relevance for E – i.e. it is not causally neutral.

In addition to determining whether C is causally relevant to E, we may want to describe *how* relevant C is to E. As in Suppes' theory, it is possible to have causally relevant factors with small roles. One method Eells gives for measuring the significance of a factor X for a factor Y is:

$$\sum_i Pr(K_i)[Pr(Y|K_i \wedge X) - Pr(Y|K_i \wedge \neg X)], \tag{2.3}$$

where this is called the average degree of causal significance (ADCS).

Unlike Suppes, Eells notes that with $X_{t'}$ and Y_t, where $t' < t$ and these are particular times, the factors being held fixed may be at any time t'', earlier than t, including between t' and t as well as earlier than t'. Note also that since we are describing particular times, we can account for the fact that causal relevance may change over time.

For example, smoking in a forest may cause a forest fire. However, it is highly unlikely that a lit match at time t' caused a fire at time t if $t' \ll t$.

16 For further discussion, see [22, 26, 23].

Token level causation

Token claims depend on their context. For example, in one scenario we asked whether a lit match was the cause of a house fire on a particular occasion. Regardless of the general, type-level, relationship between lit-matches and house fires we need to know more about the particular situation to determine whether it was the cause of the fire on that particular occasion. Token causation is used to analyze the causes of a particular event, and allows for the possibility that a type-level positive cause of an event may be a negative token-level cause. This type of analysis is Eells's major contribution in [27].

The general form of the question looked at here is: what is the significance of x's being of type X for y's being of type Y, where events x and y are specified by their locations in time and space (which may include intervals of time) as well as the properties of these locations (i.e. they may be thought of as a set of coordinates plus the factors of those coordinates). These questions may be answered by "because of," "despite," or "independently of," corresponding to positive, negative, and neutral causal factorhoood as we saw earlier.

Eells begins by looking at *probability trajectories*, the main idea being that we can study the probability of y being of type Y over time. Then, we define that y is of type Y *because of* x if the following four conditions apply:

1. The probability of Y changes at the time of x;

2. Just after x the probability of y is high;

3. The probability is higher than it was before x; and

4. The probability remains high until the time of y.

In general, we may summarize the four possible relations as:

Despite: y is Y *despite* x if the probability of Y is lowered after x_t,

Because: y is Y *because of* x if the probability of Y is increased after x_t and remains increased until y_t,

Autonomously: y is Y *autonomously of* x if the probability of Y changes at x_t, this probability is high, but then decreases before y_t, and finally,

Independently: y is Y *independently of* x if the probability if Y is the same just after x_t as it is just before x_t,

where x_t and y_t are the respective times of those events. Then, x is *causally relevant* to y if Y happened either because of or despite x. As Eells states, this is only the basic idea of token-level causation, and we need to look more at x (previously we considered only its time) as well as hold fixed the background contexts as we did with type-level causes.

Eells describes two sets of factors that must be held fixed. The first category consists of:

> Factors such that they are actually exemplified in the case in questions, their exemplifications are token uncaused by x being X and they are type-level causally relevant to y's being Y during the context determined by how things are before they occur;

and the second,

> Factors such that they are actually exemplified in the case in question, their exemplifications are token uncaused by x being

> X and they interact with X with respect to Y in the context determined by how things are before x_t.

These factors may occur at any time before y_t. The causal background context is obtained by holding positively fixed all factors of these two kinds. However, these cases do not improve the classification of all relationships. Take one example described by Eells. We have a patient who is very ill at time t_1. She is likely to survive until t_2 but not until a later t_3. Now, assume that at t_1 a treatment is administered that is equally likely to kill the patient as the disease is. Now, at t_2 a completely effective cure is discovered and administered and the only remaining chance of death is due to the first ineffective treatment – not the disease. However, the probability of death did not change after the first treatment, so death was token causally independent of it. But, the relation should actually be despite, as the treatment put the patient at unnecessary risk due to its severe side effects (which remain unchanged by the second treatment that cured the underlying disease).

In the example above, the second drug is causally relevant to Y (survival) and is not caused by the administration of the first drug. When we hold fixed the second drug being given, using the first kind of factor described, again the first drug has no effect on the probability of Y. Using the second kind of factor has no effect in this case, as the two drugs do not interact, so the probability of survival after the first drug does not change dependent on the presence or absence of the second drug.

To summarize, Eells describes two main sorts of causation: type-level and token-level and then presents methods of characterizing their relationships based on probabilities. For the first type of causation, causes

either positively or negatively produce their effects in all background contexts, where these contexts include all events earlier than the final effect. For the second, token level causation, Eells presents a method to probabilistically analyze what role a cause played in an effect in a particular instance.

3

CURRENT WORK IN CAUSAL INFERENCE

Recent efforts in causality are found primarily in the following areas: the development of statistical techniques for description and inference, logics for description and analysis of causal relationships, and experimental work that applies these methods to data in a variety of domains.

3.1 CAUSAL INFERENCE ALGORITHMS

Despite the many causal assumptions made in the sciences, comparatively little work has been done to examine the meaning of these assumptions and how we may go about making inferences from experimental data. The two main areas of work have been in: (1) characterizing what can be learned from statistical data and how we may learn it, and (2) the development of a statistical theory of counterfactuals that supports queries on known models and determination of the "actual cause" in such cases. In both cases, the theories are technically probabilistic, but causal relationships are generally deemed deterministic, where the probabilities are due to the limits of what we may observe.

3.1.1 *Bayesian networks: Spirtes, Glymour and Scheines*

The primary method for causal inference has been through the use of graphical models [72, 102, 36]. The main work in this area has been by Spirtes, Glymour and Scheines (hereafter SGS) as described in [114, 120]. Their method is an explanation of the types of causal structures that can be inferred based on the assumptions made and data that is available. The technique takes a set of statistical data and outputs sets of directed graphs depicting the underlying causal structures. Temporal asymmetry is not assumed by their inference algorithm.

The first part of this theory utilizes directed graphs to describe independence assumptions. The directed acyclic graphs (DAGs), called Bayesian networks (BNs), are used to represent probability distributions and causal structures. The probability distributions are introduced into the graph by the Markov condition, or the notion of d-separation.[1] D-separation describes the set of independencies in a DAG in terms of whether, for two vertices X and Y, there is some set of vertices Z blocking connections between them in some DAG G. If so then X and Y are d-separated by Z in G. That is, if there is a DAG: $X \to Z \to Y$, then the only directed path between X and Y is blocked by Z, and thus X and Y are d-separated by Z in this graph. This independence is written as: $X \perp\!\!\!\perp Y|Z$ (X and Y are independent conditional on Z). These conditions are given a causal interpretation with the *causal Markov condition*, which implies the same independencies as d-separation.

1 In the DAG, these methods are equivalent.

Definition 3.1.1. The *causal Markov condition* is that: A variable is independent of all of its non-descendants conditional on all of its direct causes (those that are connected to the node by one edge)

The direct causes, however, need not be direct in space or time. As in a Markov process, where the future state depends only on the current state and not any prior past states, the causal Markov condition (CMC) states that information about a variable is found only in its direct causes – not its effects or any indirect causes. This relates to Reichenbach's common cause principle, described in Section 2.3.1, where two events are causally connected if one causes the other or if there is another event that is a common cause of both. With CMC, if two events are dependent and neither one is a cause of the other, then there must be some common causes in the set of variables such that the two events are independent conditional on these common causes.

In a causal graph with a set of variables V, two vertices are connected with an arrow if one is a direct cause of the other, relative to V. We note that the statement "relative to V" means that the causal graphs are not necessarily complete, there may be causes of some of the variables or variables intermediate between cause and effect that are left out. However, the graphs are assumed to be complete in that all *common causes* of variables are included, and that all causal relationships among the variables are included in the graph. The intention of the causal graph is that it show for possible ideal manipulations (where an alteration directly affects only one variable, with all other changes being a result of the change in the single altered variable), what other variables may or may not be

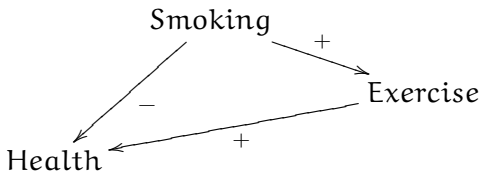


Figure 3.1.: Faithfulness example.

affected. That is, if one variable is an effect of some cause, changing the effect will have no bearing on what happens to the cause.

The inference of causal structures relies on two more assumptions: faithfulness and causal sufficiency. *Faithfulness* assumes that *exactly* the independence relations found by d-separation in the causal graph, hold in the probability distribution over the set of variables. This requirement implies that the independence relations obtained from the causal graph are due to the causal structure generating it. If there are independence relations that are not a result of CMC, then the population is unfaithful. The idea of faithfulness is ensuring that independencies are not from coincidence or latent variables, but from some structure.

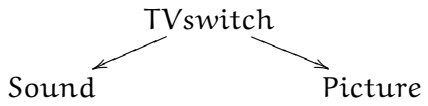
In an example given by Scheines [114], suppose we have the graph in figure 3.1. Then, smoking is negatively correlated with health, but positively correlated with exercise, which is in turn positively correlated with health. Then, it is possible to have a distribution generated by that structure such that the positive effect of exercise (via smoking) on health exactly balances the negative effect of smoking on health leading to no association between smoking and health. In that case, the population would be *unfaithful* to the causal graph generating it.[2] *Causal sufficiency* assumes that the set of measured variables includes all of the common

2 Note that this is precisely the same graph as in figure 2.2, illustrating Simpson's paradox.

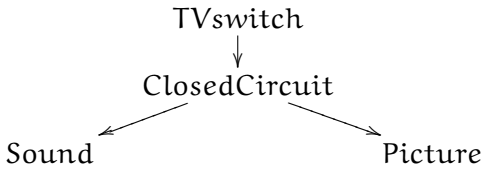
causes of pairs on that set. This notion differs from that of completeness in that we are assuming that there is a causal graph that includes these common causes and that these common causes are part of the set of variables measured.

The primary objections to (and problems with) this theory hinge on the argument that the preceding assumptions do not normally hold and are thus unrealistic. Usually, the objection is to CMC. This is perhaps the most debated portion of the theory, criticized heavily by Cartwright and defended by Hausman and Woodward [53, 13, 12]. Cartwright's main argument against CMC is that common causes do not always screen off their effects. One example given [120] is that of a television with a switch that does not always turn the TV on (See figure 3.2). But, when the TV does turn on, both sound and picture are on. Given that the sound has turned on, even after knowing that the switch is turned on, we know more about whether there will be a picture than we would if we did not know that the sound was on. It would seem that the picture is not independent of the sound, violating CMC as there is no arrow between picture and sound, and their earlier common cause fails to screen them off from one another. The second objection to the SGS method is with the faithfulness condition. One problem with this and the conditional independence stipulations in general are that they only hold when the relationship is exact but it is not possible to verify the exact independence from finite sample data [59]. Thus, the argument goes, we must be finding approximate independence, but that has no meaning in the SGS algorithm.

The result of the SGS method is a set of graphs that all represent the independencies in the data, where the set may contain only one graph



(a) Common cause, no screening off.



(b) Common cause, screening off.

Figure 3.2.: Screening off example.

in some cases when all assumptions are fulfilled. However, when using these graphical models there is no natural way of representing or inferring the time between the cause and the effect or a more complex relationship than just one node causing another at some future time. Following the use of Bayesian networks, dynamic Bayesian networks (DBNs) [39] were introduced to address the temporal component of these relationships. DBNs extend BNs to show how the system evolves over time. For this purpose, they generally begin with a prior distribution (described by a DAG structure) as well as two more DAGs: one representing the system at time t and another at $t + 1$, where these hold for any values of t. The connections between these two time slices then describe the change over time. As before, there is usually one node per variable, with edges representing conditional independence. While DBNs are a compact representation in the case of sparse structures, it can be difficult to extend them to the case of highly dependent data sets with thousands of variables, none of which can be eliminated [93].

Recent work by Langmead et al. [71] has described the use of temporal logic for querying pre-existing DBNs, by translating them into structures that allow for model checking. This approach allows the use of known DBNs for inference of relationships described by temporal logic formulas. However, only a subset of DBNs may be translated in this way [70], and thus the benefit of this approach (as opposed to one where the model inferred already allows for model checking or where we test formulas directly in data) is limited.

3.1.2 *Structural Equations: Judea Pearl*

Pearl's work on causality addresses three main areas: how causality may be inferred both with and without the aid of temporal information [102], how to define a formal theory of counterfactuals using structural models [102], and finally, how to determine the actual cause of an effect [47, 46]. The basic idea in Pearl's work is that there are functional relationships between variables and that causation is a method of encoding the behavior of the system under interventions, where interventions are manipulations of the functional relationships, or mechanisms [101]. Then, the goal of the causal inference is to be able to predict the effect of interventions on the system. Pearl's work on causal inference bears many similarities to that by SGS, so here we summarize only the structural equation model and method of determining actual causes.

Structural Equation Model

Pearl presents a structural-model semantics of counterfactuals [101, 102], where causal models are defined as deterministic and the probabilities come from background conditions. In this work, Pearl says that the role of a causal model is to encode the truth value of sentences relating to causal relationships. These sentences have three types: action sentences (B will be true if we do A), counterfactuals (B would be different if it were not for A), and explanations (B occurred because of A).

Then, a causal model in the structural model semantics is defined as a triple, $M = \langle U, V, F \rangle$ where:

1. U is a set of background variables (determined by factors outside the model);

2. V is a set of endogenous variables (determined by variables in the model - i.e. in $U \cup V$); and

3. F is a set of functions where each $f_i \in F$ is a mapping from $U \cup (V \setminus V_i)$ to V_i s.t. the set F is a mapping from U to V.

Each f_i gives the value of V_i given the values of all other variables in $U \cup V$, and the set F has a unique solution $V(u)$, as the system is acyclic.

The causal model M can be associated with a directed graph $G(M)$ such that each node corresponds to a variable in V and the directed edges point from members of PA_i (parents of i) and U_i toward V_i. In this model, these graphs are called *causal diagrams*. Essentially, the graph identifies the background and endogenous variables directly influencing each V_i. Note that the parent variables are only those in V, as the background variables are not always observable, though the set may be extended to

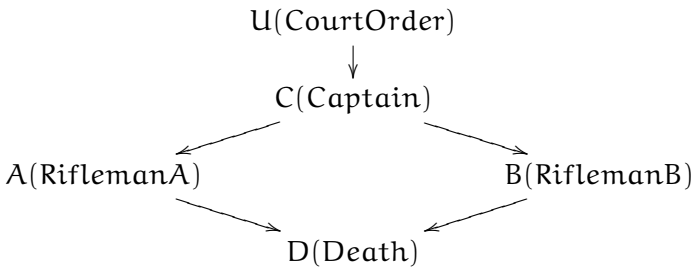


Figure 3.3.: Firing squad example.

include those that *are* observed. The variables in V are those that we can potentially influence, while those in U are outside the system. In figure 3.3, the background variable would be U (Court Order), with all other variables being endogenous. The corresponding causal model is:

- $C = U$,
- $A = C$,
- $B = C$, and
- $D = A \vee B$.

The counterfactuals are interpreted in response to alterations to these equations. For example, the counterfactual "the value that Y would have had, had X been x," where Y is a variable in V and $X \subset V$ is interpreted as denoting the potential response $Y_x(u)$. It is then possible to assess, as Lewis did, whether the effect would still happen in a model where the cause did not. This can also be generalized in terms of probabilistic systems.

Finding the Actual Cause

Much as Eells gives a theory of token causation, Pearl formalizes the notion of what he calls the "actual cause." That is, the cause of an effect

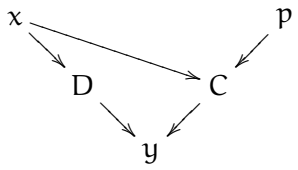


Figure 3.4.: Desert traveler example.

on a particular occasion. For example, "Socrates drinking hemlock was the actual cause of Socrates death" versus "drinking hemlock causes death" [102]. Pearl refers to token-level causes as "actual causes" and type-level causes as "general causes." In contrast to the philosophical approaches that treat each type as a separate species of causal claim [27], Pearl's structural account treats them in the same manner, where the only difference is in the supporting information needed.

In the firing squad example of the previous section (figure 3.3), what was the *actual cause* of death? In the standard counterfactual account we would find that had A not shot the prisoner, B would have and thus A's shot is not the cause of death and neither is B's (since had B not shot the prisoner, A would have). As described in prior sections describing Lewis' counterfactual account of causation, this is an example of overdetermination [74]. Similarly, we can remember the problem of preemption, where if rifleman A had moved a bit closer to the prisoner, then A's shot may hit him before B's does (assuming they were both acting on a court order). In that case, A should be the actual cause of death, as B is *preempted* by A. Lewis deals with this by stating that c causes e if there is a causal chain from c to e (defined as before, where each member is counterfactually dependent on the prior link).

According to Pearl, the real difference is structural. This is illustrated by the example in figure 3.4 [102]. There, a traveler has two enemies. One

poisons his canteen (p) and the other one, not knowing about this, shoots the canteen, emptying it (x). The traveler later dies, but who actually caused his death? The intermediate variables C and D represent cyanide intake and dehydration respectively, and y denotes death. Then, the values of each variable are given by:

$$\begin{aligned} c &= p \wedge \neg x, \\ d &= x, \text{ and} \\ y &= c \vee d. \end{aligned}$$

Simplifying, we find that:

$$y = x \vee (p \wedge \neg x) \equiv x \vee p \tag{3.1}$$

However, Pearl argues here that the equations on either side of the $\equiv$ are structurally different. That is, $x \vee p$ is symmetric, whereas $x \vee p \wedge \neg x$ means that if x is true, p has no effect on y or any intermediate variable. Thus, Pearl says, this asymmetry is what allows us to conclude that x is the actual cause of death.

The basic ideas of the account Pearl proposes come from Lewis' counterfactual account [74] and Mackie's INUS conditions [83]. Here, Pearl addresses the problems found with counterfactual analysis, mainly that it ignores sufficiency, and fails in cases of pre-emption and overdetermination. The main component of this account is the idea of sustenance, which combines necessity and sufficiency of causation while also taking into account structural information.

First, *dependence* is intended to capture the idea of necessity. That is, y is dependent on x in u means that when X's value is altered from x to x', Y's value changes from y to some other y'. This addresses the necessity of x for maintaining the value of Y at y. Second, *production* refers to the ability of a cause to create an effect in a case where neither are present (this relates to the notion of sufficiency). That is, if X's value (in u) is modified to x from x', where Y's value is y', Y will then take on value y. Finally, sustenance combines ideas from both dependence and production.

Definition 3.1.2. x *sustains* y in u relative to W, where W is a set of variables in V and w, w' are specific realizations of these variables, iff: $X(u) = x$, $Y(u) = y$, $Y_{xw}(u) = y$ for *all* w and $Y_{x'w'}(u) = y' \neq y$ for *some* $x' \neq x$ and some w'.

This is a weaker version of necessity in that Y need only differ in the absence of x under one condition, however it is also a stronger version of sufficiency, in that Y must maintain its value y under *any* w. Note that this argument implies that there is some $w = w'$ such that x is both necessary and sufficient for y.

The concept of sustenance is central to Pearl's method for finding the actual cause, called *causal beams*. A causal beam is a new model created by removing all parents except for those sustaining their children. That is, the parents that remain are those that are sufficient for maintaining the value of their children – regardless of how the other parents are set. Then, the other parents are set to some other value. The causal beam allows explanations for actual events under a hypothetical "freezing" of variables. This freezing may be at the actual values of the variables

(natural beam) or at some nearby values (causal beam). Then, an event $X = x$ is an actual cause of an event Y if it changes when X is not x at the actual values of the variables. If the change only happens when values are removed from their actual values, then x is a contributory cause of y.

A few counterexamples to Pearl's theory of actual (or token) causation are given by Menzies [89]. The types of examples given are those that are generally given against theories of token causation that allow transitivity. While some examples – such as a switch that causes a train to go on one of two tracks (where it arrives at its destination regardless of the track) – will find erroneous causes using three variables (i.e. the switch will be the cause of arrival, even though arrival does not depend counterfactually on the switch's position), this can be remedied by using more detailed variables (i.e. one for each train track). Conversely, counterexamples can be found where adding an extra variable creates anomalous results. The argument against these examples is that they use unnatural models that do not reflect how we would normally reason about the cases.

3.2 GRANGER CAUSALITY

Another statistical method, applied primarily in economics, was developed by Granger to take two time series and determine whether one is useful for forecasting the other [44]. While this approach does not attempt to relate to standard notions of causality (rather it proposes a new definition that is most similar to correlation), it is widely used in many of the same applications as the causal inference approaches. Further, it is one of the few methods that explicitly include temporal information, so it

will later be included in our experimental comparisons. Here, pairwise causality is defined by [45]:

Definition 3.2.1. With Ω_t being all available (non-redundant) knowledge at time t, Y_t *Granger causes* X_{t+1} if $P(X_{t+1} \in A|\Omega_t) \neq P(X_{t+1} \in A|\Omega_t - Y_t)$ where A is some set of observations.

That is, the information contained in Y_t provides information on X_{t+1} that is not contained in the rest of the set. Here there is an assumption of temporal priority between cause and effect, but no mention as to whether the probability is higher or lower in the absence of the cause – only that it is different (note that this is similar to Eells's definition of causal relevance). Similarly to the other methods described, there is no notion here of how much of a difference Y_t makes to X_{t+1} and whether there are better predictors or other pieces of information that may be added to Y_t to improve its use as a predictor of X_{t+1}. Further, there is no intrinsic method of representing complex factors such that their causal roles may be inferred automatically from the data. For example, we may want to test not just whether there is a relationship between unemployment and a bull market, but perhaps:

$$(a \wedge b)Uc \leadsto^{\geqslant t_1, \leqslant t_2}_{\geqslant p} d,$$

which could be interpreted to mean that after a bear market (a) and increasing unemployment (b) persist *until* unemployment reaches 20% (c), then within 1 (t_1) to 2 (t_2) months, there will be a bull market (d) with probability p.

In practice Granger causality is frequently tested using linear regression and testing whether use of the information in the possible cause leads

to a smaller variance in the error term than when this information is omitted. An extension, proposed by Chen et al. [17], allows analysis of an arbitrary number of time series as well as nonlinear models. There they introduced the Conditional Extended Granger Causality Index (CEGCI), where multiple time series are analyzed. That is, with three time series, A, B, and C, to determine whether A is causally relevant to C we look at the prediction error of C given only B versus that given $A \wedge B$.

Building on this, recent work by Eichler and Didelez [32] focuses on time series and explicitly capturing the time elapsed between cause and effect. They define that one time series causes another if an intervention on the first alters the second at some later time. That is, there may be lags of arbitrary length between the series, and they find these lags as part of the inference process. While it is possible to also define the variables in this framework such that they represent a complex causal relationship as well as the timing of the relationship, the resulting framework still does not easily lead to a general method for testing these relationships.

3.3 CAUSALITY IN LOGIC

One motivation for the use of causal reasoning in logic has been due to its role in diagnosing causes of system malfunctions based on symptoms (visible errors), referred to as *fault diagnosis* [2, 7, 16, 81, 104]. In this and other cases, what is desired is a framework in which it is possible to reason about changes in state due to actions and causal dependencies among actions. In particular, there has been a focus on reasoning about the indirect effects (ramifications) of actions. That is, how to take into account

the effect of an action and propagate its changes on the world [41]. The difference between standard notions of implication and the terminology proposed here is that one action *causing* another means that the first is responsible for the second, rather than the second simply happening some time after the first in a reliable manner [100]. This difference is why it is desirable to be able to describe causal relationships, so it is possible to reason about the results of actions upon the system. The meaning of causality in the majority of the logics described below relates primarily to dependency and the ordering of events – not the philosophical meaning it was given in the other described approaches.

3.3.1 *Situation Calculus*

One of the most influential works on the problem of determining the effects of actions is by McCarthy and Hayes [85]. In that work they introduce the situation calculus as a method of reasoning about causality, ability and knowledge. They attempt to bridge philosophical representations of the world with a logical representation using automata. The situation calculus is a logic that allows specification of *actions*, *situations* (sequences of actions), and *fluents* (things that may have changing truth value dependent on the particular state). The concern of that paper was to enable a computer program to decide that a particular strategy will achieve a given goal. Despite this advance many open problems remained. One that was introduced in the original paper [85] and subsequently studied by many others is the *frame problem*: how to succinctly specify which fluents will change as a result of a particular action. Another problem,

known as the *ramification problem*, was later introduced in [35]. Here, the goal is to find not just the direct consequences of an action but also those that are indirect (i.e. secondary and other effects).

In order to solve these problems, a number of modifications to the situation calculus have been proposed. One method by Lin [79, 80] introduces the predicate $\mathtt{Caused}(p, v, s)$, which is true if fluent p is *caused* to have truth value v in situation s. One of the central ideas in this work is that a fluent's value persists unless it is caused to be otherwise. That is, if something is caused to be true, it will remain true unless there is another action to make it false. Note that this differs from [77] in that it can represent ramifications.

Most recently, Hopkins and Pearl [56] have proposed a framework drawing on earlier work on structural models [46] as well as the work described above. In this work, it is shown that counterfactuals may be modeled using the situation calculus, however one must still specify all dependencies, including those of counterfactuals. Here, a causal model is a situation calculus specification of the system (including preconditions of actions, etc.) and a potential situation and, as in the other theories described, one may test whether a formula (here, it may be given a counterfactual interpretation) holds given the constraints on execution of the system (i.e. action preconditions, etc.).

3.3.2 *Modal Logic*

Another approach to the ramification problem uses modal logic. Work by McCain and Turner [84] focuses on determining a set of possible next

states after an action is performed. That is, they propose a formalization of the effect of performing a specific action, where background knowledge (constraints) is given in terms of "causal laws". This set of states is given using a fixpoint formulation, with the causal laws represented using an extension of S5 modal logic.[3] Note that the laws here are known, though both cause and effect may be arbitrary logical formulas.

Giordano et al. [41] use a subset of PDL (propositional dynamic logic) and introduce the © modality to express causality. Here, truth values of formulas change depending on the actions that are performed. In this logic, it is possible to make statements such as "after all terminating executions of a, p will hold" (written $[a]p$). With this, they allow the expression of ramifications of an action – effects that were not directly caused, but that follow from a causal rule and action. Similar to the successor state axiom, they assume persistency: that is, from one state to the next a fluent is assumed to persist as long as it does not lead to an inconsistency. Similarly to the situation calculus and the "caused" predicate introduced by Lin [79], actions are methods of changing the truth value of states, which are sequences of actions. The main difference is that the approach by Giordano et al. does not allow the use of the contrapositive of causal rules for making inferences. In comparison to the approach of McCain and Turner, it is possible to reason about sequences of actions, rather than single actions, though the causes and effects in this case are simple conjunctions of events.

3 This method was later adapted to the situation calculus by Lifschitz in [78]. A logic of universal causation (UCL) was developed by Turner [125] extending the work of McCain and Turner.

3.3.3 *Interval Logic*

The main method of representing causal relationships using temporal logic has been through the use of interval logic. With this, events are viewed not as time points but rather as having durations. The first logic explicitly taking time into account in order to analyze causality and address the frame problem came from McDermott [86], where he introduced a first-order temporal logic. In that work, causality was defined as being between events and other events or between events and facts, where one causes another if the first is always followed by the second. To describe this, the following predicate was introduced: $(\texttt{ecause}\ \texttt{p}\ e_1\ e_2\ \texttt{rf}\ \texttt{i})$, meaning e_1 is always followed by e_2 after a delay interval $\texttt{i}$ unless $\texttt{p}$ becomes false before $\texttt{i}$ ends. The point $\texttt{rf}$ denotes when interval $\texttt{i}$ begins (at the beginning of e_1 (denoted $\texttt{rf} = 0$), at the end, or at any time in between). Similarly to the notion of persistency described by Giordano et al. [41], McDermott introduces the notion that a fact may persist from one timepoint with a certain lifetime. Here, the primary goal of causal reasoning is for the purpose of planning, that is, reasoning about what is currently true and what will be true in order to determine what actions should be taken to achieve a particular goal.

Later work by Halpern and Shoham [48, 116] introduced a logic of continuous time, where rather than a state (or point in time) satisfying a formula, we write that a pair of states (forming a closed interval) satisfy the formula. In that work two relations corresponding to type and token causation were introduced – "x causes y" and "x (actually) caused y" – where x and y are propositions (events, properties, facts, etc.).

The truth value of a "causes" statement is determined by whether the statement is contained in the background causal theory (a set of logical formulas comprising knowledge of causal relationships). The case of actual causation is similar, but here the background contains rules with actual times (being the actual times of the events) while type level causal rules contain relations between generic times (i.e. 5 and 6 versus t and $t + 1$). To be an actual causal rule, the model of the actual scenario must also contain the causal part of the rule and there must not exist another rule in the theory that leads to the same effect. That is, it must be the only possible cause of the effect.

3.4 EXPERIMENTAL INFERENCE

In this section we review the current state of causal inference in terms of three main types of data: high-throughput biological experiments (such as gene expression microarrays), neural spike trains, and financial data (such as stock returns). The methods use some of the theoretical ideas described above, but as many of those do not allow inference of causal relationships without knowing at least some of the structure a priori, a number of other methods have been employed.

3.4.1 *Biological experiments*

In biology, the most recent work has been done in applying notions of causality to the problem of determining relationships among genes (usually from microarray data). To our knowledge, all current methods seek to

infer "causal networks" – graphs where an edge $A \to B$ means "A causes (or regulates) B" – or associations between individual causes and effects. A primary use of these networks is in linking genetic factors to diseases. Recent techniques used for inferring and modeling causality amongst genes include: Granger causality [92], Bayesian networks [38, 128], mutual information [3] and likelihood-based approaches [113]. Each method begins with pairwise correlations across the entire time series, connecting them to form graphs of networks. However, it can be difficult to see how the network describing one set of experiments differs from that of another (say, two cancer patients). One recent method [95] begins with a correlation network and transforms it into one that includes causation. The partially directed network allows the visualization of multiple relationship types simultaneously, as well as the identification of hub nodes. The general approach provided in that work is meant to be applicable in biological, financial and medical settings; however it does not easily lead to the probabilistic rules that are useful when extending the method to financial data.

3.4.2 *Neural spike trains*

Recent advances, such as the development of micro-electrode arrays (MEA) have resulted in much data on the activities of neurons over time [9]. This has led researchers to attempt to take this time series data (detailing the firings of neurons) and determine the underlying structure: which neurons are causing which others to fire. A primary method choice has been one based on Granger causality [21, 54, 61]. With this, work

has focused on pairwise relations between neurons. Using this type of causality, as described earlier, there is no notion of spuriousness or levels of spuriousness. In highly connected graphs, such as those representing neurons, it can be difficult to determine the genuine causes.

3.4.3 *Finance*

Financial applications, primarily applied to stock market data, have generally focused on finding correlation, not causation. Methods used for this purpose focus on clustering the data to find stocks behaving similarly, using tools such as correlation matrices [62, 94]. In recent years, there has been an effort to correlate the movement of stock prices with news events using keywords or classifications of news stories and press releases [40, 88, 90, 115]. Some of these focus on the task of characterizing news stories, using for instance their content and tone. To our knowledge, the only use of causality in finance has been in the application of Granger causality [45]. None of the previous methods attempt to find causal relationships, or result in a way of characterizing interactions between financial and news events. Similarly, the only use of temporal logic has been in maintaining and querying financial databases [15], not inferring relationships.

4

DEFINING THE OBJECT OF ENQUIRY

4.1 PRELIMINARIES

Our focus is on proposing a new method for causal inference, but before we can do so we must first discuss what we mean, in this work, by "causal", and thus what we will be inferring. Now that we have discussed a number of theories of causality and methods used for inference, we can describe the target of our investigation: what causes we will infer, and how they relate to the types of relationships we have described so far. When we consider inference or the definition of causality, it is important that the meaning we ascribe to the term "causal" has a basis in prior work in this area, particularly that of philosophy. In our everyday lives we often say that things have been *caused* to happen, but what does this actually mean? Looking at biomedical research, there are frequent reports of genes or environmental factors *causing* (or being *responsible* for) cancer, but the term "causes" is taken for granted without any discussion of this terminology and when we can and cannot infer such a relationship. For instance, in biology it is common to perform just a single experiment where a gene is suppressed (knocked-out), and then it is tested whether a given observable trait (phenotype) is present in the absence of the knocked-out gene. If it is not, then the usual explanation given is that the gene causes the trait. However, a number of possible explanations are

consistent with this result. Beyond this, if the phenotype is not absent it does not necessarily mean that it is not caused by the gene. There may be other causes of the trait or the relationship may be more complex than the pairwise one studied. More thought must be given to such causal claims, especially when they are to be used for diagnosis and treatment of life threatening diseases. Thus, before we can suggest any methods for finding causes, we must be clear about exactly what we are finding and what can be done with this information. In this chapter, we will describe the target of our causal investigation. We will give sufficient but not necessary conditions for a causal relation, and thus not all actual causes will fit our definitions.

4.1.1 *What is a cause?*

We must first distinguish between the metaphysical concern of what a cause *is* and the epistemological concern of what can be *known* about causes. When we use "causality" in this work our aim is to predict and explain phenomena and we will focus on practical definitions that help achieve this by identifying what is and is not causal. We do not attempt to define what it *means* for something to be a cause or suggest that our definitions are for what *is* causal. That is, there may be genuine causes that do not fit our criteria, as these are not necessary conditions for causality.

Many theories refer to "event A" causing "event B", but we do not necessarily want to imply that A is an event in the usual sense of something that happens or that occurs at a particular time and place. For example,

we want to allow that a cause could be a property ("the scarf being red caused the bull to charge") – and not necessarily the event of a change in property (i.e. the scarf did not become red, it had the property of being red before and after the charging bull). Thus we are not restricting what types of things may *be* causal. We are agnostic as to what sorts of things have the capability to cause others (e.g. we make no claims about mental causation, causation by omission, and so on), we are restricting only what sorts of things we will look for and consider as potential causes. In this work we will only describe and infer those that may be represented by some proposition or logical formulas. We will refer to these alternately as formulas and factors, noting now that they may be properties, facts, mental states, and so on.

Let us look at a few examples of what we mean by these logical representations. One relationship we will be able to represent is: "a student not doing homework and not going to class until he fails multiple classes causes the student to be expelled from school." (See equation 4.3 for the formal representation of this relationship) In this case, there is no single event causing expulsion, nor is it caused by a conjunction of events. Here we have properties and actions of a student that must continue for some period of time – long enough for a third property to become true. While it is perhaps also true that failing multiple classes causes expulsion, representing the relationship in this way gives more insight into exactly how one must fail in order to be expelled. Also note that this is not equivalent to representation by a causal chain, as we would generally not suggest that not doing homework and not attending class (where this could be satisfied by missing one class and one homework, maybe due to illness), causes failure of multiple classes as in many cases students

skip both with no ill effects. It is also possible that the probabilities of expulsion differ between the potential causes. Further, we note that "not doing homework" is neither an event nor an action, though it is a property of the student that can be easily represented ($\neg homework$). This is also an illustration of why we allow causation by omission (it is possible) and how simple it is to represent logically.[1] Something we will not be able to represent in the logic we will use is a case involving durations. For example, "holding a lit match to curtains for 30 seconds causes house fires." We note that if a cause cannot be represented in this manner, we do not say it is not causal but rather that it is outside the scope of this method and may be inferred by other means.

4.1.2 *How can we identify causes?*

Given a number of possible explanations, or causes, of a phenomenon, how can we determine which are actually causal? For example, we may have data on patients (including their age, whether or not they smoke, and so on) and their current state of health. We make judgements about which variables we should include in a study based on whether we think they can have an impact on health. It is unlikely that we would consider whether the patient was born on an even day or an odd day, but why? Our common sense tells us that this has no bearing on health, but what about cases where one can have no such intuition? For this reason we have two main criteria that help us weed out non-causes from causes. This is not to say that these are features essential to actually being a cause,

1 We will not specifically discuss omissions any further than we have in Chapter 2, as in logical formula, they only amount to negations of properties.

but rather that they are exhibited in enough cases (and cases of interest to us) that they are useful as indicators for causality.

First, we will stipulate that a cause must precede its effect in time.[2] This is in keeping with the philosophical foundations described in Chapter 2, particularly that of Hume and Suppes.[3] While it may be possible for a cause and effect to be simultaneous, in that case we will not be able to identify which is the cause and which is the effect from only observing them.[4] In general, the cases in which we want to make inferences are those consisting of observations of a system over time, so we already have information on the temporal order of events and should make use of this information. While other methods for inference (such as that of Pearl and SGS) do not require timecourse data and infer the direction of the causal relationship as well as its existence, this comes at the expense of much stronger claims about the way the data have been generated and the conditional independence of cause and effect is estimated. We are generally interested in cases involving distributions where few, if any, of their assumptions hold: not all common causes have been measured, the data is quite noisy, relationships are not necessarily linear, and common causes may not fully screen off their effects.

We know that it is possible to make useful inferences in such messy cases with relationships more complex than "event c causes event e." This is something untrained juries do every day. Similarly, doctors manage to diagnose patients when given inaccurate, incomplete and conflicting

2 We do not consider the possibility that a cause could be later than its effect.

3 For more discussion of the direction of the causal relationship and the direction of time, see [108, 98, 51].

4 While there is no inherent reason that a cause and effect could not be at the same time, this has not proven important in our applications. Further, we assume that such a case is a result of the timescale of the measurements being taken and were the measurements to be made on a finer timescale, we would find the cause earlier than the effect.

information. While the success of both of these types of inference relies on extensive prior beliefs and background knowledge (as well as common sense), one might also imagine that much of this may be amenable to automation given enough data (and data of the right type). One bottleneck is the preponderance of counterexamples and arguments against causal inference and theories of causality that do not act as expected in all cases – including those of far fetched examples. The question here should be: why would we expect our method to perform better than a human would?[5] If we can make inferences in even most of the cases that can be handled by a human examining all information manually, then we will consider the method a success. We propose that our standards for causal inference not be held restrictively higher, but rather focus on what can be learned "beyond a reasonable doubt."

Second, a cause can be identified by its ability to make its effect more probable. This is a standard feature of probabilistic theories of causality, such as that of Suppes [124]. Leaving aside the question of negative causation (a cause inhibiting its effect),[6] there may be cases where a cause is so weak that the difference it makes to the probability of its effect is not perceptible, though it is still a cause (perhaps there are other factors required for it to cause the effect or it is stronger in some small set of cases while being weak in general). In that case, we can defend our assumption by noting that if the cause has so little influence on the effect, it will not be helpful for either of our stated purposes (prediction and explanation), so there must be other causes that account better for the effect and it would be more fruitful to first explore those.

5 We assume no human is infallible.
6 Negative causes can be defined in terms of making the negation of the effect more likely.

What we mean by "causal", then, is implicit in the way we identify causality. We have said that, at least some of the time, causes are things that precede their effects and that increase the probability of their effects. Thus, causes are also helpful for prediction and explanation of their effects. This is what will be meant by the terms "causal" and "causes" throughout the rest of this work, albeit with some qualifications.

4.1.3 *Requirements for a definition of causality*

We have described two features that aid in identification of causal relationships and are ready to look in more detail at what is needed in terms of both these features and their representation.

Probability & Time

While we described the temporal priority condition and its motivation, simply stating that the cause is earlier than the effect is not enough; we must be able to represent the *amount* of time between cause and effect. Consider what happens if our inference focuses solely on the production of networks representing conditional independencies. Here, if we have two data sets with the same conditional independencies between variables (represented by edges connecting appropriate vertices), but where the relationships occur over varying time scales, we will not see any difference between them. If the information gleaned from a system ends at the presentation of the system as a network – the inference of relationships where we know only that the cause is earlier than the effect – we lose vital temporal information.

For example, cigarettes in the UK have warnings such as "smoking kills" and "smoking causes fatal lung cancer" in large type. Without any other details, we must base our actions on implicit assumptions: that smoking will cause death before something else would, that it will always cause death, and that it will cause our death in particular. In and of itself "smoking kills" is no more informative than "birth kills" as with probability one, everyone who is born dies. Now imagine one pack of cigarettes says "smoking kills: within 2 years" and another says "smoking kills: 80 years from now." In this case one would likely choose the second package. We need to be able to represent, explicitly, how long after the cause the effect will happen. This will be useful as well when attempting to apply our inferred relationships to specific, or token, cases. For example, if a person begins smoking and then dies the day after, we would likely say that there must be some other cause of death. Without any further temporal information in our causal relationship, however, we cannot capture that intuition.

Note that we have still made no reference to the probability of death nor to the duration that one must smoke for death or cancer to occur.[7] The first case described above could be a 0.01% chance with the latter being 50%. This additional information and the way that it will affect our decision making process shows the need for a more detailed description of causality. That is, when we describe a causal relationship we need to be able to describe its probability, the time over which it takes place, and whether there are other events and properties required for the effect to be caused.

7 While durations are not considered in this work, they are an important aspect that should be explored in future studies.

Expressing Causality

We have described what we mean by causality and aspects of the relationship that are important, but how should we represent this relationship? Rather than a verbose English sentence, can we compactly describe a relationship such as in the examples above? For instance, we want to be able to make statements such as "not doing home work and not attending class until multiple classes are failed will lead to expulsion within two semesters with probability greater than .6". Further, we also want to be able to test whether such an assertion is true.

A natural method for reasoning about such information is by using a probabilistic temporal logic. While inference methods such as those based on graphical models, as well as nearly all theories of causality, allow causes and effects to be defined arbitrarily by their users, this does not easily lead to methods for specifying and testing these arbitrarily complex relationships. If, for example, we are testing "smoking causes lung cancer in a minimum of 15 years with probability 0.5", there would be no convenient way of expressing this using an arrow between two nodes (as is done with graphical models), and in the absence of some standardization, methods for testing this would need to be written for each individual case. By formulating our relationship using a well defined logic, we can avail ourselves of pre-existing methods for testing these properties and develop general methods to extend these as needed. Here we will briefly discuss how the problem of causal inference relates to that of model checking in order to provide context for the following definitions. Inference will be discussed in depth in Chapter 5.

We may generalize our problem as follows: given a set of time series data representing a system in which we hypothesize there may exist a causal structure, we seek to infer the underlying relationships characterizing this structure. When we observe a system over time, what we see is one possible path through the system. For example, we can take barometer measurements over time and record the weather conditions, the day of the week and the season. Thus at one time point we might observe that on a Tuesday in the winter, the barometer is falling rapidly and there is a storm and at another time point we might see that on a Thursday in the summer the barometer is steady and there is no rain. These collections of propositions specify two possible states the system can be in: {(barometer-falling-rapidly, storm, Tuesday, winter), (barometer-steady, ¬rain, Thursday, summer)}. Our set of observations gives the frequency with which the system occupies these states. Note that we may not observe all states, only some subset of those that are possible. As our observations have a time order, we have also observed the transitions between states of the system, as well as the frequency of each transition. A state is simply a collection of properties that are true and a transition between two states, s_1 and s_2, means that it is possible for the system to be in an s_1 state at some time t_1 and to then be in an s_2 state at time t_2, where $|t_2 - t_1| = 1$ time unit, and where this unit size is defined by the scale of our measurements. The probability of this transition can be inferred (increasingly accurately as our number of observations tend toward infinity) based on the frequency with which it is observed in the data. Thus, we can potentially reconstruct or redescribe this structure consisting of states and transitions between them from the data. However, in practice this problem is not so trivial. It is possible,

though, to query the data directly. Then we can ask questions about what properties a structure or a set of observations satisfies. In our case, we can find out whether it satisfies various causal relationships. Solving problems like these is precisely the goal of model checking: verifying whether a system satisfies some specified properties. Thus if we specify our causal relationships and define causality in this way, our problem becomes one of model checking.

4.2 A LITTLE BIT OF LOGIC

Before describing causal relationships in terms of logical formulas, we will give a brief introduction to temporal logic and the particular logic used. For a more in depth introduction, see Appendix A. Modal logic was introduced in order to describe "modes of truth." That is, a formula might be "necessary" or there might simply be some possible world in which it holds, in which case it is "possibly" true. This means that the formula might very well be false in the current world, while still being possibly true. Temporal logic, introduced by Prior [105, 106], modified modal logic to include *when* formulas must hold, or be true. For example, we can specify whether some property must be true at the next point in time or simply at some point in the future. In branching time logics, such as computation tree logic (CTL) [18], the future may be along any number of possible paths, thus we can also express whether the property should hold for *all* possible paths through the system, or whether it is enough for there to *exist* a path where it is true. We will use a probabilistic extension of this logic, probabilistic computation tree logic (PCTL), as introduced

by Hansson and Jonsson [49], as we also want to capture the probabilistic nature of our system.[8]

First, we begin with a probabilistic version of a Kripke structure [69, 19], also called a discrete time Markov chain (DTMC). This is a directed graph with a set of states, S, that are labeled with the properties true within them via a labeling function. This function maps states to a set of atomic propositions of the system. In the example of section 4.1.3, this set would be: {rain, storm, barometer-falling-rapidly, barometer-steady, Thursday, Tuesday, summer, winter}, and a state could be labeled with any combination of these. Note that if a state is not labeled with a proposition, such as $\mathtt{rain}$, it is considered to be labeled with its negation (i.e. $\neg\mathtt{rain}$). It is possible to make the labeling probabilistic, so that "barometer falling" may be false due to the barometer being broken with some small probability. There, the two separate probabilities (one due to the actual probability of the system, and the other due to our lack of knowledge about the system), would be explicitly captured. In our case, we do capture both probabilities, but they are combined into one measure: the transition probability. We also have an initial state from which we can begin a path through the system. Finally, we have a transition function that defines, for each state, the set of states that may immediately follow it as well as the probability of each of these transitions. This is a total transition function, which means that each state has at least one transition to itself or another state in S with a non-zero probability. The sum of the transition probabilities from any state is 1, meaning that at each time point a transition must be made – the system cannot remain in the

8 Another possibility would be to use a logic such as UTSL [129], which incorporates statistical hypothesis testing, and extend this for multiple hypothesis testing.

state without making a transition. More formally, we have a structure $K = \langle S, s^i, L, \mathcal{T} \rangle$ and a set of atomic propositions AP, where:

S is a finite set of states;

s^i is an initial state;

$L : S \to 2^{AP}$ is a state labeling function; and

$\mathcal{T} : S \times S \to [0, 1]$ is a transition function such that:

$$\forall s \in S \sum_{s' \in S} \mathcal{T}(s, s') = 1.$$

We also use $\mathtt{labels}(s)$ to denote the labels of a particular state, s.

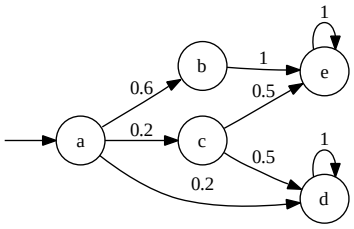
Then, the types of formulas that can be expressed in PCTL are path formulas and state formulas. State formulas express properties that must hold within a state, such as it being labeled with certain atomic propositions (e.g. is a state s labeled with $\mathtt{rain}$?), while path formulas refer to sequences of states along which a formula must hold (e.g for some sequence of states, will it eventually $\mathtt{rain}$?). Valid formulas are defined as follows.

1. All atomic propositions are state formulas.
2. If f and g are state formulas, so are $\neg f$, $f \wedge g$, $f \vee g$, and $f \to g$.
3. If f and g are state formulas, and t is a nonnegative integer or ∞, $f \mathcal{U}^{\leqslant t} g$ and $f \mathcal{U}^{\leqslant t} g$ are path formulas.
4. If f is a path formula and $0 \leqslant p \leqslant 1$, $[f]_{\geqslant p}$ and $[f]_{>p}$ are state formulas.

The second item above says that we can combine and negate state formulas to make new formulas, with $\neg, \wedge, \vee$, and $\to$ defined in the usual

manner as: negation, conjunction, disjunction and implication. In the third item, we have the until ($\mathcal{U}$) and unless ($\mathcal{U}$) operators. In this context, "until" means that one formula must hold at every state along the path until a state where the second formula becomes true. The formula above, $f\mathcal{U}^{\leqslant t}g$, means that f must hold until g holds at some state, which must happen in less than or equal to t time units. Unless is defined the same way, but with no guarantee that g will hold. If g does not become true within time t, then f must hold for a minimum of t time units. Finally, we can add probabilities to these until and unless path formulas to make state formulas. For example, $[f\mathcal{U}^{\leqslant t}g]_{\geqslant p}$ (which may be abbreviated as $f\mathcal{U}^{\leqslant t}_{\geqslant p}g$), means that with probability at least p, g will become true within t time units and f will hold along the path until that happens. This until formula with its associated probability defines a state formula. The probability of the formula is calculated by summing the probabilities of the paths from the state, where a path's probability is the product of the transition probabilities along it.

For example, let us take the following structure.



Recall that each state has at least one transition. This means that paths are infinite sequences of states, written $\sigma \equiv s_0 \to s_1 \to \cdots s_n \cdots$. However,

we can look at the prefix, say of length n, of path σ. This is denoted by $\sigma \uparrow n$ and defined by:

$$\sigma \uparrow n = \sigma = s_0 \to s_1 \to \cdots s_n.$$

Then the probability measure for a path (denoted by μ_m) is the product of the transition probabilities. For the prefix above, this is: $\mathcal{T}(s_0, s_1) \times \cdots \times \mathcal{T}(s_{n-1}, s_n)$. Now, looking at the structure above, let $\sigma = a \to b \to e \to \cdots e \cdots$, and let us take $\sigma \uparrow 2$. The probability of this path is then 0.6×1. Then, we can take the set of paths of length two, from a to e. There are two such paths: one through b and one through c. The probability of this set of paths is the sum of the individual path probabilities: $0.6 \times 1 + 0.2 \times 0.5 = 0.7$. Then, for a particular state, a probabilistic formula such as $[f]_{\geqslant p}$ is satisfied if the sum of the path probabilities of the set of paths satisfying the formula is at least p. A structure K satisfies a state formula if s^i satisfies it.

We will also make use of the standard modal operators as shorthand for their PCTL equivalents. That is, we can define PCTL operators analogous to the path operators A ("for all paths") and E ("for some future path") and temporal operators G ("holds for entire future path") and F ("eventually holds").

- $Af \equiv [f]_{\geqslant 1}$,
- $Ef \equiv [f]_{>0}$,
- $Gf \equiv f\mathcal{U}^{\leqslant\infty}false$, and
- $Ff \equiv true\ \mathcal{U}^{\leqslant\infty}f$.

One final operator we will need is "leads to," as described by Hansson and Jonsson [49]:

$$f \leadsto^{\leqslant t}_{\geqslant p} g \equiv AG[(f \rightarrow F^{\leqslant t}_{\geqslant p} g)]. \tag{4.1}$$

This means that for every path from the current state, if we are in a state where f holds then through some series of transitions taking time $\leqslant t$, with probability p, we will finally reach a state where g holds. One difference here is that as defined in equation (4.1), leads-to considers the case where f and g are true at the same state as one that satisfies this formula. We will stipulate that there must be at least one transition between f and g. In addition to being important for our temporal priority condition for causality, this is also in keeping with how one naturally reasons about the term "leads to." The expectation if someone says "one thing led to another" is that there was some sequence of events connecting "one thing" and "another" and that they are not co-temporary. We may also wish to write:

$$f \leadsto^{\geqslant t_1, \leqslant t_2}_{\geqslant p} g, \tag{4.2}$$

which is interpreted to mean that g must hold in between t_1 and t_2 time units with probability p. If $t_1 = t_2$, this means it takes exactly t_1 time units for g to hold. We show in Appendix C.2 that this minimum time can be added to the leads-to operator.

Let us return to our prior example, where "a student not doing homework and not going to class until he fails multiple classes causes the student to be expelled from school", and see how this may be represented

as a PCTL formula. The propositions are doing homework (h), class attendance (c), failure of two or more classes (f), and expulsion (e). Then, the relationship is:

$$[(\neg h \wedge \neg c) U^{\leqslant \infty}_{\geqslant p_1} f] \rightsquigarrow^{\geqslant 1, \leqslant t}_{\geqslant p_2} e, \tag{4.3}$$

where t is the maximum amount of time it will take for expulsion to occur.

4.3 TYPES OF CAUSES AND THEIR REPRESENTATION

We will now define four main types of causes in terms of logical formulas and discuss how they relate to the probabilistic theories of causality, described in Chapter 2.3.

4.3.1 *Prima facie causes*

According to how we have formulated how to identify causes, we will give the basic conditions for causality. For some c and e, for us to identify c as a possible cause of e, c must be temporally prior to e and must raise the probability of e. Prima facie causes are those that satisfy these basic requirements. Recall that when we describe some cause c and effect e, that both c and e may be arbitrarily complex logical formulas. Below and in the following examples we will refer just to c and e and note now that there are no conditions on them other than that they must be valid PCTL state formulas.

First, we specify the temporal priority condition of the causal relationship in terms of the time that elapses between cause and effect, rather than the occurrence times of the cause and effect. If c occurs at some time t and e occurs at a later time t', we characterize the relationship by the time that elapses between them, which is $|t' - t|$. If we want to state that after c becomes true, e will be true with probability at least p in $|t' - t|$ or fewer time units – but with at least one time unit between c and e – we write:

$$c \leadsto^{\geqslant 1, \leqslant |t'-t|}_{\geqslant p} e.$$

That is, there is a window of time in which e may occur. Note that satisfying this formula requires there is at least one and potentially any number of transitions between c and e, as long as the sum of probabilities of the paths between c and e taking at least one time unit is at least p. The transitions are assumed to each take one time unit, but there is no restriction on the definition of a time unit. If we only want to say that c is earlier than e, the lower bound will be 1 and the upper bound ∞. In some cases, we will have domain specific knowledge and will want the amount of time between cause and effect to be in terms of a known period of time. In that case, the bounds on the second condition (1 and ∞) can be replaced with any arbitrary t_1 and t_2 where $1 \leqslant t_1 \leqslant t_2 \leqslant \infty$, with $t_1 \neq \infty$.

Then, the probabilistic nature of the relationship between cause and effect can be described in terms of the probability of reaching c and e states and of the paths between c and e states. We need to specify that c must occur at some point and that the conditional probability of e given

c is greater than the marginal probability of e. We now define *prime facie*, or potential, causes, as shown below.

Definition 4.3.1. c is a *prima facie* cause of e if there is a p such that the following conditions all hold:

1. $F^{\leqslant\infty}_{>0}c$,
2. $c \rightsquigarrow^{\geqslant 1,\leqslant\infty}_{\geqslant p} e$, and
3. $F^{\leqslant\infty}_{<p}e$.

These conditions state that we will reach a state where c is true (beginning from the initial state of the system) with non-zero probability and that the probability of reaching a state where e is true (within the time bounds) is greater after being in a state where c (probability $\geqslant$ p) is true than it is by simply starting from the initial state of the system (probability $<$ p).

For example, take the structure in figure 4.1. We will use the term "causal structure" in this work to denote just such a structure, where this is the underlying one governing the behavior of the system. In general our goal is to infer its properties from the data (observations of the system moving through these possible states over time), but for the moment let us assume that it is given. Note that unlike the causal models previously described, such as Bayes nets, the arrows between states in these structures have no causal interpretation. They only imply that it is possible to transition from the state at the tail to the state at the head with some non-zero probability (which is used to label this edge). Note also that there may be multiple states with the same labels. For example, there may be two states labeled with identical sets of propositions, but that are reached by different paths and which have different paths possible

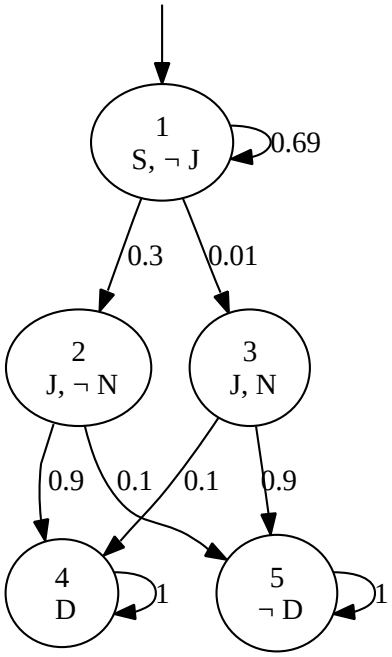


Figure 4.1.: Example structure containing states with their labels and transitions with their associated probabilities. s_1 is the initial state.

from them. It follows then that we can look at properties that are true for every state with a certain label, as well as those that hold only when other sets of conditions leading up to (or occurring after) those states are true. Thus, this type of model fundamentally differs from Bayesian networks, where each variable generally has one node with incoming and outgoing edges (and lack thereof) representing (in)dependencies. In this chapter, we will use the convention that diagrams with circular nodes represent such structures, while those with no borders around the nodes illustrate only temporal ordering (with the node at the tail being earlier than that at the head and the length of the arrow being proportional to the amount of time between the nodes) and some probabilistic dependence.

This example will be discussed in full when we look at token causality in Chapter 6, but for the moment let us say that we aim to find the

cause and probability of death (D) for a suicidal person (S) who jumps from a building (J) that may (N) or may not ($\neg$N) have a net and who may also survive this jump ($\neg$D). The states are numbered so we can refer to them, and labeled with non-negated or negated literals, e.g. $S, J, N, \neg N, D,$ and $\neg D$. The transitions in the graph are also labeled with their probabilities. Now let us say $c = J \wedge \neg N$ and $e = D$ and find out whether c is a prima facie cause of e. We know that this meets the first condition of definition 4.3.1, since s_2 satisfies c, and the probability of reaching s_2 from s_1 is greater than zero.[9] Then, the probability of e for the system is calculated using the approach outlined in Appendix C.2 (See theorem C.2.2 and the related algorithms C.1 and C.2). We find $P = \{s_4\}$, $Q = \{s_5\}$ and $R = \{s_4\}$. The probability of e for the structure, represented

9 $\mathcal{T}(s_1, s_2) = 0.3$, but the probability of reaching s_2 from s_1 is greater than this due to the cycle at s_1. Remember that we are looking at the probability of reaching s_2 at any time. That means we can visit the cycle once, then transition to s_2, or visit twice before transitioning, and so on.

by the probability of $F^{\leqslant\infty}e$, is given by $P(1,\infty,s_1)$, as s_1 is the initial state of the system.

$$
\begin{aligned}
P(1,\infty,s_1) =& \mathcal{T}(s_1,s_1)P(0,\infty,s_1) + \mathcal{T}(s_1,s_2)P(0,\infty,s_2) \\
&+ \mathcal{T}(s_1,s_3)P(0,\infty,s_3); \\
P(0,\infty,s_2) =& P(\infty,s_2) = \mathcal{T}(s_2,s_4)P(\infty,s_4) + \mathcal{T}(s_2,s_5)P(\infty,s_5); \\
P(0,\infty,s_1) =& P(\infty,s_1) = \mathcal{T}(s_1,s_1)P(\infty,s_1) + \mathcal{T}(s_1,s_2)P(\infty,s_2); \\
P(0,\infty,s_3) =& P(\infty,s_3) = \mathcal{T}(s_3,s_4)P(\infty,s_4) + \mathcal{T}(s_3,s_5)P(\infty,s_5) \\
&+ \mathcal{T}(s_1,s_3)P(\infty,s_3); \\
P(\infty,s_4) =& 1, \text{ since } s_4 \in R'; \\
P(\infty,s_5) =& 0, \text{ since } s_5 \in Q'; \\
P(\infty,s_2) =& 0.9 \times 1 + 0.1 \times 0 = 0.9; \\
P(\infty,s_3) =& 0.1 \times 1 + 0.9 \times 0 = 0.1; \\
0.31 \times P(\infty,s_1) =& 0.3 \times 0.9 + 0.01 \times 0.1; \\
P(1,\infty,s_1) =& 0.69 \times \frac{0.271}{0.31} + 0.3 \times 0.9 + 0.01 \times 0.1 \approx 0.87.
\end{aligned}
$$

Thus, the probability of e is < 0.88 and ≈ 0.87. Finally, the probability of $c \rightsquigarrow^{\geqslant 1,\leqslant\infty} e$ is exactly 0.9 (there is only one path from a state where c holds to a state where e holds and it is the transition between states s_2 and s_4). Thus, since $0.9 > 0.88$, c being prior to e raises the probability of e and it is a *prima facie* cause of e.

Equivalence to probabilistic theory of causality

Our conditions, stated in definition 4.3.1, are based on Suppes' conditions for probabilistic causality (definition 2.3.2) and we can show that these conditions are in fact equivalent. First, let us recall the two sets of

conditions. Recall that Suppes' notation A_t and $B_{t'}$, where $t' < t$ only implies that B occurs earlier than A, not that t and t' refer to specific times. We are implicitly summing over all t, considering any scenario where B is before A.

Suppes' conditions for *prima facie* causality (denoted SC):

1. $P(E_t|C_{t'}) > P(E_t)$,
2. $t' < t$, and
3. $P(C_{t'}) > 0$.

Our conditions for *prima facie* causality (denoted LC):

1. $c \overset{\geqslant 1, \leqslant \infty}{\underset{\geqslant p}{\rightsquigarrow}} e$,
2. $F^{\leqslant \infty}_{>0} c$, and
3. $F^{\leqslant \infty}_{<p} e$.

Theorem 4.3.1. *The conditions of Suppes (SC) – temporal priority and probability raising (where the cause occurs with non-zero probability) – are satisfied if and only if the formulas of our conditions for prima facie causality (LC) are satisfied. That is,* $SC \iff LC$.

We begin by showing $LC \to SC$ and then show $SC \to LC$.

Proposition 4.3.1. $LC \to SC$

Proof. We assume that $c = C$, $e = E$ and that we have a structure, $K = \langle S, s^i, L, \mathcal{T} \rangle$, representing the underlying system governing the occurrences of these events. We also assume that states in K that satisfy c and e are labeled as such. If $t' < t$ in SC, we assume that in K there will be at least one transition between an event at t' and one at t. That is, the timescale

of K is as fine as that of Suppes and vice versa. Further, we assume that the probabilities of Suppes' formulation and those in K come from the same source and thus if represented correctly, $P(E)$ in SC is equal to $P(e)$ in LC.

Condition 1 $P(E_t|C_{t'}) > P(E_t)$

By definition of $F^{\leqslant\infty}_{<p}e$, $P(E_t)$ – the probability of E occurring at any time, denoted t – is less than p. Recall that the probability of a path is the product of the transition probabilities along the path, and the probability of a set of paths is the sum of their individual path probabilities. For a structure to satisfy this formula, the set of paths from the start state that reach a state where e holds must be less than p, and thus the probability of reaching a state where e holds in this system is less than p. Thus,

$$P(E_t) < p.$$

Now we must show $P(E_t|C_{t'}) \geqslant p$. That is, the probability of E_t is greater given that C has occurred at some time t' prior to E. We will now show that this conditional probability is greater than or equal to p if:

$$c \overset{\geqslant 1, \leqslant\infty}{\underset{\geqslant p}{\rightsquigarrow}} e \tag{4.4}$$

is satisfied.

The probability p_1 of a transition from state s_1 to state s_2 that labels the edge between them,

$$s_1 \xrightarrow{p_1} s_2,$$

is the conditional probability:

$$P(s_{2,t+1}|s_{1,t}), \tag{4.5}$$

the probability of reaching s_2 one time unit after s_1. Then, for a path:

$$s_1 \xrightarrow{p_1} s_2 \xrightarrow{p_2} s_3,$$

we can calculate the probability, given s_1, of reaching s_3 (via s_2) within two time units:

$$P(s_{3,t+2}, s_{2,t+1}|s_{1,t}) = P(s_{3,t+2}|s_{2,t+1}, s_{1,t}) \times P(s_{2,t+1}|s_{1,t}), \tag{4.6}$$

and since s_3 and s_1 are independent conditioned on s_2 this becomes:

$$P(s_{3,t+2}, s_{2,t+1}|s_{1,t}) = P(s_{3,t+2}|s_{2,t+1}) \times P(s_{2,t+1}|s_{1,t}). \tag{4.7}$$

Note that the probabilities on the righthand side are simply the transition probabilities from s_1 to s_2, and s_2 to s_3 (since there is one time unit between the states, they can only be reached via a single

transition). Thus, the conditional probability is precisely the path probability:

$$P(s_{3,t+2}, s_{2,t+1}|s_{1,t}) = p_2 \times p_1. \tag{4.8}$$

Then, if we have a set of paths from s_1 to s_3, the conditional probability $P(s_3|s_1)$ is the sum of these path probabilities. For example, we may have the following paths:

$$s_1 \xrightarrow{p_1} s_2 \xrightarrow{p_2} s_3, \text{ and}$$
$$s_1 \xrightarrow{p_3} s_4 \xrightarrow{p_4} s_3,$$

in which case:

$$P(s_{3,t+2}|s_{1,t}) = P(s_{3,t+2}, s_{2,t+1}|s_{1,t}) + P(s_{3,t+2}, s_{4,t+1}|s_{1,t}), \tag{4.9}$$

and from equation (4.8) this becomes:

$$P(s_{3,t+2}|s_{1,t}) = p_2 \times p_1 + p_4 \times p_3, \tag{4.10}$$

the sum of the individual path probabilities. Let us now say that s_1 is labeled with c and s_3 is labeled with e, these are the only c and e states in the system, and there are no other paths between the states taking less than or equal to 2 time units. Then, this probability we have computed is in fact the probability of:

$$c \overset{\geqslant 1, \leqslant 2}{\rightsquigarrow} e, \tag{4.11}$$

since the probability of reaching s_3 during a window of time simply means looking at the set of paths reaching s_3 during that window. Similarly, to find the probability of:

$$c \overset{\geqslant 1, \leqslant \infty}{\rightsquigarrow} e, \tag{4.12}$$

we must consider the set of paths from states labeled with c to those labeled with e taking at least 1 time unit. Since there can be cycles in our graph, calculating the probability associated with a leads-to formula with an infinite upper time-bound requires a slightly different method. This is described in detail (and proven correct) in Appendix C.2.

When this is calculated to be at least p, then:

$$P(E_t | C_{t'}) \geqslant p, \tag{4.13}$$

and since

$$P(E_t) < p, \tag{4.14}$$

we have:

$$P(E_t | C_{t'}) > P(E_t). \tag{4.15}$$

Thus Condition 1 is satistfied.

Condition 2 $t' < t$

In LC condition (1), we state:

$$c \overset{\geqslant 1, \leqslant \infty}{\underset{\geqslant p}{\rightsquigarrow}} e. \tag{4.16}$$

That means that there is at least one transition (with a transition taking a nonzero amount of time), between c and e. This means that c must be earlier than e and we satisfy the second condition of SC (temporal priority).

Condition 3 $P(C_{t'}) > 0$

By definition of $F^{\leqslant \infty}_{>0} c$, we satisfy condition (3) of SC. If a structure K satisfies this formula it means that, from its starting state, c will be reached with non-zero probability and thus $P(C) > 0$.

Thus, if the three logical formulas (LC) are satisfied, so are Suppes' conditions (SC) for *prima facie* causality and thus LC $\rightarrow$ SC. □

Proposition 4.3.2. SC $\rightarrow$ LC

Proof. We begin with the same assumptions as for the LC $\rightarrow$ SC case. We also assume that the system of SC is first-order Markovian.

Conditions 1 and 3 $c \overset{\geqslant 1, \leqslant \infty}{\underset{\geqslant p}{\rightsquigarrow}} e$ and $F^{\leqslant \infty}_{<p} e$

Let us denote the probabilities of Suppes' conditions by:

$$P(E_t | C_{t'}) = p', \text{ and} \tag{4.17}$$

$$P(E_t) = p'', \tag{4.18}$$

where we recall that $p' > p''$. From condition (2) of SC, we also know that $C_{t'}$ is earlier than E_t, i.e. $t' < t$. Then, the conditional

probability in equation 4.17 represents the probability of E at any time after C has occurred. Again, we have assumed the same granularity of time in both sets of conditions, and thus if C is earlier than E in SC, there is at least one transition between a state where c holds and one where e holds. That is, applying the same reasoning as we did earlier, C can cause E any number of time units after C occurs. Thus we can show that the probability $P(E_{t'}|C_t)$ is the probability of the set of paths from states where C holds to states where E holds. That is, it is the μ_m measure. In the previous section we showed that the path probabilities yield the conditional probabilities, now we must show that the conditional probability yields the path probability and thus the μ_m-measure for our leads-to formula. We have two cases to consider. First, if there is one time unit between t and t′, i.e. $t = t' + 1$, then

$$P(E_t|C_{t'}) = p', \tag{4.19}$$

where for all states, s where C holds, there is a transition to some state s′ where E holds such that $\mathcal{T}(s, s') \geqslant p'$.

In the second case, if $t' > t + 1$, then in the path from C to E there will be at least one other state s″ between the C and E states (called s and s′ as before). Let us say there are two time units between C and E. We can then rewrite our probability:

$$P(E_t|C_{t'}) = \sum_{s'' \in S} P(E_t, s''|C_{t'}) \tag{4.20}$$

$$= \sum_{s'' \in S} P(E_t|C_{t'}, s'') \times P(s''|C_{t'}), \tag{4.21}$$

where S is the set of all scenarios at t''. Since we have assumed the system to be first-order Markovian, we know that time t and t' are independent given t''. Thus,

$$P(E_t|C_{t'}) = \sum_{s'' \in S} P(E_t|s'') \times P(s''|C_{t'}). \tag{4.22}$$

We have now reduced the problem to the first case, and each of the conditional probabilities represent transitions from one time unit to the next, and may be replaced as such:

$$P(E_t|C_{t'}) = \sum_{s'' \in S} \mathcal{T}(s, s'') \times \mathcal{T}(s'', s'). \tag{4.23}$$

Thus this is the sum of the probabilities of the set of paths from s for which E is true in two time units. This is easily extended to any arbitrary t.

This corresponds to the probability of:

$$c \overset{\geqslant 1, \leqslant \infty}{\rightsquigarrow} e. \tag{4.24}$$

Then, since there are p' and p'' such that

$$c \overset{\geqslant 1, \leqslant \infty}{\rightsquigarrow} e \tag{4.25}$$

holds with probability p' and $F^{\leqslant \infty} e$ with probability p'', we can set $p = p'$ and satisfy both conditions (1) and (3) of LC.

Condition 2 $F^{\leqslant \infty}_{>0} c$

If $P(C_{t'}) > 0$ it means that if we represent the system as a proba-

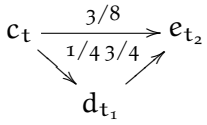
bilistic Kripke structure, it will be possible to reach a state where C is true with non-zero probability and thus K satisfies $F^{\leqslant\infty}_{>0}c$.

Thus if all conditions in SC are satisfied, so are those in LC and SC $\rightarrow$ LC. □

We have proven LC $\rightarrow$ SC and SC $\rightarrow$ LC and thus we conclude SC $\Leftrightarrow$ LC.

4.3.2 *Insignificant causes*

As we saw in section 2.3, many of these *prima facie* causes will not be the true causes of their effects, but will only appear to be so. For example, let us say that the relationships are as follows:



Here c can cause e at t_2 either directly, or through d: 1/4 of the time c will cause d at t_1, 3/8 of the time it will cause e at t_2 and 3/8 of the time it does nothing. We will assume that the probability of other things causing e is much lower than 1/4. How can we determine what causes e?

We could use Suppes' method, calling the earliest cause that can account for an effect genuine and all others spurious [124]. In this case, that would mean d is a spurious cause of e, as c comes earlier and accounts for e exactly as well as d does. We can also adjust this condition, as Suppes does, to account for the fact that in many cases $P(e|c \wedge d)$ will not exactly equal $P(e|c)$. Thus we have the notion of ε-spuriousness, stipulating

that the difference d makes to this probability is less than some small ε. However, we still have the problem that as long as we find one such c for which the relevant conditional probabilities are not exactly equal, d will be labeled an ε-spurious cause (not to mention the question of what value of ε is appropriate). What if there is a set C containing a thousand other such c for which d's contribution is greater than ε? Should we still call d spurious? If we recall our stated purposes (prediction and explanation), it would seem that we should not, as d will still be useful, at the very least, for predicting e. Secondly, in this case, d is actually a cause of e. In fact, it brings about e more quickly than c does (one time unit as opposed to two) and its direct influence is larger than c's (when c causes e directly, the probability of this is 3/8, while the probability of d causing e without any intermediaries is 3/4). Further, there may be other $c \in C$ that occur at some time between the times of d and e. Using Suppes' conditions, we only consider factors earlier than d that account better for the effect, and none of these would be considered. There is no scenario in which c could be considered spurious, as there are no earlier events that could remove its effectiveness in predicting e.[10]

Another approach, that of Eells, is to compute the average significance of a cause for its effect. That is, not to look for any single more powerful cause, but rather to measure overall how well the cause predicts the effect. As we saw in Section 2.3.3, Eells's average degree of causal significance (ADCS) (Equation (2.3.3)) addresses some of the issues raised with Suppes' definitions. First, Eells considers events that occur at any time prior to the effect (versus only those prior to the potential cause). Second, instead of looking for single causes that account better for the effect, Eells considers

10 This is remedied in part by Suppes' introduction of directness (Definition 2.3.2).

contexts (comprised of sets of all relevant factors, held fixed in all possible configurations). The result is not a partition into sets of genuine and spurious causes, but rather a quantity denoting how significant each cause is for its effect. This is desirable as in the case above with thousands of other factors, both c and d would have high significance in these contexts. Further, unlike Suppes, the value for this ADCS can be positive or negative, allowing for the possibility of negative causal significance (without defining negative causation separately). We face the problem that since there are a large number of such contexts 1) it is rare to have enough data to see them occur with enough frequency to be meaningful, and 2) testing all such contexts is a non-trivial computational task. If each background context occurs with nonzero probability, we will have 2^n such contexts, where n is the number of relevant factors. In our examples, where we may have data for thousands of genes, it is not possible to construct such a set of contexts (let alone to do so for each possible cause whose significance we aim to compute). We also have the same problem as with ε-spuriousness – determining which values of the ADCS should be considered significant.

Taking inspiration from both of these methods, we proceed as follows. First, we note that spuriousness implies falsity,[11] and will refrain from using this terminology. We do not intend to imply that the prima facie causes we abandon are necessarily false, only that they are of low import. Adopting the language of Eells, we will now discuss how to determine which of the *prima facie* causes are *insignificant*. When testing for insignificance of some particular c for some particular e in the context of our set of

11 The Merriam-Webster dictionary gives three definitions of spurious, one of which is: "outwardly similar or corresponding to something without having its genuine qualities" [1].

logical formulas and structure, we examine all the other states from which we may transition to e states. The primary idea is to determine whether these other states (with their associated labels) may tell us more about e than c does. Recall that states are labeled with formulas true within them. Initially this begins with all states being labeled with propositions, but states can be labeled with arbitrarily complex formulas, such as the leads-to ones described earlier. These other factors may occur at any time prior to e, but must themselves be prima facie causes of the effect.[12]

That is, we compute the average difference in probabilities for each prima facie cause of an effect in relation to all other prima facie causes of the effect. To test if a particular c is insignificant as a cause of a particular e, we begin with X being the set of prima facie causes of e. Then, for each $x \in X \setminus c$, we compute the predictive value of c in relation to x. We look at the probability of e after $c \wedge x$ versus after $\neg c \wedge x$. If these probabilities are very similar, then c might be an insignificant cause of e. As noted earlier, there may only be one such x, while there may be a number of other x's for which there is a large difference in the computed probabilities. Note further that the relationships between c and e and x and e have

12 We do not include factors that are independent of or negatively correlated with e (or which have probability zero). If e and a factor x are independent, then if we compute the difference $P(e|c \wedge x) - P(e|\neg c \wedge x)$, we will find this is equal to $P(e|c) - P(e|\neg c)$, and x does not change the significance of c for e. Then, we have the case when x is negatively correlated with e. Recall that when computing ε_{avg} we are testing whether there is another factor that better explains the effect. Intuitively, such x's should not be able to make c spurious. Thinking of what kinds of factors may account better for an effect, this could be due to a common cause of x, c and e (with perhaps c being less frequent than x), x causing both c and e, or the case of a causal chain, where c causes x and then x causes e. In all of those cases, though, x will be at least a *prima facie* cause of e, and will already be in the set tested pairwise with c. None of these can be the case, though, if e is negatively correlated with x. However, it is possible that a factor x that is negatively associated with an effect may cause it in conjunction with some other set of factors. Note though that if x causes e in conjunction with c, then the righthand side of the difference will still be quite small, as x cannot cause e in the absence of c, while $c \wedge x$ can together cause e. If x causes e in conjunction with some other factors (not including c) then it alone still cannot better account for e than c does.

associated windows of time in which either c or x may cause e. We omit the subscripts for ease below, but when calculating these probabilities these will be constraints on the instances of each formula that will be considered. For further details on these timings, see Chapter 5.2.1. With

$$\varepsilon_x(c, e) = P(e|c \wedge x) - P(e|\neg c \wedge x), \tag{4.26}$$

we compute:

$$\varepsilon_{avg}(c, e) = \frac{\sum_{x \in X \setminus c} \varepsilon_x(c, e)}{|X \setminus c|}. \tag{4.27}$$

For each prima facie cause, we have now assessed its average potency as a predictor of its effect. Finally, we use this ε_{avg} to determine c's significance.

Definition 4.3.2. A prima facie cause, c, of an effect, e, is an *ε-insignificant cause* of e if $|\varepsilon_{avg}(c, e)| < \varepsilon$.

The set X being comprised of all prima facie causes of e (aside from the c being tested) means that its components may occur at any time prior to e (they can be before, after, or at the same time as c), and may be causes of or caused by c. Some will turn out to be causally intermediate, that is, effects of c and causes of e. Such intermediate factors are not customarily held fixed, as it is assumed that doing so will lead to the erroneous conclusion that c does not cause e.[13] First, without background knowledge, we do not know what c's causes or effects might be (aside

13 One exception is that Cartwright suggests holding fixed effects of c that were not actually caused by c on that particular occasion ([11], p 95–96. Also see [28]). It is unclear how one might glean such information.

from, at this stage, the *prima facie* ones). It would not make sense to not condition on a factor that later turns out to be an insignificant effect of c, because we at first thought it might be a genuine effect. Secondly, identifying this set of actual causes and effects of c means that at one point we had to have some base level of background knowledge, or the argument becomes circular. We want to be parsimonious in our assumptions and choose to not suppose any knowledge of these actually relevant factors and thus take the set of prima facie causes, meaning the set of "possibly relevant" factors.[14]

What does $\varepsilon_{avg}(c, e)$ mean? If it is positive, this is saying that c being true has positive influence (proportional to the magnitude of ε_{avg}) on e. When ε_{avg} is negative, this means c not being true tells us more about e's occurrence than c being true does. Small values of ε_{avg} may mean that c is simply a statistical artifact. In other cases, c may indeed be a real cause, but one that only makes a small difference to e. In both cases, c will be discarded as a cause of e: in the first case because it is not a real cause and in the second because despite being a real cause, it makes little difference (and will not be particularly useful for prediction or explanation). Note that if ε_{avg} is exactly equal to zero, one cannot conclude that c neither increases nor decreases the probability of e, given any other prima facie cause of e. It is possible that c's positive and negative influences exactly canceled out.

Similarly, we do not require context unanimity (i.e. that c must raise or lower the probability of e in every context). Context unanimity is a

14 We will see examples later of such structures where a cause may be mistaken for being genuine based on it causing the effect by causing a cause of the effect. We are able to rule out this mistaken cause by conditioning on all prima facie causes of the effect (including the incorrect cause's direct effects).

common feature of probabilistic theories of causality, with Eells going so far as to say that "average effect is a sorry excuse for a causal concept".[15] However, we are not assembling the full set of contexts,[16] and will argue that it does not make sense to require that c raise e's probability with respect to each of the other prima facie causes of e. This condition would mean that if we have two causes of an effect where each is active just in the case that the other is not (e.g. an exclusive or), we would find neither to be a cause, as neither would raise the probability with respect to the other. Now, even if we could construct all of the background contexts, we would still argue against context unanimity. It does indeed seem sensible that if c's role changes, then there must be other factors that determine whether it will be positive or negative. However, similarly to our reasoning for averaging over other factors rather than looking for only one that made a cause insignificant, there may be only one such factor for which c does not raise the probability of e, but a multitude of others for which it does. In that case we can still be secure in the fact that c raises the probability of e, based on our averaging. For each case in which it has negative significance, there must be another positive case that offsets the amount of negative influence. Thus if c has an ε_{avg} that exceeds ε, then it means that it has an overall significantly positive role and if we do not know the other factors that make a difference, our best estimate is to say that c causes e. On the other hand, if $|\varepsilon_{avg}(c, e)| < \varepsilon$, then knowing c alone does not help us predict e. This is not to say that c does not cause e (we may find other conditions that in conjunction with c are significantly

15 [28], p 54.

16 This is an ideal that may someday be possible, but is currently unachievable. It is not clear that if all causally relevant factors were included and we had all the data in the world, we would find each context occurring more than once. As each context is specified by more features, they become narrowed down to the point of fully specifying individuals.

positive causes), but rather that c alone is not significant for e and does not give us enough evidence to say that e will occur if we know c. This is another reason why we eschew the language of spuriousness as, among other possible explanations, an ε-insignificant cause may be part of a significant cause, and it is nonsensical to say that a significant cause may be comprised of spurious causes. However, it is possible that causes that are insignificant on their own, are, together, significant.

As a simple example of testing for insignificance, let us say we have data on smoking (S), yellow stained fingers (YF) and the incidence of lung cancer (LC) in people who smoke and have stained fingers. Let us now assume that smoking and staining of fingers both occur prior to the development of lung cancer. Then, we are likely to find both S and YF as prima facie causes of LC. However, if we now look at $P(LC|S \wedge YF) - P(LC|S \wedge \neg YF)$, testing YF's contribution to LC, we will likely find this difference to be nearly zero (accounting for the possibility that there may be some other reason for stained fingers that is also a cause of lung cancer). This scenario is shown in figure 4.2. In that structure (where transition probabilities from a state sometimes add to less than one, indicating states not shown, but none of these hidden states are labeled with S, YF or LC), we find:

$$\varepsilon_S(YF, LC) = P(LC|S \wedge YF) - P(LC|S \wedge \neg YF) = 0.85 - 0.75 = 0.10, \text{ and}$$
$$\varepsilon_{YF}(S, LC) = P(LC|S \wedge YF) - P(LC|\neg S \wedge YF) = 0.85 - 0.01 = 0.84.$$

We should thus call YF an ε-insignificant cause of LC as we only have two possible causes and $\varepsilon_S(YF, LC)$ is very small. We see that $\varepsilon_{YF}(S, LC)$

is much higher and it is not insignificant. This example leads us to the question: "What should we call non-insignificant prima facie causes?"

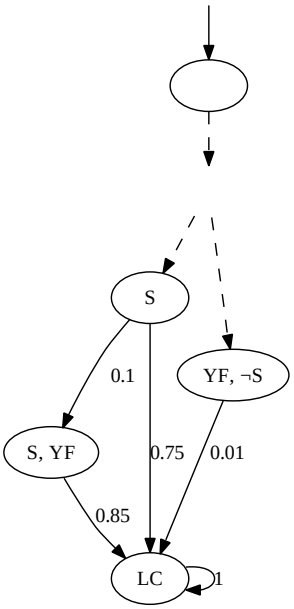


Figure 4.2.: Smoking (S), Yellow Fingers (YF) and Lung Cancer (LC).

4.3.3 *Just so causes*

It is possible that some of our insignificant causes will actually be genuine and that some of the non-insignificant causes will actually be spurious. This is due to a number of factors, such as the choice of ε (we could be too strict or too lax) as well as the sample from which we calculate the probabilities not being representative of the actual distribution of the data. What, then, are we claiming? If a prima facie cause is not ε-insignificant, what is it? We have good reason to believe that these are genuine and that the ε-insignificant ones are spurious, but at the moment, without further investigation, the degree of this belief is proportional to our confidence

in the completeness of our structure and the statistical methods (to be described later) for determining the best value for ε. For that reason, we refer to the prima facie causes that are not ε-insignificant as "just so" (or ε-significant) causes. These are useful for prediction and could potentially also be used for influencing and controlling the system. However, in order to say whether any of these are genuine causes – more than just "beyond a reasonable doubt" – we would need to conduct such experiments where we attempt to alter the system's behavior by altering the cause. In some areas this can be carried out, but there are other situations where it is not possible to conduct just any controlled experiment we come up with (due to ethical, financial or other restrictions). In these cases we have still made valuable inferences if we have identified these just so causes, and may still continue collecting observational data to see if they may be refuted.

4.3.4 *Genuine Causes*

Thus far we have provided primarily negative definitions, focusing on what *cannot* be inferred or determining what is *not* truly causal. We now turn our attention to positive claims: what are genuine causes and under what conditions are just so causes genuine? In one sense, if we have the correct underlying structure, it does not matter what is and is not causal. We can see exactly how the system works, what the probabilities of any possible transitions are and thus decide how we can change as well as predict the system's behavior in the future. After all, is it not our goal to understand precisely these inner workings from observations? This position is perhaps unsatisfying philosophically, as we are then claiming

that there is no need for the notion of causality in a system that is correctly specified. Note however that in most cases we will have only data from which to determine whether various logical formulas are true – we will not be given or attempt to infer this underlying structure.

Assumptions

While we do want to understand how the system works, we also recall that the transitions do not need to be causal, since a transition that leads to the occurrence of an effect can in fact decrease its probability. All is not lost, as we are also seeking a compact representation of the system using logical formulas. If we want to explain the occurrence of an effect, we do not necessarily want to fully specify our system, including information that is only marginally relevant.[17] We want only the parts of it that tell us the most about the effect. For example, if we have a structure such as in figure 4.2, what is the best way to predict whether someone will get lung cancer or explain the occurrence of lung cancer? Then, when is that explanation the genuine cause? That is, what needs to be true about the structure in order for the best explanation to be the genuine one?

First, we note there can be a number of genuine causes of an effect. Genuine does not mean that something is "the" cause or that "is x a genuine cause of y" is a true or false question. Rather, particularly in the case of probabilistic causality, we think of this as being along a continuum, with genuine causes as those which are most descriptive of the system. There may be weak genuine causes as well as strong ones, and our statistical tests will rarely divide the possible causes evenly into sets of

17 As will become clear when we discuss inference, it is easier to find the formulas best characterizing the system than it is to infer the entire structure.

spurious and genuine causes. It is much more likely that our statistical values will be continuous, and we will have to choose a threshold at which we will call something significant. Second, the primary idea is that a genuine cause is one that can be manipulated in order to bring about the effect. That is, in other cases we may be observing an anomaly but if the cause is genuine, then forcing the cause to be true should make the effect true (with the associated probability). What assumptions do we need to make about the system in order to make such a statement?

From our prior definitions we know that just so causes are candidates for genuine causality (though again it is possible that there are others if we have chosen too high a value for ε). Remember that we are currently concerned with describing causal relationships relative to some structure, K, using logical formulas (conditions for successful inference as well as discussion of how to infer the relationships from data will be dealt with in the next chapter). As before, we assume that this structure is the correct one underlying the system. This assumption implies that any states reached with non-zero transition probabilities in K are indeed possible states of the system, and that the transition probabilities are the actual ones of the system. When we say "system" we do not necessarily mean that K contains every part of the whole. If we were looking at the human body, K could be a model of the workings of the elbow or digestive tract, and not the entire body. However, we cannot simply amputate the elbow from the body and understand how it works, we must keep some information from the larger whole in order to create a self contained system for study. What does it mean for a system to be "self contained" and why is this desirable?

This means that all common causes are included. If they are not, then the just so causes we find may simply be indicators for genuine causes. For example, take the case of smoking, yellowed fingers and lung cancer. We can likely agree that smoking (in some way) causes both yellowed fingers and lung cancer, and that the two effects play no causal role in either smoking or with each other. Then, suppose we have only measured yellowed fingers and lung cancer, and have done so over time. It is possible then that someone who has smoked enough to yellow their fingers will be more likely to develop lung cancer than if this was not the case. It is also possible that the yellowing will show up before the cancer. In that case, assuming there were some other factors also tested, we would find that yellow fingers just so cause lung cancer. While this is not the actual causal relationship, yellowed fingers could still be a useful predictor of lung cancer and we may be able to use this information as an indicator of the stage of the disease.

The primary question that we are led to is the following: How can we know that we have enumerated all common causes? This question is at the crux of the issue, as knowing that common causes are included means being able to take pairs of formulas and say whether they have a common cause. Problematically, this approach implies for the theory to have background knowledge – otherwise the algorithms would have no starting point. Indeed, Nancy Cartwright summarizes this situation "no causes in, no causes out" [11]. This view is shared by Pearl [18] and SGS among others [127, 109]. But how can we acquire this "old" causal knowledge in the first place? One strategy Cartwright gives is to conduct perfectly

18 Pearl suggested "Occam's razor in, some causes out."([102], p 60) Cartwright is in fact against this simplicity assumption.([11], p 72)

randomized experiments, thus controlling for all unknown factors. The fact that we only need to include common causes, and not every other cause of the effect is also touted as greatly simplifying matters [99], but it still does not address how we can build this background causal knowledge. There must be a level of causes that are not subject to these inference rules, on which all other inferences can be based. This is dissatisfying, as there is no well defined base case for this recursion. Must everything come from physical laws? If not, it seems that there must be some matter of opinion or belief in the causal relationships. That is, if we cannot build these relationships from some lower level laws, it must be that asserting there is no common cause of two factors is really saying that I *believe* that there is no common cause of them or that I have *confidence* that the experiment was conducted such that all common causes were included. We suggest that we be more explicit about the amount of intuition and judgement required. In many cases we have observational data, whose collection we did not control. It is obvious that if we do not measure smoking, we will not discover that smoking causes lung cancer, but whether yellow fingers are genuine causes of lung cancer (as that data set would suggest), depends on whether we already know enough about lung cancer to know that it and yellowed fingers have a common cause. If we already have that information, it seems unlikely that we would conduct such a study, but perhaps only maintain a strong belief that such is the case while continuing to examine alternatives.

Thus, a just so cause *is* genuine in the case where all of the outlined assumptions hold (namely that all common causes are included, the structure is representative of the system and, if data is used, a formula satisfied by the data will be satisfied by the structure). Our *belief* in

whether a cause is genuine, in the case where it is not certain that the assumptions hold, should be proportional to how much we believe that the assumptions are true.

4.4 DIFFICULT CASES

In this section we will discuss common counterexamples for theories of causality. Most of these are handled correctly by our definitions but a few are not.

4.4.1 *Determinism*

Let us revisit our example of Chapter 2.3.2. Here we have Bob and Susie, holding rocks that they intend to throw at a glass bottle. In the original problem, Bob is standing a little closer to the bottle, so Susie aims and throws earlier and their rocks hit the bottle simultaneously. We do not currently consider spatial information or include this in the logical formulas, but we can represent the system as shown in figure 4.3a. Once they decide to play the game, Bob throws first, then Susie throws at the next time unit, both of their rocks hit the glass and then the glass breaks and stays broken forever. In this deterministic case, where all state transitions have probability one, we would find that neither Susie nor Bob causes the rock to break, as the probability of the glass breaking in this system is one. Thus, we would need more information on the probability of the glass breaking on its own (perhaps due to a gust of wind) in order

to say whether either of them or their game causes the breaking of the glass.

This is perhaps a disappointing result, in that one would think that if one throws a rock and breaks a glass with probability one, then throwing a rock (or, at the very least, the rock hitting the glass) should cause the glass to break. However, we can also look at this scenario in another way. If we had no background knowledge or understanding of the relationship between rocks and glasses, how would we understand this structure? We would see that if we enter this system, the outcome will inevitably be a broken glass. There are no instances of a glass not being broken, so we cannot tell what difference Bob or Susie make to the breaking. This example illustrates the situation when we know that we do not know: we cannot postulate any causes (even prima facie ones) of the glass breaking. Without further information, we can only say that it is a property of the system that the glass always breaks. Whether this result is acceptable or not depends on how one feels about determinism, but it is not intrinsically incorrect and it may be argued that it is in fact the correct result.

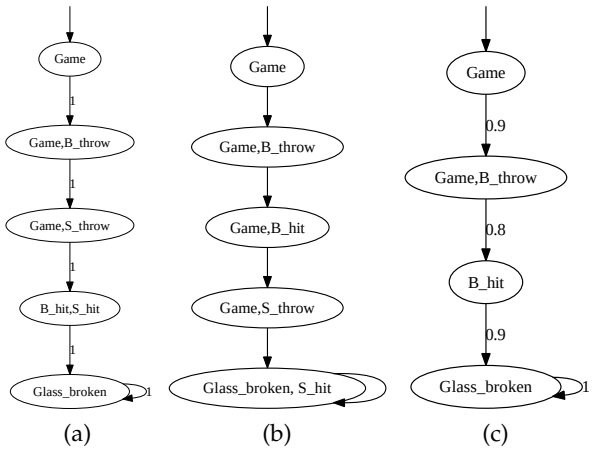


Figure 4.3.: Bob and Susie throwing rocks at a glass bottle.

4.4.2 *Overdetermination*

Continuing with Bob and Susie, let us alter the example slightly. In this case we will remove the deterministic element, and will try to find which of them breaks the glass. Since we do not give any preference to earlier explanations over later ones, we will not give any weight to the fact that Bob throws his rock earlier. Thus, both will be equally strong causes of the glass breaking, with the strength of these claims proportional to the probability with which the glass breaks when each throws. If each is as likely to throw their rock and, once thrown, equally likely to break the glass, then they are equally possible causes of the glass breaking. Unlike the results from approaches based on counterfactuals, we do not find that neither causes the glass to break.

This looks promising, so let us continue with a trickier case: preemption. We will alter the example so that Bob's rock always hits the glass before Susie's does. That is, when Susie's rock hits, the glass is already broken (see figure 4.3b). In this case, Bob's throw and his rock hitting the glass preempts Susie's from causing the glass to break. Now, if we are not looking for something that causes a glass to break within a specific amount of time, but rather any earlier cause, we would find both Bob's hitting the glass and Susie's throw equally causal. However, we can define the propositions of the system differently, looking just at whether they throw or hit, and thus find that only Bob makes a difference to the glass breaking. In any case, the important feature of this new system is that the glass is already broken when Susie's rock hits it, so we will correctly find that Susie's rock hitting the glass makes no difference to it breaking.

4.4.3 *Causal Chains*

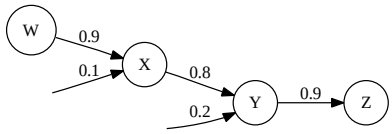
Our previous examples with Bob and Susie are also instances of causal chains. In this case, it is normally said that if we have the chain $X \to Y \to Z$, where X causes Y and Y causes Z, then X should be a cause of Z (and not a spurious one). Let us just focus on Bob for a moment. Now, take a chain going from the beginning of the game to Bob's throw to the glass breaking, and assume this is the only way the game can lead to the broken glass (since our transitions are not deterministic, there will be others, but we will assume they are to states that have no path to "glass broken") (figure 4.3c). We will further assume that the probability of a broken glass, outside this game, is very small. Then, the Game (G), Bob's Throw (T) and Bob's rock hitting the glass (H) will be prima facie causes of the glass breaking (B). To determine if H is insignificant, we can calculate $\varepsilon_{avg}(H, B)$ as follows:

$$\begin{aligned}\varepsilon_{avg}(H, B) =& (P(B|H \wedge G) - P(B|\neg H \wedge G) + P(B|H \wedge T) - P(B|\neg H \wedge T))/2 \\ &= \frac{0.9 - 0 + 0.9 - 0}{2} \\ &= 0.9.\end{aligned}$$

Note that this computation simply produces the probability of B given H, since neither T nor G can cause B except through H. Thus ε_{avg} for T and G will be undefined, as H also does not occur except through G and T and probabilities such as $P(B|\neg G \wedge H)$ will be undefined. We will not find T and G insignificant, but rather we cannot assess their roles. However, H is not insignificant, as we would correctly assume. Note that

in this example we did not include the time between the cause and effect in order to simplify matters, we only assumed that the cause is earlier than the effect.

Let us again alter this example and see what happens in the case where it is possible for members of the causal chain to occur outside the chain. That is, what happens if it is possible for Bob's rock to hit the bottle without being thrown by Bob. This does not make much sense in this example, but as a general case, it is quite possible that an element of the chain can have multiple causes. Take the following case:



We will again assume that the probability of Z is very low. Then as before, W, X and Y are prima facie causes of Z, because $P(Z|Y) = 0.9$, $P(Z|X) = 0.72$ and $P(Z|W) = 0.648$. Then,

$$\begin{aligned}
\varepsilon_{avg}(Y, Z) &= (P(Z|Y \wedge X) - P(Z|\neg Y \wedge X) + P(Z|Y \wedge W) - P(Z|\neg Y \wedge W))/2 \\
&= \frac{0.9 - 0 + 0.9 - 0}{2} = 0.9, \\
\varepsilon_{avg}(X, Z) &= (P(Z|X \wedge Y) - P(Z|\neg X \wedge Y) + P(Z|X \wedge W) - P(Z|\neg X \wedge W))/2 \\
&= \frac{0.9 - 0.9 + 0.72 - 0.09}{2} = 0.63, \text{and} \\
\varepsilon_{avg}(W, Z) &= (P(Z|W \wedge Y) - P(Z|\neg W \wedge Y) + P(Z|W \wedge X) - P(Z|\neg W \wedge X))/2 \\
&= \frac{0.9 - 0.9 + 0.72 - 0.72}{2} = 0.
\end{aligned}$$

Thus if these are the only factors in our system, we have correctly ranked Y as being the most important and X as being second in importance, which is consistent with how we would think about the problem. If you

could find out whether X is true or whether Y is true, which is more useful to know? The answer is Y. However, we do have one problem here, in that *W* is seemingly irrelevant for Z. This is actually not a problem, if we remember our goals. In the next chapter we will not be given this structure, we will have to find it. Now, *Y* having higher relevance for Z than X will give us good reason to suspect precisely the structure in this diagram. Further, we will find the relationships between *W* and X and between X and Y, recovering the true relationships of the system. Finally, if we had a number of other factors, outside this chain, against which to test *W*'s relevance, we might find it has more value to the prediction of Z. Nevertheless, we are not disturbed by the idea of *W* being insignificant for Z, even if some consider it a cause of X. It is perfectly fine to say that *W* does not cause Z, but rather causes X, which in turn causes Y, which causes Z. In fact this is a much truer representation of the system.

4.4.4 *Transitivity*

The problem of the previous section leads us directly to the question of 1) whether causality is transitive and 2) whether transitivity is captured in our representation of causal relationships. We will not address the first question, but note that as we saw earlier when looking at the counterfactual approach to causality (Chapter 2.2.2), we can arrive at anomalous results when we allow that if C causes D and D causes E, that C causes E.[19] In general our definitions, like Suppes' [124], are not transitive, but we do indirectly allow transitivity in other ways. Remember that we allow multiple transitions between cause and effect. This means that if we are

19 For further discussion on the transitivity of causation see [28].

looking at whether some C causes some E in *at least* 2 time units, then there will be at least 2 states on the path from C to E. These states may be labeled with causes of E and in fact we may see that C causes E only because it does so through those states. In the previous example, imagine if the lower bound was, instead of one time unit, three time units. Then *W* would be the only cause of Z, and we would be, indirectly, allowing transitivity. Anomalous results caused by this approach can be remedied by carrying out the analysis in a smaller time window.

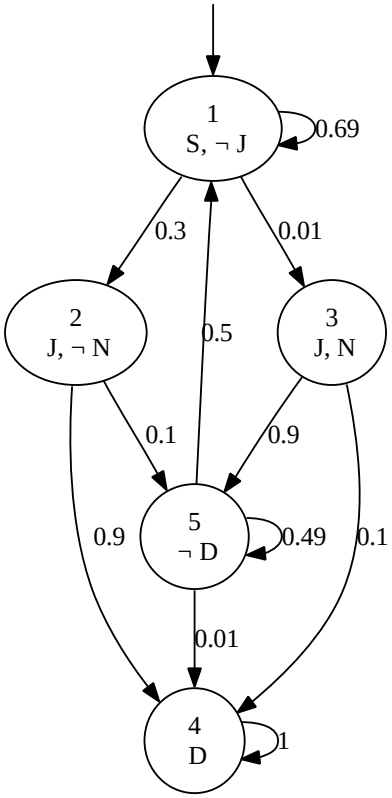


Figure 4.4.: Example structure containing states with their labels and transitions with their associated probabilities. s_1 is the initial state. This is an altered version of figure 4.1.

4.4.5 *Cycles*

One side effect of the structures we use is that due to the total transition function, we must allow cycles in the graph (either that or an infinite number of states, which we do not allow). Inference methods, such as those based on Bayesian networks, generally specify that the structure must be a DAG (directed acyclic graph), but this is limiting in that the probabilities in the graph only convey the probability of the effect happening at some point after the cause. But in many cases it is better to think of this as an ongoing situation, where there may be a number of opportunities for the cause to occur and to bring about the effect (i.e. in a window of time after the cause, where we may want to vary this window to test a number of possibilities). For example, take the structure in 4.1, where the propositions have the following interpretations: S is "being suicidal", D is "being dead" (where ¬D means alive), J is "jumps from building" and N is "net below area being jumped from" (¬N means there is no such net). In the figure shown, we have specified that a person who is dead remains dead but that one who survives an attempted suicide somehow becomes immortal and remains not dead forever. If we remove all three of the self loops, we end up with a familiar DAG and avoid such problems. However, it would be better to augment the graph as follows (shown in figure 4.4), resulting in a graph with cycles that behaves as expected. We should keep the self loop at s_4, as we do not want to allow resurrection in this example. At s_5 we should add an edge with non-zero probability to s_1, meaning that there is some probability that a person who survives a suicide attempt will become suicidal again. This

means changing the probability on the self loop, to make sure the sum of the transition probabilities at s_5 add to one. We should also allow a transition from s_5 to s_4, in consideration of the possibility that the person may die of other causes. Our goal at the moment is not to reconstruct the most accurate structure of suicide by jumping from buildings, but rather illustrate the expressiveness of PCTL as opposed to other frameworks as well as the desirability of cycles in a causal structure.

Then, if we make no other modifications, we see that given an infinite amount of time, a person who becomes suicidal will eventually succeed in their attempts. This is a result we could not achieve without allowing cycles in our graph. However, it is not particularly useful to assume that a person has infinite time to kill himself. Thus we can add an upper bound and ask whether a person who becomes suicidal will successfully kill themselves within x units of time. Such a question is easily represented and posed in the framework presented, but has no analogue in the graphical models previously described.

5

INFERRING CAUSALITY

In the previous chapter we defined the kinds of causes we will aim to identify. Now we will discuss the details of how to go about testing these in a set of data. We will begin by considering the set of causal hypotheses, the format of the data and the satisfaction of formula in time series data before discussing significance testing and in particular the choice of threshold for determining the significance of causes. Finally, we will examine theoretical issues such as the complexity of the testing procedures.

5.1 TESTING PRIMA FACIE CAUSALITY

5.1.1 *The set of hypotheses*

The hypotheses are a set of formulas representing causal relationships. Each of these is of the form:

$$c \rightsquigarrow^{\geqslant r, \leqslant s} e, \tag{5.1}$$

where c and e are PCTL state formulas, $1 \leqslant r \leqslant s \leqslant \infty$ and $r \neq \infty$. These will be tested to find those that meet the conditions for *prima facie* causality, as defined in Chapter 4. To form this set, the simplest case is

when we have some knowledge of the system and either explicitly state the formulas that may be interesting or use our background information to generate the set. For instance, we may have data for a set of neurons firing over time, where we know the time window between one neuron firing and it causing another to fire, and we are only interested in simple relationships between individual neurons. In the next type of case, we may have information on the timing, but not about the complexity of the relationships. Here we could choose to generate increasingly large formulas and stop at some predefined size, or determine this threshold based on whether or not the quality of causal relationships is continuing to increase. In the case of limited data, we could begin by determining what types of formulas may be found (at satisfactory levels of significance) using this data based on formula size, length of the time series, and the number of variables. Finally, when the related timing is unknown, we can simply generate formulas with various associated time windows, testing which are most significant.

5.1.2 *The set of data*

Testing the set of hypotheses for *prima facie* causality means testing whether the relationships are satisfied by the data, and with what probabilities. We must first relate the observational data to logical formulas. We assume that the data consists of a series of time points, with measurements of variables or the occurrence of events at each. For instance, a subset of one data set (which may have any number of time points and variables) might look like:

	t_1	t_2	t_3
a	1	0	1
b	0	0	1
c	1	1	0

Here we have observations of three variables at three time points. When specifying the logical formula denoting causal relationships in this system, the set $\{a, b, c\}$ will contain the atomic propositions. In this case, occurrence of a proposition is denoted by 1, and non-occurrence by 0. We see that at t_1, a and c are true. Another way of describing this is to say that the system is in a state where a and c are true. Each observation yields a state the system can occupy, and the temporal order of these observations shows possible transitions between states. We assume that there is some underlying structure, which may be very complex and include thousands of states, and we are observing its behavior over time.

Note that we can have two types of data. In the first, we observe one long sequence of times. In that case, we see one of many partial runs of the system. The second type is a group of (usually shorter) observation sequences (also called traces in model checking). Cases like this arise in medical domains, where we have sets of patients observed over time. While one long run may initially seem equivalent to many shorter runs, there are some important distinctions. To understand these, let us consider an example. Assume the underlying structure is as shown in figure 5.1. Say we then observe the sequence $P, Q, S, T, ..., T, S, ...$ and do not know the underlying model (as this is normally the case). If we see only this one trace (beginning from the start state, s_1), we will never see the transition from s_1 to s_3 (i.e. P, R) and will not know that it is possible. However, if we have a large set of short traces, then as the size

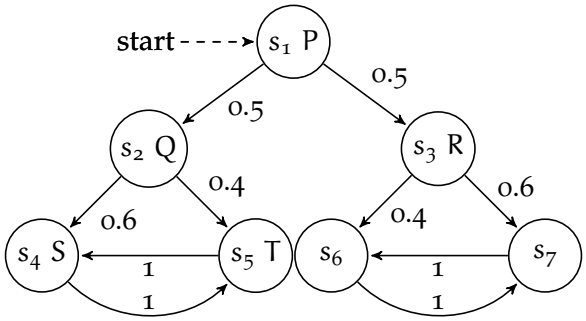


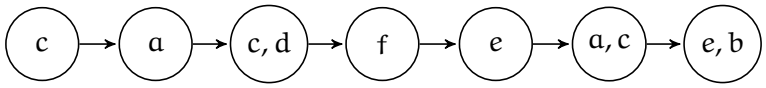
Figure 5.1.: Example of a probabilistic structure that might be observed.

of this set increases, we will get closer to observing the actual transition probabilities. That is, half the traces will begin with P, Q and the other half with P, R.

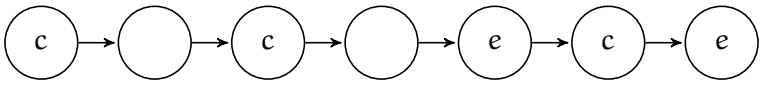
Note that in practice many systems, such as biological systems, have cyclic patterns that will repeat over a long trace. That is, these systems can be modeled as recurrent Markov processes. While we may only observe the start state once, we will see other states and transitions repeated multiple times. With a recurrent Markov process, we can then infer properties from one long trace. However, when the system is non-recurrent, inference may require a set of traces sampled from a population. If the properties of interest are related to the first few timepoints and do not occur again, then we will not be able to infer these from a single trace. In cases where one has control over the data collected, it is worth noting the differences between the two types.

5.1.3 *Intuition behind procedure*

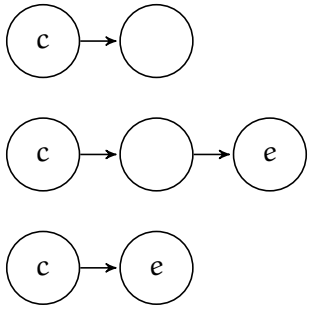
Before discussing how to test logical formulas in data, let us consider an example to see the general idea behind this. Say we are testing $c \leadsto^{\geqslant 1, \leqslant 2}_{\geqslant p} e$ for some p and we observe the sequence c, a, cd, f, e, ac, eb. We may represent this as:



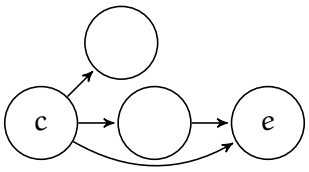
Now we must determine whether the probability of the formula, given what we have observed, is at least p. Thus the part of this sequence that we are interested in is:



Since we do not know the underlying structure and are not trying to infer it (consider that for a set of 1000 genes that are only on or off, there are 2^{1000} possible unique states and there may be multiple states with the same labels) we observe all instances of c as possibly leading to e, regardless of what the current underlying state is (there may be no path, or there could even be a deterministic path). At this point, we consider any timepoint (and thus state) where c is true, to be identical. That means that the above sequence looks like the following set of paths:



and will seem to be generated by the following (partial) structure:



The probability of the leads-to formula we are testing is the probability of the set of paths leading from c to e (within the specified time limit), which is defined by how frequently we observe those paths from c. Thus when we have a trace of times labeled with c and e, and a formula $c \rightsquigarrow^{\geqslant 1, \leqslant 2} e$, the probability of this formula is the number of time points labeled with c where e also holds in at least one and fewer than two time units, divided by the number of time points labeled with c. In this example, the probability is estimated to be $2/3$. The trace is then said to satisfy the formula $c \rightsquigarrow^{\geqslant 1, \leqslant 2}_{\geqslant p} e$ if $p \leqslant 2/3$.

5.1.4 *Satisfaction of a formula*

Testing the set of hypotheses for *prima facie* causality means testing whether each relationship is satisfied by the data, and with what probability. Now we move to the general case, where we begin with either one long time series or a set of shorter ones, and a set of formulas to be tested.[1] The satisfaction and probability of PCTL formulas relative to a trace consisting of a sequence of ordered timepoints (with either measurements at every point in time, for some granularity of measurement, or time indices of the measurements such that we can compute the time

1 For an introduction to the problem of runtime verification, see [73].

between two measurements) for any PCTL formula is as follows.[2] Note that our approach differs from others using PCTL [14], as we must deal with 1) long traces that cannot be broken up into shorter traces based on knowledge of the start state 2) short traces whose first observations vary and are not indicative of the start state. Thus, we cannot use the usual approach of computing the probability of a formula by finding the number of traces satisfying the formula.

We assume that propositions are events that either occur or do not or are otherwise known to be either true or false at every time point along the trace. Each timepoint is considered to be initially labeled with the atomic propositions true at that time. Assume that t is a time instant in the observed trace.[3] From these propositions, we may define more complex state and path formulas, which describe properties true at a particular instant (a state) or for a sequence of times (a path).

1. Each atomic proposition is a state formula.

An atomic proposition is true at t if it is in $L(t)$ (the labels of t).

2. If g and h are state formulas, so are $\neg g$, $g \wedge h$, $g \vee h$, and $g \rightarrow h$.

If a time point t does not satisfy g, then $\neg g$ is true at t. If both g and h are true at t, then $g \wedge h$ is true at t. If g is true or h is true at t, then $g \vee h$ is true at t, and if $\neg g$ is true at t or h is true at t, then $g \rightarrow h$ is true at t.

3. If f and g are state formulas, and $0 \leqslant r \leqslant s \leqslant \infty$ with $r \neq \infty$, $f \mathcal{U}^{\geqslant r, \leqslant s} g$ and $f \mathcal{U}^{\geqslant r, \leqslant s} g$ are path formulas.

2 For the case of a set of traces, the frequencies below simply refer to the frequencies in the combined set of time points.

3 In the unlikely event that a structure is given, this procedure is unnecessary and one may proceed with the algorithms of Hansson and Jonsson [49], with the modified version of leads-to. However, we remind the reader that it is unlikely to begin with a structure, and attempting to infer one may introduce further errors.

The "until" path formula $f\mathcal{U}^{\geqslant r,\leqslant s}g$ is true for a sequence of times beginning at time t if there is some $r \leqslant i \leqslant s$ such that g is true at $t+i$ and $\forall j : 0 \leqslant j < i$, f is true at $t+j$. The "unless" path formula $f\mathcal{U}^{\geqslant r,\leqslant s}g$ is true for a sequence of times beginning at time t if either $f\mathcal{U}^{\geqslant r,\leqslant s}g$ is true beginning at time t, or $\forall j : 0 \leqslant j \leqslant s$, f is true at $t+j$.

4. If f and g are state formulas, then $f \rightsquigarrow^{\geqslant r,\leqslant s} g$, where $0 \leqslant r \leqslant s \leqslant \infty$ and $r \neq \infty$ is a path formula.

We now treat leads-to formulas separately. Recall that leads-to was originally defined using $F^{\geqslant r,\leqslant s}e$, where the associated probability of the leads-to is that of the F part of the formula. Thus the computed probability will be that of e occurring within the window r–s after any timepoint, while we actually want the probability of e in the window r–s after c. Note that when checking formulas in a structure, we do not have this difficulty, as we are computing the probabilities relative to particular states. However, when we must check formulas in traces (when a model is not given or will not be inferred), we do not know which state a timepoint corresponds to and thus we can only compute the probability relative to a trace. Thus, the formula $f \rightsquigarrow^{\geqslant r,\leqslant s} g$ is true for a sequence of times beginning at time t if f is true at t and there is some i, where $r \leqslant i \leqslant s$, such that g is true at $t+i$. Note that when $r = 0$, this reduces to the usual case of leads-to with no lower bound.

5. If f is a path formula and $0 \leqslant p \leqslant 1$, $[f]_{\geqslant p}$ and $[f]_{>p}$ are state formulas.

The probabilities here are in fact conditional probabilities. There are two primary cases. For $[f\mathcal{U}^{\geqslant r,\leqslant s}g]_{\geqslant p}$ the probability p' associated with the

data is estimated as the number of time points that begin paths satisfying $f\mathcal{U}^{\geqslant r,\leqslant s}g$ divided by the number of time points labeled $f \vee g$. The formula $[f\mathcal{U}^{\geqslant r,\leqslant s}g]_{\geqslant p}$ is satisfied by the trace or set of traces if $p' \geqslant p$. In the case of a $\mathcal{U}$ formula, the probability is estimated the same way as for the preceding case, except that we consider the timepoints beginning paths satisfying $f\mathcal{U}^{\geqslant r,\leqslant s}g$ (which includes paths where f holds for s time units, without g later holding). For a leads-to formula, $h = f \rightsquigarrow^{\geqslant r,\leqslant s} g$, the probability is estimated as the number of time points that begin sequences of times labeled with h, divided by the number of time points labeled with f. Thus, the probability of $f \rightsquigarrow^{\geqslant r,\leqslant s} g$ is the probability, given that f is true, that g will be true in between r and s units of time.

See Appendix D.1 for algorithms.

Let us see that this formulation yields our desired result. That is, let $p = P(g_{t'}|f_t)$, where $t + r \leqslant t' \leqslant t + s$. Dropping the time subscripts for the moment, by definition we have:

$$P(g|f) = \frac{P(g \wedge f)}{P(f)}.$$

Recall that the probabilities are in fact frequencies, so that the probability $P(x)$ of some formula x, is the number of time points labeled with x divided by all time points. Thus, using $\#x$ to denote the number of time points with some label x, and T to denote the total number of timepoints, we find:

$$\begin{aligned} P(g|f) &= \frac{(\#g \wedge f)/T}{\#f/T} \\ &= \frac{(\#g \wedge f)}{\#f}. \end{aligned}$$

Thus, we return to our previous statement, which is that the probability of such a formula is the number of states beginning paths satisfying the leads-to formula, divided by the number of states satisfying f.

The *prima facie* causes are those in the set of hypotheses where the associated probability, computed from the data, is greater than the probability of the effect alone and where the relationship satisfies our other conditions – namely that the time lag is such that c is prior to e.

5.2 TESTING FOR SIGNIFICANCE

In the previous chapter we defined a new value, ε_{avg}, that indicates how significant a *prima facie* cause is for its effect. We then defined that a cause is ε-insignificant if its $|\varepsilon_{avg}| < \varepsilon$, for some small value of ε. Now we will discuss the computation of ε_{avg} as well as how to find an appropriate value for ε.

5.2.1 *Computing* ε_{avg}

Let us recall the definition for ε_{avg}. With X being the set of *prima facie* causes of e, we compute:

$$\varepsilon_{avg}(c, e) = \frac{\sum_{x \in X \setminus c} \varepsilon_x(c, e)}{|X \setminus c|}, \tag{5.2}$$

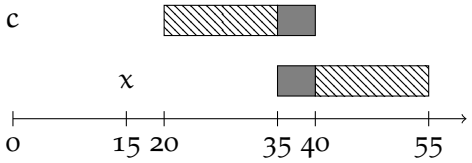
where:

$$\varepsilon_x(c, e) = P(e|c \wedge x) - P(e|\neg c \wedge x). \tag{5.3}$$

Recall that each c and x are of the form $c \leadsto^{\geqslant s, \leqslant t} e$ and $x \leadsto^{\geqslant s', \leqslant t'} e$. That means that $c \wedge x$ refers to c and x being true such that e could be caused in the appropriate intervals. That is, $P(e|c \wedge x)$ means $P(e_A|c_B \wedge x_C)$ where the time subscripts are constrained by:

$$B + s \leqslant A \leqslant B + t, \text{ and}$$
$$C + s' \leqslant A \leqslant C + t'.$$

As before, we compute this probability with respect to a probabilistic structure or set of data using the satisfaction rules described earlier. For example, if part of the observed sequence is c at time 0 and x at time 15, where $s = s' = 20$ and $t = t' = 40$, then e must occur in the overlap of these windows, shown in gray below.



Thus, this will be considered an instance of $(c \wedge x) \leadsto e$ if there is an observation e_A such that: $20 \leqslant A \leqslant 40$ and $35 \leqslant A \leqslant 55$. That is, there must be an instance of e between times 35 and 40. If e was true at $A = 10$, then only c would have been true before e, while if e was true at $A = 50$, then c's time window to cause e would be over.

Recall that the probabilities here come from frequencies of occurrence in the data. The computation is exactly as described for leads-to formulas in the previous section. Thus, remembering that there are time subscripts and constraints on the time of e:

$$P(e|c \wedge x) = \frac{\#(e \wedge c \wedge x)}{\#(c \wedge x)}, \tag{5.4}$$

and

$$P(e|\neg c \wedge x) = \frac{\#(e \wedge \neg c \wedge x)}{\#(\neg c \wedge x)}, \tag{5.5}$$

where these refer to the number of paths where e holds after $c \wedge x$ (or $\neg c \wedge x$) holds, in the appropriate time window, divided by the number of paths where $c \wedge x$ (or $\neg c \wedge x$) holds.

We now give one algorithm for computing

$$\varepsilon_x(c, e) = P(e|c \wedge x) - P(e|\neg c \wedge x) \tag{5.6}$$

relative to a trace T. As before, we note that c and x have corresponding formulas $c \leadsto^{\geqslant r, \leqslant s} e$ and $x \leadsto^{\geqslant r', \leqslant s'} e$. We assume all times satisfying c, x and e are already labeled. Then $c \wedge x$ refers to c and x holding such that either could be a cause of x. Thus the primary task of the algorithm is to identify instances of $c \wedge x$ that fit these criteria, and then to identify instances of e that fall in the overlap of the time windows from these instances. Similarly, for $\neg c \wedge x$, we find instances of x where there is no overlapping window with an instance of c.

Algorithm 5.1 $\varepsilon_x(c, e)$

$c_T = \{t : c \in labels(t)\}$, $x_T = \{t : x \in labels(t)\}$, $e_T = \{t : e \in labels(t)\}$
$W = W' = \emptyset$
$E = E' = 0$
for all $t \in c_T$ **do**
  if $\exists t' \in x_T : [t + r..t + s] \bigcap [t' + r'..t' + s'] \neq \emptyset$ **then**
    $W = W \bigcup \{(t, t')\}$
  end if
end for
for all $t' \in x_T$ **do**
  if $\nexists t \in c_T : [t + r..t + s] \bigcap [t' + r'..t' + s'] \neq \emptyset$ **then**
    $W' = W' \bigcup \{t'\}$
  end if
end for
for all $(t, t') \in W$ **do**
  if $\exists t'' \in e_T : t'' \in [t + r..t + s] \bigcap [t' + r'..t' + s']$ **then**
    $E++$
  end if
end for
for all $(t') \in W'$ **do**
  if $\exists t'' \in e_T : t'' \in [t' + r'..t' + s']$ **then**
    $E'++$
  end if
end for
return $\frac{E}{|W|} - \frac{E'}{|W'|}$

In summary, we begin with a set of *prima facie* causes (identified by generating or otherwise specifying some set of potential relationships and testing which of these satisfy the conditions for *prima facie* causality relative to the given data) and then compute the average significance for each of these causes, yielding a set of ε_{avg}'s. We must now determine what value of ε to use when determining which of these causes are significant.

5.2.2 *Choice of* ε

One can determine an appropriate threshold through simulation (creating data with a structure similar to the real data of interest), or examining the hypotheses manually. However, since we are generally testing large datasets for which there are a multitude of hypotheses, we may use the resulting empirical statistics to our advantage. We assume that while there may be many genuine causal relationships in some tested set, these are still relatively few compared with the total number of hypotheses tested. The key observation here is that this relatively large number of non-causes will provide a background (or control) against which we may compare the true causes.

We accept that all thresholds have tradeoffs: if ε is too low, we will call too many causes significant (making false discoveries), while if ε is too high we will call too many causes insignificant (leading to false negatives). For our purposes, we concentrate on the first case. While we may miss some causes, we aim to be confident in those identified. The priorities of other users may vary, and they may wish to focus on identifying the full set of causes (at the expense of some of those identified being spurious). Many statistical methods exist for both purposes. We will concentrate on controlling the false discovery rate (FDR), which is the number of false discoveries as a proportion of all discoveries. In our case this will be the number of non-causes we call significant as a proportion of all causes deemed significant. The basic idea is that when doing many tests we are more likely to see results that seem significant, just by chance. To control for this, we generally compute some statistic (such as a p-value) for each

hypothesis, and compare these against the distribution expected under the null hypothesis. For a particular value of this statistic, we accept a hypothesis (rejecting the null hypothesis), if this value is significant when compared with the null hypothesis and after accounting for the number of tests being conducted. To determine the null hypothesis in our case, we would need to know how the ε_{avg}'s would be distributed if there were no genuine causal relationships. As an alternative, methods using empirical nulls allow us to estimate these from the data. These methods are particularly suited to cases with many tests and few true positives, as they allow better estimation of the null.

For an introduction to multiple hypothesis testing and false discovery rate control, see Appendix B. It is assumed that the reader is familiar with the goals and procedures of these methods, so we will only discuss the case of the empirical null here.

Computing the fdr

The basic idea of this approach is that we assume our data mostly fit a null model, where there are no causal relationships, with deviations from this model indicating true causal relationships. This assumption implies that we expect the computed ε_{avg}'s to follow a normal distribution, with the z-values calculated from these ε's having a mean of zero and a standard deviation of one. These ε's (even with no true causal relationships in the system) are not all equal to zero due to correlations from hidden common causes and other factors influencing the distributions, such as noise. The distribution of ε tends toward a normal due to the large number of hypotheses tested.

When there are causal relationships in the system, then there are two classes of ε's: those corresponding to insignificant causes (which may be spurious or too small to detect) and those corresponding to significant causes (which may be genuine), with the observed distribution being a mixture of these classes. Then, p_0 and $p_1 = 1 - p_0$ are the prior probabilities of a case (here, a causal hypothesis) being in the "insignificant" or "significant" classes respectively, with these probabilities distributed according to an underlying density. Since the insignificant class is assumed to be much larger than the significant class, and normally distributed, we can identify significant causes by finding these deviations from the normal.

For this purpose, we will calculate the local false discovery rate (fdr). Instead of computing p-values for each test and then determining where in the tail the cutoff should be after correcting for the many tests conducted, as is done when controlling the false discovery rate (FDR), this method instead uses z-values and their densities, to identify whether, for a particular value of z, the results are statistically significant after taking into account the many tests. We will use an empirical Bayesian solution to fdr control, as described by Efron [29]. This local method is better suited to the case of many hypotheses, but the same calculations can be done with the standard tail-area FDR [30].

With N hypotheses H_1, H_2, ..., H_N we have z-values z_1, z_2, ..., z_N. The z-value, also called the standard score, is the number of standard deviations a result is from the mean. In the case of our causal analysis, these z-values are computed from the ε_{avg}'s. We begin by assuming the N tests fall into two classes, namely, interesting and un-interesting (not-null and null). We also assume the proportion of interesting cases

is small relative to N, say, 10%. These classes correspond to rejection and acceptance of the null hypothesis, with prior probabilities p_0 and $p_1 = 1 - p_0$. That is, p_0 and p_1 are the prior probabilities of a case (here, a causal hypothesis) being in the "interesting" or "uninteresting" classes respectively. The densities ($f_0(z)$ and $f_1(z)$) describe the distribution of these probabilities. When using a theoretical null, $f_0(z)$ is the standard $N(0, 1)$ density. Note that we need not know $f_1(z)$, though we must estimate p_0 (usually $p_0 \geqslant 0.9$). We define the mixture density:

$$f(z) = p_0 f_0(z) + p_1 f_1(z), \tag{5.7}$$

then the posterior probability of a case being uninteresting given z is

$$\Pr\{\mathrm{null}|z\} = p_0 f_0(z)/f(z), \tag{5.8}$$

and the *local false discovery rate*, is:

$$\mathrm{fdr}(z) \equiv f_0(z)/f(z). \tag{5.9}$$

Note that, in this formulation, the p_0 factor is ignored, yielding an upper bound on $\mathrm{fdr}(z)$. Assuming that p_0 is large (close to 1), this simplification does not lead to massive overestimation of $\mathrm{fdr}(z)$. One may also choose to estimate p_0 and thus include it in the FDR calculation, making $\mathrm{fdr}(z) = \Pr\{\mathrm{null}|z\}$. The procedure is then:

1. Estimate $f(z)$ from the observed z-values;
2. Define the null density $f_0(z)$ from either the data or using the theoretical null;

3. Calculate $\mathtt{fdr}(z)$ using equation (5.9);

4. Label H_i where $\mathtt{fdr}(z_i)$ is less than the threshold (say, 0.01) as interesting.

Then, for each prima facie cause where the z-value associated with its ε_{avg} has $\mathtt{fdr}(z_i)$ less than a small threshold, such as 0.01, we label it as a just so, or significant, cause. With a threshold of 0.01, we expect 1% of such causes to be insignificant, despite their test scores.

5.3 CORRECTNESS AND COMPLEXITY

In this section we show that the associated procedures for verifying formulas over traces yield the desired results, and analyze their computational complexity.

5.3.1 *Correctness*

Correctness of procedure for checking until formulas in traces

Theorem 5.3.1. *The satisfaction by a time point* t *of the until formula* $f\mathcal{U}^{\geqslant r,\leqslant s}g$*, where* $0 \leqslant r \leqslant s < \infty$ *is given by:*

$$\mathtt{sat}_{\mathcal{U}}(t,r,s) = \begin{cases} \mathtt{true} & \text{if } (g \in \mathtt{labels}(t)) \wedge (r \leqslant 0), \\ \mathtt{false} & \text{if } (f \notin \mathtt{labels}(t)) \vee t = |T| \vee s = 0, \\ \mathtt{sat}_{\mathcal{U}}(t+1, r-1, s-1) & \text{otherwise.} \end{cases} \tag{5.10}$$

Proof. Assume trace T, where times $t \in T$ satisfying f and satisfying g have been labeled. Then, we will show by induction that any time t will be correctly labeled by equation 5.10. By definition, a timepoint t satisfies $fU^{\geqslant r,\leqslant s}g$ if there is some $r \leqslant i \leqslant s$ such that g is true at $t+i$ and $\forall j : 0 \leqslant j < i$, f is true at $t+j$.

Base cases:

$$sat_U(t, r, 0) = \begin{cases} true & \text{if } g \in labels(t), \\ false & \text{otherwise.} \end{cases} \tag{5.11}$$

$$sat_U(|T|, r, s) = \begin{cases} true & \text{if } (g \in labels(|T|)) \wedge (r \leqslant 0), \\ false & \text{otherwise.} \end{cases} \tag{5.12}$$

Note that in the first base case, since we have already stipulated that $r \leqslant s$, we know that if $s = 0$, $r \leqslant 0$. However in the second base case we must add the condition on r. If $s = 0$, the only way the formula can be satisfied is if t is labeled with g. Similarly, if $t = |T|$, then this is the last timepoint in the trace and t can only satisfy the formula if it is labeled with g.

Inductive step: Assume we have $sat_U(n, r, s)$. Then, for $s > 0$ and $n + 1 \neq |T|$:

$$sat_U(n-1, r+1, s+1) = \begin{cases} true & \text{if } (g \in labels(n-1)) \\ & \qquad (\wedge r \leqslant 0), \\ false & \text{if } f \notin labels(n-1), \\ (sat_U(n, r, s) & \text{otherwise.} \end{cases} \quad (5.13)$$

Timepoint $n-1$ satisfies the formula if it satisfies g or if it satisfies f and the next timepoint, n, satisfies the formula. However, we assumed that we can correctly label timepoints with f and g as well as $sat(n, r, s)$.

□

Corollary. *The satisfaction by a time point* t *of the until formulas* $fU^{\geqslant r, \leqslant \infty}g$ *or* $fU^{\geqslant r, < \infty}g$*, where* $r \neq \infty$ *is given by* $sat_U(t, r, |T|)$*.*

Corollary. *The probability of the formula* $fU^{\geqslant r, \leqslant s}g$*, in a trace of times* T*, where* $0 \leqslant r \leqslant s \leqslant \infty$ *is given by:*

$$\frac{|\{t \in T : sat_U(t, r, s)\}|}{|\{t' \in T : (f \vee g) \in labels(t')\}|}. \quad (5.14)$$

Correctness of procedure for checking unless formulas in traces

Claim. *The satisfaction by a time point* t *of the unless formula* $fU^{\geqslant r, \leqslant s}g$*, where* $0 \leqslant r \leqslant s < \infty$ *is given by:*

$$sat_{\mathcal{U}}(t, r, s) = \begin{cases} \text{true} & \text{if } (g \in \text{labels}(t) \wedge r \leqslant 0) \\ & \quad \vee (f \in \text{labels}(t) \wedge s = 0), \\ \text{false} & \text{if } (f \notin \text{labels}(t)) \vee (t = |T|) \\ & \quad (\vee s = 0), \\ sat_{\mathcal{U}}(t+1, r-1, s-1) & \text{otherwise.} \end{cases} \tag{5.15}$$

Proof. Assume trace T, and times $t \in T$ satisfying f and satisfying g have been labeled. Then, we will show by induction that any time t will be correctly labeled by equation (5.15). By definition, a timepoint t satisfies $f\mathcal{U}^{\geqslant r, \leqslant s} g$ if there is some $r \leqslant i \leqslant s$ such that g is true at $t+i$ and $\forall j : 0 \leqslant j < i$, f is true at $t+j$, or if $\forall j : 0 \leqslant j \leqslant s$, f is true at $t+j$.

Base case:

$$sat_{\mathcal{U}}(t, r, 0) = \begin{cases} \text{true} & \text{if } (g \in \text{labels}(t)) \vee (f \in \text{labels}(t)), \\ \text{false} & \text{otherwise.} \end{cases} \tag{5.16}$$

$$sat_{\mathcal{U}}(|T|, r, s) = \begin{cases} \text{true} & \text{if } (g \in \text{labels}(|T|) \wedge r \leqslant 0) \\ & \quad \vee (f \in \text{labels}(|T|) \wedge s = 0,) \\ \text{false} & \text{otherwise.} \end{cases} \tag{5.17}$$

If $s = 0$, the only way the formula can be satisfied is if t is labeled with either f or g. Similarly, if $t = |T|$, then this is the last timepoint

in the trace and t can only satisfy the formula if it is labeled with f or g in the appropriate time window.

Inductive step: Assume we have $sat_{\mathcal{U}}(n, r, s)$. Then, for $s > 0$ and $n + 1 \neq |T|$:

$$sat_{\mathcal{U}}(n-1, r+1, s+1) = \begin{cases} \text{true} & \text{if } (g \in \text{labels}(n-1)) \\ & (\wedge r \leqslant 0), \\ \text{false} & \text{if } f \notin \text{labels}(n-1), \\ sat_{\mathcal{U}}(n, r, s) & \text{otherwise.} \end{cases} \tag{5.18}$$

Timepoint $n-1$ satisfies the formula if it satisfies g or if it satisfies f and the next timepoint, n, satisfies the formula. Note that we assume $s > 0$ and thus the formula cannot be satisfied by only f being true. However, we assumed that we can correctly label timepoints with f and g as well as $sat(n, r, s)$.

□

Corollary. *The satisfaction by a time point* t *of the unless formulas* $f\mathcal{U}^{\geqslant r, \leqslant \infty} g$ *or* $f\mathcal{U}^{\geqslant r, < \infty} g$*, where* $r \neq \infty$ *is given by* $sat_{\mathcal{U}}(t, r, |T|)$.

Corollary. *The probability of the formula* $f\mathcal{U}^{\geqslant r, \leqslant s} g$*, in a trace of times* T*, where* $0 \leqslant r \leqslant s < \infty$ *is given by:*

$$\frac{|\{t \in T : sat_{\mathcal{U}}(t, r, s)\}|}{|\{t' \in T : (f \vee g) \in \text{labels}(t')\}|}. \tag{5.19}$$

Correctness of procedure for checking leads-to formulas in traces

Claim. *The satisfaction by a time point* t *of the leads-to formula* $f \rightsquigarrow^{\geqslant r, \leqslant s} g$ *is given by:*

$$sat_L(t, r, s) = \begin{cases} true & \text{if } f \in labels(t) \wedge (true U^{\geqslant t+r, \leqslant s} g) \in labels(t), \\ false & \text{otherwise.} \end{cases} \tag{5.20}$$

Proof. Assume trace T, and times $t \in T$ satisfying f and satisfying g have been labeled. We have already shown that we can correctly label times that begin sequences where until formula are true, and thus we can correctly label whether a state t satisfies $true U^{\geqslant t+r, \leqslant s} g$. We have also assumed that states satisfying f are already correctly labeled with f. Thus, we can label states with the conjunction of these formulas. Thus by definition of leads-to – that g holds in the window $[r, s]$ after f – we can correctly label times with such formulas. □

Corollary. *The satisfaction by a time point* t *of the leads-to formulas* $f \rightsquigarrow^{\geqslant r, \leqslant \infty} g$ *or* $f \rightsquigarrow^{\geqslant r, < \infty} g$*, where* $r \neq \infty$ *is given by* $sat_L(t, r, |T|)$.

Corollary. *The probability of the formula* $f \rightsquigarrow^{\geqslant r, \leqslant s} g$*, in a trace of times* T*, where* $0 \leqslant r \leqslant s < \infty$ *is given by:*

$$\frac{|\{t \in T : sat_L(t, r, s)\}|}{|\{t' \in T : f \in labels(t')\}|}. \tag{5.21}$$

This case is similar to the until and unless case, with the exception that the denominator consists of the set of states satisfying f, instead of $f \vee g$.

5.3.2 *Complexity*

We will now analyze the time complexity for each of the algorithms and procedures discussed. Note that each procedure (aside from the model checking ones) assume that all states have already been labeled with the formulas of interest. The complexity of that task is not included in the other procedures since it is assumed that this is performed once, with the results saved for use in the later tasks. That is, if we at some point label a time point with a formula f, we assume that when we are later interested in when f is true, we can reuse that result.

Complexity of model checking over traces

We begin with a trace of times, T. First, the complexity of labeling times along a trace with a proposition is proportional to the length of the time series, which is also denoted by T. Then, assuming states are labeled with f and g, labeling the sequence with $\neg f$, $f \vee g$, $f \wedge g$ and $f \rightarrow g$ is also of time complexity $O(T)$.

Next we have until, unless and leads-to path formulas, and finally the computation of the probabilities of these formulas. For an until or unless formula, such as $f U^{\geqslant r, \leqslant s} g$, the worst case for a single timepoint is when $r = 0$ and involves checking the subsequent s timepoints and thus for $s \neq \infty$, the worst case complexity for the entire sequence is $O(Ts)$, while for $s = \infty$, it is $O(T^2)$. However, these formulas naively assume all

timepoints are labeled with f and thus all $t \in T$ are candidates for starting such a path. Thus instead of T, the formulas should use T', the number of states labeled with f. For a leads-to formula, $f \rightsquigarrow^{\geqslant r, \leqslant s} g$, the complexity for a single timepoint is $O(|s - r|)$, where $s \neq \infty$. Where $s = \infty$, this is again $O(T)$. As for the until/unless case, if we assume all timepoints are labeled with f, then the complexity for a trace is $O(T|s - r|)$ or $O(T^2)$, though in practice most times will not be labeled with f and thus these will be significantly reduced.

Once states have been labeled as the start of path formulas or with the appropriate state formulas, computing the probability for a state formula is $O(T)$.

For any formula f, the worst case complexity of testing f in a trace T, assuming that the subformulas of f *have not* already been tested, is thus $O(|f| \times T^2)$, where $|f|$ is the length of the formula.

Complexity of testing prima facie causality

For a single relationship, $f \rightsquigarrow^{\geqslant r \leqslant s} g$, again assuming times satisfying f and g are labeled as such, we must simply compute the probability of this formula along the trace ($O(T)$) and compare this with the computed probability of $F^{\leqslant \infty} g$ (also $O(T)$). Thus for M relationships the complexity is $O(MT)$. In the case where we have N possible causes of N effects, then this case has complexity $O(N^2T)$.

Complexity of computing ε_{avg}

Assuming timepoints are already labeled with c, e and x, computing $\varepsilon_x(c, e)$ has complexity $O(T)$. Thus, in the worst case, computation of one $\varepsilon_{avg}(c, e)$ is $O(NT)$, where there are N causes and M effects and

all N causes are prima facie causes of e. To compute the significance for each cause of e we must repeat this N times so the complexity is $O(N^2T)$. Finally, repeating this for all M effects, we find the complexity is $O(MN^2T)$. In the case where the causes and effects are the same (say when testing relationships between pairs of genes or neurons), then $N = M$ and the worst case complexity is $O(N^3T)$.

5.4 OTHER APPROACHES

Alternatively, one could begin by inferring a model. Then, satisfaction of formulas and algorithms for model checking are exactly that of [49]. However, model inference is a difficult task and may include development of phenomenological models and abstractions of the true structure. When inferring a model, there is again the problem of a non-recurrent start state. The probabilities inferred will not necessarily correspond to the true probabilities of the structure. Further, it is unknown whether, as the number of observations tends towards infinity, the inferred model approaches the true structure. Thus, since we are generally interested in a set of properties that is small relative to the size of the underlying structure, we focus on inferring the correctness of those properties. In other cases, for a relatively small structure one may wish to begin by inferring a model.

6

TOKEN CAUSALITY

In this chapter, we relate our theory (developed in the preceding chapters) for general, type-level, cases to particular, token-level, instances. We begin in section 6.1 with a discussion of what constitutes a token-level case and how these differ from type-level ones, before reviewing some of the ways token causality has been reasoned about. Then, in section 6.2 we formulate a new approach to this problem, showing how to use previously inferred type-level causes and token level observations to rank possible token-causes of an effect. In section 6.3, we illustrate the approach by working through two examples. Finally, in section 6.4 we turn our attention to the analysis of cases that have proven difficult for approaches to token causality.

6.1 INTRODUCTION TO TOKEN CAUSALITY

6.1.1 *What is token causality?*

Thus far we have developed a new approach to recognizing and inferring general, type-level, causal relationships. However, in many cases we want to find the cause not of a *kind* of event, but of a *particular* one that actually

occurs at some point in time and space.[1] When we are assigning credit or blame, such as in legal or moral cases, or determining why a patient is ill, we are seeking token-level relationships explaining particular instances. Type-level relationships relate to general, statistical, properties from many repeated observations of a system or population, and allow us to *predict* the occurrence of the effect if the cause were to happen. Token-level relationships relate to single occurrences or individuals and help us to *explain* the occurrence of something that has already happened.

One particularly important use of token causality is in diagnosing patients. While there are general relationships between diseases and symptoms (and between risk factors and diseases), each patient must be understood individually in order to determine their best course of treatment. For example, when a patient arrives with a cough, his doctor's initial hypothesis may be that he has the common cold. However, patients can be queried and further examined using medical tests and by reviewing their prior medical history. Thus, after finding out that the patient also has a fever, shortness of breath, and chest pain, the doctor may update her original hypothesis and order a chest x-ray to confirm the diagnosis of pneumonia. It is important to note the distinction between the type (general) and token (singular) cases. While the type-level relationships provide initial hypotheses, these are confirmed or rejected based on the token-level information, relating the current symptoms and past medical

1 The definition of an "event" is inherently ambiguous and it is unclear at what point something goes from being a single event to being a sequence of events. For the moment we will stick to the extreme cases, where the distinction is clear. However, it should be noted that we do not assume that an event be instantaneous in time. While an effect could potentially (but not necessarily) be instantaneous, we consider the token-level event (or occurrence) to include the actual occurrence of both the cause and the effect. Thus, since we assume a cause (other than ones simultaneous with their effects, which we have ignored) precedes its effect in time, the entire event must have a non-trivial duration.

history to known relationships that indicate factors that could cause these symptoms. In short, token-level causality relates to the question of "why" on a particular occasion and type-level causality relates to the question of "why" in general.

6.1.2 *Why do we need a notion of token causality?*

Once we have a set of type-level causes, our work is not done. When we want to find out who caused a car accident, why a house caught on fire, or what made a person ill, knowing the type-level causes of accidents, fires and illness may give us some hypotheses, but these relationships alone are not enough for us to determine the token-level causes. We might believe at first glance that our type-level causes can explain these observances, but while a type-level cause can indicate that a token-level case is *likely* to have a particular cause, it does not necessitate this. Note also that a token-level case may correspond to multiple type-level relationships. Bob's death can be a token of "death", "death caused by cancer", "death caused by lung cancer", "death of a 77-year old man" and so on.

For example, going back to the case of diagnosing a patient, if a doctor suspects a patient has a particular illness, she may try to show how the patient's history and symptoms fit with the known course of the suspected illness – and conversely, the doctor would likely come up with the potential diagnosis by observing the similarity of the patient's case to a known disease. However, conflating this correlation between type and token with the *necessity* of a token relationship following from a type one is akin to conflating causation and correlation. An extreme

example of this in action would be choosing a treatment based entirely on the general, population-level, causes of a patient's symptoms. We would have type-level relationships indicating causes of, say, coughs and chest pains, or headaches and fatigue, with treatment for these based on what has proven effective over the entire population. Then, each individual's treatment would be based not on his actual disease, but on what worked in the population for treating the general cause of these symptoms. For example, a patient with chronic fatigue syndrome might be treated for depression, as this could be a more common explanation for the same symptoms. While we may have certain hypotheses, based on known type-level relationships, we must also be willing to abandon these hypotheses in the face of evidence against them. Thus, we need a way of reasoning about single cases that takes this into account, allowing us to use knowledge gleaned from type-level relationships while admitting the possibility that the sequence of events may be entirely different in token cases.

Discrepancies between type and token arise in two primary scenarios. First, if the sample from which the type-level relationships are inferred differs from that of the token case, the single case causalities will differ. Unless we have background knowledge or may experiment on or otherwise probe the system, we may not be able to identify such a discrepancy. Let us say we learn that smoking causes lung cancer within 10 to 20 years and then see a patient who smoked and developed lung cancer within 10 to 20 years. However, this patient happens to have a genetic mutation such that smoking lowers his risk of lung cancer, and in fact it was his exposure to radon during his career as an experimental physicist that

caused his lung cancer.[2] In this case, if we know nothing of the connection between radon exposure and lung cancer, our token inferences will be incorrect. However, in order to know that it is possible for smoking to be prevented from causing cancer, we must have previously had data for groups of people with the same mutation as well as for people repeatedly exposed to radon. Our explanations are only as good as the current state of knowledge, so we must either have previous type-level data that supports these claims or we must be able to use some background knowledge to rule out type-level causes and guide the exploration of new relationships.

The second case where type and token may differ is when a less significant or even an insignificant (using our terminology from Chapter 4, referring to the relative magnitude of the related ε_{avg} significance scores) type-level cause token-causes the effect. In this case, even without background information, the situation is amenable to automated inference. It is problematic only when a more significant cause also occurs or when we have incomplete knowledge of what occurred. For instance, one highly significant cause of chickenpox is close contact with someone who is infected. Another less likely cause is the chickenpox vaccine, which usually prevents the illness but causes it in a small percentage of people vaccinated. Now consider the case where we know that a person received the vaccine and then developed chickenpox, but we do not know whether she was exposed to anyone with the illness. Depending on the probability that she came into contact with an infected person, given that she now

2 This is also an example of the mechanism connecting cause and effect failing: here smoking was prevented from causing cancer by the genetic mutation. Another case could be a gun that is fired, but which was unloaded before firing.

has the illness, it is possible that we might find both the potential causes (exposure and the vaccine) equally significant at the type-level.

The method developed in this chapter will help assemble and understand the facts surrounding a token event by relating these to type-level relationships. We are not trying to develop a metaphysical theory of token causality, and make no claim as to whether one can ultimately be reduced to the other. Rather we suggest that there is a need for rigorous methods for relating general properties to singular cases. Our focus here is on developing a methodology, much as we did for type-level causes, that can be used to automatically analyze token causes. In some cases the result will be that the most significant type-level cause is the most significant token-level cause, but to arrive at that answer we need to relate our observations to the previously determined type-level causes. Since the relationships inferred are logical formulas with time constraints between cause and effect, we will need to determine whether what was seen constitutes an instance of each known relationship. If we do not have the truth value for all propositions at all times, then we will calculate the probability, given the observations, of the token case satisfying the logical formulas associated with each causal relationship. Then, with a set of possible type-level relationships that could explain the token case, we will rank their significance for the token case. As before, we will not attempt to find the true cause of every effect in an error-free manner. Instead we will determine the most probable causes given the type-level inferences and token-level observations.

6.1.3 *How can we reason about token causality?*

Most of the prior work in this area has addressed the metaphysics of token causality, leaving open the practical problem of how to reason about such cases in an automated algorithmic way. Among philosophers, there is no consensus as to how to combine type- and token-level information: we may learn type-level claims first and then use these to determine token level cases [127]; the type-level relationships may follow as generalizations of token-level relationships [52]; or they may be treated as entirely different sorts of causation [27]. One algorithmic exception is Pearl's work on the "actual cause", which attempts to link type-level structural models with counterfactual analysis of token-level cases. However, among other problems, since the underlying models and theory were not explicitly developed for cases involving time, and allow for inference from non-temporal data, we cannot avoid cases such as smoking at 10am causing lung cancer at 2pm. We could only exclude this case manually using background knowledge, but this becomes more difficult as we must decide at what point the event of smoking should be considered to fulfill the relationship (e.g. a week, a month, or a year before lung cancer).

Eells's view is also described in section 2.3.3 of Chapter 2. The counterfactual approach of Lewis, described in section 2.2.2 primarily applies to token-level cases. Pearl's theory of the "actual cause" is discussed in Chapter 3.

Given the relative sparsity of algorithmic methods for token-level inference (compared with those for type-level inference), it would be useful to be able to repurpose some of the metaphysics for our epistemic ends. While the approaches highlighted provide solid ground on which to determine whether something is a token cause, they may not be practical (due to requirements of knowledge or computationally infeasible calculations)

if translated directly. Yet we can take inspiration from these accounts, adapting them to suit the needs of methodology Regardless of which theory is correct (leaving aside what it means to be "correct" here, as we are still interested only in what can be learned – not the underlying fact of what *is*), and whether one (a type or token level claim) is necessary or sufficient for the other, we can make some token-level claims by using our type-level inferences as support. We will use the strength associated with our type-level causes to assess the strength of the token-level claims.

One way of relating these two levels of causality is by using the connecting principle, introduced by Elliot Sober [119]. Sober introduces a notion of support, where this is a numerical quantity whose value indicates how likely it is that a particular type-level cause token-caused a particular, actually occurring, effect. The support of the token hypothesis (such as that Bob's smoking caused his lung cancer) is proportional to the strength of the corresponding type-level relation (such as smoking causes lung cancer). The connecting principle is stated as follows [119]:

Definition 6.1.1 (Connecting Principle)**.** With C being a causal factor with magnitude m for producing E in population P, the support for the hypothesis that C actually occurring at t_1 caused E to occur at t_2 – given the type-level relation between C and E and the fact that instances of C and E token-occurred in population P – is the magnitude m associated with the type-level relationship. This is written as:

$$S\{C(t_1) \texttt{ token caused } E(t_2) | C(t_1) \texttt{ and } E(t_2) \texttt{ token occurred in } P\} = m.$$

The value of this support can range from -1 to $+1$, as this is the range of m. Here C and E are types of causes and effects and the time-indices

indicate the token events that occur at particular places and times, represented by t_i. The measure of m used by Sober is Eells's ADCS (average degree of causal significance). Restated, this is:

See Chapter 2 Section 2.3.3 for a discussion of the ADCS.

$$m = \sum_i [P(E|C \wedge K_i) - P(E|\neg C \wedge K_i)] \times P(K_i), \tag{6.1}$$

where the K_is are the background contexts and this measurement denotes the magnitude of causal factor C for effect E in population P. As before, the background contexts (each denoted by K_i) are formed by holding fixed all factors in all possible ways. We can see now that just as we replaced the ADCS with ε_{avg}, that ε_{avg} will be our value of m (and that our earlier arguments against context unanimity will apply here as well).

See Chapter 4 for a discussion of ε_{avg}.

For a particular token case, according to Sober, the relevant population means using whatever is known about the case. So, if a person's age and weight are known, then the population is one comprised of individuals with those properties. If less is known, perhaps only that he is a U.S. citizen, then the relevant population is U.S. citizens. However, in practice we will not have arbitrarily specific type-level relationships that will allow us to take advantage of all available information. Further, truly using all information about the token-case will result in a population of size one (the single case under study). Thus we note that the likelier case is that we will have separate structures (and/or type-level relationships) representing different populations (e.g. one for, say, middle aged smokers and another for elderly non-smokers) or the features that would define someone or something as being part of a population could simply be propositions.[3] Thus we may still have varying results based on the finer

3 Note that in this case, saying that something is true for a population, where the population is defined by properties $p_1, p_2 ... p_n$ means testing whether, in addition to the formulas

details, but whether and how these are used depends on both the available type and token-level information.

The main principle here is that a known type-level relationship between some c and e is good evidence for c causing e, if we see that both c and e have occurred. Clearly, the type-level relationship alone is not enough, the relation must actually be instantiated. In both Sober's method and ours, though, the type-level causes are precisely such due to their frequency of observation in some population. That is, if we find that 80% of people who develop disease X die shortly after, then we have reason to believe that if we observe a new patient who contracts X and dies, this is another instance of the disease being fatal.

6.2 FROM TYPES TO TOKENS

We now turn our attention to formulating a new approach. The problem we aim to solve is one where we have inferred some type-level relationships (using the method discussed in Chapters 4 and 5) and are attempting to explain the occurrence of an effect using this type-level information and knowledge of the token-level event (which consists of a sequence of times, with propositions true at those times). A token-level hypothesis is that, given a type level relationship such as $c \leadsto^{\geqslant r, \leqslant s}_{\geqslant p} e$ and the satisfaction of c and e (remember these are logical state formulas) based on the token-level observations, an instance of c token-caused an instance of e. The result of the procedure will be a ranking of the type-level causes (possible explanations for the token-level effect) using a measure

for the causal relationships, $p_1 \wedge p_2 \wedge \cdots p_n$ holds. For example, instead of a structure representing the functioning of a bull market, we could have relationships where causes are of the form $c \wedge b$, where c is any state formula, and b denotes a bull market.

of their significance combined with their probability of token occurrence. Thus we again do not partition our possible explanations into causes and non-causes, but rather quantitatively assess their significance, with those having the highest values of the measure being likelier explanations for the token case.

In section 6.2.2, taking inspiration from Sober's work, we will define a measure (called support) of the significance of a token-level cause for a particular type-level instance. Since we rarely have complete knowledge of a scenario (in the case of diagnosis, we cannot do all medical tests and patient histories contain many omissions) we allow for the possibility that we may only have evidence pointing towards the cause's occurrence. Thus, this support weights a measure of the type-level significance by the probability of the relationship having occurred in the token-case, given the observations. In Chapter 4, we introduced a new measure for type-level significance (called ε_{avg}), which is the average difference in probability a cause makes to its effect given, pairwise, all other prima facie causes of the effect. In the case where we know the truth value for all propositions (and thus whether or not a particular cause occurred in such a way that it could have caused the effect), the support for a token-level hypothesis that actually occurs will be exactly equal to the associated $\varepsilon_{avg}(c, e)$.

In section 6.2.3, we discuss the computation of the probability of a cause in detail. There we may use either a structure (a probabilistic Kripke structure as described earlier) or the original data used for the type-level inference along with the token-level observations to find the probability of any cause having occurred, given the observations. Note that the probabilities do not directly relate to Eells's probability trajecto-

ries (discussed earlier in this chapter and in Chapter 2). While we are determining probabilities, and in theory we could calculate the probability that a causal relationship is satisfied at each time instant given the observations up until that time, these would not produce the same results as Eells's analysis, since the probabilities would still be based on the type-level distributions.[4]

Before we can assign support to causes, we must first determine which hypotheses should be examined. Since there can be a number of significant causes and multitudes of insignificant ones, we need a way of systematically exploring these that allows for differences between the type and token level, while remaining computationally feasible. In section 6.2.1 we recognize that since the measure of support defined will be larger for actually occurring genuine and just so causes than for insignificant causes, we can begin by testing which of these occurred in the token case. If none occurred or we cannot determine their truth value, then we can calculate the probabilities for these significant type-level causes token-occurring and test whether any insignificant type-level causes token-occurred. In the case of diagnosis, this would mean first testing which significant causes of a patient's symptoms occurred. Then, if none are satisfied by the patient's history, we may examine less likely relationships. Finally, in section 6.2.4, we bring all of these pieces together, and discuss the procedure for taking type-level relationships and a token-level observation and assigning support to the potential causes.

4 For example, Eells's discusses the case of a squirrel kicking a golfball and raising the probability of it going into the hole based on the exact way it was kicked. We cannot account for the individual squirrel kicking the ball in a way that was different from squirrels in general.

6.2.1 *What can be a token cause?*

We start with the question of selecting the hypotheses to examine further. First, we note that an insignificant type-level cause can be a token-level cause. In fact, a token-level cause does not have to be even a *prima facie* type-level cause. Think of the case of seatbelts and automobile deaths. While in general seat belts help prevent deaths from automobile accidents, there are cases (though rare) where seat belts in fact cause death (for example, via chest or neck injury). We want to be able to consider such cases, and not immediately rule out factors that are not causes at the type level.

At this point it seems like we may have to enumerate every conceivable potential cause of the effect, an inefficient and possibly hopeless pursuit. However, let us recall that we are calculating the support of token causal claims with the presumption that we are interested in those with high levels of support. If two possible token causes took place on a particular occasion and one is a type-level genuine cause while the other is a type-level insignificant cause, the more likely explanation for the effect is that it was token caused by the type-level genuine cause. That is, if we have a number of token causal hypotheses, those with the highest support will be those with the highest value for ε_{avg} – our just so or genuine causes. Thus, if we know that a just so cause of the effect in question took place, we do not need to examine any insignificant or non-prima facie causes of the effect, as the only other causes that may have higher significance for the effect are other just so or genuine ones. If none of the just so

or genuine causes occurred, then at that point we would go down the hierarchy of all possible explanations.

We will illustrate this with the following scenario. A student, Alice, achieved a perfect score on her exam. Alice says that this was because she remembered to wear her lucky sweater. Bob disagrees and told Chris that Alice must have studied a lot. If Alice then confirms that she did spend a lot of time studying, what should Chris believe? We may safely assume that studying is a genuine cause for success on an exam, and that any role played by sweaters is negligible. If we put aside any individual prior beliefs Chris might have about the impact of sweaters and luck on exam success, then he would not continue to ask about increasingly unlikely factors once he knows that Alice has studied for the exam (i.e. a type-level genuine cause has token occurred). However, if Alice said that she had not studied, then Chris may be more willing to accept the sweater hypothesis. In a slightly trickier case, Alice might have said that she did well because she had a cup of coffee before the exam. It is possible that in general coffee has a minor impact on studying, but that for Alice, it helps her concentrate and enables her to perform better on her exams (or might have some placebo effect, since she believes it will work). However if we do not have type-level information about Alice's past history and what affects her grades, we could only assess the situation using our general type-level relationships. We cannot account for varying information or beliefs between individuals in a system. One could potentially extend this approach to include prior beliefs, using these to weight the support of a hypothesis for each individual. However, it is unclear whether support should vary between individuals. Certainly it would not change the fact

of what *actually* caused the effect, but could be important in cases where type-level information is not all public and scattered across individuals.

While an insignificant cause can be a token-level cause, we begin by assessing the just so and genuine causes. We are considering the most plausible hypotheses first, and may miss atypical cases where there is an unusual explanation for the effect and both it and a genuine cause token-occur. However, if these more plausible causes are found not to have caused the effect, then we will go back to the set of all possible causes, using the facts we have about the situation to narrow these to a smaller set of those that are satisfied by the data, and then assess the support for each of these. In this way we can use token events to find causal relationships we may have missed. If there are a number of token-level instances where the only possible cause is one we previously deemed insignificant, then we must reevaluate this assertion.

6.2.2 *Support of a causal hypothesis*

We now turn our attention to reformulating Sober's connecting principle for our purposes. Recall that we have type-level relationships of the form:

$$c \leadsto^{\geqslant r, \leqslant s}_{\geqslant p} e, \tag{6.2}$$

where c and e are PCTL state formulas, $1 \leqslant r \leqslant s \leqslant \infty$, $r \neq \infty$, and p is a probability. Unlike in Sober's examples, we will not always know if c is true, and may only have evidence pointing toward this. For example, e could be related to a particular illness, and we might have symptoms and

a medical history that make c seem likely, but we will not know for sure whether c is true or false. Thus, we cannot simply use the strength of the type-level relationships, but rather must weight these by the probability that they token-occurred.

Our notation for token cases will be as follows. First, a token case (also referred to as an event) is defined by a sequence of times and propositions true at those times. Thus when we refer to the effect having "token-occurred" we mean that the PCTL state formula that represents the effect (e) is satisfied at some actual time (t_2), and represent this by e_{t_2}. Then, our token-level causal hypothesis will be that c (where there is a type-level relationship between c and e as described in formula (6.2)) at a time t_1, where $t_2 - s \leqslant t_1 \leqslant t_2 - r$, caused e_{t_2}. We will write this hypothesis as:

$$c_{t_1} \rightsquigarrow e_{t_2}. \tag{6.3}$$

This is not a PCTL leads-to formula, but rather denotes that c at time t_1 (with associated constraints on t_1) "led-to" e at time t_2.

Thus the support that we aim to compute is $S(c_{t_1} \rightsquigarrow e_{t_2})$, which we will define as:

$$S(c_{t_1} \rightsquigarrow e_{t_2}) = S(c_{t_1} \rightsquigarrow e_{t_2} | c_{t_1}, e_{t_2}) \times P(c_{t_1}, e_{t_2}). \tag{6.4}$$

That is, we are computing support for the hypothesis that c_{t_1}, where $t_1 \in [t_2 - s, t_2 - r]$, token-caused e_{t_2}. This is equal to the support for this hypothesis given the evidence that c token-occurred at t_1 and e token-occurred at t_2 (meaning that times t_1 and t_2 satisfy these logical

formulas), multiplied by the probability of this evidence (which is simply the probability that c and e token-occurred at these times). In the case where we know that they have token-occurred (by determining that the formulas have been satisfied), then this reduces to the case outlined by Sober [119], where the probability of the evidence was always assumed to be one.

Let us look in detail at the components of equation (6.4). First, we define that:

$$S(c_{t_1} \rightsquigarrow e_{t_2} | c_{t_1}, e_{t_2}) = \varepsilon_{avg}(c, e), \tag{6.5}$$

meaning that the support for the token-level hypothesis given the evidence of the token-occurrence of c and e in such as way as to satisfy the corresponding type-level relationship, is exactly the strength of the type-level causal relationship, which we previously computed to be $\varepsilon_{avg}(c, e)$.

Recall that with X being the set of prima facie causes of e, this is defined as:

$$\varepsilon_{avg}(c, e) = \frac{\sum_{x \in X \setminus c} \varepsilon_x(c, e)}{|X \setminus c|}, \tag{6.6}$$

where:

$$\varepsilon_x(c, e) = P(e|c \wedge x) - P(e|\neg c \wedge x). \tag{6.7}$$

Note that there are still time windows associated with the relationships between c and e and between x and e, and that when calculating the

This measure is introduced in Chapter 4, with more detail on its computation given in Chapter 5.2.1.

probabilities, these windows constrain the instances of each formula that will be considered.

Then, to determine $P(c_{t_1}, e_{t_2})$, we first note that we are calculating this probability relative to our observations (a sequence of times with the propositions true at each time, denoted by $\mathcal{V}$).[5] Thus, if both c_{t_1} and e_{t_2} are satisfied by this sequence of observations, their probability will be one (as they actually occurred). Since we assume that we always know that e_{t_2} is true (since we are attempting to explain it), the relevant probability is that of c_{t_1}, which is $P(c_{t_1}|\mathcal{V})$.[6] We now redefine support.

Definition 6.2.1 (Support for a token-level cause)**.** Assume that: there is a type-level relationship between c and e of the form $c \leadsto^{\geqslant r, \leqslant s}_{\geqslant p} e$; e token-occurred at time t_2; the token-level observations are V; the probability, given the data, that c token-occurred at time t_1 where $t_1 \in [t_2 - s, t_2 - r]$, is $P(c|\mathcal{V})$; and $\varepsilon_{avg}(c, e)$ is the strength of the type-level relationship between c and e. Then the support for the hypothesis that c_{t_1} (where this denotes only c's occurrence in the relevant time range) token-caused e_{t_2} (where this hypothesis is written $c_{t_1} \rightsquigarrow e_{t_2}$) is:

$$S(c_{t_1} \rightsquigarrow e_{t_2}) = \varepsilon_{avg}(c, e) \times P(c_{t_1}|\mathcal{V}). \tag{6.8}$$

5 Note that this set of observations may be quite large, with many facts being irrelevant. For example, when explaining a death, the day of the week on which the person died is unlikely to have any bearing on the cause of death. Yet, there may be causes that while insignificant, do have some small impact. Note though that if a number of these insignificant causes together have a meaningful impact on the probability of c, then their conjunction will be a genuine or just so cause, so we only need to concern ourselves with the case where the cause makes a very small difference. Since together these insignificant causes must still be insignificant, a likely heuristic approach is to limit the knowledge used to events that are part of causes and effects of c.

6 We cannot yet disentangle the probability of something actually occurring from the probability that it is known that it actually occurred. If we know that something occurred, we say its probability is one. However, even if it did actually occur, if we do not know this fact, its probability will not necessarily be one in our system.

We assume that if the type-level relationship is related to a particular population, then e and c must token-occur in that same population. However, given the ambiguity of the definition of a population, it is likelier that one will define the properties related to the population by additional propositions as part of the cause c.

6.2.3 *Calculating the probability of a cause token-occurring*

Assume that there is a type-level causal relationship between some c and some e (which are each logical state formulas) such that c causes e in between r and s time units, the effect occurs at time t, our token-level observations are $\mathcal{V}$ and we are now attempting to calculate the probability of c occurring at time t', where $t' \in [t-s, t-r]$. To calculate the probability of a particular cause (c) token-occurring, we could go back to our original data, using frequencies (calculating the frequency of sequences where the evidence holds). However if we have or have inferred the structure of the system, we may use that as follows.[7] First note that we are computing the posterior probability of c having occurred at a time where it could have caused the effect (i.e. the logical formula c being satisfied by a particular time point). Our evidence is a sequence of observations, comprised of a set of time-ordered facts about the scenario ($\mathcal{V}$). It will be easier to later represent the probability of $\neg c_{t'}$ than $c_{t'}$ and thus we are now interested in:

$$P(c_{t'}|\mathcal{V}) = 1 - P(\neg c_{t'}|\mathcal{V}). \tag{6.9}$$

7 The same procedure may be used with a set of time series data in the same way we tested prima facie causality in Chapter 5.

Since

$$P(\neg c_{t'}|\mathcal{V}) = \frac{P(\neg c_{t'} \wedge \mathcal{V})}{P(\mathcal{V})}, \tag{6.10}$$

we see that:

$$P(c_{t'}|\mathcal{V}) = 1 - \frac{P(\neg c_{t'} \wedge \mathcal{V})}{P(\mathcal{V})}. \tag{6.11}$$

Then, note that the facts we have about the current scenario will be time-indexed such that we have facts at times t_0, t_1 and so on (where these are ordered observation times). These facts constrain the set of states our system has occupied (assuming our model of the system is correct, or our data is representative of the system). If q is true at $t = 3$ then at t_3 the system must be in a state labeled with q. Let us now construct the set F where each $f_i \in F$ is the conjunction of facts that are known to be true at time i, for $i \in [0..t]$, where time zero is the beginning of the token event and the effect e occurred at time t. When for a particular i there are no known facts of that time then $f_i = true$. Otherwise, a particular f_i might be something like $(asbestos \wedge smoking)$.

Remember that there is relationship such as:

$$c \rightsquigarrow_p^{\geqslant r, \leqslant s} e, \tag{6.12}$$

between c and e (and c and e are themselves logical state formulas) where we assume $s \geqslant r$, e is true at time t, and that we are computing $P(c_{t'})$, where $t' \in [t-s, t-r]$. Then, when computing the numerator of equation (6.11) we add to our set F: $\{\neg c \in f_i : t-s \leqslant i \leqslant t-r\}$. For both numerator and denominator, we proceed in the same manner, with

the only difference being the addition of $\neg c$ to the f_is of the numerator. The negated c means that c did not occur in such a way as to satisfy the formula representing the relationship between c and e. Thus we are calculating the probability of c not having happened during that time window, given e's occurrence and all other known facts about the case.

Claim. *With* $K = \langle S, s^i, L, \mathcal{T} \rangle$ *being the structure representing the system, and where states satisfying each* $f_j \in F$ *have been labeled as such and all states are labeled with* true*, then for* $0 \leqslant t < \infty$*, the probability (denoted* $\mu_m^t(s_0)$*) of the set of paths beginning in* s_0 *where each* $s_j \models_K f_j$*, and the paths are of length* t*, is given by the following recurrence, where we begin with* $j = t$ *and* $s = s_0$*:*

$$P(j, s) = \begin{cases} 1, & \text{if } j = 0 \text{ and } f_{t-j} \in \mathit{labels}(s); \\ 0, & \text{if } f_{t-j} \notin \mathit{labels}(s); \\ \sum_{s' \in S} \mathcal{T}(s, s') \times P(j-1, s'), & \text{otherwise.} \end{cases} \tag{6.13}$$

Proof. For the set of states s and integer time t, take $\Pi(t, s_0)$ to be the sequences of states $s_0 \to s_1 \to \ldots \to s_t$, beginning in s_0 and where, for all j from 0 to t, $s_j \models_K f_j$. Then, by definition

$$\mu_m^t(s_0) = \sum_{s_0 \to s_1 \cdots \to s_t \in \Pi(t, s_0)} \mathcal{T}(s_0, s_1) \times \cdots \times \mathcal{T}(s_{t-1}, s_t). \tag{6.14}$$

We will show by induction that the recurrence of (6.13) satisfies this equation.

Base case ($j = 0$): According to the recurrence in (6.13), $P(0, s_0) = 1$ if $s_0 \models_K f_0$. By definition, the μ_m-measure of a path of one state is 1, so $\mu_m^0(s_0) = P(0, s_0) = 1$. If $s_0 \not\models_K f_0$ then $P(0, s_0) = 0$. Note that the formula for μ_m above only considers paths such that each $s_i \models_K f_i$. Thus, since $s_0 \not\models_K f_0$, by definition $s_0 \notin \Pi(0, s)$. Adding zero leaves both μ_m and P unchanged and thus they are still equivalent.

Inductive step: If we assume $P(j-1, s_1) = \mu_m^{j-1}(s_1)$ then we must show $P(j, s_0) = \mu_m^j(s_0)$.

By definition:

$$\mu_m^j(s_0) = \sum_{s_0 \to \cdots \to s_j \in \Pi(j, s_0)} \mathcal{T}(s_0, s_1) \times \cdots \times \mathcal{T}(s_{j-1}, s_j). \quad (6.15)$$

This can be rewritten:

$$\mu_m^j(s_0) = \sum_{s_1} \mathcal{T}(s_0, s_1) \times \sum_{s_1 \to \cdots \to s_j \in \Pi(j-1, s_i)} \mathcal{T}(s_1, s_2) \times \cdots \times \mathcal{T}(s_{j-1}, s_j). \quad (6.16)$$

However, we assumed $P(j-1, s_1) = \mu_m^{j-1}(s_1)$, and since by definition:

$$\mu_m^{j-1}(s_1) = \sum_{s_1 \to \cdots \to s_j \in \Pi(j-1, s_i)} \mathcal{T}(s_1, s_2) \times \cdots \times \mathcal{T}(s_{j-1}, s_j), \quad (6.17)$$

we find:

$$\mu_m^j(s_0) = \sum_{s_1} \mathcal{T}(s_0, s_1) \times P(j-1, s_1). \quad (6.18)$$

When $s_o \models_K f_o$, this is equal to the third item of our recurrence:

$$\sum_{s' \in S} \mathcal{T}(s, s') \times P(j-1, s').$$

When $s_o \not\models_K f_o$, both μ_m and P are zero.

□

Note that our times begin at $t = 0$, upon entry to the start state of the system. Thus, we are computing the probability of the set of paths from that start state such that each state s_i satisfies the corresponding f_i. This means that with $\mathcal{V}$ being our time-indexed evidence, including $\neg c$ at the appropriate times, the recurrence above yields the probability $P(\neg c \wedge \mathcal{V})$ in the case where $t \neq \infty$. However, since we know that e has occurred at some actual time t, the path from s_i must be of length t and is thus finite. For the denominator of equation (6.11), $P(\mathcal{V})$, we repeat the same procedure, with F modified such that it does not include $\neg c$ as it did for the numerator. Thus, following this procedure we have calculated $P(c_{t'}|\mathcal{V})$ for a particular potential cause c of effect e, with evidence $\mathcal{V}$. Note that if observation does not begin at the start state of the system, the procedure may be easily used with a trace or set of traces, as is done in the case of inferring type-level relationships. It may be helpful at this point to consider an example. Turn to Appendix E.1 to go through the calculations for a particular cause and effect.

6.2.4 *Procedure for assigning support to causes*

Recall that we have sets of type-level genuine, just so, and insignificant causes of the token-effect in question. In order to determine the support for each (as defined in section 6.2.2), we must first ascertain – using the facts about the situation – which of these occurred. When we do not have enough information to determine if one has occurred, we use the above procedure to determine its probability using our observed evidence. Recall that the support for each hypothesis is the previously computed ε_{avg}, weighted by the probability of the cause occurring given the observations. That is, the largest possible value of the support for a token hypothesis is its associated ε_{avg} (since the probability can be at most one). If any genuine or just so type-level causes have occurred, this means that they will have the highest values of this support. As our goal is to find the likeliest causes (those with the most support) we can begin by taking these sets and testing whether any of their members are true on the particular occasion.

That is, with C being the set of just so and genuine causes of the effect, e, and F being the set of time indexed propositions, we test whether each $c \in C$ is true on this occasion given the facts. Let us recall the types of formulas and discuss their truth values:

1. Each atomic proposition is a state formula.

2. If g and h are state formulas, so are $\neg g$, $g \wedge h$, $g \vee h$, and $g \rightarrow h$.

An atomic proposition, g, is true at time t if it actually occurred at t. Conversely, $\neg g$ is true at t if g is not true at t. Then, $g \wedge h$ is true at t if

both g and h are true at t; $g \vee h$ is true at t if at least one of g or h is true at t and $g \rightarrow h$ is true at t if at least one of $\neg g$ or h is true at t.

3. If f and g are state formulas, and $0 \leqslant r \leqslant s \leqslant \infty$ with $r \neq \infty$, $fU^{\geqslant r,\leqslant s}g$ and $fU^{\geqslant r,\leqslant s}g$ are path formulas.

The path formula $fU^{\geqslant r,\leqslant s}g$ is true for a sequence of times, beginning at time t if there exists an $r \leqslant i \leqslant s$ such that at time $t+i$ the state formula g is true and $\forall j : 0 \leqslant j < i$ the state formula f is true at $t+j$. The path formula $fU^{\geqslant r,\leqslant s}g$ is true for a sequence of times beginning at time t if either $fU^{\geqslant r,\leqslant s}g$ is true beginning at t or $\forall j : 0 \leqslant j \leqslant s$, f is true at $t+j$.

4. If f and g are state formulas, then $f \rightsquigarrow^{\geqslant r,\leqslant s} g$, where $0 \leqslant r \leqslant s \leqslant \infty$ and $r \neq \infty$ is a path formula.

For consistency with the type-level case, we treat leads-to formulas separately. The formula $f \rightsquigarrow^{\geqslant r,\leqslant s} g$ is true for a sequence of times beginning at time t if f is true at t and there exists an i, where $r \leqslant i \leqslant s$, such that g is true at $t+i$.

Finally,

5. If f is a path formula and $0 \leqslant p \leqslant 1$, $[f]_{\geqslant p}$ and $[f]_{>p}$ are state formulas.

In the token case, these state formulas are true at time t if there is a sequence of times, beginning at t that satisfy the path formula f.

Following this formulation, we may identify if any $c \in C$ is true on the occasion in question, in which case its support is simply the associated $\varepsilon_{a\nu g}$ value. However, if this set is empty – either none occurred or we do not have enough information to determine whether any occurred –

we must then calculate their probabilities, as described in the previous section. Note that we cannot assume that if the probability of a genuine or just so cause is non-zero, then the support for the corresponding token hypothesis will be greater than for any insignificant causes. We did not test whether any insignificant causes actually occurred, so it is possible that for a genuine cause, its probability is low enough that despite its higher value for ε_{avg}, an actually occurring (probability = 1) insignificant cause has a larger value for the support. In the case where there are many insignificant causes, testing whether each occurred may be computationally intensive. It is possible to define a threshold such that if the support for a cause is below it, insignificant and other causes are examined.

In any case, we begin with the probabilities, and thus support, for all genuine and just so causes. When these values are very low or zero, we must examine the other potential explanations: our previously discarded insignificant causes, and perhaps those that are not even prima facie causes. Further, it is possible that a negative cause (one that normally prevents the effect) actually was the token cause. After examining all of these, the final result is a set of possible explanations ranked by their support, with those having the highest values being the preferred explanations for the effect.

6.3 WHODUNIT? (EXAMPLES OF TOKEN CAUSALITY)

6.3.1 *The return of Bob and Susie*

Let us now return to a simple example to see how this approach works out in practice. We will again take the example of Bob and Susie, who are each armed with rocks that they may throw at a glass bottle. Now let us say we have already found one type level genuine cause (with all other causes being insignificant) of such a bottle breaking in this system. This relationship is represented by:

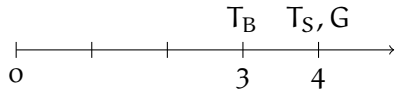
$$T \overset{\geqslant 1, \leqslant 2}{\underset{\geqslant p_1}{\rightsquigarrow}} G. \tag{6.19}$$

That is, throwing (T) a rock from a certain distance causes the glass to break (G) in greater than or equal to one time unit, but less than or equal to two time units, with at least probability p_1. Since we have found this to be a type-level cause, we have the associated value of ε_{avg} for the relationship: $\varepsilon_{avg}(T, G)$.

Now, on this particular occasion we aim to analyze, we will start with the following facts:

1. Bob threw his rock at time 3;
2. Susie threw her rock at time 4;
3. The glass broke at time 4;
4. The only genuine cause of a broken glass is that in formula (6.19).

As a timeline:



For each proposition (B, S, G), we have marked its time of occurrence (with T_B denoting T being true due to Bob's throw and T_S denoting T being true due to Susie's throw). Our type level relationship says that if T is true at some time t then it can lead to G being true at time $t+1$ or $t+2$. The facts we begin with are that T is true at $t=3$ and at $t=4$. We first test whether our type level relations token-occurred. For T_B to satisfy the causal formula of (6.19), G would need to be true at $t=4$ or $t=5$. G is true at $t=4$ and thus T_B can be considered as a possible token-cause of G. Now, for T_S to be a token cause of G, G would need to be true at $t=5$ or $t=6$. However, G is true at $t=4$, which means this causal relationship did not occur, and T_S is not a possible token cause (since it could not lead to G at the time at which G actually occurred). Thus in this case our only potential token cause is T_B, and the support for this token cause will be $\varepsilon_{avg}(T, G)$. Note that while in our system T_B must have caused G, the support for the hypothesis that T_B token-caused G is not one. If T had an ε_{avg} of one, meaning that it is the only type-level cause of the effect and no other factors make a bit of difference, then the support would be one.

6.3.2 *The case of Ronald Opus*

Take the following example, a paraphrased and condensed version of one presented by Don Harper Mills [126]:

A medical examiner viewed the body of Ronald Opus and determined that he died due to a gunshot wound to the head. It was found that he had jumped from a high building, intending to commit suicide, as was revealed in the note he had left. However, a few feet into his jump he was shot and instantly killed. Further, there was a net erected for window washers just two floors down from the roof, and it is assumed that this net would have prevented the completion of the suicide. It was also assumed that neither the jumper nor the shooter was aware of the net. Accordingly, the medical examiner ruled this a homicide, as Ronald would not have died had he not been shot.

It turned out that directly across the street from where Ronald was shot, an old couple had been arguing. The husband had threatened his wife with a shotgun, but due to his anger he could not hold the gun straight. Thus when he pulled the trigger, he shot the jumper across the street. The man and his wife insisted that they did not know that the gun was loaded and that he was merely threatening the woman, as he frequently did, with an unloaded shotgun. Since he had no intention of killing anyone, it seemed that the shooting of the jumper was an accident (as the gun had been accidentally loaded).

However, there was a witness who had seen the couple's son load the shotgun a few weeks prior. Their son, upset that his mother had cut him off financially, and knowing that his

> father threatens his mother with an unloaded gun when he is upset, had loaded the gun with the expectation that his father would shoot her. Thus it now seems the son is responsible for Ronald's death.
>
> Upon further investigation, it was revealed that their son, the same Ronald Opus, had become quite distraught about his situation, and his failure to get his mother murdered. In fact he was so upset that he jumped off the building across the street, but was killed on his way down by a shotgun blast through the window. The case was ruled a suicide.

Our goal is to use the facts about the case to determine whether the death should be ruled a murder, accident, or suicide. Unlike our previous examples, where the answers were intuitively obvious, the result is not immediately clear when we attempt to reason about this case. It is even more difficult to try to understand it using an automated method, where there is no room for intuition and background knowledge. This brings us to our first obstacle: throughout the example there is reference to what the father knew or what Ronald knew. However, we have not given any way to denote what a person knew, only the facts of the case. That is, we allow for the possibility that someone may commit a murder by shooting a person with a gun – even if they do not know that the gun they have shot is loaded. In most cases, we would reason about such a scenario as an accident.[8] Further, we have no method for representing intentions,

8 Interestingly, this is not always true in legal cases. For example, if in the course of committing a crime, one has a gun, the punishment automatically increases. If the gun is wielded or fired, it increases even more. The Supreme Court recently ruled that even in cases where the gun is fired accidentally, and is not known by the firer to be loaded, the increased automatic punishment still applies [20].

outside of adding these as propositions in our causal formulas. Despite the limitations of our method, we will attempt to dissect this example and make a ruling in order to show that we can make some advances even in such difficult cases.

Let us begin by summarizing what is known. At some time, a few weeks before the man and his wife had the fatal argument, Ronald was seen loading his father's gun. During the course of the argument, a suicidal Ronald jumped off a building across the street. At the same time, his father pulled the trigger, firing the fatal shot through an open window. As a timeline this is:

Note that we have taken these facts about the case for granted. We have not examined who the witness was who saw Ronald load the gun, nor the possible motives of this person. It is possible that this witness was actually the husband or wife or a third party who wanted to protect the husband. We have accepted that the husband was once again threatening his wife, with no intention of killing her on this particular occasion, despite his pulling the trigger. Further, we have believed that both the husband and wife are truthful about not knowing whether the gun was loaded – and not knowing that Ronald was outside the window. We have also assumed that neither Ronald nor his father knew about the net outside. We will accept these facts about the case in order to somewhat simplify our difficult task, but the reader should keep in mind what we have assumed and how the scenario would change if these assumptions did not hold. We will continue to omit time indices, as the case is already

quite complex, but remember that the events and causal relationships have associated times. We assume that the occurrences fit within the known time windows for any relevant type-level relationships.

Recall that when looking at the support for a hypothesis, we compute the support for c token causing e given that c and e token-occurred in population P, where there is a type-level relation between c and e in P. Our first task is to identify the type-level causes of death, whose support we will then compute. However, before we can do that we must identify the relevant populations. If we assume that the mother's actions (being the subject of threats) did not contribute to her son's death, we then have to consider populations related to the remaining two people. We will assume that we have access to precisely the populations and type-level relationships that we desire. Ronald's father frequently threatens his wife with an unloaded gun, and is part of a population of people who frequently wield unloaded guns and for whom shooting a gun has a very low probability of death. In fact, in that population, shooting is not a type level (positive) cause of death, and there are possibly no instances of a gun being loaded within it (population F).

Ronald, on the other hand, is part of a population of people who are homicidal (population H) since he was plotting the murder of his mother, and later suicidal as well (population SH) as he was distressed about the failure of his plans. In such populations, one would likely think that shooting a gun is a type level cause of death. Unfortunately, someone in

these populations did not shoot the gun. Thus, if our known type-level causes are:

$$(\text{jump} \wedge \text{no net}) \rightsquigarrow \text{death (in S), and} \tag{6.20}$$

$$\text{loaded gun} \cup (\text{shoot} \wedge \text{loaded gun}) \rightsquigarrow \text{death (in SH)}, \tag{6.21}$$

then, we still have no type-level causes that actually occurred. What we can do then is assess the strength of other, possibly insignificant, causal relationships: loading a gun in population H (perhaps loading a gun that someone else shot, or a gun that is later shot), jumping from a building with a net in population SH, and shooting a gun in population F. That is,

$$\text{load gun} \rightsquigarrow \text{death (in H)}, \tag{6.22}$$

$$\text{jump} \wedge \text{net} \rightsquigarrow \text{death (in SH), and} \tag{6.23}$$

$$\text{shoot} \rightsquigarrow \text{death (in F)}. \tag{6.24}$$

Now, as before, we begin by testing which of these occurred on the occasion in question. Then, we will use the associated ε's to determine the support for each. First, all three relationships are satisfied by the known facts. Next, for the ε's, it seems that sensible that jumping with a net rarely results in death and that the support for the relationship in (6.22) is likely quite high, and certainly much higher than the ε associated with (6.24). While the details may change based on the length of time between the loading and shooting (accounting for the fact that a gun may become unloaded by someone other than the initial loader), the ranking of these possible causes should persist, with loading a gun being more

significant than shooting a presumably unloaded gun and jumping out a window onto a net.

The intuition behind using two populations is that each man had different intentions and different sets of knowledge, and thus the same action by both would correspond to different rules and probabilities. Ronald knew that his father would hold and possibly pretend to shoot the gun, and loaded it with the intent to kill his mother. That is, Ronald had full knowledge of the scenario surrounding the gun, which is taken into account somewhat by noting that he is part of population H. His father assumed that the gun was in the same state where he left it, and would behave as it had in the past, namely, that it would not be loaded when he pulled the trigger. Reasoning about the father as part of population F, as he continued to act as part of F, captures this in a crude way.

Since Ronald loading has the highest level of support, it is the likeliest token cause of his death. However, the judgement on whether this corresponds to a suicide, homicide or accident goes beyond the reasoning we can do here. In order to do that we would need rather convoluted relationships such as "shooting a loaded gun with homicidal intensions leads to death by homicide."

6.4 DIFFICULT CASES

We will now look at a few classic scenarios that have been difficult to reason about in the token case. Since the examples are generally abstract and we are primarily looking at the reasoning behind them, details such

as the time subscripts of events have been omitted in some cases where the problem is clear without them.

6.4.1 *Overdetermination*

Symmetric case

We begin with symmetric overdetermination, where two known type-level causes of an effect both occur in the token case such that either could have caused the effect. Let us discuss Bob and Susie one last time. Recall that we have two people, each armed with a rock, which they may throw at a glass bottle. Let us say that Bob is standing a little closer to the bottle than Susie is. So, Susie aims and throws (S_t) her rock a little earlier than Bob does (B_t), but their rocks hit the glass simultaneously, breaking (G) it shortly after impact. That scenario may correspond to the following type-level relationships:

$$B_T \rightsquigarrow^{\geqslant 1, \leqslant 2}_{\geqslant p_1} G, \text{ and} \tag{6.25}$$

$$S_T \rightsquigarrow^{\geqslant 3, \leqslant 4}_{\geqslant p_2} G, \tag{6.26}$$

where people of type Bob, who stand closer to the bottle in this game, are represented by B and the relationship in (6.25) and those of type Susie, who stand further from the bottle, are represented by S and the relationship in (6.26). The facts are:

1. Susie threw her rock at $t = 1$;
2. Bob threw his rock at $t = 3$;

3. The glass broke at $t = 4$; and

4. The only significant causes of a broken glass are those in formulas (6.25) and (6.26).

We analyze this scenario as follows. First, note that both B_T and S_T occurred in such a way as to satisfy the formulas in (6.25) and (6.26). For B_T at time 3 to cause G, G would have to occur between time 4 and 5, which it did, and for S_T at time 1 to cause G, it would have to occur between times 4 and 6, which is also true. Thus we know not only that the probability of each potential cause given the evidence is one, but also that each occurred at such a time as to fulfill the corresponding token-level relationships. Then the support for B_T and S_T causing G will be the computed ε_{avg}'s. If these are equal, the support for either as the token-cause of the glass breaking will be the same. However, if Susie's aim is better, her value of ε_{avg} will be larger and thus the support for her breaking the bottle higher. Note that in that case we would not say that Bob's throw did not cause the glass to break, but only that there is more support (proportional to the difference in probability) for S_T than B_T causing G. Note that in practice, if instead of children throwing rocks we had the possible culprits for a patient's heart failure or carriers of an infectious illness, it is desirable to be able to identify multiple potential causes, with their associated weights.

Asymmetric case

In the previous case, either rock being thrown could have been the cause of the bottle breaking. Now we perturb this scenario slightly to make it asymmetrical, and an example of preemption. In this case, Bob throws

his rock a bit earlier than Susie throws hers, so his rock hits and breaks the glass before hers does. Now the bottle is already broken when Susie's rock hits it. We should deem Bob's throw the cause of the glass breaking. If S_T occurs at such a time that it could have caused G (according to the inferred rules), then we have no way to account for the fact that the bottle is already broken – we cannot augment the type-level relationships with our observations to give further constraints. However, since there is a small window of time in which a rock hitting a bottle can cause the bottle to break, if we can model the events more finely using variables such as B_H and S_H to denote whether the corresponding rocks have hit the bottle, then we can correctly handle this case. If in practice we find incorrect diagnoses using our inferred type-level causes, we can take this as an indication that these are too coarsely grained to capture the details of the system, and we should go back and look for relationships with more detail and at a finer timescale. This has traditionally been a difficult case for methods that look for the earliest cause that accounts for the effect. In those cases, if Susie throws earlier than Bob, but is standing further away, so that her rock still hits after the glass is broken, we incorrectly find that since she threw the first rock, she caused the bottle to break.

It is important to note that the difficulties in our case are due to not modeling the events finely enough and not being able to account for observations that are outside the causal formulas. Had we not observed the rocks hitting the bottle, the idea that either throw could have caused the glass to break would be acceptable. The contradiction is that we cannot augment the type-level relationships with our observations of further constraints. We could look for more specific type-level relationships,

using the rock hitting the bottle as an event, or specifying that the rock hits an unbroken bottle.

6.4.2 *The hard way*

Another set of examples all have the same structure, but highlight various features of the problem as well as the differing intuitions one may have in each case. In these types of scenarios a cause of the effect occurs, followed by an event that usually makes the effect less likely, but which in this particular case seems to bring about the effect. Thus the effect occurred, but happened "the hard way" – with most odds stacked against it. While there are a number of examples of this type, we will look at the three most widely used, and introduce one of our own.

Sherlock Holmes and the Boulder

We begin with an example by Good [43], with some modifications by Hitchcock [55]. Sherlock Holmes takes a walk below a cliff, where his nemesis, Moriarty, is waiting for him. Moriarty has set a boulder on the edge of the cliff so that when he sees Holmes walk past, he will push the boulder off the edge, giving him a 90% chance of killing Holmes. Holmes's loyal companion Watson, however, sees what Moriarty is plotting and decides to push the boulder out of Moriarty's hands. Just as Holmes walks below them and Moriarty is about to push the boulder, Watson runs over and pushes it first, trying to aim it in another direction. This random push, since Watson is unable to see Holmes and be sure that the boulder is aimed away from him, has a 10% chance of killing Holmes.

In an unfortunate turn of events, the boulder falls directly on Holmes, killing him.

The type level relationships are that a boulder pushed by an enemy (E) is a type-level cause of death by boulder (D), with probability 0.9:

$$E \rightsquigarrow_{\geqslant 0.9} D,$$

while a boulder pushed by a friend (F) is likely not a type-level cause of death, but this depends on the probability of death by boulder. That is,

$$F \rightsquigarrow_{\geqslant 0.1} D,$$

but we do not know $P(D)$, so we are not sure whether the probability of death is raised. Let us say the probability of death by boulder (this includes boulders pushed as well as those falling from above) is lower than 0.1. Then, F is a prima facie cause of D. It may still be an insignificant cause of D but the value of ε_{avg} is quite likely to be positive. Now, one type-level cause of death has occurred: F. We find that the support for F as a token cause of death is positive, but probably small (since F actually occurred, the support is precisely the earlier computed ε). Nevertheless, no other causes occurred, so we are left with F as the only possibility. However, as Hitchcock [55] notes, we are comparing the probability of death when pushed by Watson to the probability of death when not pushed at all, not to the probability when the boulder is pushed by Moriarty. This is why, while we may not think Moriarty caused Holmes's death, we find that relative to no push, he raised the probability of death. In the traditional counterfactual approaches, we would have reasoned

that if Watson had not pushed the boulder, Moriarty would have, and thus Watson did not cause the death since Holmes would have died anyway. This seems plausible, but it requires us to accept that Moriarty's push was inevitable.

In the other case, $P(D)$ is somewhere in between, perhaps 0.5 (maybe there are frequently falling boulders). Now, we have one type-level cause of death, but it did not actually occur, since Moriarty was prevented from pushing the boulder. Thus, we must examine other possible causes, testing relationships comprised of the facts known about the scenario. The primary possibility is that F led to D. However, F lowers the probability of death (it is a *negative* cause of death by boulder). Thus, the computed $\varepsilon_{avg}(F, D)$ will be negative, and the support for a boulder pushed by Watson as the token-cause of Holmes's death is negative. What does it mean for the actual cause to have negative support? Remember, we are implicitly, through the ε_{avg}, comparing one possible cause to others. In this case we have found that at the type level, F is a negative cause of death. When a negative cause is the token-cause of death, we could potentially give it the interpretation of "despite", as a negative cause usually has the opposite outcome. In the case where the actual cause is only insignificant, it is not clear that we can use the same despite interpretation, but perhaps just that the effect was unlikely. Now, we could go further and say that witnessing a foe attempting to push a boulder is a cause of a friend rushing in and pushing it instead. That is, since it is known that pushes by foes are more deadly than pushes by friends, a friend is likely to attempt to save the person by pushing the boulder themselves. However, we do not automatically say that if X caused Y and Y caused Z then X caused Z, so it would not change

the cause of death in this case. We would only say that witnessing Moriarty about to push the boulder caused Watson to push the boulder, and Watson's pushing the boulder caused Holmes's death. Whether Moriarty caused Holmes's death by causing Watson to push the boulder is a separate issue that is relevant primarily in cases of assigning blame, but we will not discuss this here.

The plant and the defoliant

This example, due to Cartwright [10], has the same structure as the case of Sherlock Holmes, but may lead to a different interpretation. Nancy wants to get rid of some poison oak in her garden, so she sprays it with a defoliant that will cause the plant to die with probability 0.9. However, even after a few months, the plant is still alive. Let us say that the probability of plant death in the absence of a defoliant is only 0.1. We will shift our attention to survival, though, so this case better parallels that of Holmes. Thus, spraying the plant with the defoliant (D) leads to survival with probability 0.1, whereas the probability of survival if no action is taken is 0.9. In the previous case, the probability of death from Moriarty's push was 0.1, whereas the probability of death if no action was taken on Moriarty's part was 0.9.

In the case of the defoliant, it does not make much sense to ask what caused the survival, as the plant was alive both before and after the spraying. When reasoning about death, the system (Holmes or the plant) is in a different state after some action is taken, so we aim to determine why the system has changed states. The fact that we do not usually ask what caused a state to persist is likely responsible for our thought that the spraying did not cause the survival, while we would agree that it

can at least be argued that Watson caused Holmes's death, regardless of $P(D)$.[9] That said, we can in fact reason about the case as before and compute the support of spraying with defoliant (D) as a token-cause of survival (S). Note:

$$P(S|D) < P(S).$$

Thus D is not a prima facie cause of survival, but is in fact a negative cause of survival. As in the case of Holmes, the computed ε_{avg} will be negative and the support for D as a token cause of S will be negative. We can again give this the interpretation of: the plant survived despite the spraying of the defoliant (a type-level negative cause of survival), which is consistent with one's intuitions about the problem. Note that this hypothesis is only tested due to our knowledge of the problem – it is unlikely that we would wonder what caused survival and if we did, there are likely genuine causes (sunlight, water, etc) which did occur and thus we would not automatically examine an insignificant cause such as the defoliant. Despite that, we are still able to test this hypothesis and arrive at an answer that is consistent with intuition.

The golfer and the squirrel

We now look at a classic example, due to Rosen [110], with some modifications by Salmon [112] and Hitchcock [55].

9 Another way of explaining this is that the situation would correspond better to the Holmes case if there was a 99% effective defoliant, thus using the weaker one relative to the stronger version can be argued to have caused the survival of the plant [55]. Alternatively, this case can be understood in terms of capacities. While the push of a boulder is capable of causing death, the spray of a defoliant does not seem to have the needed capability to cause survival.

> Alvin, a slightly above average golfer, tries to make a birdie on a particular hole. He hits the ball and it looks like he might just make the shot. However, a mischievous squirrel comes by and kicks the ball. Usually when such squirrels kick golf balls they lower the probability of making birdies. However, in this case the squirrel kicked the ball right toward the hole and Alvin made the shot.

Now, the question is: what caused the birdie? The squirrel's kick (K) lowered the probability of a birdie (B) but it seems to have caused it. Recalling the work of Eells, which we discussed in Chapter 2.3.3, we could analyze this example using its probability trajectory. In that way, we distinguish between the general properties of squirrels and golf balls and how this particular squirrel affected the probability of this particular golf ball going into the hole. However, it is unlikely that without extensive background knowledge we could ever know the true probability trajectory – that is, the probability of a birdie at each moment in time from Alvin's shot to the actually occurring birdie. So, we will proceed as discussed above.

First, what are the relevant type-level causes of birdies? For the sake of simplicity the times of cause and effect are omitted, but we assume some known window of time after the ball is hit in which a birdie may be made (i.e. hitting the ball on Tuesday cannot cause a birdie on Friday).

We will assume that an above average golfer hitting the ball (A) raises the probability of a birdie (B): [10]

$$P(B|A) > P(B).$$

Thus, $A \rightsquigarrow B$, with some moderate probability. However, kicks by squirrels (K) lower the probability of birdies:

$$P(B|K) < P(B),$$

so K is not a prima facie cause of B and is likely a negative cause of B. In this example, we have only one type-level positive cause of the birdie: the golfer's swing. So, we will find that this swing token-caused the birdie since there are no other type-level causes of birdies. If we still want to assess the significance of the squirrel for the birdie, we will find as before that squirrels have negative significance for birdies, and the birdie occurred despite the squirrel. Note that we cannot capture the fact that this particular squirrel happened to kick this particular ball in just such a way that we know it was actually responsible for the birdie. Here we diverge from the results of Eells, who, using his probability trajectories (showing the probability of the birdie became higher after the kick and remained high until the actual birdie occurred) found that the squirrel's kick caused the birdie. Due to the probability trajectory Eells can distinguish between the general properties of squirrels and golf balls and how this particular squirrel affected the probability of this particular golf ball going into the hole.

10 Note that if Alvin was a terrible golfer, the analysis would be unchanged, with the exception that the hypothesis with the most support (Alvin or the squirrel causing the birdie) could change depending on just how bad a golfer Alvin is.

Since we do not know and cannot represent this change in probability in our method, we will only find that the birdie occurred despite the squirrel, not any contribution the squirrel made to the birdie. While this result may be problematic, it is also important to consider how often – and in what cases – we will actually know such a trajectory and get results inconsistent with our knowledge of the situation. A possible real world equivalent is: Alvin has a genetic mutation that gives him an above average probability of developing lung cancer. Once he finds out about this, he stops smoking in order to protect his lungs. Two years later, he is diagnosed with lung cancer. In general, we would say that cessation of smoking lowered Alvin's probability of lung cancer. Thus, we would say that his lung cancer was despite the fact that he stopped smoking. However, later research shows that, oddly enough, people with Alvin's mutation are more likely to develop lung cancer once they stop smoking. We can use this added information and find that stopping smoking was a positive cause of developing lung cancer. Note that our first assessment was correct as far as knowledge at the time, when we learned more about the underlying type-level relationships, we were able to better explain Alvin's condition. In the case of the squirrel, perhaps this particular squirrel was benevolent and attempting to aid the golfer. If we later obtain this information, we could find that he was a token-cause of the birdie.

A car accident and seatbelt use

We now return to our previous example of a car accident in which a seatbelt causes death, which turns out to be another example of things happening "the hard way." On Monday morning Paul drove his car to

work, wearing his seatbelt as he always does. Unfortunately, on this particular drive he was in a bit of a rush and collided head on with an ice cream truck. The collision resulted in injury to Paul's carotid artery. He later died, apparently due to this injury.

Let us make the example somewhat tricky by assuming we know only of a general type-level relationship between seatbelts and death (that is, not one involving carotid artery injury specifically). This relationship accounts for the myriad ways (including by carotid artery injury) a seat belt can cause death. However, since seatbelts generally prevent death, the associated probability will be quite low. Let us further assume that a seatbelt can only cause death in the context of a car accident. Then, we have general relationships between death (D), car accidents (C) and wearing a seat-belt (S):

$$P(D|C \wedge \neg S) > P(D|C \wedge S) > P(D), \tag{6.27}$$

and a general relationship between car accidents and death:

$$P(D|C) > P(D). \tag{6.28}$$

For ease we have omitted the time subscripts and assume the token events are within the known type-level time frames. So, this is akin to the probability of death within some window of time, versus the probability of death within some window of time given that the person has been in a car accident. While it may be that being in a car accident and not wearing a seatbelt is a significant type-level cause of death, being in a car accident and wearing a seatbelt results in a lower probability of death. However, it

is still at least a *prima facie* cause of death. It seems unlikely that seatbelt use plus a car accident should be a negative cause of death, so we will ignore that possibility.

Given we have the relationship, $C \rightsquigarrow D$ (let us say that this is all considered to be a single population, not separate populations of seatbelt users and non-users or good drivers and bad drivers), which is a significant type-level cause of death, we find that it occurred in the token case, it is the only type-level genuine or just so cause that occurred (since $C \wedge \neg S$ is false), and it has a value of support equal to the associated ε_{avg}. This means that unless $C \wedge S$ is a significant type-level cause, we would not automatically consider it as a possible explanation for the effect. Thus, regardless of whether the specific injury was caused by the seatbelt, it would still be the car accident that caused death. While this explanation is not as precise as what we may desire, note that the seatbelt injury only occurs within the context of a car accident, so we can think about this case as death by car accident, with the mechanism being carotid artery injury due to the seatbelt. That is, there is a general relationship between car accidents and death, and this relationship may be fulfilled by a variety of means (seatbelt injury, airbag injury, ejection, etc.).

As before, we may still want to test the hypothesis of $C \wedge S$ causing death (just as we tested other unlikely hypotheses outside our general algorithm in the previous sections). In this case, we see that it did occur and its support will be exactly equal to its ε_{avg}, which will be less than that of C as a cause of death. Thus what we may know to be the "actual cause" will have less support than a more general cause. This case has a slightly different structure than the previous ones, since we included the general car accident-death relationship. If we omitted this, and only

had $C \wedge \neg S$ as a type-level genuine cause, with $C \wedge S$ as an insignificant or negative cause of death, we would have precisely the same case as in the previous example, with Holmes, Watson, and Moriarty. Similarly, we could have previously looked at the probability of death when hit by a boulder (regardless of who caused it) and found that to be the cause of death. In the modified seatbelt case, where we omit the relationship $C \rightsquigarrow D$, we have no occurring significant type-level causes. Thus, we examine our insignificant and other causes. In this case, $C \wedge S$ would be found to be the only known potential cause and would again have low support. Just as before, we would say this was an unlikely occurrence, but the seeming cause of death.

7

APPLICATIONS

We will discuss two types of applications: one involving synthetically generated data, where the goal is to see how well our algorithms can recover the known causes, and the other using real data, where the goal is to find novel relationships of interest. We look first at validation on generated neuronal data, then at both validation and experimentation on financial time series. We will also compare the software implementation of our approach, called, `AITIA`, against other competing algorithms.

7.1 NEURAL SPIKE TRAINS

We begin our study of applications with the case of synthetically generated neural spike trains. The inferred relationships will be simple (one neuron causing another to fire in some pre-defined window of time), but the data will allow us to validate our algorithms in an area of interest. As discussed in Chapter 3, there has been much recent work on determining the connectivity between neurons by applying causal inference methods to spike train measurements. This is an area where timing information is a central part of the causal relationships, so it is useful to compare our approach to others that include this information to varying extents.

7.1.1 *Synthetic MEA data*

The data were created to mimic multi-neuronal electrode array (MEA) experiments, in which neuron firings may be tracked over a period of time.[1] Data was generated for five different structures, with neurons denoted by the 26 characters of the English alphabet. Each data set contained 100,000 firings generated using one of the five structures plus a degree of noise (this is a parameter that was varied). A total of 20 data sets were generated, with two runs output for each structure and each of two noise levels. The five structures (shown in figures 7.8, 7.9, 7.10, 7.11, and 7.12), include a binary tree of four levels, a chain of neurons, and so called "scatter gather" relationships in various configurations.

At each time point a neuron may fire randomly (with the probability of this happening depending on the noise level selected, with a higher noise level meaning a higher probability) or may be triggered to fire by one of its cause neurons. Additionally, there is a 20 time unit refractory period after a neuron fires and then a 20 time unit window after this when it may trigger another to fire. Consequently, our algorithm need only search for relationships where one neuron causes another to fire during a window of 20–40 time units after the causal neuron fires. That means that when testing for *prima facie* causality, the relationships will be of the form $c \rightsquigarrow^{\geqslant 20, \leqslant 40}_{\geqslant p} e$, where c and e represent the firing of individual neurons.

1 The data was provided as part of the 4^{th} KDD workshop on Temporal Data Mining. It is publicly available at: `http://people.cs.vt.edu/~ramakris/kddtdm06/`.

7.1.2 *Comparison with BNs, DBNs and Granger causality*

We compared our results to those found with the TETRAD IV [42] implementation of the PC algorithm of SGS [120], the Banjo package for DBN inference [50] and the `granger.test` function in the MSBVAR R [8] package on the same data. All algorithms tested for simple pairwise relationships between neurons, but the use of timing information varied. When possible we used the default settings for each software package.

Our algorithm tested for relationships where one neuron causes another in 20–40 time units. We then computed the empirical null from the set of ε_{avg} values using the method and R code made available by Jin and Cai [60].

TETRAD IV was given the full time series data and for each input it produced a graph with both directed and undirected edges (with the undirected edges indicating a relationship, with the algorithm unable to determine whether the nodes cause each other or have a common cause). Undirected edges were not considered to be true or false positives, they were ignored in these calculations to provide better comparison with other algorithms.

The Banjo package was used with simulated annealing, testing for links between neurons in 20–40 time units (note that this is not a window, but rather determines whether A causes B in 20 time units, 21 time units, and so on with one arrow in the graph for each of these temporal links). Based on information in the documentation and the size of the problem, the algorithm was allowed to run for 30 minutes on each input file. The algorithm output the graph with the highest score, indicating edges

Method	FDR	FNR	Intersection
`AITIA`	0.0093	0.0005	0.9583
Granger	0.5079	0.0026	0.7530
DBN	0.8000	0.0040	0.4010
PC	0.9608	0.0159	0.0671

Table 1.: Comparison of results for four algorithms on synthetic MEA data, with ours being AITIA.

between each of the neurons for each of the timepoints. The algorithm never identified relationships between a pair of neurons for the entire time window (i.e. an edge between them for each value in [20,40]), so we collapsed all inferences to be between two neurons (i.e. if there was an edge between two neurons for any value in [20,40], we called that a positive. If there were ten edges found between two neurons, that still corresponded to one relationship).

We used the `granger.test` function with a lag of 20 time units, as it is not possible to specify a window of time using this algorithm[2]. The algorithm output F-scores and p-values for each possible pairwise relationship. To determine the threshold at which a relationship was considered a positive result, we used the same false discovery control approach as was used with our own algorithm.

The results for all algorithms over all datasets (five patterns with two runs each for a low and high noise level) are as shown in Table 1. While we are primarily focused on controlling the FDR, we also include statistics for the FNR (fraction of false negatives out of all negatives – these occur when we fail to identify a causal relationship) as there is generally a tradeoff between controlling the FDR and FNR. Here we see that in fact

2 If we had used 40, then in scenarios such as A causes B and B causes C, the algorithm would be likely to find A causes C.

we have the lowest values for both the FDR and the FNR, with an FDR of less than 1% (two orders of magnitude lower than the competing approaches). Note that for all methods, the FNR is fairly low. This is due to the small number of true positives compared with the large number of hypotheses tested. Finally, since there were two runs for each embedded pattern at each noise level, we tested the robustness of our findings by calculating the size of the intersection between those runs. That is, we measure the number of significant causal relationships found in both runs, as a fraction of the size of the union of both runs. Our algorithm was the most consistent according to this measure.

Looking at the results for each algorithm, the false discovery rate for the PC algorithm is not unexpected, as the method tests for relationships between neurons, without testing the timing of that relationship. However, it is interesting to note that DBNs fared worse than Granger causality by all measures, despite the fact that Granger causality has difficulty distinguishing between mere correlations and causation. One possible reason for this is that since the graphical model methods score the entire graph, in theory they must search exhaustively over graphs, but this is not feasible, and thus heuristics must be used. While the greedy algorithms may get stuck in local maxima, simulated annealing algorithms must be stopped before they overfit the data. This overfitting is likely what happened, and why DBN methods perform better when the consensus of a set of graphs is taken [130]. On the other hand, both Granger causality and our algorithm run for set periods of time, consistently returning the same results for the same input.

We will examine in detail one of the five structures recovered. Figure 7.11 shows the true embedded structure, which is one of the most

difficult to infer, as neurons such as D and E are both highly correlated with H and I. The results for our algorithm are shown in figure 7.2, with a histogram of the computed z-values for the 641 prima facie causal hypotheses. The empirical null in this case is given by $N(-0.14, 0.39)$, so it is shifted slightly to the left of the theoretical null, and is significantly narrower. The tail of the distribution extends quite far to the right, continuing up to 8 standard deviations away from the mean (almost 20 times the empirical standard deviation). A close up of this area is shown in figure 7.3. The results obtained here are consistent with the known causal structures that were used to create the simulated data.

In figure 7.1 we compare our results on this structure with those of the other algorithms, to better visualize the false discoveries and non-discoveries made by each. Looking at the output from the Granger algorithm, we see that in this case all of the true relationships were identified, but that neurons with a common cause were found to be linked. For example, there is no causal relationship between B and C, but because they are both caused by A, the Granger test found a strong correlation between them. The results from the PC algorithm show that only one relationship, an undirected one between B and C, was found in both runs for this dataset. That means that depending on the input, entirely different relationships were found, suggesting that the algorithm is overfitting to the particular dataset, while also missing the true relationships, since the temporal component is excluded. Note that this is one of the cases where the assumptions made by the PC algorithm hold, as all common causes are measured and in the dataset, and since there are no inhibitory relationships, none could be "canceled out" by an unlucky distribution. Finally, looking at the DBN result for this dataset, shown

in figure 7.7, we see that while the correct relationships were identified (again remembering that an edge between A and B means A causes B at some specific time within 20–40 time units, but for ease of representation we are not showing the edge for each temporal relationship), there are many erroneous edges. Unlike the results of the Granger algorithm, these edges are totally unrelated to the embedded pattern (i.e. they are not misinterpreting a correlation as a causal connection). As noted above, we see that in large part these relationships are found in only one run, suggesting that the software is overfitting the particular distribution given.

7.2 FINANCE

7.2.1 *Simulated financial time series*

Data

To compare the proposed approach to existing approaches in finance, we developed a set of simulated financial time series.[3] This allowed us to embed a variety of causal relationships in the data and see how well each algorithm is able to recover these, finding the specific weaknesses of each.

To do this we used a factor model [34] that allowed two kinds of causality: one through the influence of factors on stock portfolios and the other a direct dependency between individual portfolios. Our simulated market consisted of 25 portfolios, with data generated for six scenarios during

3 The data was generated in close collaboration with researchers in mathematical finance. In particular, the methodology for simulating the data was developed by Petter Kolm, with assistance from students in NYU's mathematical finance program.

Name	One lag	Random lag	Portfolio dependency
A			
B	✓		
C		✓	
D			✓
E	✓		✓
F		✓	✓

Table 2.: Summary of datasets created. Half the portfolios in a dataset may have their factors lagged by a single amount (one lag), or half may have each individual factor lagged by a random amount in $[0, 3]$ (random lag). When dependency between portfolios is included, there are three portfolios whose return at t_i depends on the returns of another portfolio at t.

two 3001 day time periods. In each scenario factors could be shifted for each individual portfolio or there might be dependency between portfolios. We initially assume that a portfolio's return at time t depends on the values of the factors at time $t - 3$, making it possible to test whether factors may be treated as common causes of portfolio returns.

The six portfolios, summarized in table 2, contain three (A–C) with no dependency between individual portfolios, and three (D–F) where three such relationships were included. Each portfolio in the set can have its factors lagged the same amount (A,D), have half the portfolios lagged by a different amount (B,E) or have half the portfolios lagged by a random amount in the range [0,3], where each factor for a portfolio can be lagged independently of the others (C,F). To summarize, the six types of datasets generated are:

- A: All portfolios lagged $t - 3$, no dependency between portfolios;
- B: Half of the portfolios lagged $t - 1$, half $t - 3$, no dependency between portfolios;

- C: Half of the portfolios have each factor lagged by random amount between $t-0$ and $t-3$, other half $t-3$, no dependency between portfolios;
- D: All portfolios lagged $t-3$, three causal relationships between individual portfolios;
- E: Half of the portfolios lagged $t-1$, half $t-3$, three causal relationships between individual portfolios; and
- F: Half of the portfolios have each factor lagged by random amount between $t-0$ and $t-3$, other half $t-3$, three causal relationships between individual portfolios.

This means that each portfolio could have all of its factors lagged the same amount (cases A,B,D, and E) or some portfolios may have each factor lagged independently (cases C and F).

The return for portfolio i at time t is then given by:

$$r_{i,t} = \sum_j \beta_{ij} f_{j,t'} + \varepsilon_{i,t}, \tag{7.1}$$

where factor j at time t is denoted $f_{j,t}$. In case A, $t' = t-3$. In cases D, E, and F, ε is the sum of the randomly generated idiosyncratic (also called error) terms plus, in the case where portfolio i depends on portfolio k, $\varepsilon_{k,t-1}$. To construct these series, we used the Fama-French daily factors [34] from July 1963 through December 2007, and the 5×5 size/book-to-market portfolios, also generated by Fama-French [37]. For each of the scenarios A through F we constructed two return series, the first using daily returns from July 2, 1975 through May 15, 1987 and the second from April 12,

1995 through March 14, 2007 (time points 3000 to 6000 and 8000 to 11000 in the factor series respectively).

Tests & Results

We compared our algorithm with the MSBVAR `granger.test` [8] function in R, as the Granger test is a standard method used in analyzing such data. In order to assess the algorithms as well as some common assumptions and practices, we conducted a series of tests on the twelve datasets (six scenarios for two ranges of timepoints): one using the generated returns (sequences of $r_{i,t}$s as defined in (7.1)), one using the actual error terms used to construct the returns (sequences of $\varepsilon_{i,t}$s as defined in (7.1)), one using the generated returns with the known factors (sequences of $f_{j,t}$s) included to give a total of 28 variables, and finally one comprised of residuals calculated by regressing the returns on the known factors (approximating a common approach to such time series). For both algorithms we tested pairwise relationships between elements of the time series (portfolios, and in some cases factors) at lags of 1, 2, and 3 days. For our algorithm, this meant testing whether a positive/negative return for one variable caused a positive/negative return in another. Since the Granger implementation only returned the significance of a relationship between variables (regardless of whether it was positive or negative), true positives were broadly defined as being that there is a causal relationship between two variables in a certain amount of time.

The procedure for each was to define the set of causal relationships to be tested and then run each algorithm to compute the significance of each relationship in this set, resulting in a set of ε_{avg}'s for our algorithm and F-statistics with their associated p-values for `granger.test`. Then,

the empirical null hypotheses and false discovery rates for each test were computed and relationships with an fdr$<$ 0.01 called significant. For our tests we used the `locfdr` R package from Efron et al. [31] to compute the null hypothesis, while we found that due to the different distribution, the `fdrtool` package [123] provided better results for `granger.test`.

In order to compute FDR and FNR rates, we must understand what constitutes a true positive. In the simplest case, when using the generated returns, there should be no causal relationships found in scenario A, while in B and C we should find that portfolios with lags less than $t-3$ should cause those with greater lags, with the time associated with the relationship being that of the difference between the lags. In datasets D–F, our findings should be the same, with the addition of the embedded relationships between portfolios. While the way the data is generated may make it seem that the factors could cause the portfolio returns, examination of the factors reveals that this is not the case. Recall that the Fama-French factors are constructed from the stocks themselves, thus when we lag the factors this is a proxy for some portfolios responding to external influences and affecting the market factors earlier than others.

In the datasets consisting only of the error terms, we should find only the embedded relationships between portfolios (since there is no influence from factors in these time series). Similarly, when we look at the residuals, we expect that the result should ideally be the same as that for the error terms and no influence from factors should remain. However, in practice, the returns are not so cleanly split into factor/error terms and the result of regressing on the factors and removing this component is not the same as the original error terms. We confirmed that this is the case, and that the relationships between lagged and unlagged portfolios persist.

Method	FDR	FNR	Intersection
AITIA	0.0775	0.0417	0.8090
Granger	0.6547	0.0863	0.4347

Table 3.: Comparison of results for two algorithms on synthetic financial data.

This data was not used for computation of error rates. Finally, we can also include the factors in the dataset. If the factors were not derived from the stocks, we would find them to be common causes of the lagged portfolios. However, since they may be viewed as both cause and effect, it is not possible to truly determine the underlying structure in this case. For our assessment of the two algorithms we focus on the returns data (which in addition to being the most straightforward was also the one on which both algorithms performed best). We will briefly discuss the idiosyncratic (error) term data, but for the above reasons, the residual and combined portfolio/factor experiments do not lend themselves to rigorous quantitative assessment.

In Table 3 are FDR, FNR and intersection results for the generated returns. These values are across all twelve datasets (two for each scenario), and include relationships with all levels of lags. We also compare how consistent our results are by computing the intersection of relationships found in both time ranges for a particular scenario. Since the only causal relationships in the system are those we embed, the relationships found should be the same. Note that once again we have the lowest values for both FDR and FNR, as well as the most consistent results.

On the error (also called idiosyncratic) returns, the FDRs were quite high (0.827 for our method and 0.988 for Granger), owing to the fact that there are extremely few true positives (a total of 3 in each of the D–F

datasets and 0 in each of the A–C datasets) and potentially some lingering dependency between portfolios. The 18 true positives were found by both algorithms, with both having zero false negatives. The only substantial difference was in the quantities of false discoveries. While the rates were high for both algorithms, our approach made 86 false discoveries (out of 104 total) while Granger made 1442 (out of 1460).

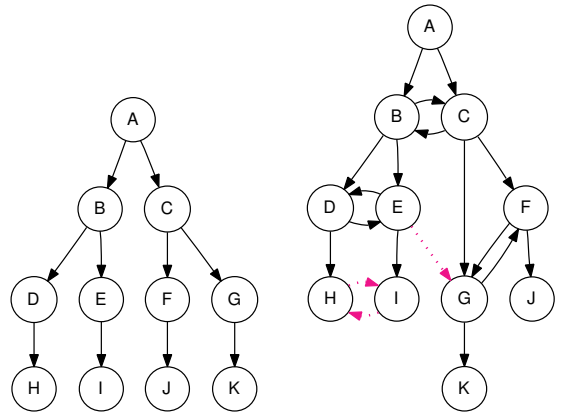
We note that while our FDR on the returns data is substantially lower than that of `granger.test`, it is still higher than our desired rate of 0.01. This is entirely due to difficulties in correctly inferring the null distribution. Since the number of true positives can be substantial (in some cases much greater than the 10% frequently assumed), we violate one of the assumptions of these methods: that our observation is mostly from the null distribution and that there are a small number of deviations from that, corresponding to non-nulls. In fact in many cases visual inspection of the graphs reveals that a human could clearly see the separation between the two classes (See examples in figure 7.4). When we allow for manual choice of thresholds, our FDR is reduced below our specified threshold (to 0.0097), with a negligible increase in false negatives (from 0.0417 to 0.0480). The results also become quite consistent (intersection of greater than 98%), meaning that the true positives are found in both runs (and that it is possible to improve results by calling significant only those causes found significant in both runs). Further work on empirical null methods will be necessary to bring automated analysis closer to this ideal.

7.2.2 *Actual financial data*

To determine how similar actual market data is to our synthetic returns, we tested our algorithm on daily stock returns using the CRSP database, downloaded through WRDS. We began with all stocks that were in the S&P 500 for the entirety of January 1, 2000 to December 31, 2007 and that remained in the S&P 500 through September 2009. Since this yielded over 2000 trading days and hundreds of stocks, we tested random subsets of 100 stocks in this set. Over the entire timecourse, we found no significant relationships (using fdr$<$ 0.01), when testing for pairwise relationships between stocks at a timescale of one day. Looking at the last 800 timepoints, we found a single significant relationship, and again found zero looking at the last 400 timepoints. In figure 7.5 we show the histograms for the test results, which illustrate that in all cases they conform closely to the null normal distribution. No significant relationships were found at longer time scales (i.e. multiple days). One explanation for the few discoveries made is that at the timescale of one day and over long periods of time, relationships between companies do not persist (and are overshadowed by market-wide factors).

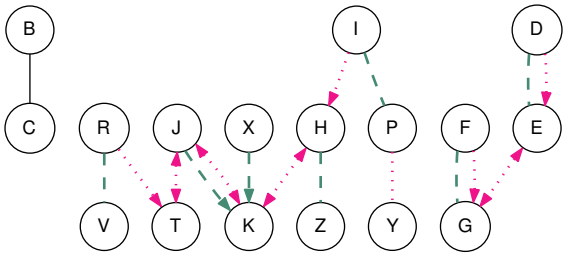
Finally, we focused on one year of trading, using the last 252 timepoints from this series. Due to the shorter time series, we examined a larger set of stocks: those that were in the S&P 500 during the 2000–2007 time period. There were 386 such stocks, and 27 significant relationships. The significant relationships are shown in figure 7.6 and are primarily of the form "a price increase in x causes a price decrease in y in exactly 1 day" (denoted by a dashed line in the figure), with a few of the form "a price

increase in x causes a price increase in y in exactly 1 day" (denoted by a solid line in the figure). Many of the causes in this set are companies involved in oil, gas and energy, while financial companies appear to be influenced by results from the technology sector.



(a) Inferred by AITIA

(b) Inferred by Granger test



(c) Inferred by PC

Figure 7.1.: Graphs representing inference results, with arrows denoting that the neuron at the tail causes the neuron at the head to fire: 7.1a Arrows denote relationships where the cause leads-to the effect in 20–40 time units, 7.1b Arrows denote that a neuron causes another to fire in 20 time units, 7.1c Arrows denote conditional dependence relationships and have the usual BN interpretation. Colored (and dashed/dotted) arrows refer to relationships that were found in one of the two runs for this parameter setting, with solid black arrows denoting relationships found in both runs. The DBN results appear separately in figure 7.7, as the graph is quite large.

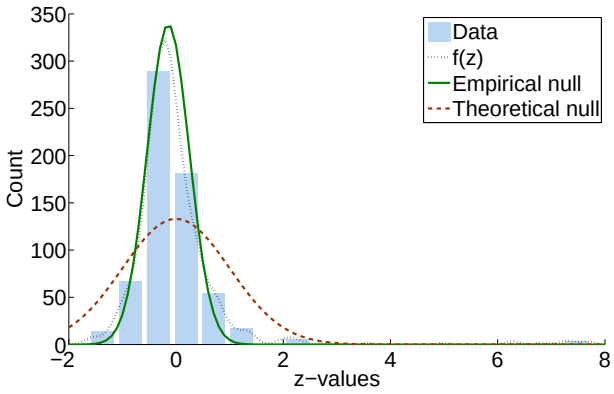


Figure 7.2.: Neural spike train example. We tested pairwise causal relationships, taking into account the known temporal constraints on the system.

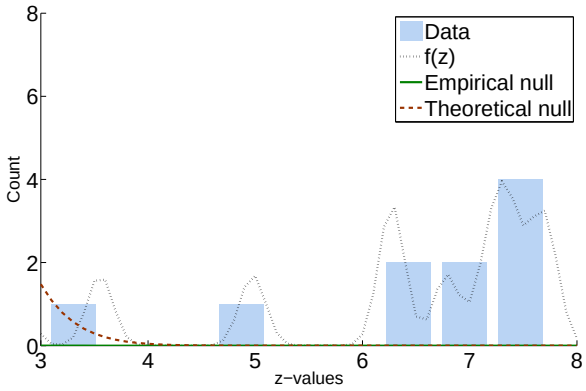
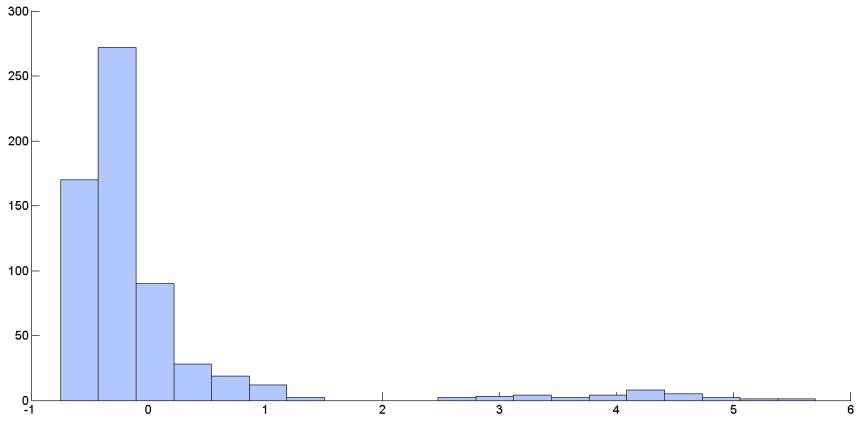
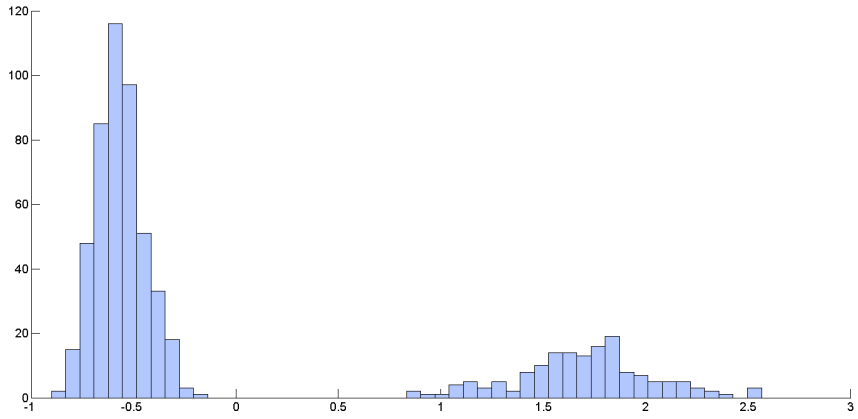


Figure 7.3.: Close-up of the tail area of Figure 7.2. The relationships in this area are exactly those of Figure 7.11.



(a) Results for one of the "C" datasets.



(b) Results for one of the "B" datasets.

Figure 7.4.: Histogram of z-values computed from the set of ε_{avg} values for two tests, using our algorithm.

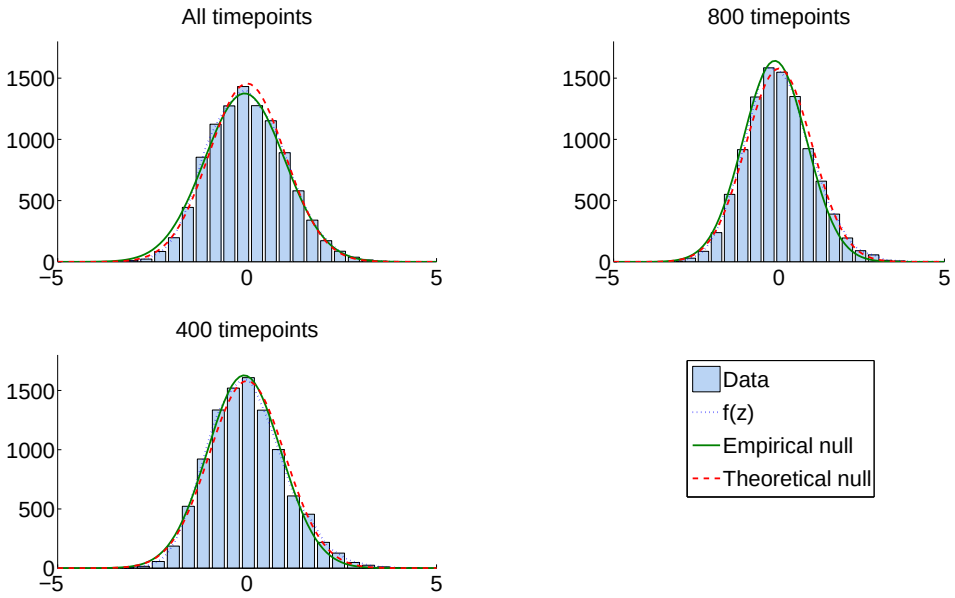


Figure 7.5.: Test results for our inference algorithm on various sized subsets of the actual market data.

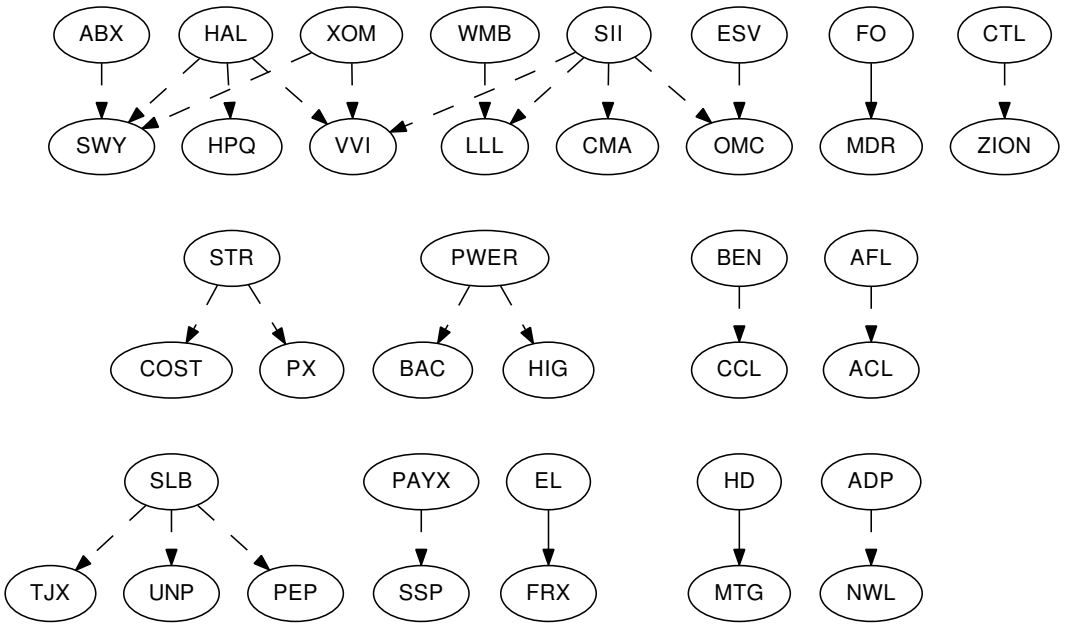


Figure 7.6.: Relationships found in one year of actual market data.

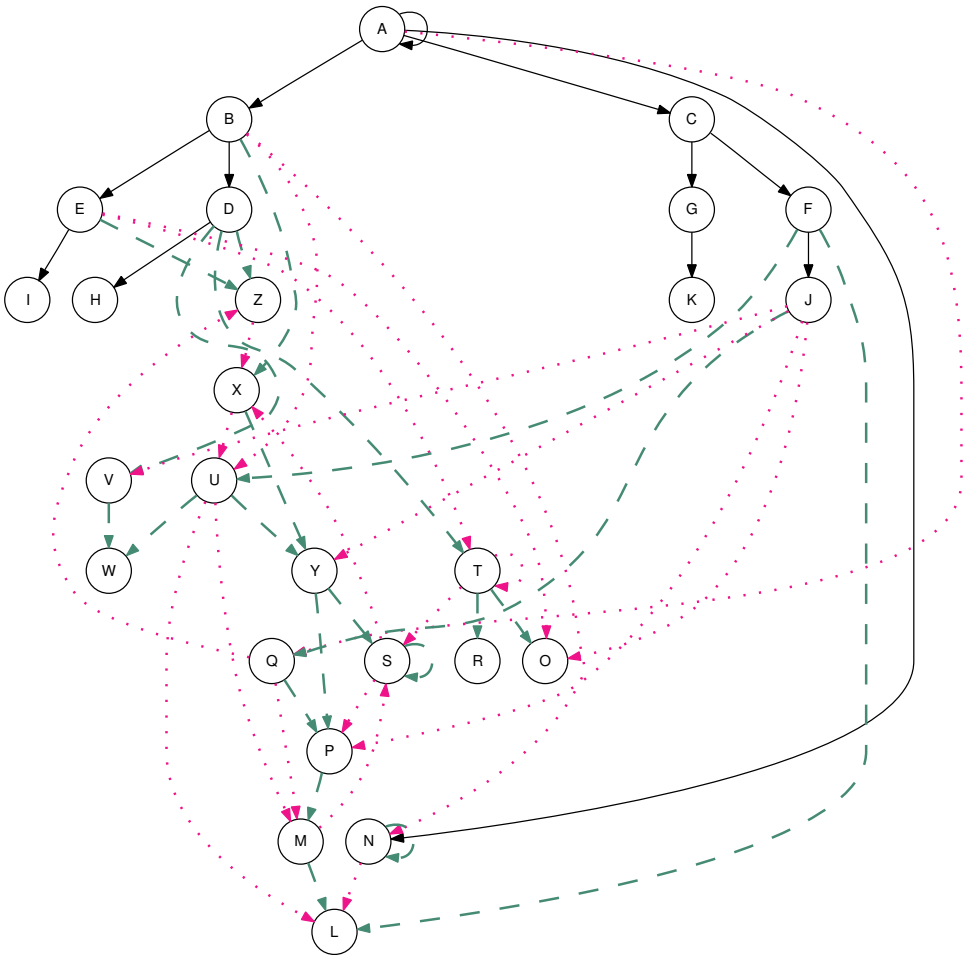


Figure 7.7.: Graph representing results from DBN algorithm on pattern 4 of the synthetic MEA data. Arrows denote relationships whether the neuron at the head causes the neuron at the tail to fire at some specific time within the range [20,40]. Colored (and dashed/dotted) arrows refer to relationships that were found in one of the two runs for this parameter setting, with black arrows denoting relationships found in both runs.

Figure 7.8.: Neuronal pattern 1.

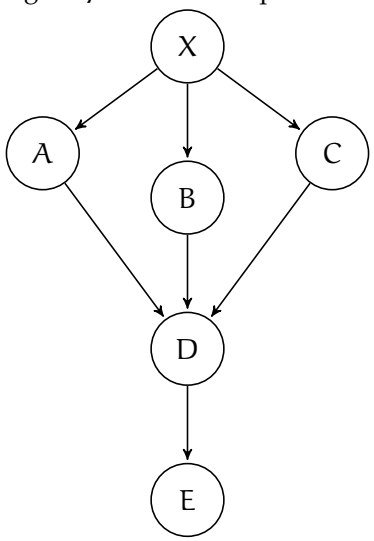


Figure 7.9.: Neuronal pattern 2.

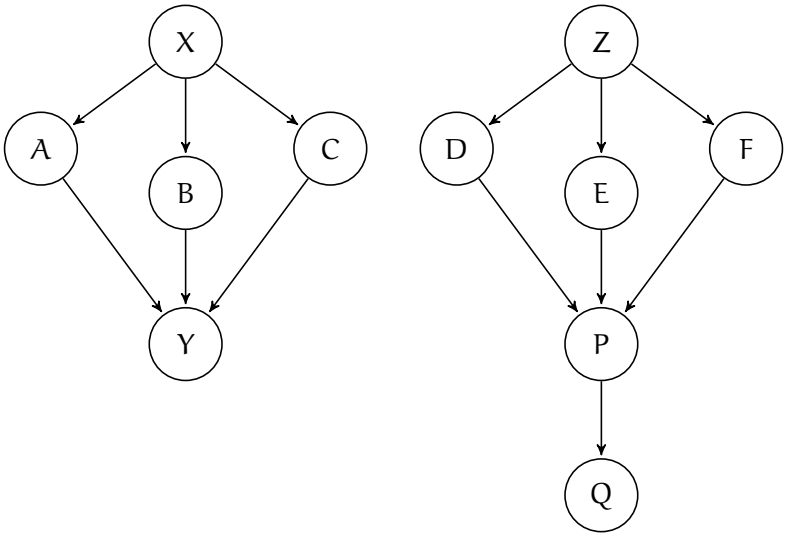


Figure 7.10.: Neuronal pattern 3.

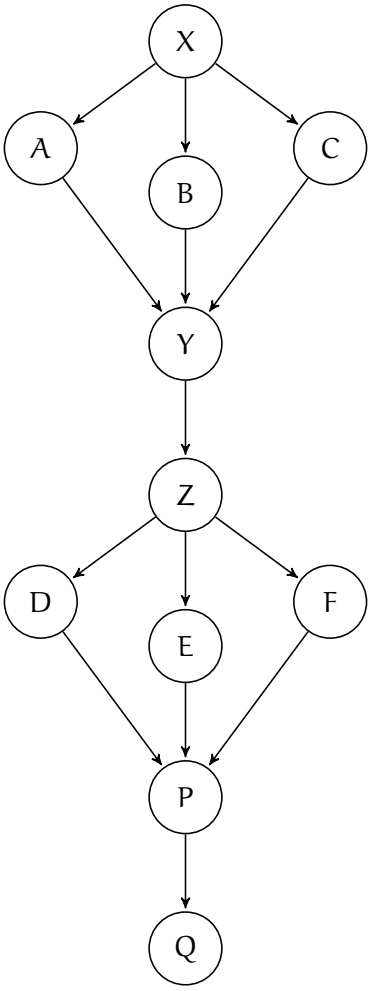


Figure 7.11.: Neuronal pattern 4.

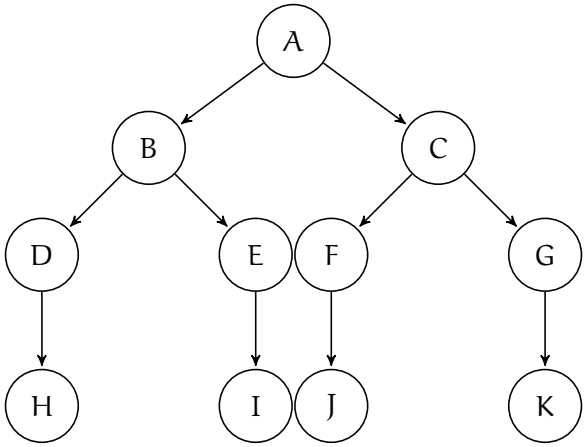
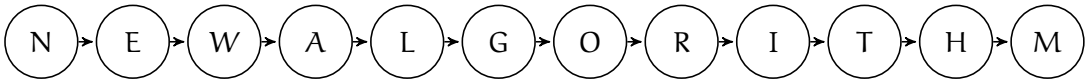


Figure 7.12.: Neuronal pattern 5.



8

CONCLUSIONS AND FUTURE WORK

8.1 CONCLUSIONS

Understanding the complex causal relationships governing why things happen is at the heart of many disciplines, including biology, finance and the social sciences. It has been difficult, though, to determine these relationships from observational data alone and the problem of formalizing the conditions for causality in terms of algorithms has only recently been addressed. However, there is a rich literature in philosophy on what constitutes a causal relationship and how these can be identified. In this dissertation, I have built on these philosophical foundations, using the fact that the conditions for identifying causality can be translated into the framework of temporal logic and model checking, and developed a powerful new approach to causal inference.

I have shown how the problem of causal inference is, in many cases, one of understanding the relationships of sequences of events over time. By translating these notions to PCTL, we allow description of vital features that have previously been left out of computational approaches to causal inference, namely the temporal component of the causal relationship as well as explicit description of the sets of conditions comprising a cause. This allows for a notion of causality that goes beyond simply "a causes b", while making sure that a and b are described in a well defined way

that allows for automated testing of the formulas in observational data. Since the rules for building such statements are given by the logic, I have described straightforward methods of computing the probability of any such statement from the data. Unlike competing methods using graphical models, this approach also allows for cycles and feedback loops. I augmented PCTL so that it is possible to reason about the truth value of formulas involving windows of time, and developed algorithms for determining the truth value of formulas in a trace (or set of traces) of data.

Many *prima facie* (potential) causes – those that precede and raise the probability of their effects – may seem to arise by chance, so I introduced a new method for computing the significance of each cause. Inspired by philosophical methods, but focused on computational feasibility as well as practical applications, the approach is to compute the average impact a cause has on its effect given (pairwise) each of the other possible causes of that effect. Once we have computed this impact, or significance score, we must determine at what level to call something causally significant. Treating the problem of weeding out insignificant causes as a multiple hypothesis testing problem using an empirical null allows us to remain neutral as to the underlying distribution of the data, and still control our false discovery rate.

This approach has been tested on synthetic data, where we can evaluate our findings against some ground truth, as well as real data where we aim to discover novel relationships. One set of generated data was created by another research group [107] to mimic the structure of neurons firing over time. In collaboration with researchers in quantitative finance, we ourselves constructed a second synthetic data set with a structure similar

to that of stock price movements. In both cases, by using the right tools for the job, namely those that can take into account the important temporal information in these examples with strong temporal dynamics, we have outperformed leading methods from both computer science and finance. We have achieved extremely low (nearly zero) false discovery rates, while succeeding in making a large number of valid discoveries.

This general approach will find applications in a wide variety of areas, from politics (where a candidate's favorability ratings can be influenced by their choice of words in speeches as well as actions such as votes), to finance (where the price movement of a stock is a result of both hidden market factors as well as the movements of other individual stocks and sectors) to computational biology (where we want to find the genes responsible for particular traits or find regulatory networks among genes). One of the most important emerging applications is in the analysis of electronic health records (EHRs), which contain valuable information on patients over long periods of time. We can use these to determine at a population level what causes a condition such as congestive heart failure and which tests should be done to predict it earlier and more accurately. This can also be used at the level of patients to determine what affects a particular person's glucose levels over time and whether her medication is effective in controlling these.

I have also shown how the type-level inferences can be used for token-level reasoning. The approach discussed allows us to take an effect whose occurrence we want to explain, and a sequence of time-indexed observations, and use these to determine the likelihood that various type-level causes are responsible the effect in the token case. Unlike other approaches in the literature, this does not require complete knowledge of

the scenario (i.e. the truth value for all propositions in the system), and outputs a score for each possible cause (rather than simply caused/did not cause). It was demonstrated that this method can handle many of the examples and counterexamples found in the philosophical literature. Specifically, the inclusion of temporal information and allowance for multiple causes of varying degrees help to provide intuitively correct answers in difficult cases. The method developed allows us to extend our general inferences to specific cases. For example, after finding causes for various medical conditions, we can then take a patient's incomplete medical history and assess her possible diagnoses. Further, we can also use this method for predictive purposes (prognosis), where the effect has not yet occurred and we want to determine its likelihood.

The research contributions of this dissertation may be summarized as follows:

- Development of philosophically sound working definitions for causality in temporal systems, which allow for cycles & feedback and explicit description of the temporal component of causal relationships.

- Efficient algorithms for type-level inference of prima facie causes from time-series data, where causal relationships are described as temporal logic formulas.

- New measure for the significance of causal relationships, which builds on work in philosophy while remaining practically applicable and yielding superior practical performance when compared with other methods.

- Method for translating type-level relationships into token-level inferences, which allows for incomplete information. The method correctly handles many of the difficult cases found in the philosophical literature.

- New approach to the epistemology of causality, an overlooked area of research in philosophy, which addresses both type and token causality.

- Adaptation of PCTL to model-checking over traces, and augmentation of path formulas to include a lower bound on timings (allowing representation of windows of time).

- Rigorous comparative analysis of the performance of many definitions of causality in several important practical domains: finance and neuroscience.

8.2 FUTURE WORK

The methodologies described in this work are only the beginning of what is needed. Here I outline a few of the most promising and pressing directions for future work. While not discussed here, there is also a need for heuristics and methods to improve computational performance. However, the algorithms are nearly all easily parallelizable.

8.2.1 *Simpson's paradox and abstraction*

A primary problem when dealing with probabilistic causality is knowing whether the relationships we observe are genuine or if they are due to a chance distribution of the variables. Simpson's paradox is when we find a correlation between variables in a set of subpopulations, but this correlation is reversed when the populations are combined. In terms of causality, we might find that a medication does not improve prognosis in a combined population of men and women, but that it is beneficial when we consider populations of just men or just women. However, since we have a notion of token causality, it is possible to use instances where we do know (or know with high probability) the true cause to refine our theory. If on many occasions the token and type-level causal relationships are in conflict, we may be observing this paradox. This directly relates to another vital question: at what level should we look at a system? Depending on our purposes (e.g. vaccine development versus public policy), the desired level of detail will differ. Biological systems may be viewed in terms of populations, individual people, organs, or cells. Abstraction, an important topic in computer science, may be able to help us determine which variables should be measured and examined more precisely.

8.2.2 *Causality in time and space*

Recent experiments have allowed measurement of neurons such that we know not only their firing times, but also their locations. Similarly,

methods for measuring gene expression have also moved towards experiments in both time and space. It is desirable then to be able to reason about spatial locality: 1) relationships in one area may not be the same as those elsewhere; 2) there may be a certain proximity required for a causal relationship to be possible. Additionally, previous philosophical theories that focused on causal processes (where we think of a causal relationship as transmission of a conserved quantity, such as momentum) stipulate both temporal and spatial locality as essential features of causality. It may be possible to bridge the gap between these and probabilistic methods by incorporating spatial information.

At the same time, we have assumed that our distributions are stationary but this will not be true in all cases. For example, in applications to finance, the relationships between companies (due to mergers and acquisitions, for one) change, as do the legal rules governing their behavior. It will be important to determine the times when the causal regime changes not just to determine when our rules stop being true, but also for inference. Further, in applications involving medical records, patients may have different stages of a disease, and the causal relationships at each stage may vary. If we assume the relationships in a long time series are stationary, but there are instead distinct periods with differing structures, we will likely fail to find the true relationships.

8.2.3 *Variable representation*

In this work it was assumed that propositions are true or not at particular, discrete, time instants. In order to define propositions and their truth

values, continuous variables had to be discretized such that they related to propositions that may be only true or false. However, variables may have durations (such that smoking for 10 years causes lung cancer), and continuous magnitudes (one gene causes another to have a certain level of regulation or smoking 10 packs a day versus 1 pack, in contrast to simply smoking versus not smoking) which may be important for practical applications. Finally, we may be more interested in an effect that persists for some time than one that is instantaneous, but as yet we also have no method for representing how long the effect lasts.

8.2.4 *Token causality*

There are a number of important future directions in the area of token causality. The problems found here are not simply theoretical, but are found in practical scenarios where we want to determine who is to blame for an accident, why a patient developed a set of symptoms, and whether a change in policy affected a company's stock prices. Thus far we have greatly simplified our task by assuming that for each possible fact, we either correctly know it to be true or false, or we do not know its truth value. However, in applications such as those in politics or finance, individuals may have varying states of knowledge about the world and some of their information may be both conflicting and incorrect. We must understand how to assess token causality in these cases. This area has similarities to argumentation and legal reasoning (where we perhaps need to understand the support for token-level hypotheses at the level of individuals), and applications in disease diagnosis, where a patient's

medical record may have incorrectly recorded tests and conflicting diagnoses and we want to find both the best explanation for their symptoms and the best course of treatment.

In the case of diagnosis of patients, symptoms may seem to rule out the true cause of the disease as a result of errors in measurement and recording of laboratory tests. However, since patients can be queried using medical tests, we could potentially suggest which additional information would most aid the diagnosis. Thus we would also like to know what information would be most useful for determining the token-cause. Further, if we can determine the actual cause in a token case, we could potentially use this information to reassess inferred type-level causes (pointing to novel relationships or the need for better inferences).

8.3 BIBLIOGRAPHIC NOTE

The work presented here has been published in various forms. An introduction to the type-level approach, as well as comparison on neuronal data with competing algorithms (see Chapters 4 and 7) was published as a conference paper:

> Samantha Kleinberg and Bud Mishra. The Temporal Logic of Causal Structures. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI-09)*, pages 303–312, Corvallis, Oregon, 2009. AUAI Press.

A popular account of the type-level approach, with extensive discussion of the future applications to healthcare and systems biology as well

as experiments involving time-series microarray data, appears in the following invited paper:

> Samantha Kleinberg and Bud Mishra. Metamorphosis: the Coming Transformation of Translational Systems Biology. *Queue*, 7(9):40–52, 2009.

The use of multiple hypothesis testing and false discovery control methods for assessing the statistical significance of causal relationships (see Chapter 5) as well as experimental results on neuronal data, microarray data, and political data were presented at the following conference and appear in the following refereed volume:

> Samantha Kleinberg and Bud Mishra. Multiple Testing of Causal Hypotheses. Canterbury, UK, September 2008. CAPITS Causality Study Fortnight.

> Samantha Kleinberg and Bud Mishra. Multiple Testing of Causal Hypotheses. In Phyllis McKay Illari, Federica Russo, and Jon Williamson, editors, *Causality in the Sciences*. Oxford University Press, 2010. (To appear).

Parts of the methodological aspects of token causality (see Chapter 6) are published as a conference paper:

> Samantha Kleinberg and Bud Mishra. The Temporal Logic of Token Causes. In *Proceedings of the 12th International Conference on the Principles of Knowledge Representation and Reasoning (KR2010)*, Toronto, Canada, May 2010. (To appear).

The methods developed in this work have also led to two pending patents:

Filed 5/21/2009 Method, System, And Computer-Accessible Medium For Inferring And/Or Determining Causation In Time Course Data With Temporal Logic (PCT Application No. PCT/US09/44862)

Filed 2/21/2010 Methods, Computer-Accessible Medium And Systems For Facilitating Data Analysis And Reasoning About Token/Singular Causality (U.S. Provisional Patent Application)

A

A BRIEF REVIEW OF TEMPORAL LOGIC & MODEL CHECKING

A temporal logic is any logic that includes modal operators allowing reasoning about *when* formulas are true. Originally introduced by philosophers, the first work in this area was in the form of tense logic, introduced by Arthur Prior in the 1960s [106]. In the 1970's Amir Pnueli built upon these ideas and provided the first introduction of temporal logic for concurrent systems in computer science [103].

An important problem in computer science is verifying the correctness of systems. However, deductive verification, proving the correctness of a system using a set of axioms and rules, is a time consuming process. It can be used in the case of systems with infinite states, though in that case it may use an unlimited amount of time and memory. Model checking imposes restrictions so that we can automate much of the process. It allows verification of finite state concurrent systems, where the system will always terminate with an answer.

A.1 TYPES OF TEMPORAL LOGIC

There are three main temporal logics: CTL* (which includes computation tree logic (CTL) [18] and linear temporal logic (LTL) [103] as subsets), the μ-calculus [33] and interval temporal logic (ITL) [91]. We will review

CTL and then its probabilistic variant, PCTL. In CTL, the future is not deterministic – there may be any number of possible future paths through time from the current state, whereas in LTL from a given state there is only one possible path through time. Probabilistic computation tree logic (PCTL) extends CTL to allow reasoning about nondeterministic systems where properties of interest may be of the form "a property will hold within some (discrete) amount of time, 99% of the time."

A.1.1 *CTL*

Temporal logics are generally interpreted over graphs called Kripke [69] structures. A Kripke structure is defined by a set of reachable states (nodes of the graph), labels that describe the properties true within each state, and a set of edges denoting the transitions of the system [19].

Definition A.1.1. Let AP be a set of atomic propositions. A Kripke structure M over AP is a four tuple $M = (S, S_0, R, L)$ where:

- S is a finite set of states,
- $S_0 \subseteq S$ is the set of initial states,
- $R \subseteq S \times S$ is a total transition relation, such that $\forall s \in S$, $\exists s' \in S$ s.t. $(s', s) \in R$, and
- $L : S \to 2^{AP}$ is a function that labels each state with the set of atomic propositions that are true within it.

A path in the Kripke structure is an infinite sequence of states $\pi = s_0, s_1 ...$ such that for every $i \geqslant 0$, $(s_i, s_{i+1}) \in R$. π^i is used to denote the suffix of path π starting at state s_i.

Formulas in CTL are composed of paired path quantifiers and temporal operators. Path quantifiers describe whether a property holds **A** ("for all paths") or **E** ("for some path") starting at a given state. The temporal operators describe where along the path properties will hold. This means that while AGf is a valid CTL formula, $AGFf$ is not, since F must be paired with one of A or E. The operators are:

- **F**: "finally", at some state on the path the property will hold;
- **G**: "globally", the property will hold along the entire path;
- **X**: "next", the property will hold at the next state of the path;
- **R**: "release" (also called weak until), for two properties, the first holds at every state along the path until a state where the second property holds, with no guarantee that the second property will ever hold (in which case the first must remain true forever);
- **U**: "until", for two properties, the first holds at every state along the path until at some state the second property holds.

The syntax of CTL is defined as follows. First, there are two types of formulas: path formulas, which are true along a specific path, and state formulas, which are true in a specific state. Then, where AP is the set of atomic propositions, the syntax of state formulas is given by:

- If $p \in AP$, then p is a state formula;
- If f and g are state formulas, then so are $\neg f$, $f \vee g$ and $f \wedge g$;
- If f is path formula, then $\mathbf{E}f$ and $\mathbf{A}f$ are state formulas.

Path formulas are specified by:

- If f and g are state formulas, then **F**f, **G**f, **X**f, f**R**g, f**U**g, are path formulas.

The basic CTL operators are: **AX**, **EX**, **AF**, **EF**, **AG**, **EG**, **AU**, **EU**, **AR** and **ER**. However each can be expressed in terms of **EX**, **EG** and **EU**:

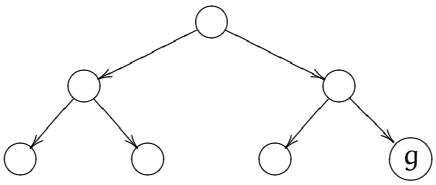
- $\mathbf{AX}f \equiv \neg\mathbf{EX}(\neg f)$
- $\mathbf{EF}f \equiv \mathbf{E}[\mathrm{True}\mathbf{U}f]$
- $\mathbf{AG}f \equiv \neg\mathbf{EF}(\neg f)$
- $\mathbf{AF}f \equiv \neg\mathbf{EG}(\neg f)$
- $\mathbf{A}[f\mathbf{U}g] \equiv \neg\mathbf{E}[\neg g\mathbf{U}(\neg f \wedge \neg g)] \wedge \neg\mathbf{EG}\neg g$
- $\mathbf{A}[f\mathbf{R}g] \equiv \neg\mathbf{E}[\neg f\mathbf{U}\neg g]$
- $\mathbf{E}[f\mathbf{R}g] \equiv \neg\mathbf{A}[\neg f\mathbf{U}\neg g]$

Figure A.1 illustrates the most common operators in terms of computation trees. Note that each tree continues infinitely beyond the states shown.

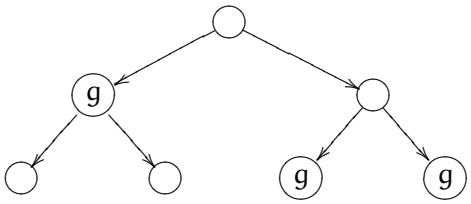
Then, the truth values of path and state formulas are represented as follows. For a state formula f, $M, s \vDash f$ means that formula f holds at state s in the Kripke structure M. For a path formula g, $M, \pi \vDash g$ means that g holds along path π in Kripke structure M.

A.1.2 *PCTL*

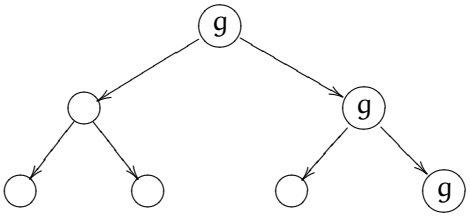
While CTL allows us to ask which properties of a non-deterministic system are possible, there are many cases in which we will want to know just how likely these properties are. In these cases, we want to be able



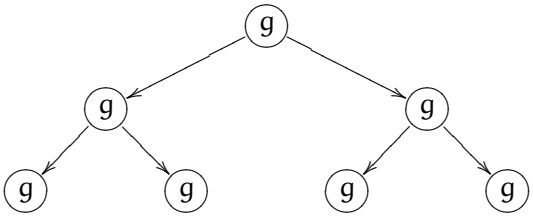
(a) **EF**g



(b) **AF**g



(c) **EG**g



(d) **AG**g

Figure A.1.: Illustrations of CTL formulas.

to represent the probability that a formula will hold. However, we also would like to specify deadlines for these properties. That is, an upper bound on how long they may take to hold with a certain probability. Hansson and Jonsson introduced probabilistic computation tree logic (PCTL) [49] to address these types of problems.

The formulas in PCTL are interpreted over probabilistic Kripke structures (also called discrete time Markov chains), where a structure is a four tuple: $K = \langle S, s^i, L, \mathcal{T} \rangle$, such that:

- S is a finite set of states;
- $s^i \in S$ is an initial state;
- L is a labeling function assigning atomic propositions (AP) to states,

$$L : S \to 2^{AP};$$

- $\mathcal{T}$ is a transition probability function, $\mathcal{T} : S \times S \to [0, 1]$ such that for all s in S:

$$\sum_{s' \in S} \mathcal{T}(s, s') = 1.$$

The formulas are comprised of atomic propositions a in the universe AP, propositional logical connectives (such as $\neg, \wedge, \vee$) and modal operators denoting time and probability. As in CTL, there are two types of formulas: path formulas and state formulas, which are defined inductively as:

- Each atomic proposition is a state formula;

- If f and g are state formulas, so are $\neg f$, $f \wedge g$, $f \vee g$, $f \rightarrow g$;
- If f and g are state formulas, and t is a nonnegative integer or ∞, then $f\mathcal{U}^{\leqslant t}g$ and $fU^{\leqslant t}g$ are path formulas;
- If f is a path formula and p is a real number with $0 \leqslant p \leqslant 1$, then $[f]_{\geqslant p}$ and $[f]_{>p}$ are state formulas.

Now we will define the truth values of formulas for specific structures in terms of their satisfaction relations. The satisfaction relation, $s \models_K f$, means that state formula f is true in state s in structure K. Then, $s \models_K a$ (state s satisfies atomic proposition a) iff $a \in L(s)$. Relations for $\neg, \wedge, \vee$ and $\rightarrow$ are then defined as usual. The path satisfaction relation, $\sigma \models_K f$ means that the path σ satisfies the path formula f in model K. Then we have the following path relations:

- $\sigma \models_K fU^{\leqslant t}g$ iff $\exists i \leqslant t$ such that $\sigma[i] \models_K g$ and $\forall j : 0 \leqslant j < i : (\sigma[j] \models_K f)$ (strong until);
- $\sigma \models_K f\mathcal{U}^{\leqslant t}g$ iff $\sigma \models_K fU^{\leqslant t}g$ or $\forall j : 0 \leqslant j \leqslant t : (\sigma[j] \models_K f)$ (weak until);
- $s \models_K [f]_{\geqslant p}$ if the μ_m-measure of the set of paths σ starting in s for which $\sigma \models_k f$ is at least p;
- $s \models_K [f]_{>p}$ if the μ_m-measure of the set of paths σ starting in s for which $\sigma \models_k f$ is greater than p.

where the μ_m-measure is the sum of probabilities over the set of paths from s that satisfy f.

One may also define analogues to the usual path quantifiers A ("for all paths") and E ("for some future path") and temporal operators F

("eventually holds"), G ("holds for entire future path"), and X ("at the next state"). These operators are defined by:

- $Af \equiv [f]_{\geqslant 1}$
- $Ef \equiv [f]_{>0}$
- $G^{\leqslant t}_{\geqslant p} \equiv f \mathcal{U}^{\leqslant t}_{\geqslant p} false$
- $F^{\leqslant t}_{\geqslant p} f \equiv true \, U^{\leqslant t}_{\geqslant p} f$
- $AGf \equiv true \, U^{\leqslant \infty}_{\geqslant 1} false$
- $AFf \equiv f U^{\leqslant \infty}_{\geqslant 1} f$
- $EGf \equiv f \mathcal{U}^{\leqslant \infty}_{>0} false$
- $EFf \equiv true \, U^{\leqslant \infty}_{>0} f$

A.2 MODEL CHECKING

In model checking, the problem is to determine which states in the system satisfy some temporal logic formula. If the initial states of the system are in that set of states, then the model satisfies the formula.

A.2.1 *CTL Model Checking*

We begin with a Kripke structure $M = (S, R, L)$ and a CTL formula f. The basic principle is that states are labeled with subformulas that are true within them, and in each iteration more complex formulas are analyzed. During the procedure there are six main cases: $f, \neg f, f \vee$

$g, \mathbf{EX}f, \mathbf{E}[f\mathbf{U}g], \mathbf{EG}f$ (as was noted before, all other formulas can be expressed in terms of those).

The rules for labeling states are as follows. a state is labeled with $\neg f$ if it is not labeled with f, a state is labeled with $f \vee g$ if it is labeled with either f or g and a state is labeled with $\mathbf{EX}f$ if the state has a transition to a state labeled with f. The final two cases are somewhat more complex. For a formula $h = \mathbf{E}[f\mathbf{U}g]$, we first find states labeled with g, and then for each such state where there is a path from those states where each state on the path is labeled with f, it is labeled with h. As the formulas are built incrementally, beginning with those of size one, the states satisfying f and g have already been labeled at this point.

Finally, for $g = \mathbf{EG}f$[1]:

Lemma A.2.0.1. $M, s \vDash \mathbf{EG}f$ *iff:*

1. $s \in S'$.
2. *There exists a path in* M' *leading from some* s *to some* t *in a nontrivial SCC*[2] *of the graph* (S', R').

where M' *is formed from* M *by removing all the states where* f *does not hold, and updating* R *and* L *accordingly.*[3]

Labeling states with some formula where all of its subformulas have already been processed takes time $O(|S| + |R|)$. So, for a formula of size $|f|$, the complexity is $O(|f|(|S| + |R|))$. This is because each iteration takes $O(|S| + |R|)$ and the algorithm begins with the innermost formula, working

1 [19], 36.
2 A strongly connected component (SCC) is a set of vertices in the graph where for each pair u and v in the component, there is a path from u to v and one from v to u. An SCC is nontrivial if it consists of more than one node, or it has one node with a self loop.
3 $M' = (S', R', L')$ where $S' = \{s \in S | M, s \vDash f\}, R' = R|_{S' \times S'} L' = L|_{S'}$.

outward, so when we get to formula f, every component of f has already been processed, and there are at most $|f|$ subformulas of f.

A.2.2 *PCTL Model Checking*

Model checking in the probabilistic case proceeds similarly to the case of CTL, labeling states with subformulas true within them, beginning with all states labeled with the propositions true within them. The labeling rules for states are:

- A state is labeled with $\neg f$ if it is not labeled with f.
- A state is labeled with $f \wedge g$ if it is labeled with both f and g.
- A state is labeled with $f \vee g$ if either f or g are in its labels.
- A state is labeled with $f \to g$ if it is labeled with $\neg f$ or with g.

As before, our other cases reduce to a small set: $f\mathcal{U}^{\leqslant t}g$, combined with $[f]_{\geqslant p}$, $[f]_{>p}$. In the case where the probability and time do not take extreme values (0 or 1 and ∞ respectively), $f\mathcal{U}^{\leqslant t}_{\geqslant p}g$ is checked as follows, with the $> p$ case being the same except that states will be labeled with the formula if the calculated probability is strictly greater than p.

Formulas using $\mathcal{U}$ can be defined in terms of $\mathcal{U}$ formulas.

As shown by Hansson and Jonsson [49]:

Proposition A.2.1. *Assume the states satisfying f and g are labeled as such. Then, for $t \neq \infty$, the μ_m measure for the set of paths σ from s for which $\sigma \vDash_K f\mathcal{U}^{\leqslant t}g$ is given by $P(t, s)$, where this is defined to be 0 if $t < 0$ and is otherwise given by:*

$$\mathcal{P}(t,s) = \begin{cases} 1 & \text{if } g \in \mathit{labels}(s); \\ 0 & \text{if } f \notin \mathit{labels}(s); \\ \sum_{s' \in S} \mathcal{T}(s,s') \times P(t-1,s') & \text{otherwise.} \end{cases} \tag{A.1}$$

The labeling algorithm follows from this recursion, labeling state s with $f U^{\leqslant t}_{\geqslant p} g$ if $P(t,s) \geqslant p$. The complexity of the resulting algorithm (remembering that all subformulas have been checked) is $O(t(|S|+|E|))$. This is the same as the earlier complexity for a CTL formula where all subformulas have been checked, with an added factor of t. Checking whether a structure satisfies formula f, as before, depends on the size of f and is at most $O(t_{max}(|S|+|E|)|f|)$, where t_{max} is the maximum time parameter in f.

Faster algorithms, linear in $|S|$, exist for the cases where p takes the values 0 or 1 regardless of the value of t. In the case where the probability is not 0 or 1, and $t = \infty$, a different approach is required, as the one given above would not terminate. In this case states in S are divided into a few subsets S_s, R, Q. S_s contains states labeled with g, Q are states not labeled with f or g as well as those from which it is not possible to reach a state in S_s, and R is the set of states for which the probability of reaching a state in S_s is 1. Then, as shown by Hansson and Jonsson [49]:

Proposition A.2.2. *Assume that in structure* K*, states satisfying* f *and* g *have been labeled with these formulas. Then, with* Q *and* R *defined above, the* μ_m*-measure for the set of paths* σ *from* s *for which* $\sigma \models_K f U^{\leqslant\infty} g$ *is given by the solution to* $P(\infty, s)$*:*

$$\mathcal{P}(\infty, s) = \begin{cases} 1 & \text{if } s \in R; \\ 0 & \text{if } s \in Q; \\ \sum_{s' \in S} \mathcal{T}(s, s') \times P(\infty, s') & \text{otherwise.} \end{cases} \tag{A.2}$$

Solving with Gaussian elimination gives a complexity of $O((|S| - |Q| - |R|)^{2.81})$. Algorithms for finding Q and R are described in Appendix C, where we reformulate until formulas to allow for a lower bound on the associated timing.

A.2.3 *Symbolic Model Checking*

The previous algorithms described used explicit representations of the Kripke structure. These explicit methods were linear in both the size of the formula and the size of the system, however this still problematic, in that the size of the transition system grows exponentially with linear growth in the number of variables (this is referred to as state space explosion). Symbolic model checking techniques allow for compact representation of large amounts of states. Binary decision diagrams, used to represent Boolean functions, can be used to encode Kripke structures to allow this symbolic model checking.

Ordered binary decision diagrams (OBDDs) are concise representations of boolean formulas that are obtained from taking binary decision trees by merging subtrees with identical structures to create a DAG. Determining the variable ordering is a difficult problem, but once an ordering is

established, the diagrams are canonical. Reducing the diagrams in this manner means that to test whether two formulas are equivalent, we can test whether their OBDDs are isomorphic. Each state of the Kripke structure is represented by a boolean formula, and then the transition from one state to another is the conjunction of their formulas.

Symbolic algorithms manipulate boolean formulas, rather than the graph-based representation of the Kripke structure. The idea is that states satisfying a CTL formula are characterized in terms of least/greatest fixpoints of sets, where a set $S' \subseteq S$ is a fixpoint of a function $\tau P(S) \to P(S)$ if $\tau(S') = S'$. The least element is the empty set (or false) and the greatest element is S (true), where these least and greatest fixpoints will correspond to properties that are eventually and always true respectively.

In the case of CTL formulas, we can represent the formulas $\mathbf{EG}f$ as $f \wedge \mathbf{EXEG}f$ and $\mathbf{EF}p$ as $p \vee \mathbf{EXEF}p$. Then, we can write $\mathbf{EF}p$ recursively as $U = p \vee \mathbf{EX}U$, which implies that $\mathbf{EF}p \subseteq U$. Then, $\mathbf{EF}p = \mu U.p \vee \mathbf{EX}U$ (where μ denotes a least fixpoint and ν denotes a greatest fixpoint). The computation is again iterative, where we begin with $U = \emptyset$, or false, and then at each iteration we have $U_i = p \vee \mathbf{EX}U_{i-1}$. Note that $\mathbf{EG}p$ would be the greatest fixpoint of a similar function, where the $\vee$ is replaced by a $\wedge$.

B

A LITTLE BIT OF STATISTICS

B.1 MULTIPLE HYPOTHESIS TESTING

When testing a single hypothesis, we can make the decision about whether to accept or reject the null hypothesis based on the likelihood that the results would occur if the null hypothesis were true. However, when we are testing multiple hypotheses at once, the likelihood that we will get such results – even under the null hypothesis – increases and we must account for this. For example, we may test the fairness of a coin by flipping it 10 times and seeing how many times it comes up heads and how many times it comes up tails. If there were nine heads, we would likely say that it is biased, as the probability of this happening when the coin is fair is ≈ 0.010, and the p-value 0.022. Frequently, a significance level of $p < \alpha = 0.05$ is sufficient to reject the null hypothesis, and thus we would call the coin unfair. In the case where we are testing 100 fair coins, we may incorrectly deem 5 unfair ($n_{tests} \times \alpha$). Thus it is necessary to account for the fact that we are performing many tests simultaneously, thus increasing our chances of seeing unlikely or anomalous behavior [122, 29, 5].

B.1.1 *Basic Definitions*

First, we define two types of error. *Type I errors* (α) are those where we reject a true null hypothesis. The *per-comparison error rate* is the probability of making such an error during each significance test. *Type II errors* (β) are those where the null hypothesis is not rejected when it should be. Whereas Type I errors mean we have made a false discovery (false positive), Type II errors mean we have missed an opportunity for discovery (false negative). While it is desirable to reduce both types of error, it may sometimes only be possible to trade one kind off against the other. The best trade-offs are judged in terms of the relative costs of these errors in a particular domain of application.

Thus, we define next the error rates over all the hypotheses being tested. The familywise error rate (FWER) is the probability of rejecting one or more true null hypotheses (i.e. the probability of having at least one Type I error), during all tests. For the FWER to approach a desired bound of $\alpha \ll 1$ we need each of the, say, n tests to be conducted with an even stricter bound, such as $\frac{\alpha}{n}$, as required by the so-called Bonferroni correction [5]. However, the FWER has low power, meaning that we have a good chance of making a Type II error [4]. Another measure, called the false discovery rate (FDR), estimates the proportion of Type I errors among all rejected hypotheses (that is, the number of false discoveries divided by the total number of discoveries). This measure results in more power than the FWER while still bounding the error. The main idea is that, if we are rejecting only a few null hypotheses, then each false discovery we make in that case is more significant than rejecting a large

number of null hypotheses and making more false discoveries. That is, in the first case, the false discoveries are a larger percentage of the overall number of discoveries than they are in the later case.

B.1.2 *Controlling the FDR*

The introduction of the FDR and procedures for controlling it are described by Benjamini and Hochberg [4]. The procedure is as follows. When testing m hypotheses, order the p-values $P_{(1)} \leqslant P_{(2)} \cdots \leqslant P_{(m)}$. Then with k selected as the largest i such that:

$$P_{(i)} \leqslant \frac{i}{m}\alpha, \tag{B.1}$$

we reject all $H_{(i)}, i = 1, 2, \ldots, k$. In the case when all hypotheses are true this controls the FWER, and otherwise controls the proportion of erroneous rejections. For independent test statistics, this procedure controls the FDR at rate α. However, it was later shown that this also holds for positively dependent test statistics and can be modified to control the FDR in other cases [5].

B.1.3 *Using an empirical null hypothesis*

In the methods described so far, it was necessary to use a theoretical null hypothesis, namely, that values have a standard normal distribution. However, this may not be appropriate for all data. It is possible, then, to take advantage of the multitude of hypotheses being tested and to determine the correct null hypotheses from the data. The use of an

empirical null hypothesis was described by Efron [29], and provides a novel empirical Bayesian solution to the problem. In that work, Efron described how one may estimate the empirical null distribution and how the choice of null hypothesis has a large impact on the discoveries made. For example, data from microarrays, financial markets, and neural spike trains may have different underlying distributions and thus their empirical nulls vary from the theoretical null in different ways [67]. In practice, most methods for inferring the null hypothesis empirically attempt to fit to the central peak of the data.

B.1.4 *Computing the fdr*

Here, we use local false discovery rate (fdr) calculations, which use densities, as our N's are large, though these methods may also be used with standard tail-area FDR methods such as that described in section B.1.2 [30]. We follow the formulation described by Efron [29].

With N hypotheses H_1, H_2, ..., H_N we have the corresponding z-values z_1, z_2, ..., z_N. These values, also called the standard score, are the number of standard deviations by which a result deviates from the mean. In the case of our causal analyses, these z-values are computed from the ε_{avg}s. We begin by assuming the N cases fall into two classes: one where the effects are either spurious or not large enough to be interesting (and thus where we accept the null causal hypotheses), and another where the effects are large enough to be interesting (and where we will accept the non-null hypotheses as true). We also assume the proportion of non-null cases is small relative to N, say, around 10%. Then, p_0 and

$p_1 = 1 - p_0$ are the prior probabilities of a case (here, a causal hypothesis) being in the "uninteresting" or "interesting" classes respectively. The densities, $f_0(z)$ and $f_1(z)$, of each class describe the distribution of these probabilities. When using a theoretical null, $f_0(z)$ is the standard $N(0, 1)$ density. Note that we need not know $f_1(z)$, though we must estimate p_0 (usually $p_0 \geqslant 0.9$). We define the mixture density:

$$f(z) = p_0 f_0(z) + p_1 f_1(z), \tag{B.2}$$

then the posterior probability of a case being uninteresting given z is

$$\Pr\{\mathrm{null}|z\} = p_0 f_0(z)/f(z), \tag{B.3}$$

and the *local false discovery rate*, is:

$$\mathrm{fdr}(z) \equiv f_0(z)/f(z). \tag{B.4}$$

Note that, in this formulation, the p_0 factor is ignored, yielding an upper bound on $\mathrm{fdr}(z)$. Assuming that p_0 is large (close to 1), this simplification does not lead to massive overestimation of $\mathrm{fdr}(z)$. One may also choose to estimate p_0 and thus include it in the FDR calculation, making $\mathrm{fdr}(z) = \Pr\{\mathrm{null}|z\}$. The procedure is then:

1. Estimate $f(z)$ from the observed z-values;
2. Define the null density $f_0(z)$ either from the data or using the theoretical null;
3. Calculate $\mathrm{fdr}(z)$ using equation (B.4);

4. Label H_i where $\texttt{fdr}(z_i)$ is less than a threshold (say, 0.10) as interesting.

C

PROOFS

C.1 PROBABILITY RAISING

Here we claim that the following conditions for probabilistic causality, are equivalent in non-deterministic cases:

$$P(E|C) > P(E) \tag{C.1}$$

$$P(E|C) > P(E|\neg C) \tag{C.2}$$

Proof. Assume and $1 > P(C) > 0$ (and thus $1 > P(\neg C) > 0$). By definition:

$$P(E) = P(E|C) \times P(C) + P(E|\neg C) \times P(\neg C) \tag{C.3}$$

$$= P(E|C) \times P(C) + P(E|\neg C) \times (1 - P(C)) \tag{C.4}$$

$$= P(E|C) + [P(E|\neg C) - P(E|C)) \times (1 - P(C)) \tag{C.5}$$

Then, if $P(E|\neg C) > P(E|C)$, it must be that $P(E|C) < P(E)$ in order to maintain the equality, and if $P(E|C) > P(E|\neg C)$, then by the same reason $P(E|C) > P(E)$. Thus, if (C.2) is satisfied, (C.1) is satisfied. Conversely, if $P(E|C) > P(E)$, then we must have $P(E|C) > P(E|\neg C)$. Thus, if (C.1) is satisfied (C.2) is satisfied and finally we conclude that (C.1) $\Leftrightarrow$ (C.2). □

C.2 LEADS TO WITH BOTH LOWER AND UPPER TIME BOUNDS

First, note that in definition 4.3.1, there is a window of time in which c leads to e. That is, in our formulation, we add a minimum time after c is true before which e is true. Here we show that it is possible to add such a lower bound. By definition:

$$f \overset{\geqslant t_1, \leqslant t_2}{\underset{\geqslant p}{\rightsquigarrow}} g \equiv AG[f \rightarrow F^{\geqslant t_1, \leqslant t_2}_{\geqslant p} g], \tag{C.6}$$

where $t_1 \leqslant t_2$. Thus, we are actually only adding a minimum time to the consequent of our conditional. If we can label states where $F^{\geqslant t_1, \leqslant t_2}_{\geqslant p} g$ is true, then we can proceed as in the algorithms of Hansson & Jonsson [49]. We first recall that this is defined as:

$$F^{\geqslant t_1, \leqslant t_2}_{\geqslant p} g \equiv \mathrm{true}\ U^{\geqslant t_1, \leqslant t_2}_{\geqslant p} g. \tag{C.7}$$

Thus we now focus on formulas of the form:

$$h U^{\geqslant t_1, \leqslant t_2}_{\geqslant p} g, \tag{C.8}$$

where for a F formula, $h = \mathrm{true}$.

Claim. *The formula* $g_1 U^{\tau_1, \tau_2}_{\geqslant p} g_2$, *where* $0 \leqslant \tau_1 \leqslant \tau_2 \leqslant \infty$ *and* $\tau_1 \neq \infty$ *can be checked in a structure* $K = \langle S, s^i, L, \mathcal{T} \rangle$, *if it can be checked when* $\tau_2 < \infty$ *(Theorem C.2.1) and when* $\tau_2 = \infty$ *(Theorem C.2.2).*

Corollary. *If a state can be correctly labeled with* $g_1 U^{\tau_1, \tau_2}_{\geqslant p} g_2$, *it can also be correctly labeled with* $f \rightsquigarrow^{\geqslant t_1, \leqslant t_2}_{\geqslant p} g$.

Since,

$$f \overset{\geqslant t_1, \leqslant t_2}{\underset{\geqslant p}{\rightsquigarrow}} g \equiv AG[f \rightarrow F^{\geqslant t_1, \leqslant t_2}_{\geqslant p} g], \tag{C.9}$$

and

$$F^{\geqslant t_1, \leqslant t_2}_{\geqslant p} g \equiv true\ U^{\geqslant t_1, \leqslant t_2}_{\geqslant p} g, \tag{C.10}$$

then let $g_1 = true, g_2 = g$ and $\tau_1 = t_1, \tau_2 = t_2$. All other components of the leads-to formula can be checked and each subformula is independent of the others. That is, if we replace $F^{\geqslant t_1, \leqslant t_2}_{\geqslant p} g$ by x in the formula above, the resulting leads-to formula can be checked. Since we show x can be checked, the entire formula can be checked.

We begin with the case where the upper bound, t_2, is non-infinite and then show how this extends to the case where it is.

Case 1: $t_2 \neq \infty$

Theorem C.2.1. For structure $K = \langle S, s^i, L, \mathcal{T} \rangle$, we begin with all states satisfying g or h labeled as such. Then for $0 \leqslant t_1 \leqslant t_2$, with $t_2 < \infty$ the μ_m-measure for the set of paths σ from s where $\sigma \models_K h U^{\geqslant t_1, \leqslant t_2} g$ is given by $P(t_1, t_2, s)$.

$$P(t_1, t_2, s) = \begin{cases} 1 & \textit{if } t_1 \leqslant 0, t_2 \geqslant 0 \textit{ and} \\ & \quad g \in labels(s); \\ 0 & \textit{else if } t_2 < 0 \textit{ or } h \notin labels(s); \\ \sum_{s' \in S} \mathcal{T}(s, s') \times P(t_1 - 1, t_2 - 1, s') & \textit{otherwise.} \end{cases} \tag{C.11}$$

Then, following this recurrence, states s will be labeled with $h U^{\geqslant t_1, \leqslant t_2}_{\geqslant p} g$ if $P(t_1, t_2, s) \geqslant p$. Now, we prove that the recurrence correctly yields the μ_m-measure.

Proof. For the set of states s and integer times t_1 and t_2 take $\Pi(t_1, t_2, s)$ to be the set of finite sequences of states $s \to \cdots \to s_i \to \cdots \to s_j$, beginning in s, such that there is some j for which $t_1 \leqslant j \leqslant t_2$, where $s_j \models_K g$ and for all i with $0 \leqslant i < j$, $s \models_K h$ and $s \not\models_K g$.

Let $\mu_m^{t_1,t_2}(s)$ be the μ_m-measure of the set of paths $\sigma \in \Pi(t_1, t_2, s)$ from s where $\sigma \models_K h U^{\geqslant t_1, \leqslant t_2} g$. Then, by definition, $\mu_m^{t_1,t_2}(s)$ is:

$$\mu_m^{t_1,t_2}(s) = \sum_{s \to s_1 \cdots \to s_j \in \Pi(t_1,t_2,s)} \mathcal{T}(s, s_1) \times \cdots \times \mathcal{T}(s_{j-1}, sj). \tag{C.12}$$

We have the following cases to consider:

Case 1: $s \models_K g$, with $t_2 \geqslant 0$ and $t_1 \leqslant 0$

Then any path σ from s satisfies $\sigma \models_K h U^{\geqslant t_1, \leqslant t_2}$. Thus, $\mu_m^{t_1,t_2}(s) = 1$.

Case 2: $s \not\models_K h$, and $s \not\models_k g$

Then for any path σ from s, $\sigma \not\models_K h U g^{\geqslant t_1, \leqslant t_2}$. Since s does not satisfy g or h, one cannot satisfy the formula by extending the path, as h must hold until g holds. Thus, $\mu_m^{t_1,t_2}(s) = 0$.

Case 3: $s \not\models g$

Here we have two sub-cases.

(a) $t_2 = 0$

Here, $\sigma \models_K h U^{\geqslant t_1, \leqslant 0} g$ iff $s \models_K g$. Thus, $\mu_m^{t_1,0} = 0$.

(b) $t_2 > 0$

In this case there must be at least two states on the path. We

can rewrite such paths $\sigma \in \Pi(t_1, t_2, s)$ in terms of a transition from s to σ' where σ' is σ after its first state. That is,

$$\sigma \in \Pi(t_1, t_2, s) \text{ iff } \sigma' \in \Pi(t_1 - 1, t_2 - 1, \sigma[1]),$$

where:

$$\sigma = \langle \sigma[0], \sigma[1], \sigma[2] \ldots \rangle.$$

Then

$$\sigma' = \langle \sigma[1], \sigma[2] \ldots \rangle.$$

Thus,

$$\begin{aligned}
\mu_m^{t_1,t_2}(s) &= \sum_{s \to \cdots \to s_j \in \Pi(t_1,t_2,s)} \mathcal{T}(s, s_1) \times \cdots \times \mathcal{T}(s_{j-1}, s_j) \\
&= \sum_{s_1} \mathcal{T}(s, s_1) \times \sum_{s_1 \to \cdots \to s_j \in \Pi(t_1-1,t_2-1,s_1)} \mathcal{T}(s_1, s_2) \times \cdots \times \mathcal{T}(s_{j-1}, s_j) \\
&= \sum_{s_1} \mathcal{T}(s, s_1) \times \mu_m^{t_1-1,t_2-1}(s_1)
\end{aligned}$$

The equation for $\mu_m^{t_1,t_2}(s)$ satisfies exactly the recurrence of equation (C.11). We conclude that due to the uniqueness of the solution to this equation, $\mu_m^{t_1,t_2}(s) = P(t_1, t_2, s)$. □

Case 2: $t_2 = \infty$

When t_2 is infinite we cannot use the recurrence of equation (C.11), as this will lead to an infinite number of computations. Since we are looking for paths of a minimum length t_1, we also cannot immediately proceed as Hansson & Jonsson [49] do. We will instead identify three sets: P, Q and R. Q is the set of states from which there is no path to g (in any amount of time) or where neither h nor g holds. P is the set of states, including those labeled with g, from which there exists at least one path to g that is shorter than t_1. Finally, R is the set of states that always reach g (i.e. $F_{\geqslant 1} g$). Note that it is possible for a state to be in both R and P, as it may have only paths resulting in reaching a state where g holds, but perhaps at least some of these may do so in fewer than t_1 time units.

We begin by decomposing K into strongly connected components (SCCs), resulting in a directed acyclic graph (DAG). We add one condition, which is that all states in an SCC must either be labeled with h or $\neg h$. Note that when testing a leads-to formula, $h = \mathit{true}$, and since all states are labeled with this, the condition is automatically met. First, we define a non-trivial SCC as one with at least one node and one edge (that is, there is one node with a self loop or there are multiple nodes). We replace non-trivial SCCs with new states that are labeled with all of the labels of the states comprising the SCC. That is, for an SCC, C, $f \in \mathit{labels}(C)$ if there is a state $s \in C : f \in \mathit{labels}(s)$. As we are checking whether a formula, g, will *eventually* hold, it is enough to know that we can reach an SCC where it holds.

Now we can partition the states into the three sets described earlier. We begin by identifying the failure states Q and inconclusive states P. Akin to the algorithm of Hansson and Jonsson, we form Q and P as follows.

Algorithm C.1 form-Q

$S_s = \{s : s \in S \text{ and } g \in labels(s)\}$
$S_i = \{s : s \in S \text{ and } h \in labels(s), g \notin labels(s)\}$
$unseen = S_i \bigcup S_s$
$fringe = S_s$
$mark = \emptyset$
$P = \emptyset$
for $i = 0$ to $|S_i|$ **do**
  if $i < t$ **then**
    $mark = mark \cup \{s : (s \in fringe \text{ and } s \text{ is an SCC}) \vee$
    $(s \in fringe \wedge \exists s' \in mark : (\mathcal{T}(s, s') > 0))\}$
    $P = P \cup fringe$
  else
    $mark = mark \cup fringe$
  end if
  $unseen = unseen - fringe$
  $fringe = \{s : (s \in unseen \wedge \exists s' \in fringe : \mathcal{T}(s, s') > 0)\}$
end for
$Q = S - (mark \cup P)$

When there are SCCs in the sets above, this means that all states in the SCC are removed when an SCC is removed from a set. Similarly, if Q contains any SCCs, we consider it to contain the set of states comprising the SCC. Then Q is equivalent to the set Q identified by the algorithm of Hansson & Jonsson [49].

Now that we have the set of states from which no success is possible, we now find those (R) for which the probability of reaching a state where g holds (i.e. a success) is 1. Here we do not concern ourselves with the amount of time as we already have P and thus know which states will not always reach g in at least t_1 time units. Now we find whether it is also possible to transition to states from which we will never reach g or

whether these P states guarantee reaching g. As we are not checking the length of the paths, we do not need to worry about termination and can proceed as Hansson and Jonsson do, without decomposing the graph into SCCs.

Algorithm C.2 form-R

$form - Q$
$S_s = \{s : s \in S \text{ and } g \in labels(s)\}$
$S_i = \{s : s \in S \text{ and } h \in labels(s), g \notin labels(s)\}$
$S_f = \{s : s \in S \text{ and } h \notin labels(s), g \notin labels(s)\}$
unseen $= S_i$
fringe $= Q \bigcup S_f$
mark $= \emptyset$
for $i = 0$ to $|S - S_s|$ **do**
  mark $=$ mark $\cup$ fringe
  unseen $=$ unseen $-$ fringe
  fringe $= \{s : (s \in$ unseen $\wedge \exists s' \in$ fringe $: \mathcal{T}(s, s') > 0)\}$
end for
$R = S -$ mark

Theorem C.2.2. For structure $K = \langle S, s^i, L, \mathcal{T} \rangle$ states satisfying g or h have been labeled as such. For $0 \leqslant t < \infty$, the μ_m-measure of the set of paths σ from s where $\sigma \models_K h U^{\geqslant t, \leqslant \infty} g$ is given by $P(t, \infty, s)$.

$$
\begin{aligned}
P(t, \infty, s) = \text{if } s \in R \text{ and } s \notin P \text{ then } 1 \\
\text{else if } s \in R \text{ and } s \in P \text{ and } t \leqslant 0 \text{ then } 1 \\
\text{else if } s \in Q \text{ then } 0 \\
\text{else if } t > 0 \text{ then} \\
\quad \sum_{s' \in S} \mathcal{T}(s, s') \times P(t-1, \infty, s') \\
\text{else} \\
\quad \sum_{s' \in S} \mathcal{T}(s, s') \times P(\infty, s')
\end{aligned}
\tag{C.13}
$$

$$
\begin{aligned}
P(\infty, s) = \text{if } s \in R \text{ then } 1 \\
\text{else if } s \in Q \text{ then } 0 \\
\text{else } \sum_{s' \in S} \mathcal{T}(s, s') \times P(\infty, s')
\end{aligned}
\tag{C.14}
$$

Proof. We have three cases to consider.

Case 1: $s \in Q$

By the definition of Q (it is not possible to reach a state where g holds), $\mu_m^{t,\infty}(s) = 0$.

Case 2: $s \in R$

(a) if $s \notin P$

By the definition of R and by s only being in R, this means that not only will a state where g holds be reached with probability 1, but that there are no paths from s where this will happen in less than t time units. Thus, $\mu_m^{t,\infty}(s) = 1$.

(b) if $s \in P$ and $t \leqslant 0$

Now that $t \leqslant 0$, we are only concerned with whether we will reach an g state - in any amount of time - i.e. at any state beginning at s. Thus, since $s \in R$, g is inevitable and $\mu_m^{t,\infty}(s) = 1$.

(c) if $s \in P$ and $t > 0$.

See case 3.

Case 3: Here we have the cases where we have not yet achieved success or failure and thus we consider transitions to the next states on the paths σ from s. The recurrences are similar to that of equation (C.11), with the difference being that once $t \leqslant 0$, if we have still not reached a success/failure state, we no longer need to keep track of t and thus use exactly the recurrence of Hansson and Jonsson in equation (C.14). If $t > 0$, then we can proceed as we did for the finite case, rewriting the paths in terms of their sequences after the first state. That is, we know that paths in this category must consist of at least two states and as before, where σ' is σ after its first state:

$$\sigma \in \Pi(t, \infty, s) \text{ iff } \sigma' \in \Pi(t-1, \infty, \sigma[1]). \tag{C.15}$$

The uniqueness of the solution for $P(\infty, s)$ was shown by Hansson & Jonsson [49]. For the cases where $P(t, \infty, s)$ is used, once we know that $P(\infty, s)$ is unique, this recurrence also has a unique solution and since the μ_m-measure satisfies the same equation, we conclude that $P(t, \infty, s) = \mu_m^{t,\infty}(s)$. □

When handling the case of an infinite upper bound on the path length Hansson & Jonsson [49] assure that their algorithm for computing the probabilities recursively will terminate by first partitioning the set of states, S, into those that guarantee success (g holds), failure (neither h nor g holds, or it is not possible to reach a state where g holds) or are inconclusive (h holds, and there is a path to a state where g holds). In determining these sets, they begin with success states and expand the "frontier" being explored by one transition during each iteration, only extending successful paths by previously unseen states. We could not do this, as we sometimes want to revisit states. This is necessary as we stipulate a lower bound on the leads-to formula, so we may extend a too-short path by visiting a cycle. However, if we do not keep track of unseen states, we again have the problem of an infinite number of computations.

Instead, we recognized that if we revisit a state, it must be due to a cycle in the graph. Further, if we know that we have visited a cycle on a path between some state s and some other state labeled with g, then we know that we can find a path of at least length t_1 between these states for any t_1, where $t_1 \geqslant 0$ and $t_1 \neq \infty$.

D

ALGORITHMS

D.1 PROBABILITY OF A PATH FORMULA IN A TRACE

D.1.1 *Leads-to*

We begin with the leads-to formula $f \rightsquigarrow^{\geqslant r, \leqslant s} g$, and timepoints $t_i \in T$ satisfying f and g labeled as such.

Algorithm D.1 $leadsto - prob$

```
F = {t_i : f ∈ labels(t_i)}
S = ∅
for all t_i ∈ F do
  if check_leadsto(g, t_{i+r}, s − r) then
    S = S ⋃ {t_i}
  end if
end for
return |S|/|F|
```

Algorithm D.2 $check\_leadsto(g, t_i, s)$

```
if g ∈ labels(t_i) then
  true
else if (s = 0) ∨ (t_i = |T|) then
  false
else
  check_leadsto(g, t_{t+1}, s − 1)
end if
```

In the worst case, f occurs at every timepoint, g never occurs and $s = \infty$. Then, for each $t \in T$, we iterate over the entire set T, making the algorithm

$O(T^2)$. However, in practice one would not consider the whole set T, but only the sets of times labeled with g.

D.1.2 *Until*

We begin with the until formula $f\mathcal{U}^{\geqslant r,\leqslant s}g$, and timepoints $t_i \in T$ satisfying f and g labeled as such.

Algorithm D.3 $until - prob$

$F = \{t_i : f \in labels(t_i)\}$
$G = \{t_i : g \in labels(t_i)\}$
$S = G$
for $t_i \in F$ **do**
  if $check\_until(r, s, t_i)$ **then**
    $S = S \bigcup \{t_i\}$
  end if
end for
return $|S|/|F \bigcup G|$

Algorithm D.4 $check\_until(r, s, t_i)$

if $(r \leqslant 0) \wedge (g \in labels(t_i))$ **then**
  true
else if $(f \notin labels(t_i)) \vee (t_i = |T|) \vee (s = 0)$ **then**
  false
else
  $check\_until(r - 1, s - 1, t_{i+1})$
end if

D.1.3 *Unless*

We begin with the unless formula $f\mathcal{U}^{\geqslant r,\leqslant s}g$, and timepoints $t_i \in T$ satisfying f and g labeled as such.

Algorithm D.5 $unless - prob$

$F = \{t_i : f \in labels(t_i)\}$
$G = \{t_i : g \in labels(t_i)\}$
$S = G$
for $t_i \in F$ **do**
  if $check\_unless(r, s, t_i)$ **then**
    $S = S \bigcup \{t_i\}$
  end if
end for
return $|S|/|F \bigcup G|$

Algorithm D.6 $check\_unless(r, s, t_i)$

if $(r \leqslant 0) \wedge (g \in labels(t_i))$ or $(f \in labels(t_i) \wedge s = 0)$ **then**
  true
else if $(f \notin labels(t_i)) \vee (t_i = |T|) \vee (s = 0)$ **then**
  false
else
  $check\_unless(r - 1, s - 1, t_{i+1})$
end if

E

EXAMPLES

E.1 CALCULATING THE PROBABILITY OF A PARTICULAR CAUSE

Take the structure in figure E.1, where X can cause Y through three paths: A,B and C. Now let us say we have a particular token instance, where we know that both X and Y occurred, and want to find the support for the hypotheses (the three possible paths to Y), given that we know X_1 and Y_3. Since we do not know whether A_2, B_2, or C_2 are true, we must first calculate these probabilities. The probability, $P(C_2|X_1, Y_3)$ is given by [1]:

$$1 - P(\neg C_2|X_1, Y_3). \tag{E.1}$$

Thus we now must compute:

$$\frac{P(\neg C_2 \wedge X_1 \wedge Y_3)}{P(X_1 \wedge Y_3)} \tag{E.2}$$

1 In this special case, we could have simplified matters with the observation that:

$$P(C \vee A \vee B|X, Y) = 1,$$

since the only paths from X to Y are through states where one of these conditions is true. These three conditions (Y, A and B) are also independent given X and Y and thus we have:

$$P(C|X, Y) + P(A|X, Y) + P(B|X, Y) = 1,$$

and thus:

$$P(C|X, Y) = 1 - [P(A|X, Y) + P(B|X, Y)],$$

where $P(A|X, Y)$ and $P(B|X, Y)$ are defined as in equation (6.11).

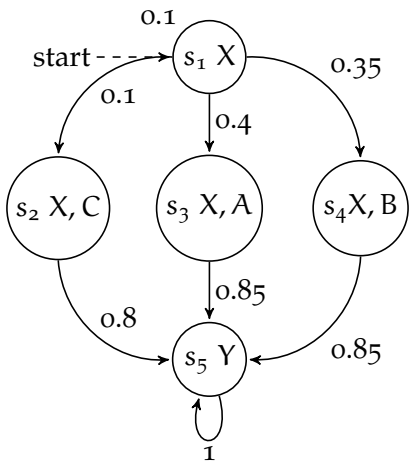


Figure E.1.: Token causality example.

First, taking the numerator, we construct K:

$$K = \{k_0 = true, k_1 = X, k_2 = \neg C, k3 = Y\}.$$

Let us also say that there is a state s_0 (not shown in Figure E.1), that is prior to s_1 and from which there is a transition to s_1 with the probability 0.1 (shown with the dashed line in Figure E.1). Then, with $t = 3$, the set of sequences, $\Pi(t, s)$, satisfying all $k_i \in K$ are:

$$s_0 \to s_1 \to s_3 \to s_5, \text{ and}$$

$$s_0 \to s_1 \to s_4 \to s_5.$$

We have two paths from X to Y that do not include a state where C is true at time 2. Thus,

$$P(3, s_0) = \mathcal{T}(s_0, s_1) \times P(2, s_1),$$

where,

$$P(2, s_1) = \mathcal{T}(s_1, s_3) \times P(1, s_3) + \mathcal{T}(s_1, s_4) \times P(1, s_4).$$

Then, we have:

$$P(1, s_3) = \mathcal{T}(s_3, s_5) \times P(0, s_5), \text{ and}$$
$$P(1, s_4) = \mathcal{T}(s_4, s_5) \times P(0, s_5).$$

In both cases,

$$P(0, s_5) = 1.$$

Substituting this value and the known transition probabilities we find:

$$P(1, s_3) = 0.85 \times 1 = 0.85, \text{ and}$$
$$P(1, s_4) = 0.85 \times 1 = 0.85,$$

thus:

$$P(2, s_1) = 0.4 \times 0.85 + 0.35 \times 0.85.$$

Finally,

$$\begin{aligned} P(3, s_0) &= 0.1 \times (0.4 \times 0.85 + 0.35 \times 0.85) \\ &= 0.06375. \end{aligned} \tag{E.3}$$

This is the numerator of equation E.2. Now we can compute the denominator similarly with:

$$K = \{k_0 = \mathrm{true}, k_1 = X, k_2 = \mathrm{true}, k3 = Y\}.$$

We have three paths satisfying our conditions:

$$s_0 \to s_1 \to s_2 \to s_5,$$

$$s_0 \to s_1 \to s_3 \to s_5, \text{ and}$$

$$s_0 \to s_1 \to s_4 \to s_5.$$

These are all three paths from s_1 (where X is true) to s_5 (where Y is true).

As before:

$$P(3, s_0) = \mathcal{T}(s_0, s_1) \times P(2, s_1),$$

but now:

$$P(2, s_1) = \mathcal{T}(s_1, s_2) \times P(1, s_2) + \mathcal{T}(s_1, s_3) \times P(1, s_3) + \mathcal{T}(s_1, s_4) \times P(1, s_4).$$

Proceeding as before, with the addition of the path through s_2:

$$P(1, s_2) = \mathcal{T}(s_2, s_5) \times P(0, s_5),$$

$$P(1, s_3) = \mathcal{T}(s_3, s_5) \times P(0, s_5), \text{ and}$$

$$P(1, s_4) = \mathcal{T}(s_4, s_5) \times P(0, s_5).$$

Once again:

$$P(0, s_5) = 1.$$

Substituting this value and the transition probabilities we get:

$$P(1, s_2) = 0.8 \times 1 = 0.8,$$
$$P(1, s_3) = 0.85 \times 1 = 0.85, \text{ and}$$
$$P(1, s_4) = 0.85 \times 1 = 0.85.$$

Thus:

$$P(2, s_1) = 0.1 \times 0.8 + 0.4 \times 0.85 + 0.35 \times 0.85,$$

and

$$\begin{aligned} P(3, s_0) &= 0.1 \times (0.1 \times 0.8 + 0.4 \times 0.85 + 0.35 \times 0.85) \\ &= 0.07175. \end{aligned}$$

Finally, substituting this and our previous result (E.3) into equation (E.2):

$$\begin{aligned} P(C_2|X_1, Y_3) &= 1 - \frac{0.06375}{0.07175} \\ &\approx 0.11. \end{aligned}$$

The support for $C_2 \rightsquigarrow Y_3$ is $\varepsilon_{avg}(C_t, Y_{t+1}) \times P(C_2)$. To find the probabilities and thus the support for A_2 and B_2 we repeat this procedure, changing the set K appropriately. As the calculations proceed in exactly

the same way, we omit these calculations but note that the probabilities are:

$$\begin{aligned} P(A_2|X_1, Y_3) &= 1 - \frac{0.03775}{0.07175} \\ &\approx 0.47. \\ P(B_2|X_1, Y_3) &= 1 - \frac{0.042}{0.07175} \\ &\approx 0.41. \end{aligned}$$

GLOSSARY

actual cause In terms of token causality, the actual cause is the cause of an effect at a particular time and place. This is how the term is used by Pearl [102]. Here it is used synonymously with genuine cause, and applies to both type and token level cases.

asymmetry (of causation) Causality is asymmetric as the fact that x causes y does not imply that y causes x. There is also temporal asymmetry, since it is generally (though not always) assumed that the cause is earlier than the effect.

background context In this work, unless otherwise stated, background contexts always refer to causal background contexts.

Bayesian network Also called a Bayes net, this is a graphical model consisting of a directed acyclic graph, where the absence of an edge represents conditional independence.

causal background context Relative to a particular effect, this is the set of all factors relevant to the occurrence of the effect. When looking at a particular cause of the effect, this may be reduced to the set of all relevant factors that are independent of the cause (This is the meaning in Eells's work. See section 2.3.3.).

causal chain A sequence of relationships, $x_1 \rightarrow x_2 \rightarrow ... \rightarrow x_n$, where each x_i occurs before and causes the next x_{i+1}.

causal fork (Salmon and Reichenbach) There are various types of forks (see conjunctive forks and interactive forks), but common to all of them is a situation where two things seem correlated due to their having a common cause.

causal relationship There is a causal relationship between x and y if either x causes y or y causes x.

causal structure In this work, the term refers to the underlying DTMC (a probabilistic Kripke structure) that is assumed to be generating the behavior of the systems we observe.

causal sufficiency A set of variables is causally sufficient if it includes all of the common causes of pairs on that set. This term is used primarily by SGS [120].

causally connected (Reichenbach [108]) Two events are causally connected if one is a cause of the other, or if they have a common cause. In mechanistic theories, two events are causally connected if they are connected by a mechanism.

causally relevant x is causally relevant to y if it is a positive cause or negative cause of y. A factor with mixed relevance may also be called causally relevant depending on the theory.

common cause z is a common cause of x and y if z is a cause of x and z is a cause of y.

common cause principle (Reichenbach) If two events are probabilistically dependent, then if there is a third event such that conditional

on this event, the first two are no longer dependent, this third event screens-off the first two, and is a common cause of both.

conjunctive fork (Reichenbach) With three events x, y and z, they form a conjunctive fork if one of the events is a common cause of the other two. That is, there is a correlation between y and z and it is fully explained once we know x, as it is a cause of both y and z.

context unanimity The notion that a causal relationship must hold in all background contexts. If x is a positive cause of y, then for context unanimity to hold, there cannot be any backgrounds in which x is not a positive cause of y (for example, one in which it is a negative or neutral cause).

counterfactual (counterfactual conditional) These are of the form: If c had not happened, then e would not have occurred.

counterfactual causal dependence With distinct events c and e, e causally depends on c if: were c to occur, e would occur as well and if c were not to occur, e would not occur either.

deterministic cause Synonymous with sufficient cause.

direct cause This is the most immediate cause of the effect. That is, it does not bring about the effect by some intermediate factor. Note that direct causes are relative to the scale at which we are viewing a system. A direct cause at one level may be an indirect cause at another. See indirect cause.

directed acyclic graph A graph with directed edges between nodes that does not contain any directed cycles. In causality this usually

refers to such a graph where a directed edge goes from the node representing a cause to the node representing an effect.

dynamic Bayesian network A graphical model that describes the dependencies of variables across time. A common implementation is one Bayesian network describing the initial state of the system, with a set of Bayesian networks (one for each time slice) and connections between them describing dependencies between variables across time.

epistemology (of causation) Epistemic theories of causality focus not on what causality is, but how we can find causal relations. (See also metaphysics (of causation)).

factor (causal) We use the term factor to mean that the causal relata could be events, properties, facts, mental states, etc. We make no claims as to what is capable of being a cause.

faithfulness (Primarily used by SGS) The independence relationships in a causal graph are exactly those of the structure generating it.

genuine cause Usually this refers to the true cause of an effect, relative to a particular theory. Note that most theories of causality do not correctly handle all possible cases and thus something may genuinely cause an effect without being labeled a genuine cause. When we refer to genuine causes, we mean something independent of theory: that objectively one thing causes another. Synonymous with actual cause.

Granger cause One time series, C, Granger causes another time series, E, if the probability of E, given lagged values of all available information including C, is statistically different from the probability of E when the information does not include the lagged values of C.

graphical model A graph where edges between nodes (which represent variables) describe conditional independencies. (See Bayesian network).

indirect cause A cause that acts through some intermediate effects. There are other events and factors between an indirect cause and its effect. See direct cause.

insignificant cause A prima facie cause cause that makes little difference to the effect. Note that some causes may seem insignificant based on particular data sets, but turn out to be genuine.

interactive fork (Salmon) In this type of fork, two processes intersect such that they are changed after their intersection and are not screened off from one another by a common cause.

intervention In interventionist theories of causality, a cause is something which may be used to alter, or manipulate, its effect. Then, under an ideal intervention (one that modifies only the cause) the effect should be modified, while the reverse is not true.

INUS condition An insufficient but non-redundant part of an unnecessary but sufficient condition.

Markov chain A sequence of states where the future is independent of past states, conditioned on the present state.

mechanism An interaction of parts that produces a particular behavior. A causal mechanism elucidates how the cause brings about the effect by detailing the processes connecting them.

metaphysics (of causation) When we discuss metaphysics of causation, we are referring to what it means for something to be a cause, what makes something an instance of causation rather than a mere correlation. Here we are trying to get at the underlying fact of what is.

mixed relevance x has mixed causal relevance for y if x raises the probability of y in some causal background contexts, and lowers it in others. This definition is primarily used by Eells.

necessary cause For an effect e and cause c, c is necessary for e if e cannot occur without c. That is, every occurrence of e is preceded by an occurrence of c: e implies c.

negative cause A cause that inhibits, or prevents, the effect. In the case of probabilistic causality, a negative cause decreases the probability of the effect.

neutral relevance x has neutral causal relevance for y if x and y are independent relative to all causal background contexts. That is, for all contexts K, $P(y|x \wedge K) = P(y|\neg x \wedge K)$.

omission (causation by) The absence of a factor bringing about the effect. For example, forgetting to water a plant can cause it to die.

overdetermination See redundant causation.

perfect fork (Salmon) In this case there is a deterministic relation between a common cause and two effects, such that the related probabilities are 0 or 1, so it is not possible to determine whether the case is a conjunctive fork or interactive fork.

positive cause A cause that brings about the effect. In the case of probabilistic causality, a positive cause increases the probability of the effect.

preemption This is a special case of redundant causation. In this case the potential causes are not symmetric, one actually occurs earlier, bringing about the effect and *preempting* the other from causing the effect. Difficulties arise when using counterfactual accounts of causation, as had the first cause not occurred, the second would have brought about the effect.

prima facie cause A seeming, or possible, cause. In probabilistic theories, this is simply one that occurs earlier than and raises the probability of the effect.

redundant causation This refers to token-level cases where multiple causes of an effect are present, and either alone would cause the effect. The term is used synonymously with overdetermination.

screening off See common cause principle.

Simpson's paradox A correlation (positive or negative) between two variables is found in a general population but one can find subpopulations such that in every subpopulation the relationship is reversed. In terms of causality, we might find that C is a positive cause of

E in a general population, but that in every subpopulation C is a negative cause of E.

spurious A spurious cause is one that may seem genuine – by appearing to raise the probability of the effect – but that is actually not causing the effect.

sufficient cause For an effect e and cause c, c is sufficient for e if c alone is enough to bring about c. That is, every occurrence of c is followed by an occurrence of e: c implies e. These are also referred to as deterministic causes. Note that this is not synonymous with causal sufficiency.

supplementary cause Two events are supplementary causes if the probability of the effect given that both have occurred is greater than the probability given either alone. This term is used mainly by Suppes [124].

temporal priority A cause must be earlier than its effect. Note that this does not always mean strictly earlier, as some theories allow cause and effect to be simultaneous.

token cause The token cause of an effect that occurs at a particular point in spacetime is the cause that also occurs and which is causally connected to this particular occurrence of the effect.

token level We sometimes refer to two levels of causality. Token level claims are those that refer to token causes. This is sometimes referred to as actual or singular causality.

transitivity (of causation) The notion that if A causes B and B causes C, that A causes C.

trumping Trumping is a case of both redundant causation and preemption. Both causes occur, and either alone would have caused the effect, but in fact the effect is due to only one of the causes. A common example is the case of a soldier who hears an order to advance from both a Sergeant and a major. He is actually obeying the superior officer: the order by the major trumps that by the Sergeant.

type level Type level claims refer to general properties between factors or events, such as that between smoking and lung cancer.

INDEX

BIBLIOGRAPHY

[1] Merriam-webster online dictionary, 2008. (Cited on page 90.)

[2] Armen Aghasaryan, Eric Fabre, Albert Benveniste, Renée Boubour, and Claude Jard. Fault Detection and Diagnosis in Distributed Systems: An Approach by Partially Stochastic Petri Nets. *Discrete Event Dynamic Systems*, 8(2):203–231, 1998. (Cited on page 50.)

[3] M. Andrecut and SA Kauffman. A simple method for reverse engineering causal networks. *J. Phys. A: Math. Gen*, 39(46):L647–L655, 2006. (Cited on page 56.)

[4] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. (Cited on pages 233 and 234.)

[5] Yoav Benjamini and Daniel Yekutieli. The Control of the False Discovery Rate in Multiple Testing under Dependency. *Annals of Statistics*, 29(4):1165–1188, 2001. (Cited on pages 232, 233, and 234.)

[6] PJ Bickel, EA Hammel, and JW O'connell. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175):398, 1975. (Cited on page 21.)

[7] Maroua Bouzid and Antoni Ligeza. Temporal Causal Networks for Simulation and Diagnosis. *Proceedings of the second IEEE In-*

ternational Conference on Engineering of Complex Computer Systems, ICECCS, 96:458–465, 1996. (Cited on page 50.)

[8] Patrick Brandt. MSBVAR R package version 0.4, 2009. (Cited on pages 187 and 194.)

[9] Emery N. Brown, Robert E. Kass, and Partha P. Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience*, 7:456–461, 2004. (Cited on page 56.)

[10] Nancy Cartwright. Causal Laws and Effective Strategies. *Nous*, 13(4):419–437, 1979. (Cited on page 177.)

[11] Nancy Cartwright. *Nature's Capacities and Their Measurement*. Oxford University Press, 1994. (Cited on pages 92 and 100.)

[12] Nancy Cartwright. Against Modularity, the Causal Markov Condition, and Any Link Between the Two: Comments on Hausman and Woodward. *The British Journal for the Philosophy of Science*, 53(3):411–453, 2002. (Cited on page 40.)

[13] Nancy Cartwright. What Is Wrong with Bayes Nets? *Probability Is the Very Guide of Life: The Philosophical Uses of Chance*, 2003. (Cited on page 40.)

[14] Kenneth Chan, Iman Poernomo, Heinz Schmidt, and Jane Jayaputera. A Model-Oriented Framework for Runtime Monitoring of Nonfunctional Properties. *Lecture notes in computer science*, 3712:38, 2005. (Cited on page 117.)

[15] Rakesh Chandra. Managing Temporal Financial Data in Extensible Databases. In *Proceedings of the 19th VLDB Conference*, 1994. (Cited on page 57.)

[16] CS Chao, DL Yang, and AC Liu. An Automated Fault Diagnosis System Using Hierarchical Reasoning and Alarm Correlation. *Journal of Network and Systems Management*, 9(2):183–202, 2001. (Cited on page 50.)

[17] Yonghong Chen, Govindan Rangarajan, Jianfeng Feng, and Mingzhou Ding. Analyzing multiple nonlinear time series with extended Granger causality. *Physics Letters A*, 324:26–35, 2004. (Cited on page 50.)

[18] Edmumd M. Clarke, E. Allen Emerson, and A. Prasad Sistla. Automatic Verification of Finite-State Concurrent Systems Using Temporal Logic Specifications. *ACM Transactions on Programming Languages and Systems*, 8(2):244–263, 1986. (Cited on pages 68 and 219.)

[19] Edmund M. Clarke, Orna Grumberg, and Doron A. Peled. *Model Checking*. MIT Press, 1999. (Cited on pages 69, 220, and 227.)

[20] *Dean v. United States*. 556 U.S. __ (2009). (Cited on page 166.)

[21] Mingzhou Ding, Yonghong Chen, and Steven L. Bressler. Granger Causality: Basic Theory and Application to Neuroscience. *Arxiv preprint q-bio/0608035*, 2006. (Cited on page 56.)

[22] John Dupre. Probabilistic Causality Emancipated. *Midwest Studies in Philosophy*, 9:169–175, 1984. (Cited on page 31.)

[23] John Dupre. Probabilistic Causality: A Rejoinder to Ellery Eells. *Philosophy of Science*, 57(4):690–698, December 1990. (Cited on page 31.)

[24] Dorothy Edgington. On Conditionals. *Mind*, 104(414):235–329, 1995. (Cited on page 13.)

[25] Ellery Eells. Cartwright and Otte on Simpson's Paradox. *Philosophy of Science*, 54(2):233–243, June 1987. (Cited on page 21.)

[26] Ellery Eells. Probabilistic Causality: Reply to John Dupre. *Philosophy of Science*, 54(1):105–114, March 1987. (Cited on page 31.)

[27] Ellery Eells. *Probabilistic Causality*. Cambridge University Press, 1991. (Cited on pages 30, 32, 45, and 143.)

[28] Ellery Eells and Elliott Sober. Probabilistic Causality and the Question of Transitivity. *Philosophy of Science*, 50(1):35–57, 1983. (Cited on pages 92, 94, and 107.)

[29] Bradley Efron. Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Association*, 99(465):96–105, 2004. (Cited on pages 126, 232, and 235.)

[30] Bradley Efron. Size, Power, and False Discovery Rates. *Ann. Statist*, 35(4):1351–1377, 2007. (Cited on pages 126 and 235.)

[31] Bradley Efron, Brit Turnbull, and Balasubramanian Narasimhan. locfdr: Computes local false discovery rates. R package, 2008. (Cited on page 195.)

[32] Michael Eichler and Vanessa Didelez. Causal Reasoning in Graphical Time Series Models. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, 2007. (Cited on page 50.)

[33] E. Allen Emerson. Model checking and the Mu-calculus. In *DIMACS Series in Discrete Mathematics*, pages 185–214. American Mathematical Society, 1997. (Cited on page 219.)

[34] Eugene F. Fama and Kenneth R. French. The Cross-Section of Expected Stock Returns. *Journal of Finance*, 47(2):427–465, 1992. (Cited on pages 191 and 193.)

[35] Joseph J. Finger. Exploiting Constraints in Design Synthesis. Master's thesis, Stanford University, 1987. (Cited on page 52.)

[36] David A. Freedman. On specifying graphical models for causation, and the identification problem. Technical report, U.C. Berkeley, 2003. (Cited on page 37.)

[37] Kenneth R. French and Eugene Fama. Fama french - data library. (Cited on page 193.)

[38] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, 7(3-4):601–620, 2000. (Cited on page 56.)

[39] Nir Friedman, Kevin Murphy, and Stuart Russell. Learning the Structure of Dynamic Probabilistic Networks. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 139–147, 1998. (Cited on page 41.)

[40] Gabriel P. Fung, Jeffrey X. Yu, and Wai Lam. Stock prediction: Integrating text mining approach using real-time news. *Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003*

IEEE International Conference on, pages 395–402, 2003. (Cited on page 57.)

[41] Laura Giordano, Alberto Martelli, and Camilla Schwind. Ramification and Causality in a Modal Action Logic. *Journal of Logic and Computation*, 10(5):625–662, 2000. (Cited on pages 51, 53, and 54.)

[42] Clark Glymour, Richard Scheines, Peter Spirtes, and Joseph Ramsey. TETRAD IV software, 2004. (Cited on page 187.)

[43] Irving J. Good. A Causal Calculus (I). *British Journal for the Philosophy of Science*, XI(44):305–318, 1961. (Cited on page 174.)

[44] Clive W.J. Granger. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica*, 37(3):424–438, 1969. (Cited on page 48.)

[45] Clive W.J. Granger. Testing for Causality: A Personal Viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980. (Cited on pages 49 and 57.)

[46] Joseph Y. Halpern and Judea Pearl. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4):843–887, 2005. (Cited on pages 42 and 52.)

[47] Joseph Y. Halpern and Judea Pearl. Causes and Explanations: A Structural-Model Approach. Part II: Explanations. *The British Journal for the Philosophy of Science*, 56(4):889–911, 2005. (Cited on page 42.)

[48] Joseph Y. Halpern and Yoav Shoham. A Propositional Modal Logic of Time Intervals. In *Proceedings 1st Annual IEEE Symp. on Logic in Computer Science, LICS'86, Cambridge, MA, USA, 16–18 June 1986*, pages 279–292. IEEE Computer Society Press, Washington, DC, 1986. (Cited on page 54.)

[49] Hans Hansson and Bengt Jonsson. A Logic for Reasoning about Time and Reliability. *Formal Aspects of Computing*, 6(5):512–535, 1994. (Cited on pages 69, 73, 117, 136, 224, 228, 229, 239, 243, 244, 247, and 248.)

[50] Alexander J. Hartemink et al. Banjo: Bayesian Network Inference with Java Objects. http://www.cs.duke.edu/ amink/software/banjo/, 2008. (Cited on page 187.)

[51] Daniel M. Hausman. *Causal Asymmetries*. Cambridge University Press, New York, 1998. (Cited on page 62.)

[52] Daniel M. Hausman. Causal Relata: Tokens, Types, or Variables? *Erkenntnis*, 63(1):33–54, 2005. (Cited on page 143.)

[53] Daniel M. Hausman and James Woodward. Independence, invariance and the causal Markov condition. *The British Journal for the Philosophy of Science*, 50(4):521–583, 1999. (Cited on page 40.)

[54] Wolfram Hesse, Eva Möller, Matthias Arnold, and Bärbel Schack. The use of time-variant EEG Granger causality for inspecting directed interdependencies of neural assemblies. *Journal of Neuroscience Methods*, 124(1):27–44, 2003. (Cited on page 56.)

[55] Christopher Read Hitchcock. The Mishap at Reichenbach Fall: Singular vs. General Causation. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 78(3):257–291, 1995. (Cited on pages 174, 175, and 178.)

[56] Mark Hopkins and Judea Pearl. Causality and Counterfactuals in the Situation Calculus. *Journal of Logic and Computation*, 17(5):939, 2007. (Cited on page 52.)

[57] David Hume. *A Treatise of Human Nature*. Prometheus Books, 1992. (Cited on page 9.)

[58] David Hume. *An Enquiry Concerning Human Understanding*. Dover Publications, 2004. (Cited on page 13.)

[59] Paul Humphreys and David Freedman. The Grand Leap. *The British Journal for the Philosophy of Science*, 47(1):113–123, March 1996. (Cited on page 40.)

[60] Jiashun Jin and T. Tony Cai. Estimating the Null and the Proportion of non-Null effects in Large-scale Multiple Comparisons. *Journal of the American Statistical Association*, 102:495–506, 2006. (Cited on page 187.)

[61] Maciej Kamiński, Mingzhou Ding, Wilson A. Truccolo, and Steven L. Bressler. Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological Cybernetics*, 85(2):145–157, 2001. (Cited on page 56.)

[62] Dong-Hee Kim and Hawoong Jeong. Systematic analysis of group identification in stock markets. *Physical Review E*, 72(4):46133, 2005. (Cited on page 57.)

[63] Jaegwon Kim. Causes and Events: Mackie on Causation. *The Journal of Philosophy*, 68(14):426–441, July 1971. (Cited on page 11.)

[64] Samantha Kleinberg and Bud Mishra. Multiple Testing of Causal Hypotheses. Canterbury, UK, September 2008. CAPITS Causality Study Fortnight.

[65] Samantha Kleinberg and Bud Mishra. Metamorphosis: the Coming Transformation of Translational Systems Biology. *Queue*, 7(9):40–52, 2009.

[66] Samantha Kleinberg and Bud Mishra. The Temporal Logic of Causal Structures. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI-09)*, pages 303–312, Corvallis, Oregon, 2009. AUAI Press.

[67] Samantha Kleinberg and Bud Mishra. Multiple Testing of Causal Hypotheses. In Phyllis McKay Illari, Federica Russo, and Jon Williamson, editors, *Causality in the Sciences*. Oxford University Press, 2010. (To appear). (Cited on page 235.)

[68] Samantha Kleinberg and Bud Mishra. The Temporal Logic of Token Causes. In *Proceedings of the 12th International Conference on the Principles of Knowledge Representation and Reasoning (KR2010)*, Toronto, Canada, May 2010. (To appear).

[69] Saul Kripke. Semantical Considerations on Modal Logic. *Acta Philosophica Fennica*, 16(1963):83–94, 1963. (Cited on pages 69 and 220.)

[70] Christopher J. Langmead. Towards Inference and Learning in Dynamic Bayesian Networks using Generalized Evidence. Technical Report CMU-CS-08-151, Carnegie Mellon University, 2008. (Cited on page 42.)

[71] Christopher J. Langmead, Sumit K. Jha, and Edmund M. Clarke. Temporal Logics as Query Languages for Dynamic Bayesian Networks: Application To D. Melanogaster Embryo Development. Technical Report CMU-CS-06-159, Carnegie Mellon University, 2006. (Cited on page 42.)

[72] Steffen L. Lauritzen. Causal Inference from Graphical Models. In *Complex Stochastic Systems*, pages 63–107. Chapman and Hall/CRC Press, 2001. (Cited on page 37.)

[73] M. Leucker and C. Schallhart. A brief account of runtime verification. *Journal of Logic and Algebraic Programming*, 78(5):293–303, 2009. (Cited on page 116.)

[74] David Lewis. Causation. *The Journal of Philosophy*, 70(17):556–567, October 1973. (Cited on pages 13, 14, 45, and 46.)

[75] David Lewis. Postscripts to "Causation". *Philosophical Papers*, 2, 1986. (Cited on pages 15 and 16.)

[76] David Lewis. Causation as Influence. *The Journal of Philosophy*, 97(4):182–197, April 2000. (Cited on page 15.)

[77] Vladimir Lifschitz. Formal theories of action. *Readings in nonmonotonic reasoning*, pages 410–432, 1987. (Cited on page 52.)

[78] Vladimir Lifschitz. Situation Calculus And Causal Logic. In Anthony G. Cohn, Lenhart Schubert, and Stuart C. Shapiro, editors, *KR'98: Principles of Knowledge Representation and Reasoning*, pages 536–546. Morgan Kaufmann, San Francisco, California, 1998. (Cited on page 53.)

[79] Fangzhen Lin. Embracing Causality in Specifying the Indirect Effects of Actions. In Chris Mellish, editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1985–1991, San Francisco, 1995. Morgan Kaufmann. (Cited on pages 52 and 53.)

[80] Fangzhen Lin. Embracing Causality in Specifying the Indeterminate Effects of Actions. In *AAAI/IAAI, Vol. 1*, pages 670–676, 1996. (Cited on page 52.)

[81] Jan Lunze and Frank Schiller. An example of fault diagnosis by means of probabilistic logic reasoning. *Control Engineering Practice*, 7(2):271–278, 1999. (Cited on page 50.)

[82] John Leslie Mackie. Causes and Conditions. *American Philosophical Quarterly*, 2(4):245–264, 1965. (Cited on page 11.)

[83] John Leslie Mackie. *The Cement of the Universe*. Clarendon Press, 1974. (Cited on pages 10 and 46.)

[84] Norman McCain and Hudson Turner. A Causal Theory of Ramifications and Qualifications. In Chris Mellish, editor, *Proceedings of*

the Fourteenth International Joint Conference on Artificial Intelligence, pages 1978–1984, San Francisco, 1995. Morgan Kaufmann. (Cited on page 52.)

[85] John McCarthy and Patrick J. Hayes. Some Philosophical Problems from the Standpoint of Artificial Intelligence. *Machine Intelligence*, 4(463-502):288, 1969. (Cited on page 51.)

[86] Drew McDermott. A temporal logic for reasoning about processes and plans. *Cognitive Science*, 6(2):101–155, 1982. (Cited on page 54.)

[87] Michael McDermott. Redundant Causation. *The British Journal for the Philosophy of Science*, 46(4):523–544, December 1995. (Cited on page 14.)

[88] Grant McQueen and V. Vance Roley. Stock Prices, News, and Business Conditions. *Review of Financial Studies*, 1993. (Cited on page 57.)

[89] P. Menzies. Causal Models, Token Causation, and Processes. *Philosophy of Science*, 71:820–832, 2004. (Cited on page 48.)

[90] Marc-Andre Mittermayer and Gerhard F. Knolmayer. NewsCATS: A News Categorization and Trading System. *Proceedings of the Sixth International Conference on Data Mining*, pages 1002–1007, 2006. (Cited on page 57.)

[91] Ben Moszkowski. A Temporal Logic for Multilevel Reasoning about Hardware. *Computer*, 18(2):10–19, 1985. (Cited on page 219.)

[92] Nitai D D. Mukhopadhyay and Snigdhansu Chatterjee. Causality and pathway search in microarray time series experiment. *Bioinformatics*, 23(4):442, 2007. (Cited on page 56.)

[93] Kevin Murphy and Saira Mian. Modelling Gene Expression Data using Dynamic Bayesian Networks. Technical report, University of California, Berkeley, CA, 1999. (Cited on page 41.)

[94] Jukka-Pekka Onnela, Kimmo Kaski, and Janos Kertész. Clustering and information in correlation based financial networks. *The European Physical Journal B-Condensed Matter*, 38(2):353–362, 2004. (Cited on page 57.)

[95] Rainer Opgen-Rhein and Korbinian Strimmer. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology*, 1(1):37, 2007. (Cited on page 56.)

[96] Richard Otte. A critique of suppes' theory of probabilistic causality. *Synthese*, 48(2):167–189, 1981. (Cited on page 25.)

[97] Richard Otte. Probabilistic Causality and Simpson's Paradox. *Philosophy of Science*, 52(1):110–125, March 1985. (Cited on page 21.)

[98] David Papineau. Can We Reduce Causal Direction to Probabilities? *Philosophy of Science Association*, 2:238–252, 1992. (Cited on page 62.)

[99] David Papineau. *Stochastic Causality*, chapter Metaphysics over Methodology–Or, Why Infidelity Provides No Grounds To Divorce Causes from Probabilities. University of Chicago Press, 2001. (Cited on page 101.)

[100] Judea Pearl. Embracing causality in default reasoning. *Artificial Intelligence*, 35(2):259–271, 1988. (Cited on page 51.)

[101] Judea Pearl. Reasoning with Cause and Effect. In *IJCAI*, pages 1437–1449, 1999. (Cited on pages 42 and 43.)

[102] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000. (Cited on pages 37, 42, 43, 45, 100, and 258.)

[103] Amir Pnueli. The temporal logic of programs. In *FOCS*, pages 46–57. IEEE, 1977. (Cited on page 219.)

[104] David Poole. Representing diagnosis knowledge. *Annals of Mathematics and Artificial Intelligence*, 11(1):33–50, 1994. (Cited on page 50.)

[105] Arthur N. Prior. *Time and Modality*. Oxford University Press, 1957. (Cited on page 68.)

[106] Arthur N. Prior. *Past, Present and Future*. Oxford University Press, USA, 1967. (Cited on pages 68 and 219.)

[107] Naren Ramakrishnan, P.S. Sastry, K.P. Unnikrishnan, , and R. Uthurusamy. 4th KDD Workshop on Temporal Data Mining, 2006. (Cited on page 209.)

[108] Hans Reichenbach. *The Direction of Time*. Dover Publications, 2000. (Cited on pages xii, 18, 20, 62, and 259.)

[109] James M. Robins and Larry Wasserman. *Computation, causation, and discovery*, chapter On the Impossibility of Inferring Causation from Association without Background Knowledge, pages 305–322. AAAI Press / The MIT Press, 1999. (Cited on page 100.)

[110] Deborah A. Rosen. In defense of a probabilistic theory of causality. *Philosophy of Science*, pages 604–613, 1978. (Cited on page 178.)

[111] Bertrand Russell. *Human Knowledge: Its Scope and Limits*. Simon and Schuster, 1948. (Cited on page 10.)

[112] Wesley C. Salmon. Probabilistic Causality. *Pacific Philosophical Quarterly*, 61(1):50–74, 1980. (Cited on page 178.)

[113] Eric E. Schadt, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj Guhathakurta, Solveig K. Sieberts, Stephanie Monks, Marc Reitman, Chunsheng Zhang, Pek Y. Lum, Amy Leonardson, Rolf Thieringer, Joseph M. Metzger, Liming Yang, John Castle, Haoyuan Zhu, Shera F. Kash, Thomas A. Drake, Alan Sachs, and Aldons J. Lusis. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37:710–717, 2005. (Cited on page 56.)

[114] Richard Scheines. An Introduction to Causal Inference. *Causality in Crisis*, pages 185–99, 1997. (Cited on pages 37 and 39.)

[115] Young-Woo Seo, Joseph Giampapa, and Katia Sycara. Financial News Analysis for Intelligent Portfolio Management. Technical Report CMU-RI-TR-04-04, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, January 2004. (Cited on page 57.)

[116] Yoav Shoham. *Reasoning about change: time and causation from the standpoint of artificial intelligence*. MIT Press, 1988. (Cited on page 54.)

[117] Edward H. Simpson. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2):238–241, 1951. (Cited on page 20.)

[118] Brian Skyrms. *Causal Necessity*. Yale University Press, 1980. (Cited on page 21.)

[119] Elliott Sober and David Papineau. Causal Factors, Causal Inference, Causal Explanation. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 60:97–136, 1986. (Cited on pages 144 and 153.)

[120] Peter Spirtes, Clarke Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000. (Cited on pages 37, 40, 187, and 259.)

[121] Robert Stalnaker. A Theory of Conditionals. *Studies in logical theory*, 2:98–112, 1968. (Cited on page 13.)

[122] John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440, 2003. (Cited on page 232.)

[123] K. Strimmer. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics*, 24(12):1461, 2008. (Cited on page 195.)

[124] Patrick Suppes. *A probabilistic theory of causality*. North-Holland, 1970. (Cited on pages 22, 23, 24, 63, 88, 107, and 265.)

[125] Hudson Turner. A Logic of Universal Causation. *Artificial Intelligence*, 113(1-2):87–123, 1999. (Cited on page 53.)

[126] Wikipedia. Ronald opus — wikipedia, the free encyclopedia, 2008. [Online; accessed 2-December-2008]. (Cited on page 164.)

[127] James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, USA, 2005. (Cited on pages 100 and 143.)

[128] Changwon Yoo, Vesteinn Thorsson, and Gregory F. Cooper. Discovery of Causal Relationships in a Gene-Regulation Pathway From a Mixture of Experimental and Observational DNA Microarray Data. In *Proceedings of Pacific Symposium on Biocomputing*, 2002. (Cited on page 56.)

[129] Hakan L.S. Younes and Reid G. Simmons. Statistical probabilistic model checking with a focus on time-bounded properties. *Information and Computation*, 204(9):1368–1409, 2006. (Cited on page 69.)

[130] Cunlu Zou and Jianfeng Feng. Granger causality vs. dynamic Bayesian network inference: a comparative study. *BMC bioinformatics*, 10(1):122, 2009. (Cited on page 189.)